Beyond a snapshot:

Using machine learning to monitor and adaptively support organic chemistry students' mechanistic reasoning over time

Kumulative Dissertation

zur Erlangung des Doktorgrades

der Naturwissenschaftlichen Fachbereiche

(Fachbereich 08 – Chemie und Biologie)

der Justus-Liebig-Universität Gießen

Vorgelegt von

Paul P. Martin

Gießen, 2025

Die in dieser Dissertation dokumentierten Forschungsergebnisse wurden zwischen Juni 2022 und März 2025 am Institut für Didaktik der Chemie der Justus-Liebig-Universität Gießen unter der wissenschaftlichen Betreuung von Frau Prof. Dr. Nicole Graulich erarbeitet.

The research documented in this dissertation was conducted between June 2022 and March 2025 at the Institute of Chemistry Education at Justus-Liebig-University Giessen under the academic supervision of Prof. Dr. Nicole Graulich.

Ich danke meiner wissenschaftlichen Betreuerin Frau Prof. Dr. Nicole Graulich für die Möglichkeit, Teil ihrer Arbeitsgruppe zu sein, für das mir entgegengebrachte Vertrauen bei der Erforschung der hier dargelegten Thematik sowie für ihre fortwährende Unterstützung und die wertvollen, konstruktiven Diskussionen. Vielen Dank für die jahrelange vertrauensvolle Zusammenarbeit und die bereichernden Erlebnisse im Forschungsalltag und darüber hinaus.

I sincerely thank my academic supervisor Prof. Dr. Nicole Graulich for the opportunity to be part of her research group, for the trust she placed in me while exploring the topic of this dissertation, and for her continuous support and insightful, constructive discussions. I deeply appreciate our years of trustful collaboration and the enriching experiences in research and beyond.

"Given the far-reaching influence of [mechanistic reasoning], it is important to understand how and why students' [mechanistic reasoning] develops and changes over time."

(Adapted from Limeri et al., 2020, p. 2)

Limeri, L. B., Carter, N. T., Choe, J., Harper, H. G., Martin, H. R., Benton, A., & Dolan, E. L. (2020). Growing a growth mindset: Characterizing how and why undergraduate students' mindsets change. *International Journal of STEM Education*, *7*, 1-19.

Table of Contents

I. Abstract ix

II. Zusammenfassung xi

III. List of publications xiii

1. Synopsis: Beyond a snapshot–Using machine learning to monitor and adaptively support organic chemistry students' mechanistic reasoning over time 1

1.1 Introduction 1
1.2 Research areas 2
1.3 Theoretical framework: Essential epistemic heuristics for mechanistic reasoning in chemistry 7
1.4 Methodological framework: Computational grounded theory 10
1.5 Research objectives 12
1.6 Study context 13
1.7 Data analysis 17
1.8 Results 19
1.9 Discussion 34
1.10 Conclusion 40
1.11 Outlook and future work 40
1.12 References 47

2. When a machine detects student reasoning: A review of machine learning-based formative assessment of mechanistic reasoning 57

3. Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry 81

4. Beyond language barriers: Allowing multiple languages in postsecondary chemistry classes through multilingual machine learning 119

5. Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence 137

6. To what extent can we trust a machine? Establishing trustworthiness guidelines to evaluate machine learning models across chemistry contexts 155

7. Revealing rubric relations: Investigating the interdependence of a research-informed and a machine learning-based rubric in assessing student reasoning in chemistry 203

8. Mechanistic reasoning about alternatives: How students reason about plausible and implausible mechanisms in chemistry 245

9. Tracing trajectories: Using machine learning to map how chemistry students' mechanistic reasoning develops over time 271

10. To what extent is adaptive support equitable? Exploring the impact of machine learning-assisted adaptive support on students' mechanistic reasoning 323

11. Appendix 349

12. Acknowledgments 351

I. Abstract

Science education seeks to equip students with the skills to engage with the natural and engineered world. Mechanistic reasoning is essential for achieving this goal because it enables students to identify, describe, or analyze the underlying mechanisms of observed phenomena to explain, predict, or control their outcomes. However, students often use rote memorization or pattern recognition when identifying mechanisms, rather than engaging in mechanistic reasoning. As a result, they may redescribe mechanisms without identifying cause-and-effect relationships or view mechanisms as linear sequences of unrelated steps. To address this, explicitly supporting students' mechanistic reasoning in science instruction is essential. Nonetheless, research shows that uniform support is less effective in guiding students' mechanistic reasoning than adaptive support. Yet, adaptively supporting students' mechanistic reasoning is challenging due to the time-consuming process of manual coding. Furthermore, our understanding of how mechanistic reasoning develops over time and interacts with other cognitive and affective variables remains limited, complicating the design of adaptive support.

This dissertation addresses these challenges from a methodological and pedagogical perspective. From a methodological perspective, it explores the potential of unsupervised machine learning (ML) to capture the heterogeneity of undergraduate organic chemistry students' mechanistic reasoning about alternative reaction products. It also examines how supervised ML can automatically categorize this heterogeneity over time using a fine-grained scoring rubric. From a pedagogical perspective, this dissertation investigates how students integrate essential epistemic heuristics for mechanistic reasoning—operationalized through their use of granularity, disciplinary core ideas, and causality—using quantitative methods. In detail, it explores how these heuristics evolve, how they relate to other cognitive and affective variables, and how they are influenced by adaptive support.

Our findings indicate that unsupervised ML techniques can detect significant heterogeneity in how students demonstrate mechanistic reasoning. Specifically, our analysis reveals that students integrate diverse disciplinary core ideas, interconnected at varying levels of granularity and causality, into their mechanistic reasoning. The extent to which students integrate these heuristics depends on the specific task affordances. Building on these findings, we developed a supervised ML model that automatically evaluates students' mechanistic reasoning in line with these heuristics based on a fine-grained 24-category rubric. This automated scoring model allowed us to monitor and adaptively support students' mechanistic reasoning beyond snapshots in time. A quantitative analysis of the collected data revealed a statistically significant relationship between students' use of granularity and causality. Integrating disciplinary core ideas related to energy contributed to more complex mechanistic reasoning, along with high chemical concept knowledge and confidence when solving organic chemistry exercises. However, while many students improved their mechanistic reasoning due to the adaptive support over time, not all students benefited equally.

Overall, this dissertation contributes to the existing literature by advancing the methodological rigor and pedagogical utility of ML in monitoring and supporting students' mechanistic reasoning in undergraduate science courses. By combining the strengths of human and machine analysis, it demonstrates how to thoughtfully integrate ML methods into research processes to generate evidence-based insights into students' mechanistic reasoning. Leveraging ML facilitated large-scale analysis of students' mechanistic reasoning over time, guiding the design of adaptive support tailored to their mechanistic reasoning skills.

II. Zusammenfassung

Naturwissenschaftliche Bildung zielt darauf ab, Lernende zu befähigen, sich fundiert mit Phänomenen aus Natur und Technik auseinanderzusetzen. Mechanistisches Begründen trägt zu diesem Ziel bei, da es Lernenden ermöglicht, die zugrunde liegenden Mechanismen beobachteter Phänomene zu identifizieren, zu beschreiben oder zu analysieren, um Beobachtungen zu erklären oder vorherzusagen. Allerdings nutzen Lernende beim Beschreiben von Mechanismen häufig nicht mechanistisches Begründen, sondern Auswendiglernen oder Mustererkennung. Dadurch identifizieren sie Mechanismen, ohne Ursache-Wirkungs-Beziehungen zu erkennen. Darüber hinaus betrachten sie Mechanismen als lineare Abfolgen zusammenhangsloser Schritte. Um diesen Herausforderungen entgegenzuwirken, ist die Unterstützung des mechanistischen Begründens in der naturwissenschaftlichen Bildung essenziell. Forschungsergebnisse zeigen jedoch, dass uniforme Unterstützung das mechanistische Begründen weniger effektiv als adaptive Unterstützung fördert. Die Umsetzung longitudinaler adaptiver Unterstützung des mechanistischen Begründens ist allerdings herausfordernd, da die manuelle Kodierung von Freitext-Antworten zeitaufwendig ist. Zudem sind unsere Erkenntnisse darüber, wie sich mechanistisches Begründen longitudinal entwickelt und wie es mit anderen kognitiven und affektiven Variablen interagiert, noch begrenzt. Dies erschwert die Gestaltung adaptiver Unterstützung.

In dieser Dissertation werden die genannten Herausforderungen sowohl aus methodischer als auch aus fachdidaktischer Perspektive adressiert. Aus methodischer Sicht wird das Potenzial unüberwachten maschinellen Lernens (ML) untersucht, um die Heterogenität des mechanistischen Begründens über alternative Reaktionsprodukte von Studierenden der Organischen Chemie zu analysieren. Zudem wird erforscht, inwiefern überwachtes ML genutzt werden kann, um diese Heterogenität longitudinal mit einem ausdifferenzierten Kodierungsschema automatisch zu kategorisieren. Aus fachdidaktischer Sicht untersucht diese Dissertation quantitativ, wie Studierende zentrale epistemische Heuristiken für mechanistisches Begründen, erfasst durch ihre Nutzung von Granularität, fachlichen Kernideen und Kausalität, integrieren. Dabei wird untersucht, wie sich diese Heuristiken über die Zeit verändern, wie sie mit anderen kognitiven und affektiven Variablen zusammenhängen und wie sie durch adaptive Unterstützung beeinflusst werden.

Unsere Ergebnisse zeigen, dass unüberwachte ML-Techniken große Heterogenität im mechanistischen Begründen von Studierenden erkennen können. Insbesondere integrieren Studierende verschiedene fachliche Kernideen auf unterschiedlichen Granularitäts- und Kausalitätsebenen in ihr mechanistisches Begründen. Das Ausmaß, in dem Studierende diese Heuristiken integrieren, hängt von den spezifischen Anforderungen der jeweiligen Aufgabe ab. Aufbauend auf diesen Erkenntnissen wurde ein überwachtes ML-Modell entwickelt, das das mechanistische Begründen von Studierenden anhand dieser Heuristiken automatisch analysieren kann, wobei ein ausdifferenziertes Kodierungsschema mit 24 Kategorien genutzt wurde. Das überwachte ML-Modell ermöglichte eine Lernverlaufsdiagnose des mechanistischen Begründens mit einer großen Stichprobe. Eine quantitative Analyse der dabei gesammelten Daten zeigt eine statistisch signifikante Korrelation zwischen der von Studierenden genutzten Granularitäts- und Kausalitätsebenen. Die Integration fachlicher Kernideen zum Thema Energie, ein hohes chemisches Fachwissen sowie gesteigertes Vertrauen beim Bearbeiten organisch-chemischer Aufgaben förderten ein komplexeres mechanistisches Begründen. Während viele Studierende durch die adaptive Unterstützung Fortschritte zeigten, profitierten nicht alle gleichermaßen davon.

Insgesamt trägt diese Dissertation zu bestehender Forschung bei, indem sie methodische Ansätze und fachdidaktische Anwendungen von ML zur Erfassung und Förderung des mechanistischen Begründens in den Naturwissenschaften weiterentwickelt. Durch die gezielte Verknüpfung menschlicher und maschineller Analysen zeigt sie, wie ML sinnvoll in Forschungsprozesse integriert werden kann, um fundierte Erkenntnisse bezüglich des mechanistischen Begründens von Studierenden zu gewinnen. Der Einsatz von ML ermöglichte eine longitudinale Studie zur Untersuchung der Entwicklung des mechanistischen Begründens von Studierenden. Die Studie liefert wertvolle Erkenntnisse für die Gestaltung adaptiver Unterstützung, die gezielt an das mechanistische Begründen von Studierenden angepasst ist.

III. List of Publications

Papers in peer-reviewed journals

Martin, P. P., Kranz, D., & Graulich, N. (2024). Revealing rubric relations: Investigating the interdependence of a research-informed and a machine learning-based rubric in assessing student reasoning in chemistry. *International Journal of Artificial Intelligence in Education*, 1-39. https://doi.org/10.1007/s40593-024-00440-y

Martin, P. P., & Graulich, N. (2024). Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence. *Computers and Education: Artificial Intelligence*, *7*, Article 100265. https://doi.org/10.1016/j.caeai.2024.100265

Martin, P. P., & Graulich, N. (2024). Beyond language barriers: Allowing multiple languages in postsecondary chemistry classes through multilingual machine learning. *Journal of Science Education and Technology*, *33*(3), 333-348. https://doi.org/10.1007/s10956-023-10087-4

Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2024). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*, *61*(8), 1757-1792. https://doi.org/10.1002/tea.21903

Martin, P. P., & Graulich, N. (2023). When a machine detects student reasoning: A review of machine learning-based formative assessment of mechanistic reasoning. *Chemistry Education Research and Practice*, *24*(2), 407-427. https://doi.org/10.1039/D2RP00287F

Papers submitted to peer-reviewed journals

Martin, P. P., Kubsch, M., Yik, B. J., Burlingham, B. T., & Graulich, N. (under review). Tracing trajectories: Using machine learning to map how chemistry students' mechanistic reasoning develops over time. *Journal of the Learning Sciences.*

Martin, P. P., Kubsch, M., Yik, B. J., Burlingham, B. T., & Graulich, N. (under review). Adaptive, but equitable? Exploring the impact of machine learning-based adaptive support on educational debt in undergraduate chemistry. *Science Education.*

Martin, P. P., Dood, A. J, Watts, F. M., Shultz, G. V., & Graulich, N. (under review). To what extent can we trust a machine? Establishing trustworthiness guidelines to evaluate supervised machine learning models across chemistry contexts. *International Journal of Artificial Intelligence in Education.*

Kranz, D., Martin, P. P., Schween, M., & Graulich, N. (under review). Should we scaffold it? – Analysing the effect of task format and scaffolding on students' learning gain. *Chemistry Education Research and Practice.*

Papers in non-peer-reviewed journals

Braun, I., Martin, P. P., & Graulich, N. (2025). Von Biokunststoffen bis zum Verbrennungsmotor: Wie FERMI-Aufgaben Bewertungsprozesse im Chemieunterricht anstoßen können. *Unterricht Chemie, 205*, 36-39.

Martin, P. P., & Graulich, N. (2024). Lehre in der organischen Chemie individualisieren. *Nachrichten aus der Chemie*, *72*(3), 8-11. https://doi.org/10.1002/nadc.20244141003

Book chapter

Martin, P. P., Yik, B. J., Burlingham, B. T., & Graulich, N. (accepted). Mechanistic reasoning about alternatives: How students reason about plausible and implausible mechanisms in chemistry. In R. W. Bachtiar, M. Haskel-Ittah, & N. Graulich (Eds.), *Mechanistic Reasoning in Science Education.* Springer.

Conference proceedings

Martin, P. P., Kubsch, M., Yik, B. J., Burlingham, B. T., & Graulich, N. (2025). Modellierung von Lerntrajektorien beim KI-gestützten adaptiven Lernen in der Organischen Chemie. In H. van Vorst (Ed.), *Entdecken, lehren und forschen im Schülerlabor: Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung Bochum*. GDCP.

Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2024). Tiefgreifende Analyse von Argumenten in der Organischen Chemie mit maschinellem Lernen. In H. van Vorst (Ed.), *Frühe naturwissenschaftliche Bildung: Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung Hamburg* (pp. 74–77). GDCP.

Kranz, D., Martin, P. P., Schween, M., & Graulich, N. (2024). Analyse des Effekts verschiedener Aufgabenformate und Strukturierung auf den Lernerfolg. In H. van Vorst (Ed.), *Frühe naturwissenschaftliche Bildung: Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung Hamburg* (pp. 366–369). GDCP.

Martin, P. P., & Graulich, N. (2024). Mit Computerintelligenz Lehr-Lern-Prozesse gestalten: Anwendung von ChatGPT zur Unterrichtsplanung. In I. Rubner, T. Wilke, S. Syskowski, D. Ditter, N. ter Horst, A. Tschiersch, & D. Weiser (Eds.), *DiCE 2023 Conference Proceedings* (pp. 1–5). Thüringer Universitäts- und Landesbibliothek Jena. https://doi.org/10.22032/dbt.59418

Oral presentations

Martin, P. P., Kubsch, M., Yik, B. J., Burlingham, B. T., & Graulich, N. (2025, February 20–21). Who is supposed to benefit from adaptive support? *Austauschtreffen Chemie- und Physikdidaktiken Gießen und Berlin*, Berlin, Germany.

Martin, P. P. (2025, February 12). Insights into chemistry education research: Applying AI to analyze students' learning in organic chemistry. *FCI-Stipendiatentreffen*, Mainz, Germany.

Martin, P. P., & Graulich, N. (2024, September 18–20). Sprachbarrieren überwinden: Mit künstlicher Intelligenz Sprachenvielfalt im Chemieunterricht ermöglichen. *40. Fortbildungs- und Vortragstagung der GDCh-Fachgruppe Chemieunterricht: Praxis und Forschung kollegial reflektieren*, Regensburg, Germany.

Martin, P. P., & Graulich, N. (2024, September 12–14). Wenn Maschinen Denkprozesse analysieren: Anwendung von maschinellem Lernen in chemiedidaktischer Forschung und Praxis. *ROChET-Netzwerktreffen*, Essen, Germany.

Martin, P. P., Kubsch, M., Yik, B. J., Burlingham, B. T., & Graulich, N. (2024, September 9–12). Modellierung von Lerntrajektorien beim KI-gestützten adaptiven Lernen in der Organischen Chemie. *GDCP-Jahrestagung 2024: Lernen, lehren und forschen im Schülerlabor*, Bochum, Germany.

Martin, P. P., & Graulich, N. (2024, July 28–August 1). Beyond language barriers: Allowing multiple languages in postsecondary chemistry classes through multilingual machine learning. *Biennial Conference on Chemical Education 2024: Distilling Solutions for Chemical Education*, Lexington, KY, United States.

Martin, P. P., Kubsch, M., Yik, B. J., Burlingham, B. T., & Graulich, N. (2024, July 28–August 1). Mapping chemistry students' learning paths when reasoning about reaction mechanisms. *Biennial Conference on Chemical Education 2024: Distilling Solutions for Chemical Education*, Lexington, KY, United States.

Martin, P. P., & Graulich, N. (2024, May 23). Beyond language barriers: Allowing multiple languages in postsecondary chemistry classes through multilingual machine learning. *Symposium of the RSC Chemistry Education Research Group on AI in Chemistry Education Research* [Online].

Martin, P. P., & Graulich, N. (2023, November 27–28). Adaptive Langzeitförderung von Argumentationskompetenzen in der Organischen Chemie durch maschinelles Lernen. *Tagung der Joachim-Herz-Stiftung: Adaptives Lernen und KI in der schulischen und beruflichen Bildung: Potenziale und Herausforderungen technologiegestützten Lehrens und Lernens*, Hamburg, Germany.

Martin, P. P., & Graulich, N. (2023, September 18–19). Data-driven analysis of mechanistic reasoning in chemistry. *SOMER-Symposium: Symposium on Mechanistic Reasoning* [Online].

Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2023, September 11–14). Neue Tiefen entdecken: Analyse von Argumenten in der Organischen Chemie. *GDCP-Jahrestagung 2023: Frühe naturwissenschaftliche Bildung*, Hamburg, Germany.

Martin, P. P., & Graulich, N. (2023, March 23–24). Mit Computerintelligenz Lehr-Lern Prozesse gestalten: eine praxisorientierte Einführung in die Anwendung von maschinellem Lernen. *DiCE-Tagung 2023: Digitalization in Chemistry Education* [Online].

Poster presentations

Martin, P. P., Kubsch, M., Yik, B. J., Burlingham, B. T., & Graulich, N. (2024, September 18–20). Analyse von Lernverläufen beim KI-gestützten adaptiven Lernen in der Organischen Chemie. *40. Fortbildungs- und Vortragstagung der GDCh-Fachgruppe Chemieunterricht: Praxis und Forschung kollegial reflektieren*, Regensburg, Germany.

Martin, P. P., & Graulich, N. (2023, September 4–6). Argumentationstraining in der Organischen Chemie durch maschinelles Lernen adaptiv gestalten. *GDCh-Wissenschaftsforum 2023: Rethinking Chemistry*, Leipzig, Germany.

Kranz, D., Martin, P. P., Schween, M., & Graulich, N. (2023, September 4–6). Wie viel Struktur ist notwendig? Analyse des Effekts verschiedener Aufgabenformate und Strukturierungen auf den Lernerfolg. *GDCh-Wissenschaftsforum 2023: Rethinking Chemistry*, Leipzig, Germany.

Martin, P. P., & Graulich, N. (2023, July 9–14). Exploring new depths: Applying deep learning for the analysis of student argumentation. *Gordon Research Conference on Chemistry Education Research and Practice: Coordinating the Production and Consumption of Knowledge on Chemistry Teaching and Learning*, Lewiston, ME, United States.

Martin, P. P., & Graulich, N. (2023, March 2–3). Adaptives Lernen trifft maschinelles Lernen: Fachdidaktisches Potenzial von maschinellem Lernen. *Austauschtreffen Chemiedidaktiken Gießen und Münster*, Münster, Germany.

Kranz, D., Martin, P. P., Schween, M., & Graulich, N. (2022, September 15–17). Bedeutsame Kontraste: Untersuchung des Potenzials von Aufgabenformaten zur Förderung des mechanistischen Denkens von Lernenden in der Organischen Chemie. *38. Fortbildungs- und Vortragstagung der GDCh-Fachgruppe Chemieunterricht: Chemie nachhaltig kommunizieren*, Osnabrück, Germany.

Martin, P. P., & Graulich, N. (2022, September 15–17). Nutzung von Machine Learning für formatives Feedback. *38. Fortbildungs- und Vortragstagung der GDCh-Fachgruppe Chemieunterricht: Chemie nachhaltig kommunizieren*, Osnabrück, Germany.

Kranz, D., Martin, P. P., Schween, M., & Graulich, N. (2022, September 12–15). Bedeutsame Kontraste: Untersuchung des Potenzials von Aufgabenformaten zur Förderung des mechanistischen Denkens von Lernenden in der Organischen Chemie. *GDCP-Jahrestagung 2022: Lernen, Lehren und Forschen in einer digital geprägten Welt*, Aachen, Germany.

1 | Synopsis: Beyond a snapshot—Using machine learning to monitor and adaptively support organic chemistry students' mechanistic reasoning over time

1.1 | Introduction

An essential goal of science education is to enable students to engage with the natural and engineered world through science practices, such as developing models, constructing explanations, or building arguments (National Research Council, 2012). Mechanistic reasoning, which is required in some science practices, contributes to this goal as it helps students understand how and why natural phenomena occur (Craver, 2007; Illari & Williamson, 2012; Machamer et al., 2000). Mechanistic reasoning involves identifying, describing, or analyzing the underlying and often unobservable mechanisms that drive chemical, biological, and physical phenomena to explain, predict, or control their outcomes (Craver, 2006; Haskel-Ittah, 2023). By decoding the sequence and causality of mechanisms, mechanistic reasoning provides a tool for understanding the complexities inherent to the natural and engineered world (Russ et al., 2008).

Mechanistic reasoning is a complex practice (Craver, 2007; Illari & Williamson, 2012). It requires students to describe the target phenomenon, assess the impact of setup conditions, identify relevant entities and their properties, and analyze how these entities interact in activities to produce the observed phenomenon (Russ et al., 2008). Additionally, mechanistic reasoning involves chaining a mechanism's steps causally to make inferences about past activities that led to current entities or to predict future activities based on the presence of certain entities (Caspari, Weinrich, et al., 2018; Darden & Craver, 2002; Russ et al., 2008).

Due to these complexities, students often face challenges in mechanistic reasoning across science domains. These challenges include identifying the necessary entities required for a causal explanation, determining their properties, and explaining how interactions of entities lead to specific activities (Bachtiar et al., 2022). For instance, students frequently rely on surface-level properties to explain phenomena, rather than examining implicit properties and interactions (e.g., Bhattacharyya & Bodner, 2005; Cooper et al., 2013; Ferguson & Bodner, 2008; Graulich et al., 2019; Rotich et al., 2024). Accordingly, students describe mechanisms as linear sequences of isolated steps and neglect causal relationships among entities (e.g., Kraft et al., 2010; Perkins & Grotzer, 2005; Weinrich & Sevian, 2017). Similarly, Yan and Talanquer (2015) found that students struggle to construct causal explanations where activities arise from random interactions among entities. Besides, broader challenges in science learning also affect how students demonstrate mechanistic reasoning. For example, epistemic practices, such as the ability to apply scientific concepts or structure scientific arguments, can impact students' mechanistic reasoning (e.g., Choi et al., 2013; Duncan, 2007; Grove, Cooper, & Cox, 2012; Kranz et al., 2023; Pabuccu & Erduran, 2017).

To address these student challenges, a recent review on mechanistic reasoning in science education emphasizes the need for explicit instructional support (Bachtiar et al., 2022). Integrating mechanistic reasoning into science instruction is crucial (e.g., Scharlott et al., 2024). For example, instruction can support students' mechanistic reasoning by using case comparisons to assist students in examining implicit properties (Graulich & Schween, 2018), providing scaffolding to guide their responses' structure (Caspari & Graulich, 2019; Graulich & Caspari, 2021), or leveraging authentic data to highlight the connection between theoretical concepts and practical examples (Esselman et al., 2023; Fuhrmann et al., 2025). Also, it is suggested that adaptive formative assessments can

improve students' mechanistic reasoning (e.g., Ariely et al., 2024; Asmussen et al., 2023, 2024; Dood et al., 2020b; Lieber et al., 2022a, 2022b). However, implementing adaptive formative assessments to support students' mechanistic reasoning is resource-intensive because it requires evaluating large amounts of open-ended responses. This presents limitations when adaptive support for students' mechanistic reasoning is needed over time. Machine learning (ML) models are expected to bridge this gap by automatically analyzing students' mechanistic reasoning (Chapter 2). In summary, mechanistic reasoning is a complex practice that requires explicit support in science instruction, and ML shows promise in longitudinally monitoring and adaptively supporting students' mechanistic reasoning (Figure 1).

1.2 | Research Areas

Prior research in science education applied ML models to provide inductive, data-driven insights into students' mechanistic reasoning (e.g., Ariely et al., 2024; Haudek et al., 2015; Prevost et al., 2012; Watts et al., in print), evaluate students'

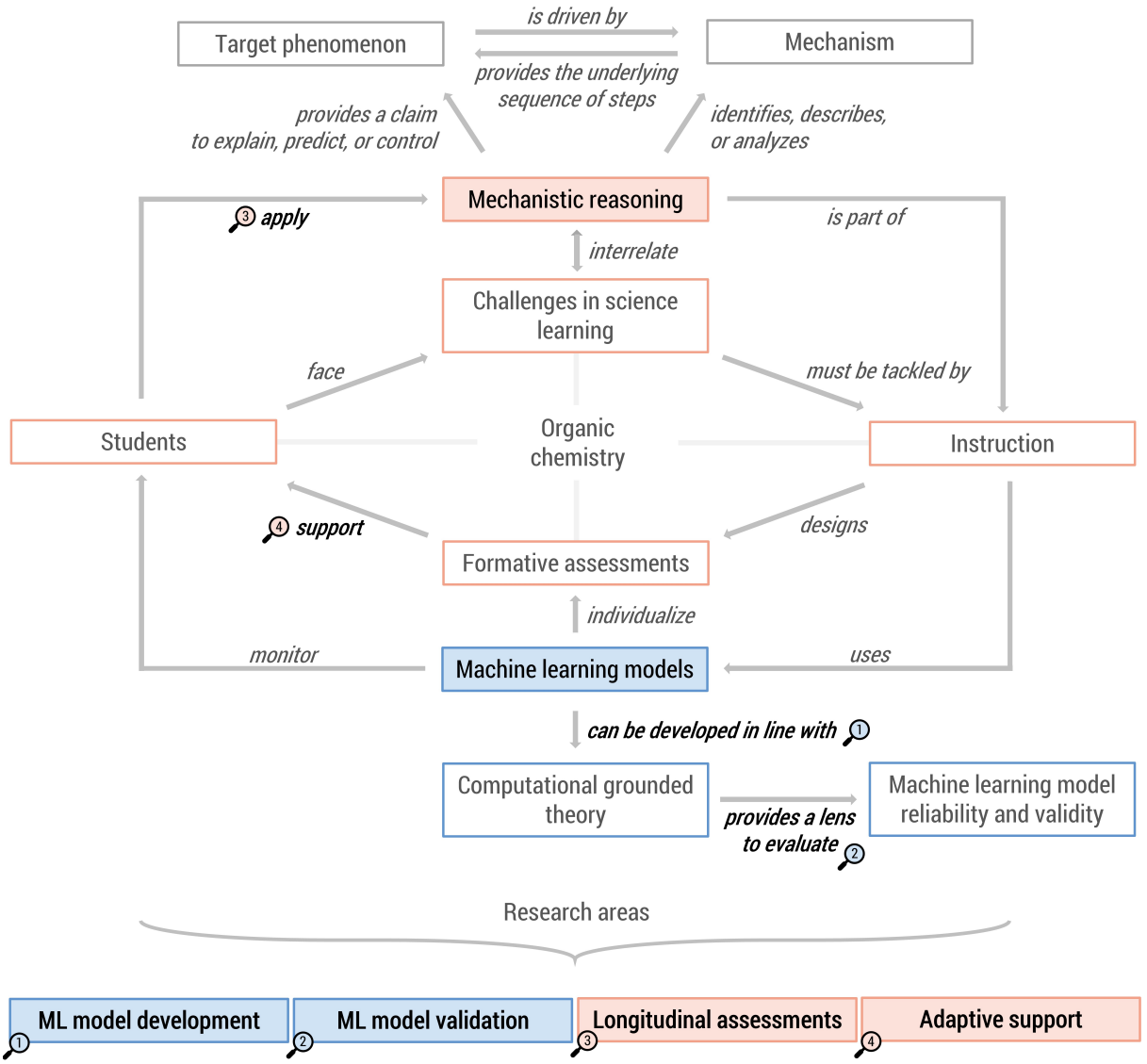


FIGURE 1 Overview of our approach to using machine learning for monitoring and adaptively supporting organic chemistry students' mechanistic reasoning. *Note:* Blue boxes represent a methodological focus and orange boxes represent a pedagogical focus. Lenses visualize our research areas.

mechanistic reasoning automatically (e.g., Dood et al., 2018, 2020a; Haudek et al., 2012; Noyes et al., 2020; Watts et al., 2022; Yik et al., 2021), implement adaptive support (e.g., Ariely et al., 2024; Dood et al., 2020b), and streamline data collection (e.g., Dood et al., 2019) (see Chapter 2 for a systematic review of ML-based formative assessment of students' mechanistic reasoning). Despite these promising ML applications, several research gaps must be addressed to monitor and adaptively support students' mechanistic reasoning over time. Our corresponding research areas can be broadly categorized into two main areas, each with two subcategories: i) the methodological application of ML, encompassing (1) ML model development and (2) ML model validation; and ii) the pedagogical implementation of ML, focusing on (3) longitudinal assessments and (4) longitudinal adaptive support (Figure 1). The following sections outline these research areas.

ML model development

Previous research has demonstrated that ML models can inductively analyze and automatically score students' mechanistic reasoning (Chapter 2). However, the achievable level of detail of these ML analyses remains unclear. For example, some studies used ML solely for binary scoring to determine whether students employ any mechanistic reasoning in their open-ended responses (e.g., Dood et al., 2018, 2019; Yik et al., 2021). Other studies increased the level of detail in automated ML analyses but still offer only coarse distinctions in students' application of mechanistic reasoning. For instance, these analyses solely determine whether students describe *what* happens or whether they provide surface-level versus deeper-level reasons for *why* a mechanism occurs (e.g., Dood et al., 2020b; Noyes et al., 2020). Due to these rather basic ML analyses, exploring how ML can be used for finer-grained scoring of mechanistic reasoning would help investigate the heterogeneity in students' mechanistic reasoning over time.

However, developing rubrics for finer-grained scoring of students' mechanistic reasoning is complex (Haudek et al., 2015), partly due to the substantial heterogeneity in mechanistic reasoning across contexts (Deng & Flynn, 2021; Watts et al., 2021). Following this, it is challenging to balance capturing detailed mechanistic reasoning patterns while maintaining applicability across contexts (Yik et al., 2021). Unsupervised ML techniques, such as clustering algorithms from the topic modeling family, have been valuable in science education research for analyzing students' open-ended reasoning inductively (e.g., Anderson et al., 2020; Sherin, 2013; Wulff et al., 2022; Wulff, Westphal, et al., 2023; Zehner et al., 2016). When carefully refined, their outputs can guide rubric development to capture the nuances of students' open-ended reasoning (Rosenberg & Krist, 2021; Sherin, 2013; Tschisgale et al., 2023). Such clustering algorithms have also been applied to analyze students' mechanistic reasoning (Ariely et al., 2024; Prevost et al., 2012; Watts et al., in print), and, among other uses, to inform rubric development (Haudek et al., 2015). The ML-based rubrics developed in previous research are context-specific and lack generalizability. Thus, investigating the use of clustering algorithms to develop a scoring rubric that reflects the heterogeneity in students' mechanistic reasoning across contexts would be beneficial.

Pre-trained, transformer-based deep neural networks for sequence classification with large language models for natural language processing—advanced ML models that can evaluate large amounts of human text—are promising in science education to perform automated analyses based on finer-grained rubrics (e.g., Dood et al., 2024; Gombert et al., 2023; Wulff, Mientus, et al., 2023). These advanced ML models hold the potential for evaluating mechanistic reasoning automatically with high accuracy to provide adaptive support

(e.g., Ariely et al., 2023, 2024). The use of these advanced ML models for evaluating students' mechanistic reasoning in chemistry has yet to be explored (Chapter 2).

Accurate ML analyses are eventually a prerequisite for reliably assigning adaptive support to students. Here, large language models like Bidirectional Encoder Representations from Transformers (BERT) were originally developed to evaluate open-ended responses in English (Devlin et al., 2018), which may compromise ML analyses in other languages. Therefore, some studies investigated the accuracy of large language models in different languages, for example, to analyze physics students' reflections in German (Wulff, Mientus, et al., 2023), physics students' application of energy concepts in German (Gombert et al., 2023), high-school science students' argumentation in Chinese (Wang et al., 2021), or biology students' explanations in Hebrew (Ariely et al., 2023). While these studies suggest that accurate ML analyses of chemistry students' mechanistic reasoning in other languages like German are feasible, it remains uncertain how well large language models can evaluate datasets containing open-ended responses that are written in different languages. Such multilingual ML analyses would allow students who learn the language of instruction as an additional language to complete assessments in their preferred language, thereby reducing linguistic barriers in science learning.

However, developing an ML model to evaluate students' open-ended reasoning automatically with high accuracy requires large amounts of student data, and collecting this data can be challenging due to resource constraints. Thus, producing synthetic data—artificially generated data that simulates authentic student responses—can improve ML model performance. Methods for producing synthetic data include using generative AI to paraphrase student responses or create novel samples (e.g., Cochran et al., 2022, 2023; Dai et al., 2023; Møller et al., 2023). While the potential for using data augmentation in a science education context is largely unexplored, it holds promise for improving ML model performance. Also, data augmentation may allow for finer-grained evaluations of students' open-ended reasoning and reduce human effort in data collection.

To address these research gaps, we explored how clustering algorithms can be leveraged to design a scoring rubric that captures the heterogeneity in students' mechanistic reasoning. Further, we developed transformer-based deep neural networks to automatically perform a fine-grained analysis of students' mechanistic reasoning (Chapter 3). In doing so, we investigated to what extent the developed ML model allows for evaluations in both English and German, as well as across these two languages, to enable students to reason in their preferred language in chemistry assessments (Chapter 4). Additionally, we explored how data augmentation can improve ML model performance and reduce the human effort required for data collection (Chapter 5).

ML model validation

So far, many studies applied well-established approaches, such as self-validation, split-validation, or cross-validation, to test ML model performance (Chapter 2). These approaches quantify machine-human score agreement to classify ML model performance, which is why additional testing is often necessary to evaluate the reliability and validity of ML models more extensively. For instance, Nehm et al. (2012) and Yik et al. (2021) demonstrated that their ML models exhibited strong generalizability across various formative assessment items. Likewise, other studies evidenced high generalizability of their ML models across courses, disciplines, or institutions (Dood et al., 2022; Noyes et al., 2020; Wulff, Westphal, et al., 2023). Although these findings suggest that ML models may be broadly applicable to evaluate students' mechanistic

reasoning, they do not address variations in ML model performance across chemistry mechanisms. This gap is significant because consistently evaluating students' mechanistic reasoning across various mechanisms is essential for implementing continuous and adaptive formative assessments with an automated ML analysis that upholds performance across chemical contexts.

Furthermore, Maestrales et al. (2021) and Yik et al. (2021) observed discrepancies between the words used by their traditional ML model to score responses and the criteria outlined in human scoring guidelines. Nonetheless, the scores assigned by an ML model matched students' responses in oral interviews (Beggrow et al., 2014; Haudek et al., 2012). These findings suggest that ML models can correctly score students' open-ended reasoning, although they may rely on different scoring criteria than humans. Whether this finding extends to more advanced ML models analyzing mechanistic reasoning is unexplored. Also, extracting the words used in machine scoring can help identify shortcuts where the ML model relied on single keywords, which may enable refining human and machine scoring guidelines to better capture the semantics of students' mechanistic reasoning. Additionally, investigating whether advanced ML models align with human scoring guidelines could foster appropriate trust in ML methodology and prevent unexpected results when it is employed over time. From a broader perspective, examining the relationship between human and machine scoring guidelines is essential for determining the extent to which ML can be integrated into traditional assessment design.

To examine ML model reliability and validity, we investigated the applicability of ML models across various reaction mechanisms, task formats, and institutions, and examined the extent to which their outputs can be transparently explained (Chapter 6). Moreover, we explored whether an ML-based rubric aligns with a research-informed counterpart (Chapter 7). These analyses provide researchers and educators with tools to assess the reliability and validity of ML models to foster appropriate trust in ML methodologies and better anticipate ML model outputs.

Longitudinal assessment

A reliable and valid ML analysis can ultimately assess students' mechanistic reasoning longitudinally. Here, previous research, particularly in chemistry education, explored the heterogeneity in students' mechanistic reasoning but provided limited insights into its development over time due to several constraints. Some studies examined students' mechanistic reasoning longitudinally across multiple semesters but focused only on specific topics, such as acid–base reactions (Crandell et al., 2019), intermolecular forces (Noyes & Cooper, 2019), or nucleophilic substitutions (Crandell et al., 2020). Other studies faced limitations, such as a small number of assessments (Lieber et al., 2022a), small sample sizes (Watts et al., 2021; Weinrich & Sevian, 2017), or cross-sectional designs (Franovic et al., 2023; Shiroda et al., 2024). As a result, our understanding of how students' mechanistic reasoning evolves and interacts with other variables remains limited.

Understanding how students' mechanistic reasoning is influenced by other variables over time is important to improve future support. For instance, research suggests that students' mechanistic reasoning is associated with their chemical concept knowledge in tasks requiring knowledge transfer, such as predicting reaction products in organic chemistry (Grove, Cooper, & Cox, 2012; Houchlei et al., 2021). Conversely, differences in chemical concept knowledge between students employing mechanistic reasoning and those relying on factual recall are not observed on easier tasks (Grove, Cooper, & Cox, 2012; Houchlei et al., 2021). These findings

underscore inconsistencies in the relationship between chemical concept knowledge and mechanistic reasoning. In addition to domain-specific knowledge, domain-general skills like scientific argumentation are hypothesized to advance mechanistic reasoning (e.g., Choi et al., 2013; Moon et al., 2016; Pabuccu & Erduran, 2017). However, evidence that these skills suffice for complex mechanistic reasoning is limited.

Mechanistic reasoning is ultimately intended to facilitate claims about target phenomena (Haskel-Ittah, 2023; Russ et al., 2008). However, evidence on the relationship between students' mechanistic reasoning and the correctness of their claims is inconclusive. Some studies report that students making correct and incorrect claims exhibit comparable mechanistic reasoning (Bodé et al., 2019), while others identified that students with correct claims demonstrate more complex mechanistic reasoning (Deng & Flynn, 2021). This inconclusive relationship underscores the need for further investigation. From a broader perspective, affective variables may also influence students' mechanistic reasoning; for example, a positive self-concept—the belief in one's ability to succeed—has been linked to productive science learning (Nieswandt, 2007). Moreover, students exhibiting complex mechanistic reasoning reported higher confidence (Gao et al., 2024).

To gain a holistic understanding of how these factors interrelate and contribute to the development of mechanistic reasoning, we modeled their interplay using ML to improve long-term adaptive support. We analyzed the longitudinal development of students' mechanistic reasoning across contexts in undergraduate organic chemistry to examine the dynamic interplay of its components (Chapter 9). Particular attention was given to the relationship between mechanistic reasoning and students' ability to correctly judge a mechanism's plausibility (Chapter 8). These insights can inform instructional strategies to better guide students' mechanistic reasoning.

Longitudinal adaptive support

Finally, research demonstrates that adaptive support can significantly improve students' mechanistic reasoning (e.g., Ariely et al., 2024; Asmussen et al., 2024; Dood et al., 2020b; Hedtrich & Graulich, 2018a, 2018b; Lieber et al., 2022a). Consequently, leveraging ML to adaptively support students' mechanistic reasoning appears effective. However, implementing adaptive support involves a tension between *convergent adaptivity*, which aims to close achievement gaps, and *divergent adaptivity*, which enables all students to progress at their own pace, thus, allowing achievement gaps to persist (Deunk et al., 2015, 2018; Smale-Jacobse et al., 2019). These different adaptivity aims underscore that implementing ML-assisted adaptive support is a complex task that requires continuous efforts to strike a balance in supporting all students.

Previous studies applied primarily divergent adaptivity, facilitated through ML, to support students' mechanistic reasoning (Ariely et al., 2024; Dood et al., 2020b). While these studies suggest—in line with the aims of divergent adaptivity—that all students could progress at their own pace, they did not explore how students' mechanistic reasoning diverges over time. Accordingly, the extent to which students of different achievement levels benefit from divergent adaptivity remains unclear. A longitudinal application of divergent adaptivity could reveal how mechanistic reasoning develops among students with varying achievement levels (Chapter 10). This could inform the design of future adaptive support to better meet the needs of all students.

In summary, several research areas emerge regarding the use of ML models to monitor and adaptively support students' mechanistic reasoning. These areas can be categorized into the methodological application and pedagogical implementation of ML related to students' mechanistic reasoning. Through this two-fold

approach, this dissertation aims to advance the methodological rigor and pedagogical utility of ML in supporting students' mechanistic reasoning (Figure 1).

1.3 | Theoretical framework: Essential epistemic heuristics for mechanistic reasoning in chemistry

Developing and validating ML models to automatically score students' mechanistic reasoning and monitor its progression over time necessitates hypotheses about the underlying structure of mechanistic reasoning. Several frameworks have been designed to capture students' mechanistic reasoning in organic chemistry (Dood & Watts, 2022), with some frameworks tailored for ML model development (e.g., Dood et al., 2020a; Raker et al., 2022). However, we chose not to adopt these frameworks for specific reasons. For example, the framework by Cooper et al. (2016) distinguishes between *mechanistic*, *causal*, and *causal-mechanistic* reasoning. While this distinction can be insightful, *mechanistic* and *causal* reasoning are often intertwined in complex chemical contexts, which requires students to employ elements of *causal-mechanistic* reasoning to explain the outcome of chemical reactions (Caspari, Kranz, et al., 2018; Talanquer, 2018; Yan & Talanquer, 2015). Similarly, the categories of *rule-based*, *case-based*, and *model-based* reasoning by Kraft et al. (2010) are conceptually rich but may be too abstract to guide adaptive support for students. A more component-based framework, such as the ones by Machamer et al. (2000) and Russ et al. (2008), can offer high accuracy in ML analysis (Watts et al., 2022); however, their binary evaluation of categories like *identifying entities* or *identifying activities* limits their utility for long-term support since it becomes difficult to assist students further who have already identified these factors in their mechanistic reasoning.

The frameworks specifically developed for ML analysis broadly categorize students' mechanistic reasoning into three or four categories (e.g., Dood et al., 2020a; Raker et al., 2022). For instance, Raker et al.'s (2022) framework consists of four hierarchically organized categories: *absent*, *descriptive*, *foundational*, and *complex* mechanistic reasoning. These categories are designed to assess students' mechanistic reasoning across different mechanisms, potentially enabling the development of more generalizable ML models. Additionally, the framework intentionally minimizes jargon, making it easier for individuals unfamiliar with mechanistic reasoning terminology to score responses. This may increase human-human inter-rater reliability and machine-human score agreement. However, the broad categorizations may lack the specificity needed to precisely identify students' strengths and challenges in adaptive support.

To overcome these limitations, we adapted Krist et al.'s (2019) *essential epistemic heuristics for mechanistic reasoning* to the context of undergraduate organic chemistry (Figure 2). These three heuristics offer a structured framework for categorizing the steps involved in mechanistic reasoning, which may enable a fine-grained ML analysis and the implementation of long-term adaptive support. The first heuristic emphasizes reasoning across scalar levels, particularly considering the scalar level below the target phenomenon. The scalar levels at which phenomena and their corresponding mechanisms occur differ across science domains, and the concept of granularity serves to identify these levels (Darden, 2002). At the lower scalar level, students must identify and unpack entities, activities, and their properties—the second heuristic. Identifying these factors relies on a solid understanding of disciplinary core ideas (Duncan, 2007). We use the term *disciplinary core ideas* to specify the entities and activities students integrated into their mechanistic reasoning; by contrast, *chemical concept knowledge* refers in our analyses to how students establish relationships among properties in tasks

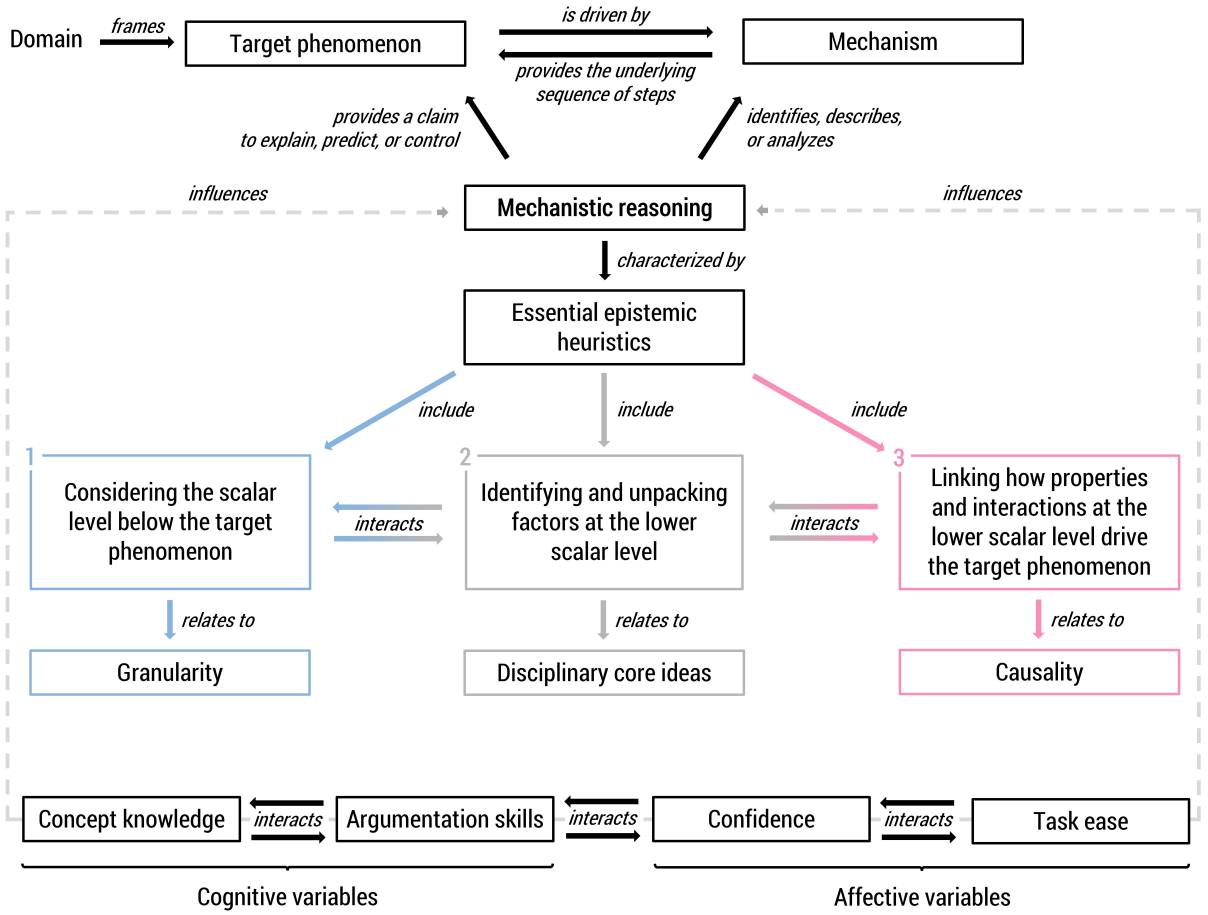


FIGURE 2 Using the *essential epistemic heuristics* adapted from Krist et al. (2019) to operationalize mechanistic reasoning and its interdependencies in science education.

that do not necessarily require mechanistic reasoning. Finally, the third heuristic requires arranging unpacked properties into a causal sequence to explain the observed phenomenon. The concept of causality evaluates how students link entities, activities, and their properties (Grotzer, 2003; Haskel-Ittah, 2024; Sevian & Talanquer, 2014).

Given this operationalization, mechanistic reasoning enables students to identify, describe, or analyze mechanisms to formulate claims that explain, predict, or control a domain-specific target phenomenon (Glennan & Illari, 2017; Russ et al., 2008). To support students' mechanistic reasoning, other variables may be critical: cognitive variables, such as chemical concept knowledge and domain-general argumentation skills, and affective variables, such as perceived confidence and task ease (Graulich, 2025). Figure 2 illustrates an overview of mechanistic reasoning and its interconnections within a broader context as elaborated in the following sections.

First essential epistemic heuristic: Reasoning across scalar levels

The first *essential epistemic heuristic* emphasizes reasoning at a scalar level below the target phenomenon, which allows students to uncover the underlying mechanism of that phenomenon (Krist et al., 2019). The specific scalar levels that must be addressed to reason about the underlying mechanism depend on the domain, context, and task. Granularity serves as a concept for assessing scalar levels in mechanistic reasoning (Darden, 2002; Deng & Flynn, 2021; Southard et al., 2017; van Mil et al., 2013). For example, scalar levels in chemistry include reactions, molecules, atoms, and electrons. In

contrast, scalar levels in molecular biology address biological systems, chemical reactions, and molecules. Some phenomena may demand more macroscopic perspectives ranging from global to organismal levels, others may require more microscopic perspectives detailing subatomic levels (Darden, 2002).

In chemical contexts, heterogeneity in students' use of granularity is apparent even within the same task (Deng & Flynn, 2021). For instance, students analyzing an acid–base reaction may focus on the scalar levels of reactions and molecules. Such reasoning may incorporate measurable properties, like equilibrium direction, pK_a values, or conjugate acid strength, or properties that can be explicitly derived from structural features, like the presence of a hydrogen atom that could potentially be abstracted. These properties of reactions and molecules reflect a *phenomenological* granularity in chemistry since they lie at the level of the target phenomenon or can be explicitly derived from structural features. By contrast, reasoning that integrates electronic properties like electronegativity or electron delocalization demonstrates *electronic* granularity as these properties pertain to the scalar levels of atoms and electrons, which lie below the target phenomenon. Examining the development of students' use of granularity in similar contexts could guide the design of adaptive support.

Second essential epistemic heuristic: Identifying and unpacking relevant factors

The second *essential epistemic heuristic* involves identifying and unpacking relevant entities, activities, and their properties at the lower scalar level (Krist et al., 2019). As the entities and activities constituting a mechanism are domain-specific, disciplinary core ideas can be used to capture students' mechanistic reasoning through this heuristic (Cooper et al., 2017; Duncan, 2007; Haskel-Ittah & Duncan, 2021; Haskel-Ittah et al., 2020).

In chemical contexts, *structural* features, such as atom types, atom connectivity, molecular shape, bonding patterns, or functional groups, primarily influence the properties of entities (Goodwin, 2008). Disciplinary core ideas derived from *structural* features can include *charged species and substrates*, *nucleophiles and electrophiles*, and *acids and bases*. Furthermore, *structural* features can also be used as models to explain energy differences (Goodwin, 2008). These energy differences can inform claims about the plausibility of mechanisms in chemistry. Consequently, *structural* features govern *energetic* features, such as equilibrium direction or reaction rate, which, in turn, influence the properties of activities (Goodwin, 2008). Disciplinary core ideas derived from these *energetic* features could encompass *thermodynamics and kinetics*. Identifying the disciplinary core ideas that advance students' mechanistic reasoning is essential for adaptive support.

Third essential epistemic heuristic: Linking properties to phenomena

After identifying and unpacking entities, activities, and their properties, these factors must be organized into a causal sequence over time or space (Krist et al., 2019). The third *essential epistemic heuristic* focuses on how students link the factors identified at a scalar level below the target phenomenon back to the original scalar level. Thus, this heuristic builds upon students' mechanistic reasoning demonstrated in the first two heuristics and emphasizes their integration. Causality can serve as a framework for evaluating how students connect the identified properties to explain, predict, or control the target phenomenon (Haskel-Ittah, 2024; Sevian & Talanquer, 2014).

Students' use of causality in mechanistic reasoning varies across domains (Grotzer, 2003; Russ et al., 2008). Some students simply list properties without making connections,

reflecting *descriptive* reasoning. Others provide surface-level explanations for why entities or activities possess certain properties, corresponding to *relational* reasoning. Still, others identify cause-and-effect relationships in a step-by-step sequence, where one cause directly leads to a single effect, demonstrating *linear causal* reasoning. In contrast, others consider multiple interacting entities within feedback loops, indicating *multicomponent causal* reasoning (e.g., Moon et al., 2016; Moreira et al., 2019; Weinrich & Talanquer, 2016). Longitudinal studies could reveal whether students advance their causality across contexts or encounter context-specific challenges.

In sum, we applied the *essential epistemic heuristics* to characterize students' mechanistic reasoning (Krist et al., 2019). The complexity of students' mechanistic reasoning can vary along these heuristics. For instance, some students may identify properties at the level of the target phenomenon while others may analyze properties at a scalar level below it. Similarly, some students may list unpacked properties while others may establish cause-and-effect relationships. Our operationalization of these heuristics provides a three-dimensional continuum for assessing the complexity of students' mechanistic reasoning regarding their use of granularity, disciplinary core ideas, and causality. The interactions between these heuristics remain largely unexplored. For instance, prior studies examined the granularity and causality of students' mechanistic reasoning in isolation (Bodé et al., 2019; Deng & Flynn, 2021). Likewise, there is limited clarity regarding which disciplinary core ideas support students' use of granularity or causality (Haskel-Ittah & Duncan, 2021).

1.4 | Methodological Framework: Computational Grounded Theory

To develop an ML model that can monitor students' mechanistic reasoning in line with the *essential epistemic heuristics*, we applied Computational Grounded Theory (CGT). CGT is a methodological framework that integrates computational tools with grounded theory's established qualitative approach to inductively generate theories about social or cognitive processes (Nelson, 2020). This synergy is particularly suited to analyzing large and unstructured datasets where manual coding alone would be infeasible. CGT seeks to increase the reproducibility of generated hypotheses, identify biases in inductive data analysis, and reduce the influence of subjective assumptions on data interpretation (Carlsen & Ralund, 2022; Nelson, 2020), aligning with recommendations to include domain-specific theoretical frameworks into ML outputs (Kubsch et al., 2023; Sherin, 2013; Zhai, 2021).

The revisited CGT framework unfolds through four interrelated steps to gain breadth and depth in data analysis: pattern detection, pattern refinement, pattern confirmation, and pattern transparency (Figure 3) (Nelson, 2020). These steps facilitate the identification of patterns in qualitative data, the inductive generation of fine-grained theories about social or cognitive processes, and the deductive verification of

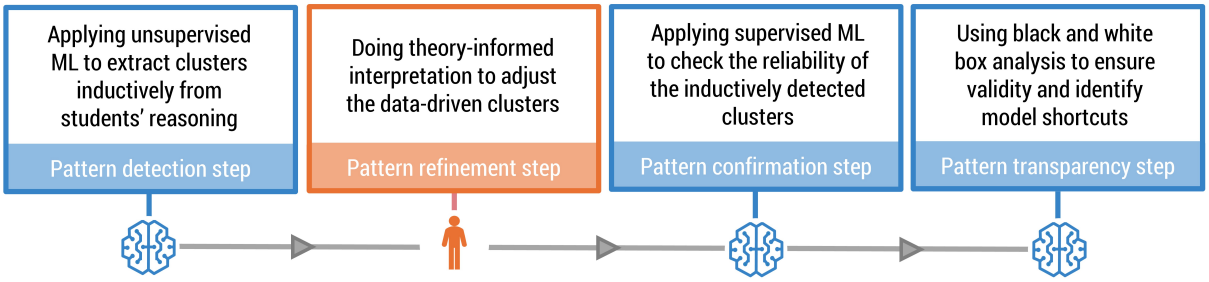


FIGURE 3 Revisited four-step computational grounded theory framework.

these theories using quantitative methods. The following paragraphs outline the steps of CGT, drawing on Nelson's (2020) discussion and our refinements as presented in Chapters 3, 6, and 7.

Pattern detection step

The first step in CGT focuses on systematically identifying patterns within large and unstructured datasets. Computational tools, such as clustering algorithms or natural language processing, can be applied to uncover recurring topics in the data. Hence, this step transforms content-rich qualitative data into preliminary categories to provide a foundation for subsequent analysis. The goal is to generate a broad overview of potential patterns without imposing preconceived hypotheses on the data, which may ensure that findings remain grounded in the data.

Pattern refinement step

In the pattern refinement step, researchers engage in an iterative process to adjust and validate the computationally detected patterns. This step uses domain expertise and qualitative coding to interpret the data-driven patterns within the research context. Beyond that, researchers may iteratively revise computational parameters, such as adjusting clustering thresholds or reconfiguring topic models, to better align a computational output with the nuances of the data. This step bridges the gap between raw computational outputs and domain-specific theoretical frameworks.

Pattern confirmation step

The pattern confirmation step focuses on the reliability and generalizability of the refined patterns. For example, supervised ML can validate refined patterns by training an ML model on datasets coded according to these patterns and testing its accuracy in identifying the patterns in novel data. By doing so, statistical measures, including machine-human score agreement, the degradation in agreement compared to human-human inter-rater reliability, or standardized mean score differences, can be used to investigate pattern reliability. Also, triangulation techniques, such as cross-validation with other datasets or populations, can assess the generalizability of the patterns. Moreover, statistical tests can confirm patterns in quantitative research. In sum, this step ensures that the patterns are not artifacts of specific computational techniques but reflect the underlying construct.

Pattern transparency step

The final step, pattern transparency, emphasizes the importance of making the analysis transparent to humans. Above all, researchers should document and share the computational workflows, scoring rubrics, and methodological decisions that shaped the analysis. Furthermore, the pattern transparency step explores the interpretability or explainability of an ML model's decision-making to enable humans to understand how the model arrives at its outputs. Explainable AI—tools that help explain the output of AI models—can increase ML transparency to foster trust in ML methodology and allow other researchers to replicate or build upon the work. Ultimately, the pattern transparency step can provide recommendations to refine the conceptualization of a construct within computational analyses and beyond.

By integrating these four steps, CGT provides a systematic yet flexible framework for applying computational techniques while maintaining the interpretive rigor of grounded theory. Its iterative nature ensures that findings are both data-driven and theoretically grounded, which offers a useful approach to understanding cognitive processes through inductive theory-generating research.

1.5 | Research objectives

Mechanistic reasoning is vital across science domains because it enables students to engage with the natural and engineered world by explaining, predicting, or controlling phenomena (Haskel-Ittah, 2023). However, its complex nature often presents challenges, resulting in substantial heterogeneity in how students apply mechanistic reasoning within and across cohorts (e.g., Crandell et al., 2019). Prior research analyzed students' mechanistic reasoning but remains fragmented regarding how students demonstrate mechanistic reasoning across contexts. This dissertation seeks to address this gap by unpacking the heterogeneity of students' mechanistic reasoning across task formats and mechanisms using unsupervised ML techniques, such as clustering algorithms. Students' heterogeneity in mechanistic reasoning subsequently informed the development of a BERT-based supervised ML model capable of automatically scoring mechanistic reasoning across mechanisms in organic chemistry. We aimed to offer a fine-grained analysis of students' mechanistic reasoning to allow for more precise longitudinal assessments and more tailored adaptive support.

With this automated analysis, monitoring and adaptively supporting students' mechanistic reasoning over an extended period was possible. Nonetheless, how to best design longitudinal support remains an open question. A deeper understanding of how mechanistic reasoning develops and interacts with other variables under adaptive support can inform future instructional practices. Thus, this dissertation examines the development of students' mechanistic reasoning over time and its relation to other variables to improve long-term adaptive support of mechanistic reasoning.

This dissertation is structured around eight hypotheses aligned with the four steps of CGT, each addressing a methodological and pedagogical focus. The subsequent sections—data analysis, results, and discussion—are organized around these hypotheses and CGT's steps, with results presented separately for the four research areas outlined in Figure 1: ML model development, ML model validation, longitudinal assessment, and longitudinal adaptive support.

Pattern detection step

1a. Methodological focus: Clustering algorithms can detect substantial heterogeneity in the extent to which undergraduate organic chemistry students demonstrate mechanistic reasoning.

1b. Pedagogical focus: Students integrate diverse disciplinary core ideas, interconnected at varying levels of granularity and causality, into their mechanistic reasoning.

Pattern refinement step

2a. Methodological focus: Data-driven ML analyses are associated with research-informed theoretical frameworks. This enables theoretical frameworks to refine ML outputs.

2b. Pedagogical focus: By refining ML outputs with theoretical frameworks, developing a broadly applicable rubric to score students' mechanistic reasoning across various task formats is possible.

Pattern confirmation step

3a. Methodological focus: Sufficient training data allows for accurate automated analysis of fine-grained mechanistic reasoning patterns while data augmentation improves ML model performance in underrepresented categories.

3b. Pedagogical focus: Students' mechanistic reasoning improves over time through adaptive support and is influenced by cognitive and affective variables, such as chemical concept

knowledge, scientific argumentation skills, perceived confidence, and perceived task ease.

Pattern transparency step

4a. Methodological focus: Accurate ML models mimic human scoring guidelines in evaluating mechanistic reasoning.

4b. Pedagogical focus: The *essential epistemic heuristics for mechanistic reasoning* are interrelated in undergraduate organic chemistry, facilitating the conceptualization of mechanistic reasoning in future analyses.

1.6 | Study Context

To monitor and adaptively support students' mechanistic reasoning over time, five objectives structured the research for this dissertation (Figure 4). Objective 1 focused on applying clustering algorithms, which are unsupervised ML models, to detect patterns in students' mechanistic reasoning. Additionally, it involved developing scoring models, which are supervised ML models, to automatically evaluate students' mechanistic reasoning based on data from previous research projects. Objective 2 involved designing task sets related to enamine hydrolysis and aldol condensation reactions—drawing from Lieber and Graulich's (2020) task format for ether synthesis and Claisen condensation reactions—that prompted students to judge the plausibility of alternative reaction products. Also, Objective 2 included creating adaptive scaffolds to guide students' mechanistic reasoning by helping them identify relevant chemical concepts within a mechanism and structure scientific arguments.

For Objective 3, these learning resources were piloted (N = 80) at a large undergraduate-only university in Western Canada. Following this pilot study, the learning resources were refined and implemented for Objective 4, which involved data collection with a larger cohort (N = 266) recruited from an Organic Chemistry II course at a public, research-intensive university in the Midwestern United States. Data collection included using a diagnostic scaffold to analyze

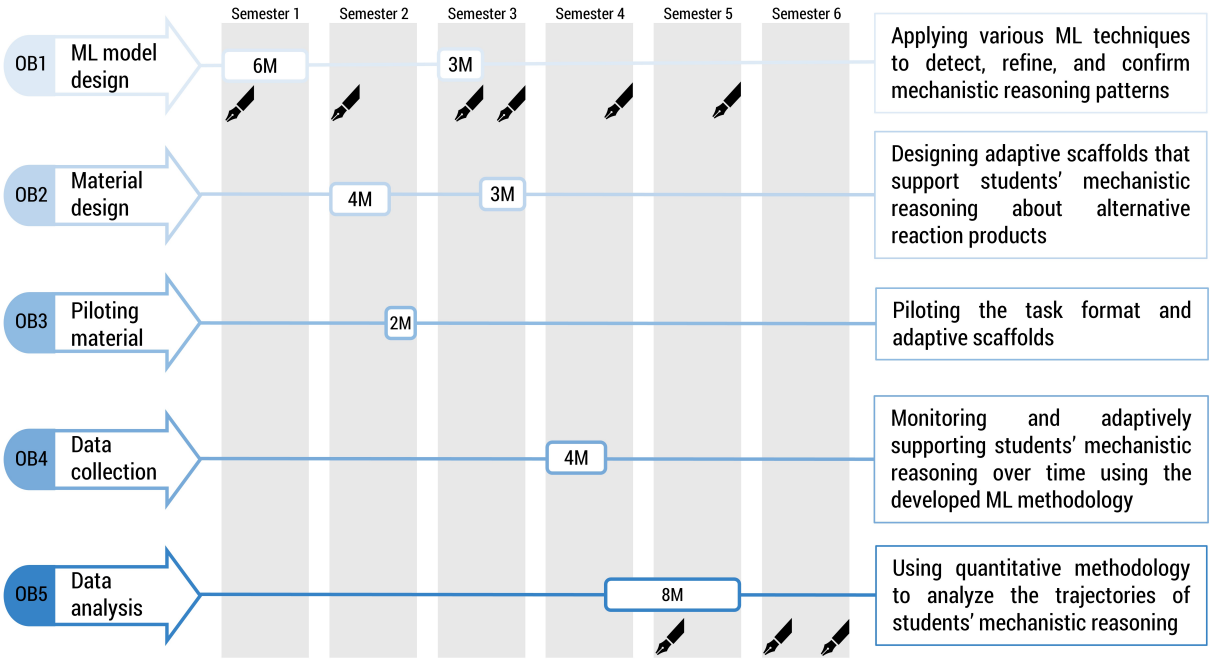


FIGURE 4 Timeline of the research project. *Note: OB* abbreviates *Objective*. A semester spans six months. Rectangles represent the periods during which the respective objectives were explored—for example, *6M* indicates a duration of six months. Pens illustrate the timeframes in which the papers presented in Chapters 2 to 10 were written.

students' current mechanistic reasoning when judging alternative reaction products, applying the ML resources developed for Objective 1 to score their mechanistic reasoning automatically, and assigning adaptive support across three subsequent time points to advance their mechanistic reasoning. Objective 5 focused on quantitatively analyzing the development of students' mechanistic reasoning over time and its interaction with other cognitive and affective variables.

Settings of data collection

For ML model development, we used data gathered at a private, research-intensive, liberal arts university in the Northeastern United States during April and May 2021 (Lieber et al., 2022a). Sixty-four undergraduate Organic Chemistry II students voluntarily participated in the study and judged the plausibility of alternative reaction products for an ether synthesis and Claisen condensation in online scaffolds. The participating students produced 1,108 open-ended mechanistic reasoning statements. For demographic information on the participating undergraduate organic chemistry students, see Lieber et al. (2022a).

To increase training data size for ML model development, we used 1,842 additional student responses from two institutions: a large undergraduate-only university in Western Canada and a research-intensive university in the Midwestern United States. Data collection in the undergraduate organic chemistry courses spanned from the Winter 2020 to the Summer 2023 semester. At the former institution, students judged the plausibility of alternative reaction products for an enamine hydrolysis and aldol condensation. At the latter institution, the research instrument included two Writing-to-Learn assignments to support students' mechanistic reasoning (Watts et al., 2020, 2022).

To ultimately monitor and adaptively support students' mechanistic reasoning, we implemented the designed scaffolds in an Organic Chemistry II course at a different research-intensive university in the Midwestern United States during the Winter 2023 semester. At the beginning of the semester, 315 undergraduate students agreed to participate in this longitudinal study. By the end of the semester, 266 students completed all four required assignments and generated 9,221 open-ended responses. Data from students who did not complete all assignments were excluded from the analysis. For detailed demographic information about the participating students, refer to Chapter 9.

Study design

The task format used in this dissertation prompted undergraduate organic chemistry students to judge the plausibility of alternative reaction products for four mechanisms (Lieber & Graulich, 2020): ether synthesis, enamine hydrolysis, aldol condensation, and Claisen condensation (Figure 5). Each mechanism contained four alternative reaction products, which were presented sequentially to the students. For each product, the students provided open-ended mechanistic reasoning statements on why the product's formation might be plausible or implausible. To scaffold students' mechanistic reasoning within this task format, a simplified version of Toulmin's argumentation pattern, the claim-evidence-reasoning model, was used (McNeill & Krajcik, 2011; McNeill et al., 2006). In this model, a claim is a statement that requires evidence to support it. The evidence can integrate domain-specific information or scientific data. Finally, the reasoning statement establishes a cause-effect relation by linking the claim and evidence. In our setting, students were prompted to claim whether the alternative reaction product is

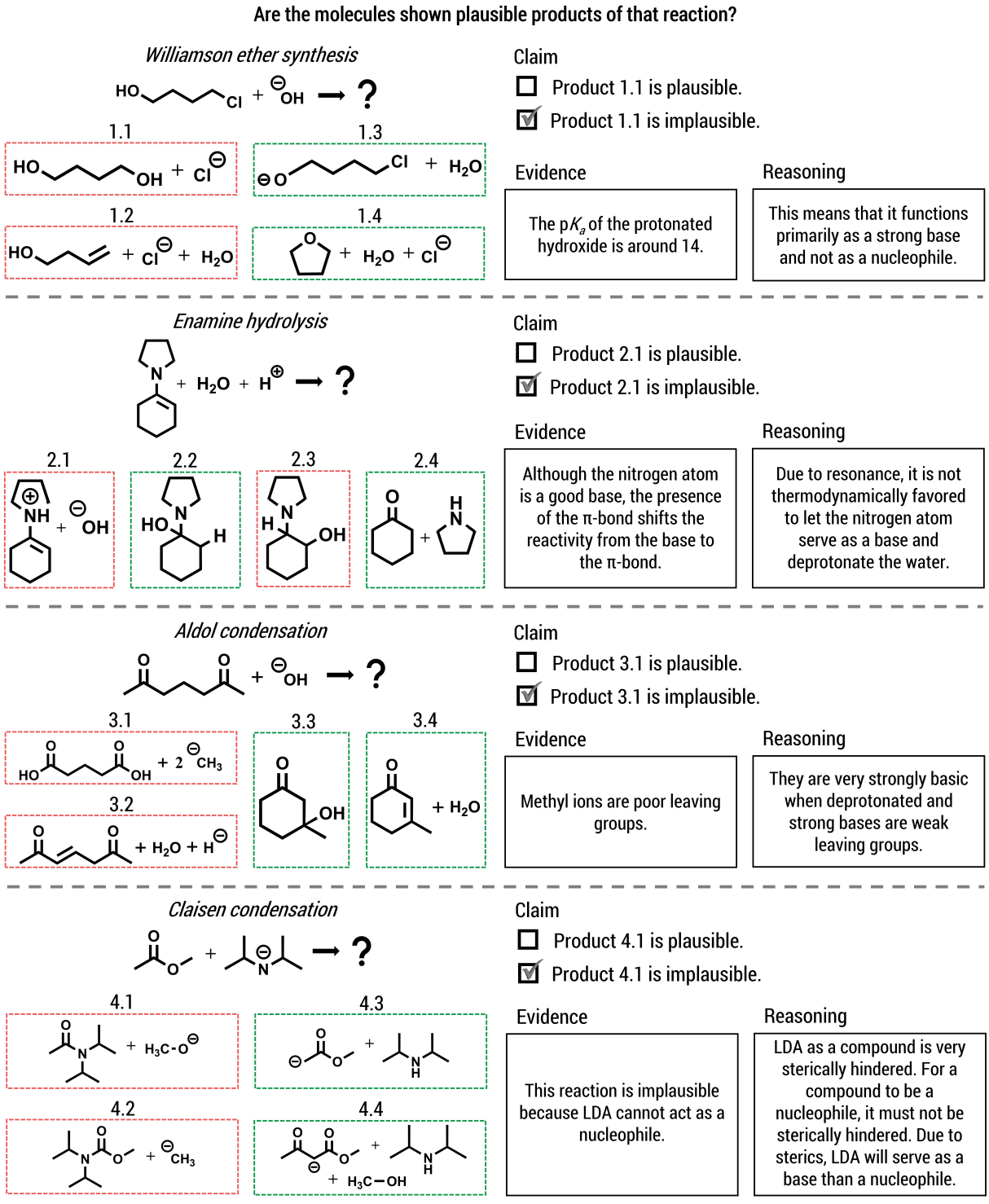


FIGURE 5 Tasks that prompted students to judge the plausibility of alternative reaction products for an ether synthesis, enamine hydrolysis, aldol condensation, and Claisen condensation. *Note:* Plausible products are highlighted in green, and implausible products are highlighted in red. Evidence and reasoning boxes include examples from student responses.

plausible or implausible, provide evidence to justify their claim, and connect the two pieces with a reasoning statement by discussing electronic, steric, or energetic effects. This task format, initially evaluated and applied for an ether synthesis and Claisen condensation by Lieber and Graulich (2020, 2022) and Lieber et al. (2022a, 2022b), was the foundation for monitoring students' mechanistic reasoning.

Within this task format, we designed and implemented four online scaffolds (Figure 6). The first scaffold—tackling the ether synthesis—included diagnostic tasks to assess students' mechanistic reasoning complexity. At this first time point, all students received the same tasks as scaffolding. This scaffolding included closed-response tasks to evaluate and guide students' skills in structuring domain-general and organic chemistry-specific arguments. Further, open-ended tasks were designed to activate students' chemical concept knowledge across four disciplinary core ideas. Following the diagnostic tasks, students judged the plausibility of alternative reaction products for the ether synthesis through open-ended items. The ML model scored students' open-ended responses automatically. Students' mechanistic reasoning at the first time point was the basis for modeling their mechanistic reasoning trajectories at this time point and for providing adaptive support at the second time point.

Automated scoring made it possible to adapt the second scaffold to match each student's current mechanistic reasoning complexity. The second scaffold included three components (Figure 7): First, adaptive process-focused feedback was provided before the students judged the alternative reaction products for the second mechanism. Then, an adaptive formative assessment with approximately 15 closed-response tasks, where task difficulty was adjusted to a student's mechanistic reasoning complexity, supported students in activating relevant chemical concepts and structuring scientific arguments. Finally, students judged the plausibility of alternative reaction products for the second mechanism—enamine hydrolysis—through open-ended tasks (Figure 5), which were again scored by the ML model. Based on this automated scoring, the subsequent scaffolds were adapted to address each student's current mechanistic reasoning complexity.

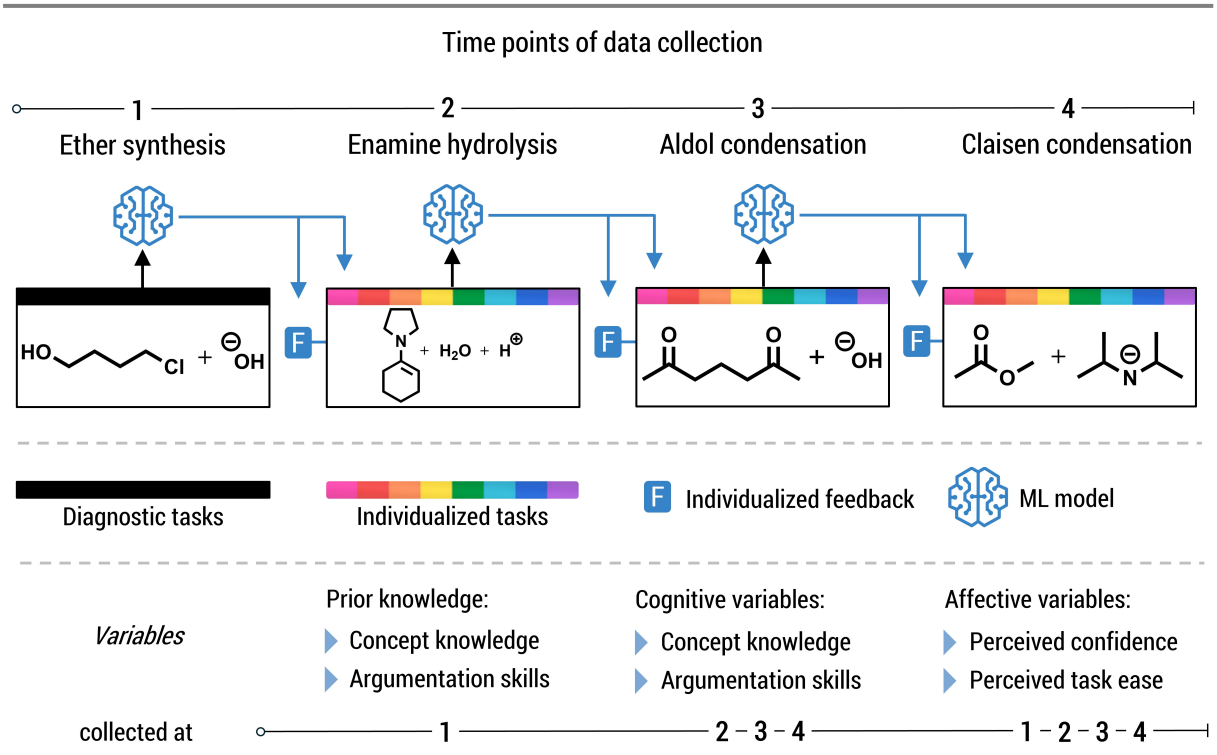


FIGURE 6 Using machine learning to support Organic Chemistry II students' mechanistic reasoning adaptively throughout one semester. [Reproduced from Martin, Kubsch, et al. (in preparation).]

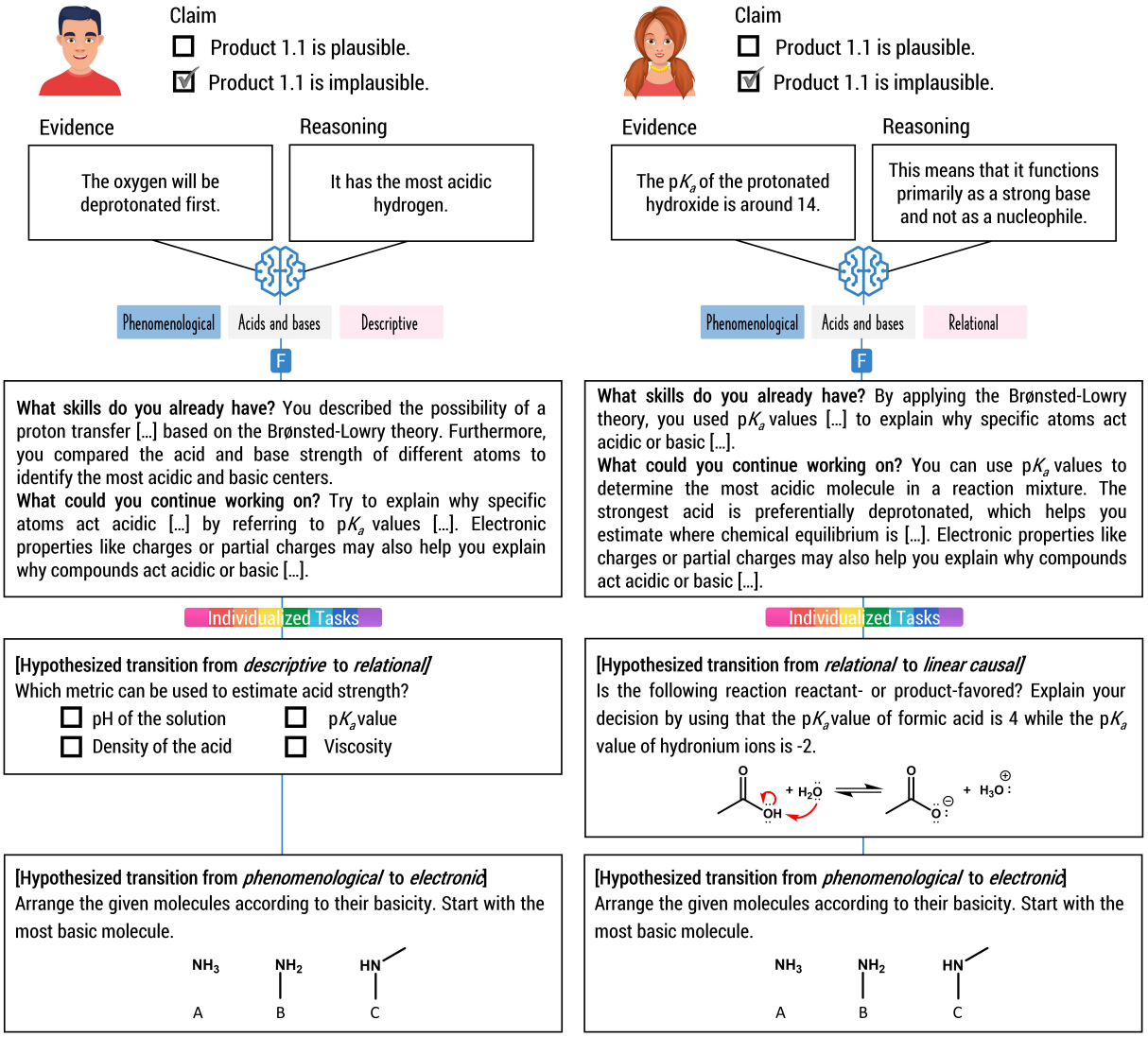


FIGURE 7 Adaptive feedback and exercises to support students' mechanistic reasoning before judging the plausibility of alternative reaction products at the study's second time point (enamine hydrolysis).

1.7 | DATA ANALYSIS

Pattern detection step

Methodological focus: To examine the heterogeneity in undergraduate organic chemistry students' mechanistic reasoning, we applied a clustering algorithm for topic modeling. However, before clustering students' open-ended mechanistic reasoning, data preprocessing was required. Therefore, we first used a large language model, *BERT-base-uncased*, to transform students' open-ended responses into contextualized embeddings—high-dimensional numerical representations of words. Next, we applied Uniform Manifold Approximation and Projection (UMAP) to reduce the dimensionality of these embeddings, which facilitates clustering (McInnes, Healy, & Melville, 2018; McInnes, Healy, Saul, et al., 2018). To identify clusters within students' mechanistic reasoning, we finally applied Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campello et al., 2013, 2015; McInnes et al., 2017). Our hyperparameter choices are detailed in Chapter 3. HDBSCAN is effective in identifying clusters of arbitrary shapes, automatically determining the number of clusters, and handling noisy data as common in students' reasoning in science (e.g., Tschisgale et al., 2023; Wulff et al., 2022; Wulff, Westphal, et al., 2023). During clustering, HDBSCAN

identifies *noise points*, which correspond to non-normative or mixed ideas. By excluding these noise points from clustering, we achieved a finer-grained analysis of students' mechanistic reasoning. At a later analysis stage, we assigned these noise points manually to a rubric category. However, we had to assign these noise points to the highest possible mechanistic reasoning complexity, rather than accounting for mixed uses of granularity, disciplinary core ideas, and causality, to ensure sufficient training data for subsequent automated analyses.

Pedagogical focus: HDBSCAN classifies each student's response into a cluster and provides representative words for every cluster, along with example responses. Based on these outputs, we assigned informative headings to each cluster to capture the key ideas in students' mechanistic reasoning. This analysis step aimed to create a broad overview of the potential patterns in students' mechanistic reasoning.

Pattern refinement step

Methodological focus: To explore the association between the clusters identified by HDBSCAN and theoretical frameworks established in chemistry education research, we statistically compared the HDBSCAN clusters with a research-informed rubric that scores for multivariate, comparative, and electronic reasoning using a χ^2 test of association (Agresti, 2012, 2018; McHugh, 2013). Details on our quantitative analysis are provided in Chapter 7. Comparing HDBSCAN clusters with a research-informed scoring rubric is valuable since interrelations between the two enhance the external validity of both approaches (e.g., Messick, 1975, 1979). In other words, interrelations between the two approaches suggest that they measure a similar construct, validating the use of theoretical frameworks to interpret unsupervised ML outputs and the application of unsupervised ML to refine and extend theoretical frameworks.

Pedagogical focus: We used the *essential epistemic heuristics for mechanistic reasoning* (Krist et al., 2019) to interpret and refine the HDBSCAN clusters. Building on these refinements, we developed a scoring rubric to evaluate students' use of granularity, disciplinary core ideas, and causality. We applied this rubric to score students' mechanistic reasoning (n = 3,123) across task formats. The analysis revealed whether students' mechanistic reasoning remains consistent or exhibits significant heterogeneity across settings.

Pattern confirmation step

Methodological focus: With the designed rubric, we developed a scoring model to evaluate students' mechanistic reasoning automatically based on their use of granularity, disciplinary core ideas, and causality. Specifically, we leveraged a pre-trained, transformer-based deep neural network for sequence classification to build the model. Model development involved applying *AutoTokenizer* to tokenize the text data and leveraging BERT variants to preprocess this data (Devlin et al., 2018). A classification head was added to predict class labels, and we established a training loop using the *AdamW* optimizer with a learning rate scheduler to train the model. To boost the accuracy of our automated analysis, we implemented three data augmentation strategies: collecting additional student data, using chatbots to paraphrase existing student responses, and prompting chatbots to generate synthetic ones. Further details on data augmentation can be found in Chapter 5. Information on the hyperparameter choices for training the scoring model is provided in Chapters 3, 4, 5, 6, and 9. We conducted all ML analysis in Python 3.9.13 (Van Rossum & Drake, 2009) using *PyTorch* as an ML framework (Paszke et al., 2019) and *SpaCy* for natural language processing (Honnibal et al., 2020).

Pedagogical focus: To investigate whether students' mechanistic reasoning benefits from adaptive support, we categorized their mechanistic reasoning into distinct trajectories using the R package *kml*, which applies a *k*-means algorithm designed to model piecewise linear trajectories (Genolini et al., 2015; Genolini & Falissard, 2010, 2011). To select the optimal trajectory number, we interpreted the trajectories qualitatively and assessed their statistical fit using parametric and non-parametric quality criteria. After thorough consideration, we selected a five-trajectory solution as it provided interpretable and distinct trajectories, and all quality criteria indicated acceptable values for this solution. We further linked these trajectories to cognitive and affective variables using a multinomial logistic regression to explore the underlying factors leading to different trajectories (Vermunt, 2010). Special attention was given to the connection between students' mechanistic reasoning and their ability to make correct claims, which was analyzed using hierarchical linear modeling (HLM) (Raudenbush & Bryk, 2002). For detailed information on trajectory modeling, refer to Chapter 9. Chapters 8 and 9 provide additional insights into our quantitative data analysis.

Pattern transparency step

Methodological focus: To assess whether our BERT-based automated scoring model mimics human scoring guidelines in evaluating mechanistic reasoning, we used an explainable AI tool: SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017). Specifically, we calculated the importance scores of the words that most significantly influenced the ML output for each response and averaged these scores within each rubric category. Gaining insights into why the automated scoring model produces specific classifications can help refine scoring models by revising the underlying rubrics and scoring guidelines (Khosravi et al., 2022).

Pedagogical focus: Ultimately, we aimed to clarify the nature of mechanistic reasoning in undergraduate organic chemistry, thereby making the construct more transparent. In particular, we explored whether the *essential epistemic heuristics for mechanistic reasoning* are statistically related in undergraduate organic chemistry. Information on our quantitative analysis is provided in Chapter 9. A deeper understanding of how to conceptualize mechanistic reasoning in this context can inform the design of future instructional support.

1.8 | Results

Pattern detection step

Methodological focus

The pattern detection step aims to identify and extract interpretable patterns from students' open-ended reasoning (Nelson, 2020). In recent years, unsupervised ML techniques, such as clustering algorithms, have proven to be valuable tools for achieving this goal in science education (e.g., Rosenberg & Krist, 2021; Tschisgale et al., 2023; Wulff et al., 2022).

***ML model development**—Using HDBSCAN to detect mechanistic reasoning patterns:* We used HDBSCAN to detect students' mechanistic reasoning patterns when judging the plausibility of alternative reaction products for an ether synthesis and Claisen condensation. HDBSCAN identified 22 distinct clusters and a noise cluster comprising non-normative and mixed ideas (Figure 8) (see Chapter 3, Section 5.1). The clusters covered interpretable and specific patterns of mechanistic reasoning as described below; their high number showcases the significant heterogeneity in how undergraduate organic chemistry students exhibit mechanistic reasoning. Methodologically, our analysis highlights the value of modeling clusters based on contextualized embeddings because these embeddings retain critical domain-specific

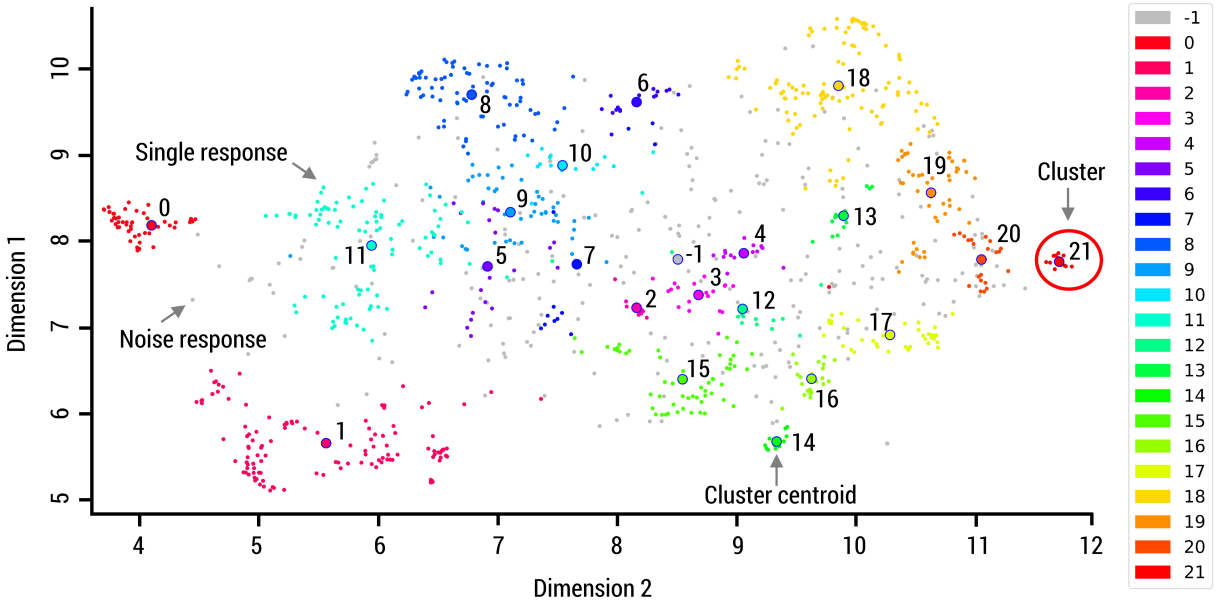


FIGURE 8 Embedding space of the extracted clusters. *Note*: Each data point in the embedding space represents a student's written response. Different colors visualize different clusters. The gray points scattered throughout the embedding space show noise responses that cannot be distinctly assigned to any cluster. The larger points, labeled as cluster centroids, represent the centers of their respective clusters. [Reproduced from Martin et al. (2024).]

information and capture the heterogeneity in students' responses.

Pedagogical focus

In our analysis, HDBSCAN categorized students' responses into clusters based on statistical similarity. Once this output is generated, interpreting the detected patterns qualitatively becomes crucial to get an overview of the topics addressed in students' responses (Nelson, 2020; Rosenberg & Krist, 2021; Tschisgale et al., 2023).

Longitudinal assessment—*Interpreting the mechanistic reasoning patterns inductively across reactions:* In a human-centered qualitative content analysis, we assigned descriptive headings to each cluster to capture the key ideas encapsulated in the detected mechanistic reasoning patterns (see Chapter 3, Table 1). Among the 22 clusters, 13 operated at a *phenomenological* granularity. Responses within these clusters highlighted *phenomenological* properties of entities or activities, such as steric hindrance, pK_a values, conjugate acid strength, equilibrium direction, reaction rate, enthalpy, or entropy. In contrast, 8 clusters reflected *electronic* granularity and incorporated properties like electronegativity, polarizability, electron delocalization, and electron-donating or -withdrawing effects. One cluster integrated *phenomenological* and *electronic* granularity by linking leaving group quality to conjugate acid strength, electronegativity, and polarizability.

Regarding disciplinary core ideas, 4 clusters focused on *charged species and substrates*, primarily emphasizing steric hindrance, leaving group quality, or substrate stability. Similarly, 4 clusters centered on *nucleophiles and electrophiles* by identifying these species, evaluating their strength, and weighing nucleophilic versus basic properties. Another 5 clusters engaged with *acids and bases* and referenced the Brønsted-Lowry or Lewis acid–base theory, identified acids and bases, estimated acid strength using pK_a values, or determined equilibrium direction based on pK_a values. Likewise, 8 clusters addressed *thermodynamics and kinetics* by discussing

energy levels, bond enthalpies, electronegativity in estimating product stability, resonance stabilization, or entropic changes. One cluster bridged the ideas of *charged species and substrates* and *thermodynamics and kinetics* by focusing on the stability of charged species based on charge distribution.

In terms of causality, 7 clusters applied *descriptive* reasoning by listing properties of entities or activities. Another 7 clusters exhibited *relational* reasoning and provided surface-level explanations for the mentioned properties. Likewise, 4 clusters engaged in *linear causal* reasoning and identified cause-and-effect relationships between multiple entities, activities, and their properties. No instances of *multicomponent causal* reasoning were identified in the data sample. Lastly, 4 clusters applied mixed levels of causality.

These findings suggest that students generally incorporated different disciplinary core ideas, interconnected at varying levels of granularity and causality, in their mechanistic reasoning, which offers data-driven support for our operationalization of the *essential epistemic heuristics* (Krist et al., 2019) (Section 1.3). HDBSCAN effectively detected these varying dimensions within students' mechanistic reasoning. However, some clusters addressed mixed levels of granularity, disciplinary core ideas, or causality, necessitating further refinements to establish discriminant categories for subsequent analyses.

What insights did the pattern detection step provide regarding mechanistic reasoning?

By extracting 22 clusters from students' responses, the pattern detection step uncovered substantial heterogeneity in how undergraduate organic chemistry students exhibit mechanistic reasoning. Broader categorizations applied in previous research may have overlooked this heterogeneity, which is why HDBSCAN proved instrumental in achieving a fine-grained analysis of students' mechanistic reasoning (hypothesis 1a confirmed). HDBSCAN identified patterns in a large dataset with scalability and efficiency that traditional qualitative methods cannot match. Thus, unsupervised ML techniques enable a large-scale analysis of students' mechanistic reasoning, which can be characterized by students' use of diverse disciplinary core ideas interconnected at varying levels of granularity and causality (hypothesis 1b confirmed). Such large-scale analyses may allow for a more comprehensive understanding of mechanistic reasoning beyond snapshots in time.

Pattern refinement step

Methodological focus

Given the novelty of advanced ML techniques, their validity in analyzing students' reasoning in science remains largely unexplored (Bejar et al., 2016; Rupp, 2018; Zhai et al., 2021). For example, findings from ML analyses may align with those of traditional research so that both support each other through methodological triangulation. Conversely, ML analyses might diverge from traditional research findings, implying that ML overlooks theory-rich insights from past educational research. For this reason, evaluating the relationship between ML analyses and traditional research methods is important to assess the validity of integrating ML into established research methodologies.

***ML model validation**—Revealing relations between an ML-based and research-informed rubric:* To evaluate the validity of pattern detection using HDBSCAN, we examined whether the data-driven analysis in a related study aligned with a research-informed scoring rubric. Specifically, we used students' mechanistic reasoning collected in the intervention study by Kranz et al. (under review) and applied the pattern detection described above. We then assessed whether the 14 HDBSCAN clusters detected within Kranz et al.'s (under review) data correspond to three research-informed

categories: *multivariate*, *comparative*, and *electronic* reasoning (Kranz et al., under review). As indicated by the significant test statistics for the three research-informed categories and their combination (Table 1), the research-informed categories and the distribution of students' responses among the ML-based clusters are statistically associated. Chapter 7 (Section 5.2) offers additional details on the associations between the clusters and the research-informed rubric.

The significant association between the research-informed and ML-based rubrics suggests that ML-based rubrics not only summarize the content of students' mechanistic reasoning but also reflect research-informed reasoning types. Hence, the ML-based rubric naturally incorporates research-informed considerations. This means that classifying a student's response into a cluster already implies the presence of a specific research-informed reasoning type. From a broader perspective, the alignment with traditional research methodologies validates the use of clustering algorithms for developing rubrics in organic chemistry. Consequently, grounding formative assessments on ML-based rubrics is feasible in organic chemistry since data-driven clusters reflect theory-driven ideas, thus, enhancing their validity.

Pedagogical focus

The alignment between ML techniques and traditional research methods highlights the usefulness of integrating established theoretical frameworks into ML outputs (Kubsch et al., 2023; Nelson, 2020; Sherin, 2013; Zhai, 2021). Theoretical frameworks help refine detected patterns to ensure that they align with specific research goals or learning objectives.

Longitudinal assessment—*Interpreting the HDBSCAN clusters through the essential epistemic heuristics:* We aligned the detected clusters with our operationalization of granularity, disciplinary core ideas, and causality via the *essential epistemic heuristics* (Section 1.3). To do so, we scored each cluster according to this operationalization. The analysis revealed that the clusters only partially reflected the *essential epistemic heuristics*, so we iteratively refined the HDBSCAN clusters in three ways: dividing clusters into subclusters when they encompassed more than one level of granularity, disciplinary core idea, or causality; merging clusters when their responses applied the *essential epistemic heuristics* similarly; and adding categories to capture mechanistic reasoning patterns not represented in the initial clusters but suggested in our operationalization of the *essential epistemic heuristics*. The number of clusters detected by HDBSCAN served as a reference for aligning the theory- and data-driven analyses of students' mechanistic reasoning since it provided an estimate of how many distinct categories could be differentiated in ML analysis.

Cluster refinements resulted in a rubric to evaluate students' mechanistic reasoning through the *essential epistemic heuristics* (Figure

TABLE 1 Results of the χ^2 tests to identify associations between ML-based and research-informed rubrics.

Research-informed category	χ^2 test				Fisher's exact test
	df	χ^2	*p*-value	Cramer's *V* [95% CI]	*p*-value
Multivariate reasoning	14	74.55	< .001	0.41 [0.37, 0.50]	< .001
Comparative reasoning	14	132.38	< .001	0.54 [0.50, 0.61]	< .001
Electronic reasoning	14	79.52	< .001	0.42 [0.37, 0.52]	< .001
Combined	98	333.39	< .001	0.32 [0.30, 0.40]	< .001

9). For the first *essential epistemic heuristic*, we categorized each mechanistic reasoning statement as having *phenomenological* or *electronic* granularity. Here, mechanistic reasoning at the level of the target phenomenon was classified as *phenomenological*; mechanistic reasoning at a lower scalar level, involving atoms and electrons, was classified as *electronic*. Regarding the second *essential epistemic heuristic*, we evaluated whether students included properties of entities, mainly *structural* features, or properties of activities, mainly *energetic* features, in their mechanistic reasoning. *Structural* features were further categorized into one of three disciplinary core ideas depending on the entities students reasoned about: *charged species and substrates*, *nucleophiles and electrophiles*, or *acids and bases*. *Energetic* features correspond to the disciplinary core idea of *thermodynamics and kinetics*. For the third *essential epistemic heuristic*, each response was scored as having a *descriptive*, *relational*, or *linear causal* level. Further details on how we used the *essential epistemic heuristics* to refine the HDBSCAN clusters are provided in Chapter 3 (Section 5.2) and Chapter 9 (Sections 2 and 4.5).

HDBSCAN was essential in rubric design as it identified mechanistic reasoning patterns across many responses and ensured that the analysis was grounded in the data. Integrating the *essential epistemic heuristics* into the analysis highlighted further areas where clusters required adjustments. This iterative refinement facilitated a fine-grained analysis of students' mechanistic reasoning and the development of a scoring rubric. For subsequent analysis, human raters scored students' mechanistic reasoning using the developed rubric.

Longitudinal assessment—*Applying the essential epistemic heuristics across task formats:* The scoring rubric was initially designed to evaluate students' mechanistic reasoning based on a single task format, where students judged the plausibility of alternative reaction products for an ether synthesis and Claisen condensation. To assess the generalizability of the scoring rubric, we tested its applicability across different task formats. Specifically, human raters applied the rubric to two additional task formats: a Writing-to-Learn activity that prompted students to explain thalidomide hydrolysis and a base-free Wittig reaction to a non-expert audience (Watts et al., 2020, 2022), and three case comparisons that required students to determine which of two reactions occurs at a higher rate based on electronic properties (Haas et al., 2024). Chapter 6 (Table 2) offers details regarding human-human inter-rater reliability.

Table 2 presents the distribution of human scores across the rubric categories for the three task

		Structural						Energetic	
Disciplinary core ideas		Charged species and substrates		Nucleophiles and electrophiles		Acids and bases		Thermodynamics and kinetics	
Granularity		Phen.	Elec.	Phen.	Elec.	Phen.	Elec.	Phen.	Elec.
Causality	Absent	0							
	Descriptive	2	3*	2	3	2	3	2	3*
	Relational	3	4	3	4	3	4	3	4
	Linear causal		5	4		4		4	5

FIGURE 9 Scoring rubric to classify students' use of granularity, disciplinary core ideas, and causality. *Note:* Numbers represent the complexity of students' mechanistic reasoning. The asterisk (*) illustrates that these categories reflect the same mechanistic reasoning pattern as both rely heavily on charges to assess product stability.

TABLE 2 Distribution of consensus scores split by dataset using the scoring rubric (Figure 9).

	Structural						Energetic	
	Charged species and substrates		Nucleophiles and electrophiles		Acids and bases		Thermodynamics and kinetics	
	Phen.	Elec.	Phen.	Elec.	Phen.	Elec.	Phen.	Elec.
Activity on judging a product's plausibility								
Absent	23							
Descriptive	127	70*	47	31	90	70	47	70*
Relational	38	32	64	74	58		82	123
Linear causal	0	36	22		20	0	22	32
Writing-to-learn activity								
Absent	130							
Descriptive	140	247*	11	21	89	121	11	247*
Relational	6	11	1	9	0		0	37
Linear causal	0	1	0		0	0	0	13
Case comparison activity								
Absent	11							
Descriptive	2	15*	108	69	0	2	5	15*
Relational	1	16	6	650	0		0	250
Linear causal	0	3	0		0	0	0	29

Note: The asterisk (*) illustrates that these categories reflect the same mechanistic reasoning pattern. Phen. = *phenomenological*, elec. = *electronic*.

formats. Notably, the distribution of human scores varies significantly among them. For judging a product's plausibility, students' mechanistic reasoning is evenly distributed across the two levels of granularity and the four disciplinary core ideas. However, rubric categories involving *descriptive* or *relational* reasoning include more responses than those involving *linear causal* reasoning. In the Writing-to-Learn activity, students' mechanistic reasoning focuses more often on an *electronic* level of granularity than a *phenomenological* one. Students are more likely to reason about *charged species and substrates* or *thermodynamics and kinetics* than about the other two disciplinary core ideas. Additionally, students' use of causality is predominantly *descriptive*. In contrast, the case comparisons have the most centralized distribution of human scores. Over two-thirds of the responses concentrate on *nucleophiles and electrophiles*, primarily with *electronic* as well as *relational* or *linear causal* reasoning. In sum, the varying distribution of students' mechanistic reasoning across task formats suggests that these tasks provide distinct affordances.

The distribution of the human scores across the three task formats demonstrates that the developed rubric can generally be applied to evaluate students' mechanistic reasoning in different tasks. However, more tailored rubrics may better capture the nuances of students' mechanistic reasoning in the Writing-to-Learn and case comparison activity. To accommodate the unique characteristics of students' mechanistic reasoning in other task formats, further data-driven analyses are needed to refine the rubric for broader use.

What insights did the pattern refinement step provide regarding mechanistic reasoning?

By refining the detected patterns, we aligned the ML output with the *essential epistemic heuristics*, which facilitated their meaningful operationalization within our organic chemistry

context. This approach is valid because data-driven ML analyses are associated with research-informed theoretical frameworks (hypothesis 2a confirmed). Without ML, integrating theory- and data-driven analyses would be challenging due to the time-intensive process of testing and refining theoretical frameworks across contexts and the limited human ability to identify nuanced patterns in large, unstructured datasets. However, the applicability of the developed rubric to other task formats was constrained as it did not fully capture the nuances of students' mechanistic reasoning in these formats (hypothesis 2b rejected). Initiating pattern detection from data collected within other task formats would be crucial to adapting our rubric to various contexts.

Pattern confirmation step

Methodological focus

To confirm that the revised patterns are not merely artifacts of the data, supervised ML can investigate whether these patterns are consistently detectable in other data samples (Nelson, 2020; Rosenberg & Krist, 2021; Tschisgale et al., 2023). High ML performance in detecting these patterns provides evidence of their reliability, demonstrating that they can be consistently recognized across data samples.

***ML model development**—Developing a model to score students' mechanistic reasoning:* To validate the refined patterns through supervised ML, we developed a BERT-based scoring model to analyze students' mechanistic reasoning when judging alternative reaction products in the ether synthesis and Claisen condensation. Due to insufficient sample responses, four of the 24 rubric categories (Figure 9) could not be automatically evaluated. By using a simplified version of the rubric with 20 categories, our scoring model demonstrated almost-perfect machine-human score agreement, achieving an accuracy of 87%, Cohen's κ of 0.86, a weighted F_1 score of 0.86, and a macro F_1 score of 0.80. These metrics point out that the refined mechanistic reasoning patterns can be consistently recognized in our data sample. The macro F_1 scores for individual rubric categories ranged from 0.50 to 1.00, indicating variability in the model's performance across categories. Predictably, the model struggled to score categories with fewer sample responses since ML is an inductive learning technique that relies on its training data. Chapter 3 (Section 5.3) provides details on the scoring model.

***ML model development**—Applying the scoring model to the German language:* One advantage of pre-trained, transformer-based deep neural networks is their flexibility in scoring students' responses in different languages by replacing the large language model. To automatically score students' mechanistic reasoning in German, we replaced English-specific BERT variants with German-specific ones and tested their accuracy in scoring German-written mechanistic reasoning. To create a dataset of German-written responses, we translated the original English-written ones using three translation tools: Google Translate, DeepL Translate, and ChatGPT (version: *GPT-3.5*). Google Translate and DeepL Translate provided more verbatim translations, closely adhering to the original English responses (see Chapter 4, Section 7.1). With Google Translate as the translation tool and the large language model *GottBERT-cased*, our scoring model achieved almost-perfect machine-human score agreement in German. The results included an accuracy of 82%, Cohen's κ of 0.81, a weighted F_1 score of 0.82, and a macro F_1 score of 0.79. Thus, our ML analysis is not restricted to the English language, although the model performed better in English than in German (see Chapter 4, Section 7.2).

To allow students to reason in their preferred language in chemistry assessments, we finally incorporated multilingual large language models into our pre-trained, transformer-based deep neural network. These models can analyze

students' open-ended reasoning simultaneously across multiple languages. We specifically evaluated the performance of our scoring model across English and German. Using *BERT-multilingual-cased*, our scoring model demonstrated almost-perfect machine-human score agreement across the two languages, achieving an accuracy of 82%, Cohen's κ of 0.80, a weighted F_1 score of 0.82, and a macro F_1 score of 0.78 (see Chapter 4, Section 7.3). This high level of accuracy highlights the potential of multilingual large language models to bridge language gaps and enhance accessibility to formative chemistry assessments for learners with diverse linguistic backgrounds.

***ML model development**—Augmenting data to boost ML model performance:* Despite almost-perfect machine-human score agreements across multiple languages, our scoring model exhibited certain limitations. First, some rubric categories had low F_1 scores, making it challenging to analyze students' mechanistic reasoning consistently with high accuracy. Second, four rubric categories could not be automatically evaluated due to insufficient training data. To address these issues, we implemented data augmentation strategies, which involve modifying or expanding existing data to increase data size and diversity for ML applications. Specifically, we compared three data augmentation techniques to boost scoring model performance and use the 24-category rubric: collecting additional student responses (Section 1.6), using chatbots to paraphrase existing ones, and employing chatbots to generate synthetic ones.

Our results show that data augmentation substantially improved scoring model performance (see Chapter 5, Section 6.1). Among the three strategies, adding chatbot-generated responses to authentic student data proved most efficient for three reasons: our scoring model trained on authentic and chatbot-generated responses achieved the highest machine-human score agreements, chatbots created sample responses in underrepresented categories where data was lacking, and chatbots reduced much of the human effort typically required for data collection. In the long term, chatbot-generated responses may offer the advantage of anticipating novel response patterns that are not yet prevalent in the data but may emerge after instructional interventions, which can advance the generalizability of scoring models in longitudinal assessments.

In practice, an optimal approach in scoring model development appears to be building an initial scoring model on authentic student data and subsequently improving performance with chatbot-generated responses, especially for underrepresented categories. However, solely relying on chatbot-generated data is problematic as such models tend to underperform when scoring authentic responses (see Chapter 5, Section 6.2). Besides, while a single data augmentation strategy improved model performance, combining multiple strategies yielded only marginal further improvements (see Chapter 5, Section 6.2).

***ML model validation**—Applying the ML model across task formats and reaction mechanisms:* As in the pattern refinement step, we applied the scoring model so far exclusively to the task where students judged the plausibility of alternative reaction products. While the model performed well within this task format, its generalizability to other task formats was not initially explored. Nevertheless, investigating the generalizability of scoring models is crucial for pattern confirmation because it ensures that the identified patterns are not just specific to a particular dataset or task format but are robust and applicable across contexts. To address this, we examined the extent to which the scoring model can automatically evaluate students' mechanistic reasoning from the Writing-to-Learn and case comparison activities, although training data from these activities was not included. The human-assigned scores analyzed in the pattern refinement step for these activities served as a

reference for assessing machine-human score agreements.

Overall, our model struggled to score students' responses from the Writing-to-Learn activity but showed better performance with the case comparisons (see Chapter 6, Section 6.1). The higher generalizability of the model to the case comparison activity can be attributed to the shared characteristics between the two task formats. Both task formats contrast reaction products and guide students to integrate electronic properties into their mechanistic reasoning. By contrast, the Writing-to-Learn activity required students to explain the structure and reactivity of molecules to a non-expert audience, so students should not rely heavily on organic chemistry jargon. Consequently, students often redescribed the given reactions without establishing links between entities, activities, and their properties (Table 2). These differences in learning objectives resulted in variations in the granularity and causality of students' mechanistic reasoning, which reduced the scoring model's generalizability to the Writing-to-Learn activity.

Nonetheless, as shown by the effectiveness of data augmentation, we incorporated students' mechanistic reasoning from one of the other two activities into the training dataset to advance our scoring model's generalizability (see Chapter 6, Section 6.2). First, including Writing-to-Learn data in the training dataset improved the scoring model's performance in the Writing-to-Learn and case comparison activity. Including case comparison data, instead of Writing-to-Learn data, in the training set only boosted performance for the case comparisons, without improving results for the Writing-to-Learn activity. Accordingly, our scoring model demonstrated better generalizability to datasets not included in the training set when these datasets involved case comparisons. By contrast, generalizability to the Writing-to-Learn activity was more limited.

Finally, using data from all three task formats improved the ML model's performance across all activities (see Chapter 6, Section 6.3). Differences in performance were still observed among the reaction mechanisms addressed by the different task formats. Difficulties in automating the analysis of certain reaction mechanisms may stem from the challenges students encounter with these mechanisms, resulting in less structured responses that are more difficult for an ML model to score. Interestingly, the three reaction mechanisms from the case comparison activity consistently performed well in our automated analysis. This suggests that the structured nature of case comparisons guides students in organizing their mechanistic reasoning, which eases subsequent ML analysis.

Pedagogical focus

Pattern confirmation can involve more than just using supervised ML to score the refined patterns automatically in novel responses. It can also include qualitative and quantitative analyses to explore how these patterns evolve over time or whether they relate to other relevant variables in a research context (Zhai et al., 2021).

Longitudinal adaptive support*—Modeling students' mechanistic reasoning trajectories:* We used our scoring model to monitor undergraduate students' mechanistic reasoning at four time points over a semester-long organic chemistry course. Particularly, our analysis scored students' mechanistic reasoning patterns related to their use of granularity (*absent* = 0, *phenomenological* = 1, *electronic* = 2) and causality (*absent* = 0, *descriptive* = 1, *relational* = 2, *linear causal* = 3), resulting in a score from zero to five with increasing mechanistic reasoning complexity (Figure 9). These scores were assigned independent of the identified disciplinary core ideas: *charged species and substrates*, *nucleophiles and electrophiles*, *acids and bases*, and *thermodynamics and kinetics*.

Using these scores, we modeled students' mechanistic reasoning trajectories and identified

five generalizable trajectories (Figure 10). Detailed quantitative analyses of these trajectories are provided in Chapter 9 (Section 5.1) and Chapter 10 (Section 3.4).

- Trajectory A (yellow): The high achievers started at a complexity level of around 3.5 and displayed a steady upward trend with minor fluctuations.
- Trajectory B (red): The moderate-to-high achievers began at a complexity level of 3.0, maintained stable mechanistic reasoning complexity, and showed a slight increase by the end of the semester.
- Trajectory C (blue): These low-to-moderate achievers began at a complexity level of approximately 2.5 and showed notable heterogeneity over time. Especially, these students faced challenges with aldol condensation but demonstrated higher mechanistic reasoning complexity in enamine hydrolysis and Claisen condensation.
- Trajectory D (green): Similarly, these low-to-moderate achievers started at a complexity level of around 2.5 and also showed notable heterogeneity over time. Students of this trajectory faced challenges with enamine hydrolysis while exhibiting higher mechanistic reasoning complexity in condensation reactions.
- Trajectory E (purple): The low achievers began at a mechanistic reasoning complexity of around 2.5 but experienced a decline during the condensation reactions, ending at a complexity level of 2.0.

By modeling interpretable, specific, and robust trajectories, this analysis confirmed the existence of meaningful patterns in students' mechanistic reasoning over time. Trajectories A and B demonstrated moderate gains in mechanistic reasoning complexity while trajectories C, D, and E exhibited greater challenges and heterogeneity, with trajectory E declining

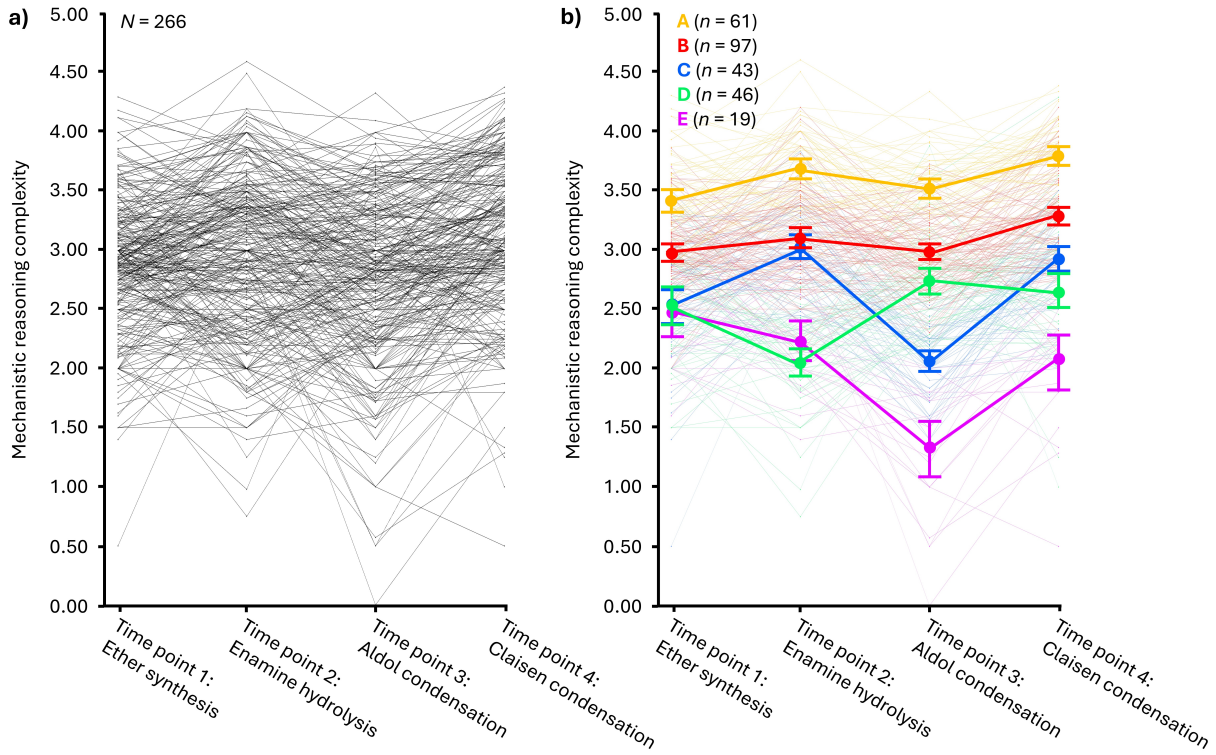


FIGURE 10 a) Students' mechanistic reasoning trajectories. b) Classifying students' mechanistic reasoning trajectories into five generalizable trajectories using a longitudinal clustering technique. *Note:* Error bars represent 95% confidence intervals adjusted for longitudinal data. [Reproduced from Martin, Kubsch, et al. (in preparation).]

significantly. Gaps in mechanistic reasoning complexity between higher- and lower-achieving students persisted throughout the semester, and in some cases, disparities widened despite adaptive support.

Longitudinal assessment—*Examining the relation between mechanistic reasoning and claim correctness:* As outlined in the theoretical framework (Section 1.3), a skill often linked to students' mechanistic reasoning complexity is their ability to make correct claims about a product's plausibility—a skill we refer to as claim correctness. Given the close connection between claims, evidence, and reasoning in our instructional setting, we hypothesized that students' mechanistic reasoning complexity is associated with their claim correctness. To test this hypothesis, we tracked students' claim correctness over time and examined its relationship to their mechanistic reasoning complexity. Figure 11 illustrates the trajectories of students' claim correctness across the four

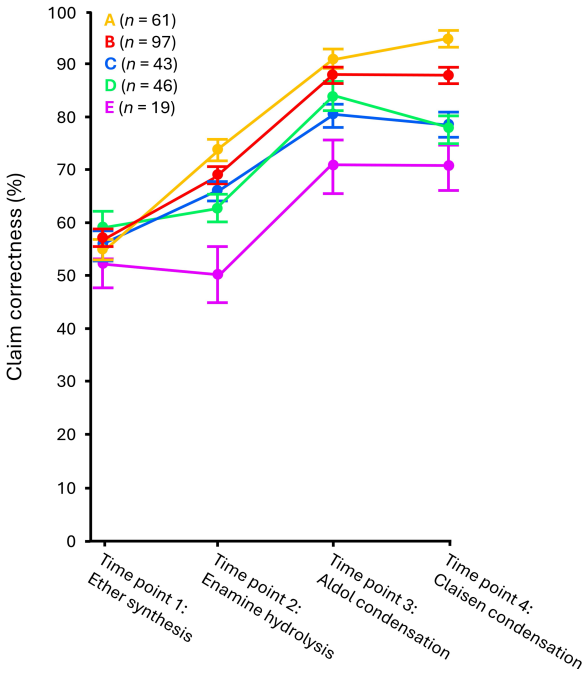


FIGURE 11 Changes in students' claim correctness over time. *Note:* The letters correspond to the mechanistic reasoning trajectories in Figure 10b. Error bars represent 95% confidence intervals adjusted for longitudinal data. [Reproduced from Martin, Kubsch, et al. (in preparation).]

reaction mechanisms. A mixed ANOVA revealed a significant association between students' mechanistic reasoning complexity and claim correctness since the trajectories modeled on mechanistic reasoning complexity also differ in their claim correctness (see Chapter 9, Section 5.2). This association was further supported by a small but significant correlation ($r = 0.178$, 95% CI [0.120, 0.236], $p < .001$), which indicates that variations in mechanistic reasoning complexity align with differences in claim correctness.

Nevertheless, the heterogeneity in students' mechanistic reasoning, particularly in trajectories C, D, and E, motivated a deeper investigation into this relationship. Using HLM, we found that students' mechanistic reasoning complexity was higher when they made correct claims about a product's plausibility in the enamine hydrolysis, aldol condensation, and Claisen condensation. However, this pattern did not hold for the ether synthesis, where the mechanistic reasoning complexity was higher when students assumed a product was plausible, regardless of the correctness of this claim. These findings, detailed in Chapter 8 (Section 6), underscore the nuanced nature of the investigated relationship. While trends can be identified, exceptions—such as the ether synthesis—highlight the context-dependency of students' mechanistic reasoning.

Longitudinal assessment—*Investigating mechanistic reasoning in a broader context:* As a final step in pattern confirmation, we explored the relationships between students' mechanistic reasoning complexity and other cognitive variables, such as students' chemical concept knowledge and scientific argumentation skills, as well as affective variables, including their perceived confidence and task ease. The χ^2 test for the multinomial logistic regression incorporating these variables was statistically significant ($\chi^2(24) = 141.82$, $p < .001$, $R^2 = .183$), indicating that the assessed variables are associated with students' mechanistic reasoning trajectories.

First, students' chemical concept knowledge varied significantly across most trajectories, except for trajectories C and D as well as C and E (Figure 12a). In contrast, students' argumentation skills showed significant differences only between trajectories A and D (Figure 12b). This finding suggests that mechanistic reasoning in undergraduate organic chemistry may rely more heavily on domain-specific chemical concept knowledge than on domain-general argumentation skills since differences in mechanistic reasoning complexity are almost similarly reflected in students' chemical concept knowledge. Furthermore, students' perceived confidence while solving organic chemistry tasks varied significantly between many trajectories (Figure 12c), with higher-achieving students reporting higher confidence. Conversely, perceived task ease differed only across three trajectories (Figure 12d), with students in the high-achieving trajectory A rating tasks as easier than some other trajectories. Chapter 9 (Section 5.5) includes post hoc tests to identify the specific mechanisms where differences between trajectories exist.

In summary, our analysis revealed that both cognitive and affective variables are associated with students' mechanistic reasoning. Since these results align largely with existing literature (e.g., Duncan, 2007; Haskel-Ittah et al., 2020) (see Chapter 9, Section 6), the observed relationship with cognitive and affective variables externally confirms the refined mechanistic reasoning patterns.

What insights did the pattern confirmation step provide regarding mechanistic reasoning?

Through data augmentation, we could develop an ML model that automatically evaluates students' mechanistic reasoning based on a fine-grained scoring rubric with high accuracy (hypotheses 3a confirmed). This model introduced a new avenue for educational research. With automated scoring, we could monitor students' mechanistic reasoning in detail over time and at scale, moving beyond a single snapshot in time. Monitoring the mechanistic reasoning of 266 students provided initial insights into how mechanistic reasoning evolves throughout a semester-long course. We found that students' mechanistic reasoning did not consistently improve over time with adaptive

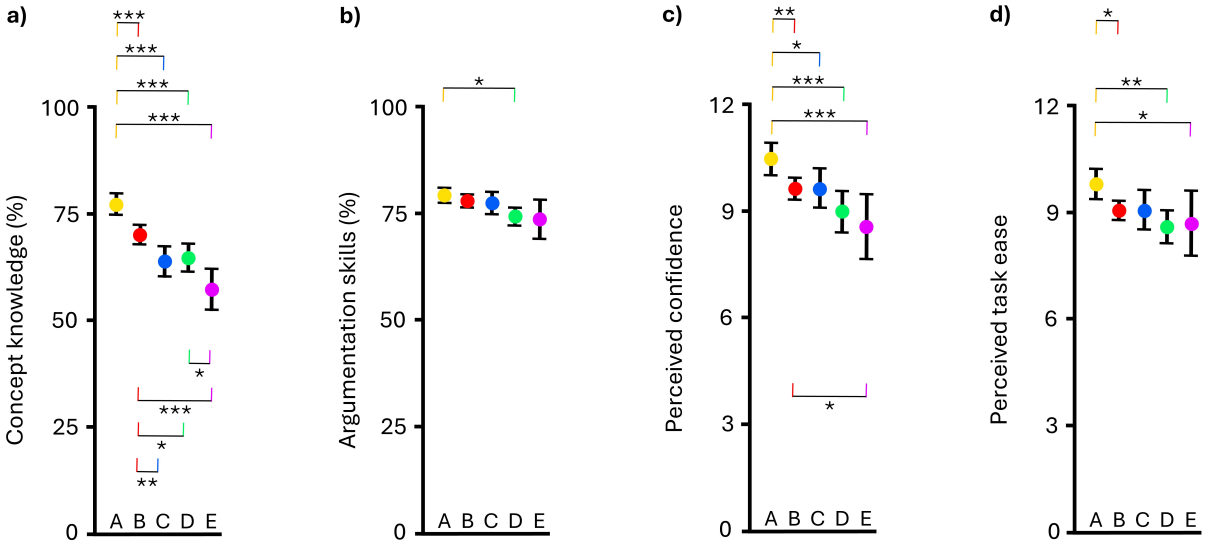


FIGURE 12 Distribution of cognitive and affective variables across trajectories. *Note:* The letters correspond to the mechanistic reasoning trajectories in Figure 10b. Error bars represent 95% confidence intervals adjusted for longitudinal data. *p < .05; **p < .01; ***p < .001 (no p-value correction applied). [Reproduced from Martin, Kubsch, et al. (in preparation).]

support (hypothesis 3b rejected). Furthermore, the ability to collect larger datasets facilitated exploring the relationship between mechanistic reasoning and other cognitive and affective variables, enhancing our understanding of the skills students must demonstrate for complex mechanistic reasoning. We found that students with higher chemical concept knowledge and perceived confidence demonstrated more complex mechanistic reasoning; this relationship did not extend to students' scientific argumentation skills or perceived task ease (hypothesis 3b rejected).

Pattern transparency step

Methodological focus

Advanced ML models often lack interpretability, making it difficult for humans to understand the decision-making behind their outputs. However, understanding why a model arrives at its decisions is crucial in educational contexts to communicate the ML output to stakeholders, such as policymakers, instructors, and students (Khosravi et al., 2022). Moreover, identifying which inputs influence specific outputs can help anticipate ML model behavior and foster trust in ML methodology (e.g., Jacovi et al., 2021). Therefore, pattern transparency becomes important when ML is applied in educational contexts, where its implications impact student lives (Lo Piano, 2020).

ML model validation*—Using explainable AI to analyze ML decision-making:* To enhance the transparency of our scoring model's decisions, we identified the words leading to specific scores using the SHAP model explainer—a method that provides insight into why an ML model generates a certain score (Lundberg & Lee, 2017). For this analysis, we used the data reported in Chapter 9 for ML model training, rather than the longitudinal data analyzed in the previous section. Figure 13 highlights the top ten words most influential for four example rubric categories. To assess the alignment between human and machine scoring, we counted the number of influential words shared between human- and machine-assigned scores for every rubric category. In some rubric categories, the scoring model and human raters did not rely on the same scoring guidelines (Figure 13c). This comparison underscores the potential discrepancies between human and machine scoring.

Besides, SHAP enabled us to identify rubric categories where scoring relied almost exclusively on a single keyword (Figure 13d). Since machine scoring is built on human scoring, this finding suggests that our human scoring was partly influenced by these shortcuts, rather than the semantics of students' mechanistic reasoning.

Pedagogical focus

A final step in increasing pattern transparency can involve reflecting on how the confirmed patterns relate (Carlsen & Ralund, 2022). This step aims to clarify the investigated construct, facilitating its future operationalization.

Longitudinal assessment*—Investigating connections between the essential epistemic heuristics:* Within this step, we investigated how the *essential epistemic heuristics* interrelate in our organic chemistry context. As an initial analysis, we examined the extent to which students' use of granularity and causality are interdependent. Figure 14 visualizes students' use of granularity and causality across the four reaction mechanisms. To quantitatively evaluate whether granularity and causality represent facets of a single skill, we conducted a discriminant function analysis. By calculating linear discriminants to maximize the separation between students' use of granularity and causality, we aimed to determine how these heuristics relate to each other. Our analysis

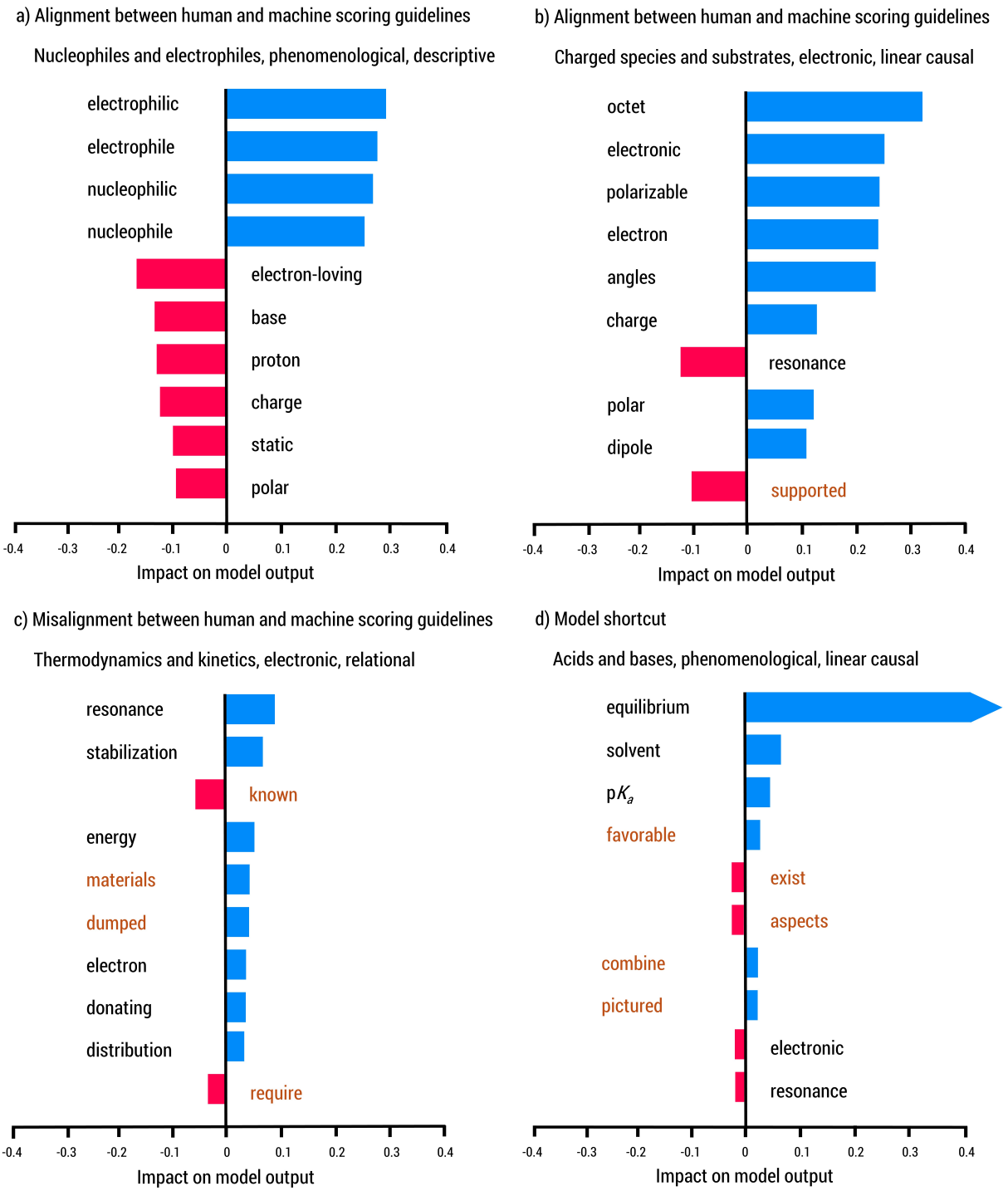


FIGURE 13 Top ten words that most significantly impact the output of the scoring model across four exemplary rubric categories, categorized as either alignment or misalignment between human and machine scoring guidelines or model shortcut. *Note:* Blue scores illustrate an increase in the likelihood of a positive classification within the relevant category when the respective word is included in a student response; red scores reduce the likelihood of a positive classification. The words highlighted in orange indicate misalignments as these words were included in the machine scoring guidelines but not in the human scoring guidelines.

revealed that only the first linear discriminant was statistically significant, showing that the variance in students' use of granularity and causality can be statistically explained with this discriminant. Notably, this discriminant explained an average of 98% of the variance in

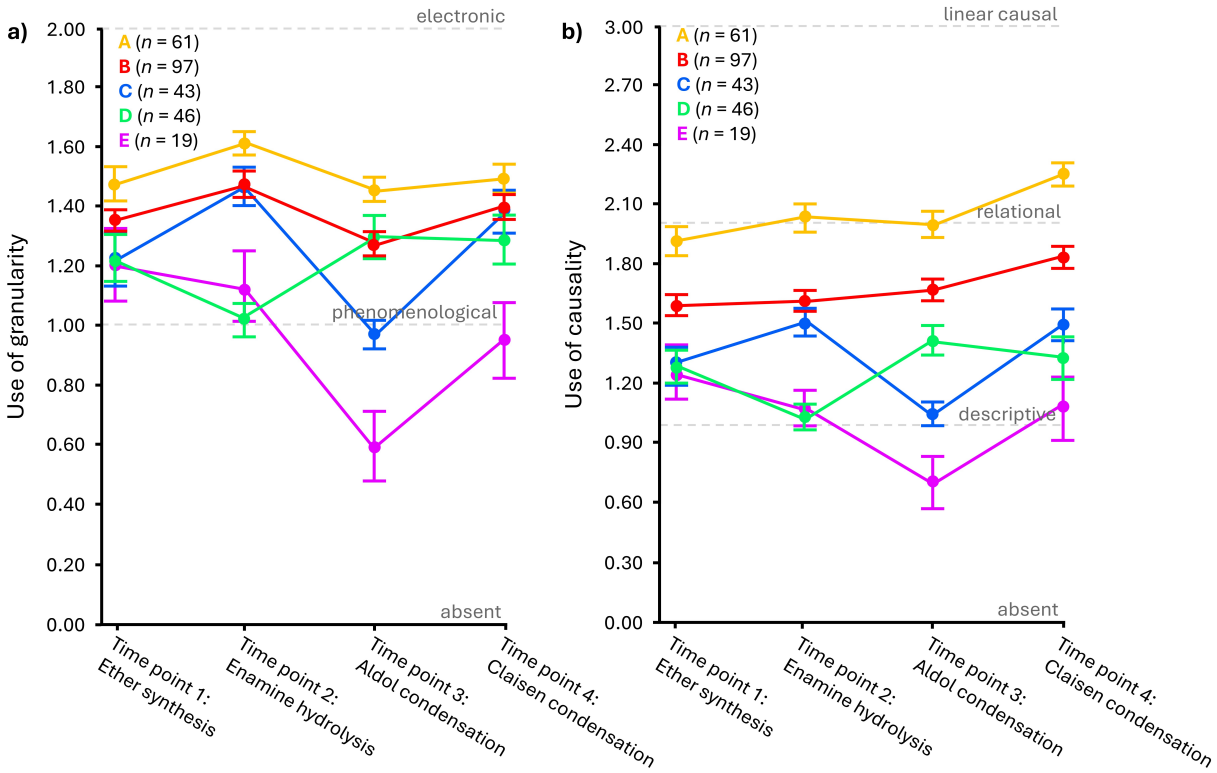


FIGURE 14 Trajectories of students' use of a) granularity and b) causality. *Note:* The letters correspond to the mechanistic reasoning trajectories in Figure 10b. Error bars represent 95% confidence intervals adjusted for longitudinal data. [Reproduced from Martin, Kubsch, et al. (in preparation).]

students' use of granularity and causality and was positively correlated with both heuristics (see Chapter 9, Section 5.3). This finding indicates that trajectory differences do not result from students having a disproportionately high or low complexity in only one heuristic. Instead, high granularity is consistently associated with high causality since both heuristics are positively correlated with the first discriminant, which accounts for nearly all the variance. A medium correlation between the two heuristics ($r = 0.478$, 95% CI [0.430, 0.523], $p < .001$) further supports this conclusion, showcasing that granularity, as the first *essential epistemic heuristic*, and causality, as the third, are interrelated.

Together, these findings indicate that variations in granularity and causality do not account for trajectory differences. To better understand these differences, we analyzed the disciplinary core ideas students applied to justify their claims: *charged species and substrates*, *nucleophiles and electrophiles*, *acids and bases*, and *thermodynamics and kinetics*. A χ^2 test revealed that students in different trajectories employed distinct disciplinary core ideas in their mechanistic reasoning ($\chi^2(12) = 111.11$, $p < .001$). To pinpoint specific differences among trajectories, we examined the standardized residuals of the χ^2 test, focusing on the disciplinary core ideas applied more or less frequently than statistically expected. Most importantly, trajectory B integrated energetic ideas more frequently than statistically expected ($n_{expected} = 87.883$; $n_{observed} = 120$; $p = .002$), whereas trajectories D ($n_{expected} = 41.677$; $n_{observed} = 21$; $p = .002$) and E ($n_{expected} = 17.214$; $n_{observed} = 4$; $p = .001$) did so significantly less. This finding underscores that incorporating thermodynamic and kinetic ideas, reflecting the second *essential epistemic heuristic*, is associated with higher mechanistic reasoning complexity, revealing how the second heuristic is linked to granularity and causality in organic chemistry. Chapter 9 (Section 5.4) includes post hoc tests to identify the mechanisms where differences between trajectories exist.

What insights did the pattern transparency step provide regarding mechanistic reasoning?

The three *essential epistemic heuristics*—considering the scalar level below, identifying entities, activities, and their properties at this lower scalar level, and linking these factors back to the original scalar level—are interrelated (hypothesis 4b confirmed). This finding improves our understanding of the nature of mechanistic reasoning in organic chemistry. The heuristics can be used to scaffold students' mechanistic reasoning, but careful attention must be given to their interconnectedness to demonstrate complex mechanistic reasoning. Although our ML model accurately evaluated students' use of these heuristics, it did not mimic human scoring guidelines (hypothesis 4a rejected).

1.9 | Discussion

Pattern detection step

Methodological focus: The aim of the pattern detection step is to identify patterns within a text corpus by leveraging exploratory analysis techniques (Nelson, 2020). Clustering algorithms, particularly those in the topic modeling family like HDBSCAN, are increasingly used in science education research for this purpose (e.g., Anderson et al., 2020; Sherin, 2013; Wulff et al., 2022; Wulff, Westphal, et al., 2023; Zehner et al., 2016). Consistent with our analysis, these applications demonstrate that clustering can generate interpretable patterns, capture critical domain-specific information, and detect significant linguistic heterogeneity. Thus, topic modeling techniques like HDBSCAN can combine the exploratory, content-sensitive strengths of qualitative research with the systematicity and scalability of quantitative methods.

However, applying clustering in educational contexts is challenging. In alignment with prior studies, we found that the optimal clustering performance relies, among others, heavily on thoughtful data preparation, including decisions such as whether to exclude low-meaning words from the analysis and how many dimensions to preserve in contextualized embeddings (Carlsen & Ralund, 2022; Tschisgale et al., 2023; Wulff et al., 2022; Zehner et al., 2016). Additionally, the choice of HDBSCAN hyperparameters, like the minimum cluster size, impacts the clustering outputs (Tschisgale et al., 2023). For instance, setting the minimum cluster size too low can lead to various small clusters, many of which may be indistinguishable from noise. Conversely, an excessively high value for the minimum cluster size may yield fewer, overly large clusters. Balancing these hyperparameters is crucial for achieving robust and meaningful cluster results (Campello et al., 2013, 2015; McInnes et al., 2017).

Given these challenges, directly clustering students' open-ended reasoning is not without critique. For instance, Ariely et al. (2024) and Watts et al. (in print) initially scored students' mechanistic reasoning in a binary way across multiple mechanistic reasoning features. Only after this step, they applied a clustering algorithm to group students' pre-scored mechanistic reasoning into more generalizable patterns. Watts et al. (in print) argue that deriving the clustering output from pre-scored responses results in more interpretable patterns that are better aligned with established theoretical frameworks.

Taken together, researchers must be cautious in pattern detection to ensure that the outputs of a clustering algorithm are not merely artifacts of specific preprocessing steps or the algorithm's parameterization (Carlsen & Ralund, 2022). Misleading clustering results can limit the detection of domain-specific nuances in the data, potentially hindering an unbiased data examination and preventing the identification of previously overlooked patterns. Hence, a thoughtful and iterative approach to hyperparameter fine-tuning, combined with

rigorous validation of clustering outputs against domain-specific criteria, is essential to ensure the robustness and interpretability of the results (Carlsen & Ralund, 2022; Tschisgale et al., 2023). Nevertheless, the subsequent pattern refinement step in CGT enables researchers to refine the detected patterns within established theoretical frameworks, which may mitigate the impact of minor classification errors.

Pedagogical focus: Besides our analysis, other studies also employed clustering algorithms, including latent class analysis, to analyze students' mechanistic reasoning (e.g., Ariely et al., 2024; Watts et al., in print). However, comparing their clustering outputs with our analysis is problematic due to differing research objectives. For instance, Ariely et al. (2024) and Watts et al. (in print) applied a binary rubric to score students' mechanistic reasoning across multiple features and clustered the pre-scored data into broader, more generalizable mechanistic reasoning patterns. Accordingly, Ariely et al. (2024) and Watts et al. (in print) merged conceptually similar responses into broader categories while our focus was on uncovering the fine-grained heterogeneity within students' mechanistic reasoning across contexts.

Earlier studies used more traditional clustering algorithms, such as *k*-means clustering, to investigate students' mechanistic reasoning (Haudek et al., 2015; Prevost et al., 2012). Here, Haudek et al.'s (2015) application of *k*-means clustering corresponds most closely to our use of HDBSCAN as both approaches aimed to capture the heterogeneity in students' mechanistic reasoning to inform rubric development. However, while our analysis spans multiple disciplinary core ideas, Haudek et al. (2015) focused exclusively on students' mechanistic reasoning about acids and bases. Specifically, their work investigated how students connect the concepts of acidity and basicity, whereas our study emphasized broader mechanistic reasoning patterns, such as students' use of granularity and causality.

By analyzing students' use of granularity, disciplinary core ideas, and causality, we observed significant heterogeneity. This finding aligns with prior research. For instance, heterogeneity in students' use of granularity—the scalar levels they address in their mechanistic reasoning—is evident within and across tasks (Deng & Flynn, 2021). Similarly, students often integrate diverse disciplinary core ideas into their mechanistic reasoning, even when working on the same tasks (Bodé et al., 2019; Cooper et al., 2017; Deng & Flynn, 2021). Additionally, it is well established that students show considerable heterogeneity in their use of causality—how they link entities, activities, and properties (e.g., Moon et al., 2016; Moreira et al., 2019; Weinrich & Talanquer, 2016). However, prior studies have examined these dimensions in isolation. By integrating them, we demonstrate that students' mechanistic reasoning varies significantly across their combinations. This suggests that the heterogeneity in students' mechanistic reasoning may be even greater than previously recognized (e.g., Crandell et al., 2019; Frost et al., 2023; Southard et al., 2017; Watts et al., 2021; Yan & Talanquer, 2015; Yik et al., 2023). The large-scale study facilitated by HDBSCAN allowed for a finer-grained analysis of the heterogeneity in students' mechanistic reasoning.

Pattern refinement step

Methodological focus: The second step of CGT focuses on qualitatively analyzing the detected clusters to confirm their plausibility, interpret their meaning, and refine their structure (Nelson, 2020). This step resonates with calls from researchers to leverage domain-specific theoretical frameworks for refining ML outputs (Kubsch et al., 2023; Sherin, 2013; Zhai, 2021). However, the validity of combining theoretical frameworks with data-driven ML outputs remains largely unexplored. Only a few studies have shown that ML-assigned codes align with students' responses in oral interviews (Beggrow

et al., 2014; Haudek et al., 2012), implying that ML can validly capture students' open-ended reasoning. By demonstrating a connection between research-informed and ML-based rubrics, we provide further empirical evidence for the association between traditional qualitative and ML-based analyses, especially when ML is integrated into the earlier stages of assessment development.

Pedagogical focus: By refining the clustering output using the *essential epistemic heuristics* (Krist et al., 2019), we operationalized them within our organic chemistry context. To our knowledge, this represents the first reported operationalization of these heuristics in chemistry. Given the high generalizability of the *essential epistemic heuristics* across domains, our operationalization holds potential for future analyses because it may facilitate comparisons of students' mechanistic reasoning across domains or everyday contexts. However, further refinements may be necessary to develop a broadly applicable rubric to score students' mechanistic reasoning across task formats.

Tracking students' mechanistic reasoning through *essential epistemic heuristics* is in line with Cooper (2020), who argues that a set of heuristics must be developed to support students' mechanistic reasoning productively since heuristics could serve as a scaffold to guide students' mechanistic reasoning across contexts. Applying a set of heuristics should begin with analyzing single phenomena, then extend to multiple phenomena within a single domain, and eventually span across domains (Cooper, 2020). By operationalizing Krist et al.'s (2019) *essential epistemic heuristics* within organic chemistry, our analysis begins to address Cooper's (2020) idea of using a set of heuristics to support students' mechanistic reasoning across multiple phenomena.

Pattern confirmation step

Methodological focus: The pattern confirmation step aims to verify that the inductively identified topics hold consistently throughout the text corpus, which serves as a final reliability check to ensure that the results are not artifacts of the clustering output (Nelson, 2020). In our automated analysis, we applied pre-trained transformer-based deep neural networks for sequence classification to assess whether the refined mechanistic reasoning patterns can be reliably identified in students' mechanistic reasoning. Our scoring model achieved almost-perfect machine-human score agreement based on a 20-category version of our scoring rubric. This finding indicates that advanced ML models can accurately analyze students' mechanistic reasoning—consistent with prior research showing that such models often outperform traditional ML approaches in science education contexts (e.g., Dood et al., 2024; Gombert et al., 2023; Wulff, Mientus, et al., 2023). Nevertheless, the performance of ML techniques is highly task-dependent (Kieser et al., 2024); for instance, Küchemann et al. (2020) found that traditional ML models outperform advanced ones in less complex scoring tasks. Therefore, the choice of which ML techniques to apply depends, among others, on the complexity of the scoring rubric and the research objectives (Kieser et al., 2024).

In addition to achieving high machine-human score agreement, our analysis shows that automated scoring can provide a level of detail that surpasses earlier ML analyses of students' mechanistic reasoning, which typically employed scoring rubrics with only two to five categories (e.g., Dood et al., 2018, 2020a; Noyes et al., 2020; Yik et al., 2021). As a result, ML techniques cannot only be used to confirm refined patterns but also to establish automated scoring models capable of analyzing students' mechanistic reasoning with fine-grained rubrics. Moreover, automated analyses are not limited to the English language, which corresponds to research that

applied ML to score students' open-ended reasoning in other languages (e.g., Ariely et al., 2023; Gombert et al., 2023; Wang et al., 2021; Wulff, Mientus, et al., 2023). Our analysis even extends these prior studies by suggesting that accurate ML analyses of students' open-ended reasoning are feasible across multiple languages. These multilingual ML techniques allow students who learn the language of instruction as an additional language to reason in assessments in their preferred language.

Despite almost-perfect machine-human score agreement, our scoring model had limitations: Some rubric categories had low F_1 scores and four categories could not be automatically evaluated due to insufficient training data. To improve scoring model performance and enable a finer-grained categorization, we used data augmentation strategies. Our analysis demonstrated that using chatbots for text data augmentation expanded the number of automatically evaluable rubric categories, balanced datasets, and improved scoring accuracy. These results align with domain-general studies showing that large language models like GPT outperform traditional data augmentation strategies (e.g., Cochran et al., 2022, 2023; Dai et al., 2023; Møller et al., 2023). Thus, augmenting authentic student data with chatbot-generated data is beneficial to improve ML model performance, which is evidenced in another study in science education (Fang et al., 2023). However, the effectiveness of text data augmentation depends, among others, on the amount of student and synthetic responses (Fang et al., 2023).

To finally evaluate the generalizability of our ML model, we tested its performance across task formats. Results revealed that our initial model exhibited poor generalizability to different task formats. This finding is consistent with Nehm et al. (2012) and Yik et al. (2021), which indicate that ML models trained on specific datasets often underperform when applied to novel datasets. However, incorporating a second context into the training set improved generalizability, even in unfamiliar task formats. This underscores the importance of including multiple contexts in training datasets when aiming to develop more generalizable ML models, which is congruent with studies by Dood et al. (2022), Noyes et al. (2020), and Wulff, Westphal, et al. (2023) that show higher generalizability when ML models are trained across courses, disciplines, or institutions. Although these findings indicate that ML models can be effectively applied to evaluate students' mechanistic reasoning across contexts, performance discrepancies persisted across the mechanisms addressed by these task formats. Observed variations may reflect students' heterogeneity in mechanistic reasoning (e.g., Crandell et al., 2019; Frost et al., 2023; Southard et al., 2017; Watts et al., 2021; Yan & Talanquer, 2015; Yik et al., 2023). Particularly, different contexts pose different challenges for students (Flynn & Featherstone, 2017; Houchlei et al., 2021), influencing the vocabulary and structure of their mechanistic reasoning (Crowder & Raker, 2024), which, in turn, affects machine-human score agreements.

Pedagogical focus: We applied the developed scoring model to monitor and adaptively support organic chemistry students' mechanistic reasoning over time. This longitudinal investigation revealed the trajectories of students' mechanistic reasoning, moving beyond a snapshot in time. By modeling students' mechanistic reasoning trajectories, we found—consistent with prior studies—that students' mechanistic reasoning improved with the implemented adaptivity (e.g., Ariely et al., 2024; Asmussen et al., 2024; Dood et al., 2020b; Hedtrich & Graulich, 2018a, 2018b; Lieber et al., 2022a). Congruent with the aims of divergent adaptivity (Deunk et al., 2015, 2018; Smale-Jacobse et al., 2019), students could progress at their own pace throughout the semester, which likely contributed to the persistence of achievement gaps. This outcome corresponds to

prior research implementing divergent adaptivity, which similarly observed gaps in students' mechanistic reasoning over time (Ariely et al., 2024; Dood et al., 2020b).

However, by examining how students' mechanistic reasoning diverged over time, we found that achievement gaps widened in part significantly due to the adaptive support. This finding suggests that the adaptivity employed in this study may have disproportionately benefited moderate- and high-achieving students and disadvantaged their lower-achieving peers. Other studies proposed that adaptivity can exacerbate achievement gaps because more advanced learning paths often target higher-order cognitive processes and lower-achieving students are less frequently assigned to them (Koedinger et al., 2023; Sbeglia & Nehm, 2024). Consequently, higher-achieving students can capitalize on these more advanced learning paths, further increasing the gap between themselves and their lower-achieving peers—a phenomenon known as the Matthew effect (Stanovich, 1986). These findings highlight that implementing adaptivity through ML is a complex process that presents challenges in ensuring equitable learning gains for all.

In addition to the varying impact of adaptive support on students' mechanistic reasoning, the trajectories also reveal different degrees of heterogeneity. Trajectories A and B, representing students with the highest mechanistic reasoning complexity, show minimal heterogeneity over time. By contrast, trajectories C, D, and E show context-dependent fluctuations in mechanistic reasoning complexity, resulting in much greater heterogeneity. Hence, lower-achieving students' mechanistic reasoning appears to follow a non-linear progression that lacks clearly defined developmental stages (e.g., Southard et al., 2017). This suggests that the development of lower-achieving students' mechanistic reasoning cannot be modeled as a stable, linear cognitive process (DiSessa et al., 2016; Hammer et al., 2005); their mechanistic reasoning complexity in one context does not reliably predict their complexity in another.

Reasons for the heterogeneity in students' mechanistic reasoning across contexts include the unique challenges posed by each reaction mechanism in organic chemistry (Deng & Flynn, 2021; Watts et al., 2021; Weinrich & Sevian, 2017). Some mechanisms are inherently more complex and require a deeper analysis of multiple steps or the application of advanced chemical concepts (Flynn & Featherstone, 2017). Due to challenges in recognizing the interconnections between these chemical concepts, especially lower-achieving students struggle to identify which chemical concepts to apply in each reaction mechanism (Bhattacharyya & Bodner, 2005; Ferguson & Bodner, 2008; Grove, Cooper, & Cox, 2012; Grove, Cooper, & Rush, 2012). As a result, lower-achieving students may struggle to activate relevant chemical concepts, leading to persistent challenges as they move between different mechanisms. From a broader perspective, these insights challenge traditional learning theories because they imply that mechanistic reasoning can be highly flexible and heterogeneous (e.g., Crandell et al., 2019; Watts et al., 2021). Following that, research on mechanistic reasoning should shift from viewing students' reasoning, ideas, or misconceptions as stable, unchangeable constructs (Hammer et al., 2005); instead, it should investigate how the context-specific use of granularity, disciplinary core ideas, and causality influences mechanistic reasoning (Graulich, 2025).

Some studies have already adopted a more flexible approach to analyzing students' learning (e.g., Crandell & Pazicni, 2023; Gao et al., 2024; Pölloth et al., 2024; Rodriguez et al., 2020). For example, Haas et al. (2024) monitored students' mechanistic reasoning across various activities while other research focused on gradually adapting task prompts to better scaffold

students' mechanistic reasoning (Cooper et al., 2016, in print; Franovic et al., 2023; Noyes et al., 2022; Zaimi et al., 2024). Studies on students' use of the resonance concept also show that students activate various ideas across tasks (Braun & Graulich, 2024) and focus on specific structural features and heuristics (Rotich et al., 2024). In line with the identified mechanistic reasoning trajectories, these studies reveal a non-linear trajectory to considering resonance in problem-solving, where students' heterogeneity is shaped by task affordances and various individual variables that influence how they perceive chemical structures and make inferences.

While this theoretical shift reveals the heterogeneity of students' mechanistic reasoning, it also presents challenges for designing effective instructional strategies. Developing adaptive support that dynamically responds to each student's unique mechanistic reasoning complexity remains a substantial hurdle. Furthermore, effort is required to create longitudinal adaptive support that integrates chemical concepts cohesively, rather than addressing them in isolated contexts. Designing adaptive support that interacts with students' mechanistic reasoning trajectories related to well-defined, domain-specific learning objectives continues to be a persistent challenge—even with ML assistance.

Pattern transparency step

Methodological focus: The pattern transparency step ultimately aims to clarify the workings of ML models to improve their interpretability. To achieve this, we employed SHAP to identify the words most influential in our scoring model's output and found that machine scoring sometimes relied on different scoring criteria than humans, resonating with other studies using post hoc explainability techniques to analyze ML decision-making (Gombert et al., 2023). On the one hand, the distinct approach of ML scoring offers advantages such as high consistency, scalability, and efficiency; on the other hand, limitations in contextual understanding and adaptability to unexpected inputs can undermine a model's transparency.

Using post hoc explainability techniques offers an additional advantage: it exposes the implicit assumptions underlying the human scores that serve as the model's training data (Fussell et al., 2024). For this reason, these techniques can help identify the biases, criteria, or shortcuts that human raters may have unconsciously applied when scoring the training data. As observed in our analysis, responses addressing *acids and bases* on a *phenomenological* and *linear causal* level heavily relied on the term "equilibrium" (Figure 13d), which reveals a shortcut in human and machine scoring. In the future, advancing the transparency of why specific responses are scored in particular ways—whether by humans or ML—could shift the focus toward the semantics of students' mechanistic reasoning, rather than the presence of certain keywords.

Pedagogical focus: To clarify the structure of mechanistic reasoning in organic chemistry, we finally explored the interrelations among the *essential epistemic heuristics* (Krist et al., 2019). Our findings reveal that students' use of granularity, reflecting the heuristic of considering the scalar level below the target phenomenon, is closely associated with their use of causality, reflecting the heuristic of linking factors back to the original scalar level. Moreover, integrating thermodynamic and kinetic ideas advanced students' mechanistic reasoning, which indicates how the heuristic of unpacking factors at the lower scalar level impacts the other two heuristics. These findings show an interdependence between the *essential epistemic heuristics* and extend prior research that often differentiated domain-general practices (Machamer et al., 2000; Russ et al., 2008) and domain-specific knowledge in mechanistic reasoning (Duncan, 2007).

Specifically, our findings suggest that domain-general heuristics operationalized through granularity and causality are deeply intertwined with domain-specific knowledge about disciplinary core ideas. This finding resonates with Haskel-Ittah et al. (2020) and Haskel-Ittah and Duncan (2021), who demonstrated that the second heuristic depends on sufficient domain-specific knowledge, which means this heuristic may not be activated in its absence. We can even identify which domain-specific knowledge is critical for complex mechanistic reasoning in organic chemistry: thermodynamic and kinetic ideas.

1.10 | Conclusion

CGT provided a methodological framework for analyzing students' mechanistic reasoning: Pattern detection uncovered fine-grained mechanistic reasoning patterns that revealed considerable heterogeneity in how students demonstrate mechanistic reasoning within and across different contexts (hypothesis 1a confirmed). Pattern refinement enabled the alignment of these data-driven patterns with an evidence-based theoretical framework—the *essential epistemic heuristics*—to situate students' use of granularity, disciplinary core ideas, and causality within a broader context (hypothesis 2a confirmed). Pattern confirmation demonstrated the applicability of these patterns across mechanisms, task formats, populations, languages, and data sources, which facilitated the development of an automated scoring model capable of accurately evaluating students' mechanistic reasoning across these patterns (hypothesis 3a confirmed). Finally, pattern transparency revealed shortcuts in human and machine scoring (hypothesis 4a rejected), opening an iterative process for continuous refinement of our analysis.

Developing an automated scoring model allowed us to longitudinally monitor and adaptively support students' mechanistic reasoning, which led to a computationally grounded theory of students' mechanistic reasoning. Guided by the *essential epistemic heuristics*, students incorporate different disciplinary core ideas, interconnected at varying levels of granularity and causality, into their mechanistic reasoning (hypothesis 1b confirmed). The application of these heuristics varies across task formats and mechanisms (hypothesis 2b rejected). Further, the heuristics are interrelated and rely on domain-specific knowledge, such as thermodynamic and kinetic ideas (hypothesis 4b confirmed). Also, students' mechanistic reasoning is tied to their chemical concept knowledge, with domain-general argumentation skills playing a less prominent role. Higher perceived confidence in solving organic chemistry tasks also correlates with more complex mechanistic reasoning. While students' mechanistic reasoning tends to improve over time, the mechanistic reasoning of lower-achieving students remains highly heterogeneous and cannot be modeled by a fixed model of stable-developing cognition (hypothesis 3b rejected). Overall, monitoring and adaptively supporting students' mechanistic reasoning beyond a snapshot in time enabled the identification of fine-grained mechanistic reasoning trajectories and a comprehensive understanding of students' flexible and individual cognitive development, thereby informing the design of future adaptive support.

1.11 | Outlook and future work

This dissertation demonstrates the application of CGT to monitor students' mechanistic reasoning over time. By tracking students' use of granularity, disciplinary core ideas, and causality, our findings provide a deeper understanding of how undergraduate students exhibit mechanistic reasoning in organic chemistry. Furthermore, we developed an automated scoring model capable of longitudinally tracking mechanistic reasoning across organic chemistry mechanisms. The

presented analysis opens new avenues for future research.

Analyzing combinations of students' use of granularity, disciplinary core ideas, and causality

We used the clusters detected by HDBSCAN to develop a data-driven scoring rubric that applies a mutually exclusive approach to evaluate students' mechanistic reasoning. Each mechanistic reasoning statement was scored as having either *phenomenological* or *electronic* granularity, using either *structural* or *energetic* disciplinary core ideas, and applying one of three causality levels. However, HDBSCAN also identified several noise points in the data, which correspond to mixed uses of granularity, disciplinary core ideas, or causality. We had to assign these noise points to the highest possible mechanistic reasoning complexity, rather than accounting for mixed uses of granularity, disciplinary core ideas, and causality in prior analyses, to ensure sufficient training data for subsequent automated analyses. For example, if a student mentioned *phenomenological* and *electronic* ideas, the ML model scored students' mechanistic reasoning so far as *electronic*. This highlights that developing scoring rubrics inductively based on students' open-ended responses relies on the available data.

Future research could bridge this lack of combinations between different rubric categories. These combinations are theoretically possible; nevertheless, they are exceedingly rare in the collected dataset. For instance, future research could analyze how students integrate *phenomenological* and *electronic* ideas, as well as *structural* and *energetic* ideas. Figure 15 illustrates students' mechanistic reasoning regarding the plausibility of the deprotonation of methyl acetate. Students' mechanistic reasoning addressed *acids and bases* at an *electronic* and *relational* level. However, notable differences arise among the three examples when examining combinations of students' use of granularity, disciplinary core ideas, and causality.

Regarding students' use of granularity, the three examples differ in both the number of unpacked properties and their scalar level. Example 1 unpacks four properties: two at a *phenomenological* and two at an *electronic* level. Example 2 unpacks a single *electronic* property since it links stability inherently to resonance. Example 3 unpacks three properties: two at a *phenomenological* and one at an *electronic* level. The variation among the three examples highlights the significant heterogeneity that remains under *electronic* mechanistic reasoning in our analysis. Future research could examine how students link *phenomenological* and *electronic* properties. Specifically, it can uncover the explanatory black boxes students create when struggling to support *phenomenological* properties with *electronic* ones (Haskel-Ittah, 2023). This includes examining whether *electronic* properties are used to explain *phenomenological* ones or merely listed alongside them (Franovic et al., 2023; Noyes et al., 2022). For instance, the electropositivity of the nitrogen atom mentioned in Example 1 is used to justify the basicity of lithium-diisopropyl amide (LDA), reflecting an alignment between the unpacked *phenomenological* and *electronic* properties. Conversely, the neutral charge of the corresponding conjugate acid is simply a byproduct of the reaction. This *electronic* property is not employed to justify any *phenomenological* property, reflecting a concurrent but isolated use of these properties. Investigating the aligned versus isolated use of *phenomenological* and *electronic* properties could further uncover how students use granularity.

Similar differences in students' mechanistic reasoning arise when examining students' use of disciplinary core ideas. Example 1 relies solely on *structural* properties to justify the basicity of LDA. By contrast, the other two examples integrate *structural* and *energetic* properties to

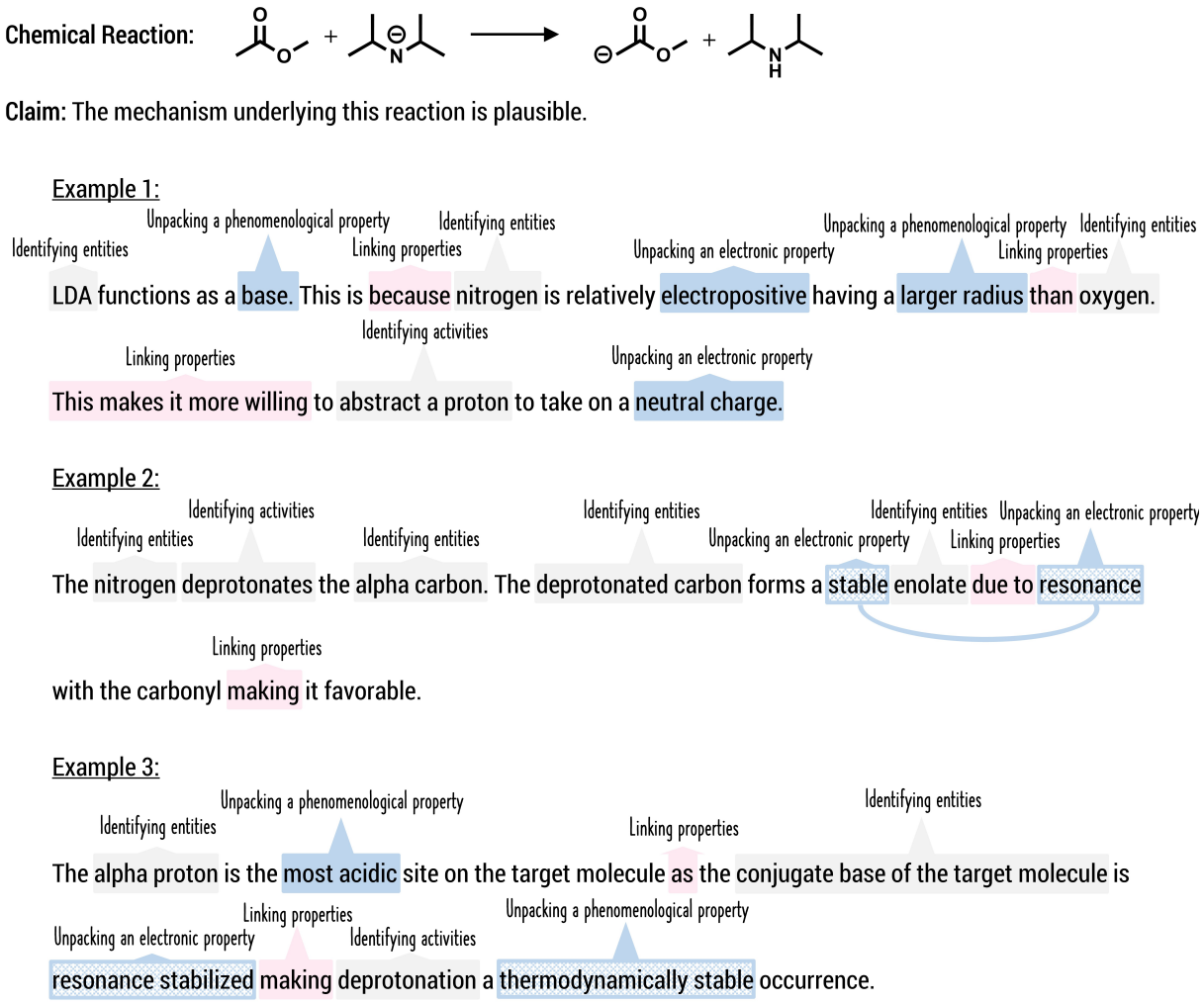


FIGURE 15 Examples of how students' mechanistic reasoning reflects mixed uses of granularity, disciplinary core ideas, and causality. *Note:* Shaded rectangles illustrate an *energetic* focus while unshaded rectangles represent a *structural* focus.

explain the acidity of the ester. Consequently, these latter two examples use the *structural* properties of entities to infer their *energetic* stability, which leads to claims about energy changes. Future research could investigate which *structural* properties are more frequently linked to *energetic* properties, whether the associated *energetic* properties pertain to an entity or an activity, and how students establish these connections (Pölloth et al., 2023). Since integrating thermodynamic and kinetic ideas into mechanistic reasoning has been shown to correlate with increased mechanistic reasoning complexity, analyzing how students connect *structural* and *energetic* properties could guide the design of scaffolds that help students make these connections.

Finally, the three examples share commonalities in their use of causality. All responses use their initial causal link to transition from the level of the target phenomenon to the scalar level below. Similarly, they employ their final causal link to connect the unpacked properties back to the original scalar level. This pattern indicates that these responses align with the *essential epistemic heuristics* and use causal links to transition between scalar levels. However, students can also establish causality in other ways, such as linking various disciplinary core ideas, even within the same scalar level, or comparing properties of entities as seen in Example 1 with its second link. Accordingly, our analysis identifies three types of causal links: transitions between scalar levels, connections

between disciplinary core ideas, and comparisons among properties. Future research could classify the types of causality integrated into students' mechanistic reasoning (Haskel-Ittah, 2024; Moreira et al., 2019). Understanding how causality is utilized in chemistry-related mechanistic reasoning could provide a basis for comparing students' mechanistic reasoning across domains.

In sum, addressing the limitation of our mutually exclusive scoring should be the next research step to move toward a more nuanced understanding of how students integrate granularity, disciplinary core ideas, and causality. Such an analysis could help develop finer-grained adaptive support, improve future scaffolding, compare students' mechanistic reasoning across domains, and assess the extent to which mechanistic reasoning and other skills like scientific literacy are associated.

Adjusting the applied scaffolding

We used the claim-evidence-reasoning model in our instructional setting to scaffold students' mechanistic reasoning (McNeill et al., 2006). However, our data-driven analysis suggests that the *essential epistemic heuristics* are more effective for analyzing students' mechanistic reasoning in our setting. Therefore, using these heuristics directly to scaffold students' mechanistic reasoning could be beneficial. Figure 16 illustrates an example of how such scaffolding might look.

The proposed scaffold asks students to identify the activities and entities in a given reaction by prompting them to focus on electron movements and identify which molecules act as acids/bases or nucleophiles/electrophiles. Hence, the first two prompts address the *essential epistemic heuristic* of considering the scalar level below the target phenomenon. Next, students are asked to unpack the properties that determine the function of molecules as acids/bases or

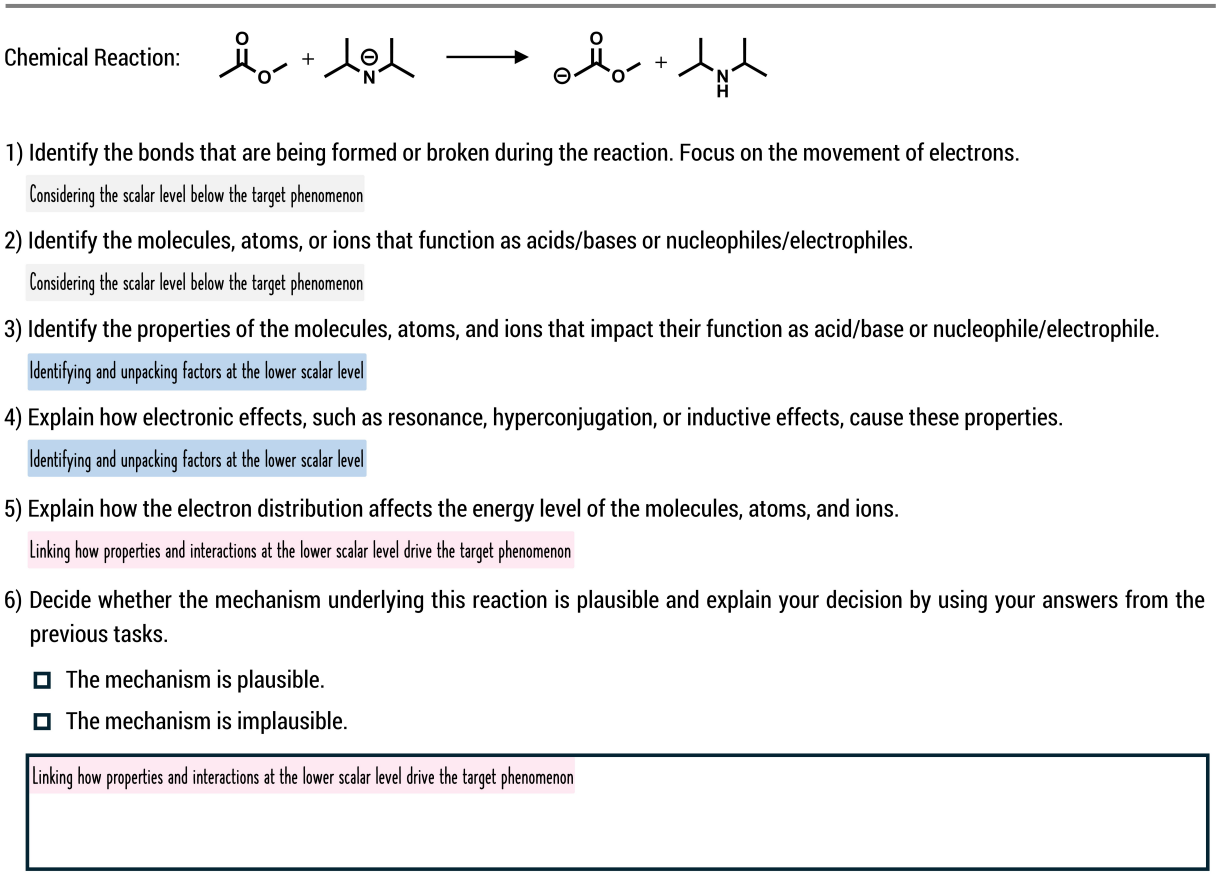


FIGURE 16 Using the *essential epistemic heuristics* to scaffold students' mechanistic reasoning.

nucleophiles/electrophiles, requiring a transition to the lower scalar level. The fourth prompt encourages students to link the unpacked *phenomenological* and *electronic* properties, which means the third and fourth prompts address the *essential epistemic heuristic* of identifying, unpacking, and linking factors at the lower scalar level. The fifth prompt asks students to explain how electronic properties affect the energy level of the involved molecules; thus, requiring them to connect these factors back to the target phenomenon as outlined in the third *essential epistemic heuristic*. As envisioned in this scaffold, linking factors back to the original scalar level in chemistry is inherently tied to connecting *structural* and *energetic* properties. This scaffold could guide students' mechanistic reasoning by facilitating transitions between scalar levels, unpacking domain-specific properties, and linking factors across and within scalar levels.

Future research could explore how the proposed scaffold affects students' mechanistic reasoning. Also, investigating ways to fade the scaffold gradually would be valuable (McNeill et al., 2006). For instance, it might be possible to combine the first and second prompts, as well as the third and fourth, over an instructional sequence and analyze how this affects students' mechanistic reasoning. Towards the end of a semester, the scaffold could be removed entirely, allowing researchers to examine students' mechanistic reasoning after scaffolding. Investigating the impact of a faded scaffold on students' mechanistic reasoning could reveal preferred durations for scaffolding, strategies for gradually removing a scaffold, and how students internalize a scaffold (Lieber et al., 2022a; Vo et al., 2024).

Investigating the development of mechanistic reasoning across domains

As proposed by Cooper (2020), applying a set of heuristics should begin with analyzing single phenomena, then extend to multiple phenomena within a single domain, and eventually span across domains. The analysis presented in this dissertation examined students' mechanistic reasoning across multiple phenomena within organic chemistry. As proposed in the previous sections, future research could provide a finer-grained understanding of students' mechanistic reasoning by analyzing combinations in their use of granularity, disciplinary core ideas, and causality. This may allow for comparing students' mechanistic reasoning across domains, such as biology and chemistry.

How mechanistic reasoning is operationalized varies slightly across domains. For example, biology often creates awareness for the first *essential epistemic heuristic* of considering a lower scalar level because the observed target phenomena are often visible while the lower scalar level deals with invisible entities (e.g., Haskel-Ittah & Duncan, 2021; Haskel-Ittah et al., 2020; Shiroda et al., 2024; Southard et al., 2017; van Mil et al., 2013). These differences in visibility emphasize the transition between scalar levels. In contrast, chemical target phenomena are at a molecular level, and the entities at the lower scalar level refer to an electronic level (e.g., Talanquer, 2018, 2022). Accordingly, the transition between these levels is often less explicit (Bachtiar et al., 2022). Nonetheless, chemistry tends to emphasize causality in mechanistic reasoning (Becker et al., 2016; Cooper et al., 2016; Crandell et al., 2019, 2020; Noyes & Cooper, 2019). For instance, *electronic* properties can be used to explain *phenomenological* properties and there is a need to connect structure and energy (Goodwin, 2003, 2008). Thus, the third *essential epistemic heuristic* of linking factors tends to be more emphasized in chemistry than in biology (Bachtiar et al., 2022).

Previous research has examined how biology and chemistry students' mechanistic reasoning differs. For example, Kohn et al. (2018a) demonstrated that students approach structure–

property–function relationships differently in their biology and chemistry courses. Some students even separated their understanding of energy concepts in these domains (Kohn et al., 2018b). In contrast, Franovic et al. (2023) reported no significant domain-specific differences in the mechanistic reasoning exhibited by students co-enrolled in biology and chemistry courses; however, these students relied on different cognitive resources in biology and chemistry. Shiroda et al. (2024) further analyzed how biology and chemistry students exhibited mechanistic reasoning across three tasks targeting different scalar levels. Students who successfully applied mechanistic reasoning in tasks involving phenomena at smaller scales were also better at applying it to larger scales, implying an association between mechanistic reasoning in biology and chemistry. However, when the scalar levels of the phenomena were too far apart, mechanistic reasoning did not necessarily translate (Shiroda et al., 2024).

These studies indicate that the relationship between students' mechanistic reasoning in biology and chemistry remains partly inconclusive. This uncertainty may stem from the reliance on broad categorizations of mechanistic reasoning (Franovic et al., 2023; Shiroda et al., 2024). Thus, future research could investigate the specific connections in students' use of granularity or causality across science domains. Such analyses are not limited to students' mechanistic reasoning in biology and chemistry but can also be applied to investigate mechanistic reasoning in physics, ecology, or life science contexts. Commonalities in students' mechanistic reasoning can ultimately be leveraged to assist students in explaining phenomena across domains while highlighting areas for advancing future interdisciplinary instructional support.

Investigating how mechanistic reasoning can contribute to scientific literacy

Mechanistic reasoning equips students with the skills to identify, describe, or analyze mechanisms to explain, predict, or control phenomena. In doing so, mechanistic reasoning establishes causal relationships across scalar levels and integrates domain-specific knowledge with domain-general strategies (Duncan, 2007; Haskel-Ittah et al., 2020; Krist et al., 2019; Russ et al., 2008). From a broader perspective, mechanistic reasoning is envisioned to foster scientific literacy (Haskel-Ittah, 2024). Although organic reaction mechanisms and everyday scientific phenomena are different domains, each rooted in different concept knowledge, it is essential to explore how mechanistic reasoning can contribute to scientific literacy.

Scientific literacy involves evaluating information to form well-supported claims about everyday phenomena, empowering individuals to make informed decisions and engage critically with scientific issues (Durant, 1994). Due to some shared characteristics, mechanistic reasoning may support students' scientific literacy (Haskel-Ittah, 2024). Just as mechanistic reasoning seeks to identify, describe, and analyze the underlying mechanisms of observed phenomena, scientific literacy demands that students gather, validate, and apply information. Similarly, just as mechanistic reasoning aims to explain, predict, or control phenomena, scientific literacy seeks to explain or evaluate sociopolitical issues. Therefore, mechanistic reasoning can help students move beyond factual recall and reliance on surface-level explanations to address contemporary sociopolitical challenges, such as evaluating ambiguous claims about climate change, vaccines, or renewable energy technologies.

Future studies could explore how students' mechanistic reasoning is associated with their ability to critically evaluate scientific claims in the media and formulate evidence-based arguments

on sociopolitical issues (Figure 17). Longitudinal tracking of the relationship between these constructs could help ensure that the identified connections are not merely artifacts of the specific context but are more generalizable. From a broader perspective, such an analysis could explore whether mechanistic reasoning is not only a tool for understanding scientific phenomena but also an epistemic practice that prepares scientifically literate individuals, equipped to navigate a complex, information-rich world.

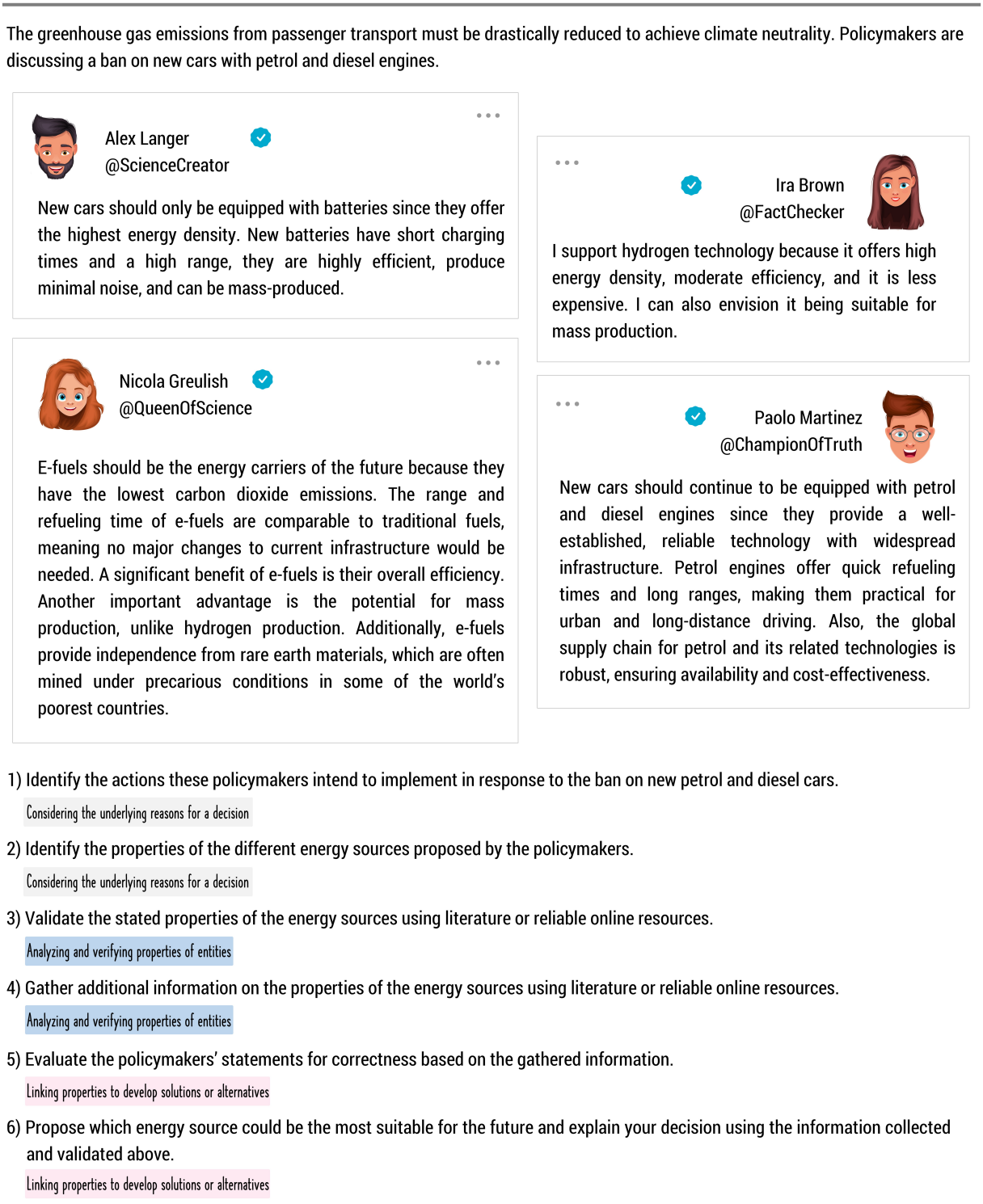


FIGURE 17 Exemplary task to assess students' scientific literacy guided by the *essential epistemic heuristics*. *Note*: The colors of the hypothesized activities correspond to those shown in Figure 16.

Developing a dashboard to guide the integration of mechanistic reasoning in teaching and learning

The findings from these analyses could ultimately be synthesized into a teaching dashboard to assist students in developing mechanistic reasoning. Generally, teaching dashboards are interactive digital tools that provide instructors with real-time data on student achievement to identify student challenges, monitor student progress, detect achievement gaps, and personalize instruction (e.g., Bayer et al., 2024; Karademir et al., 2024; Knoop-van Campen & Molenaar, 2020). Similarly, students could benefit from access to data on their progress to inform themselves about intended learning goals, areas for improvement, and future learning strategies (e.g., Reid & Drysdale, 2024).

Based on our analysis of students' mechanistic reasoning, a teaching dashboard could provide students with insights into specific learning objectives related to mechanistic reasoning and even personalize these objectives based on individual progress. Furthermore, the dashboard could visualize students' progress in mechanistic reasoning complexity and offer adaptive feedback to guide their improvement in future instruction. Additionally, the dashboard could suggest actionable steps for further progress by analyzing a student's use of granularity, disciplinary core ideas, and causality. In turn, instructors could receive cohort- and individual-level feedback on students' mechanistic reasoning complexity to adapt their instruction accordingly. For example, a dashboard may encourage instructors to revisit key chemistry concepts or address scientifically non-normative conceptions. Similarly, a dashboard could guide instructors in adopting evidence-based formative assessment items to emphasize mechanistic reasoning. A dashboard might even empower instructors to offer more individualized feedback (Karademir et al., 2024).

However, instructors' interpretations of the automated analysis will likely depend on their personal and enacted pedagogical content knowledge (Carlson et al., 2019). This raises the need to investigate the extent to which instructors adopt dashboard recommendations, especially when their pedagogical content knowledge differs from the insights provided by the dashboard. Research could uncover misalignments between what instructors prioritize in teaching and what an automated analysis emphasizes, highlighting areas where organic chemistry teaching or ML analyses could be refined. This may ultimately provide an opportunity to bridge gaps between chemistry education research and instructional practice.

References

Agresti, A. (2012). *Categorical data analysis*. Wiley.

Agresti, A. (2018). *An introduction to categorical data analysis*. Wiley.

Anderson, D., Rowley, B., Stegenga, S., Irvin, P. S., & Rosenberg, J. M. (2020). Evaluating content-related validity evidence using a text-based machine learning procedure. *Educational Measurement: Issues and Practice*, *39*(4), 53-64. https://doi.org/10.1111/emip.12314

Ariely, M., Nazaretsky, T., & Alexandron, G. (2023). Machine learning and Hebrew NLP for automated assessment of open-ended questions in biology. *International Journal of Artificial Intelligence in Education*, *33*(1), 1-34. https://doi.org/10.1007/s40593-021-00283-x

Ariely, M., Nazaretsky, T., & Alexandron, G. (2024). Causal-mechanical explanations in biology: Applying automated assessment for personalized learning in the science classroom. *Journal of Research in Science Teaching*, *61*(8), 1858-1889. https://doi.org/10.1002/tea.21929

Asmussen, G., Rodemer, M., & Bernholt, S. (2023). Blooming student difficulties in dealing with organic reaction mechanisms—An attempt at systemization. *Chemistry Education Research and Practice*, *24*(3), 1035-1054. https://doi.org/10.1039/D2RP00204C

Asmussen, G., Rodemer, M., & Bernholt, S. (2024). Stepping stones to success: A qualitative investigation of the effectiveness of adaptive stepped supporting tools for problem-solving in organic chemistry to design an intelligent tutoring system. *International Journal of Science*

Education, 1-23. https://doi.org/10.1080/09500693.2024.2361933

Bachtiar, R. W., Meulenbroeks, R. F. G., & van Joolingen, W. R. (2022). Mechanistic reasoning in science education: A literature review. *EURASIA Journal of Mathematics, Science and Technology Education*, *18*(11), Article em2178. https://doi.org/10.29333/ejmste/12512

Bayer, V., Mulholland, P., Hlosta, M., Farrell, T., Herodotou, C., & Fernandez, M. (2024). Co-creating an equality diversity and inclusion learning analytics dashboard for addressing awarding gaps in higher education. *British Journal of Educational Technology*, *55*(5), 2058-2074. https://doi.org/10.1111/bjet.13509

Becker, N., Noyes, K., & Cooper, M. (2016). Characterizing students' mechanistic reasoning about London dispersion forces. *Journal of Chemical Education*, *93*(10), 1713-1724. https://doi.org/10.1021/acs.jchemed.6b00298

Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & Boone, W. J. (2014). Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance? *Journal of Science Education and Technology*, *23*(1), 160-182. https://doi.org/10.1007/s10956-013-9461-9

Bejar, I. I., Mislevy, R. J., & Zhang, M. (2016). Automated scoring with validity in mind. In A. A. Rupp & J. P. Leighton (Eds.), *Handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 226-246). Wiley. https://doi.org/10.1002/9781118956588.ch10

Bhattacharyya, G., & Bodner, G. M. (2005). "It gets me to the product": How students propose organic mechanisms. *Journal of Chemical Education*, *82*(9), 1402-1407. https://doi.org/10.1021/ed082p1402

Bodé, N. E., Deng, J. M., & Flynn, A. B. (2019). Getting past the rules and to the WHY: Causal mechanistic arguments when judging the plausibility of organic reaction mechanisms. *Journal of Chemical Education*, *96*(6), 1068-1082. https://doi.org/10.1021/acs.jchemed.8b00719

Braun, I., & Graulich, N. (2024). Exploring diversity: Student's (un-)productive use of resonance in organic chemistry tasks through the lens of the coordination class theory. *Chemistry Education Research and Practice*, *25*(3), 643-671. https://doi.org/10.1039/D3RP00298E

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in knowledge discovery and data mining*. PAKDD 2013: Lecture Notes in Computer Science (pp. 160-172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14

Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, *10*(1), Article 5. https://doi.org/10.1145/2733381

Carlsen, H. B., & Ralund, S. (2022). Computational grounded theory revisited: From computer-led to computer-assisted text analysis. *Big Data & Society*, *9*(1), Article 20539517221080146. https://doi.org/10.1177/20539517221080146

Carlson, J., Daehler, K. R., Alonzo, A. C., Barendsen, E., Berry, A., Borowski, A., ... & Wilson, C. D. (2019). The refined consensus model of pedagogical content knowledge in science education. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 77-92). Springer Singapore. https://doi.org/10.1007/978-981-13-5898-2

Caspari, I., & Graulich, N. (2019). Scaffolding the structure of organic chemistry students' multivariate comparative mechanistic reasoning. *International Journal of Physics and Chemistry Education*, *11*(2), 31-43. https://doi.org/10.51724/ijpce.v11i2.39

Caspari, I., Kranz, D., & Graulich, N. (2018). Resolving the complexity of organic chemistry students' reasoning through the lens of a mechanistic framework. *Chemistry Education Research and Practice*, *19*(4), 1117-1141. https://doi.org/10.1039/C8RP00131F

Caspari, I., Weinrich, M., Sevian, H., & Graulich, N. (2018). This mechanistic step is *"productive"*: Organic chemistry students' backward-oriented reasoning. *Chemistry Education Research and Practice*, *19*(1), 42-59. https://doi.org/10.1039/C7RP00124J

Choi, A., Hand, B., & Greenbowe, T. (2013). Students' written arguments in general chemistry laboratory investigations. *Research in Science Education*, *43*(5), 1763-1783. https://doi.org/10.1007/s11165-012-9330-1

Cochran, K., Cohn, C., Hutchins, N., Biswas, G., & Hastings, P. (2022). Improving automated evaluation of formative assessments with text data augmentation. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial intelligence in education.* AIED 2022: Lecture Notes in Computer Science (pp. 393-405). Springer. https://doi.org/10.1007/978-3-031-11644-5_32

Cochran, K., Cohn, C., Rouet, J.-F., & Hastings, P. (2023). Improving automated evaluation of student text responses using GPT-3.5 for text data augmentation. In N. Wang, G. Rebolledo-

Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial intelligence in education*. AIED 2023: Lecture Notes in Computer Science (pp. 217-228). Springer. https://doi.org/10.1007/978-3-031-36272-9_18

Cooper, M. M. (2020). The crosscutting concepts: Critical component or "third wheel" of three-dimensional learning? *Journal of Chemical Education*, *97*(4), 903-909. https://doi.org/10.1021/acs.jchemed.9b01134

Cooper, M. M., Corley, L. M., & Underwood, S. M. (2013). An investigation of college chemistry students' understanding of structure–property relationships. *Journal of Research in Science Teaching*, *50*(6), 699-721. https://doi.org/10.1002/tea.21093

Cooper, M. M., Franovic, C. G. C., & Noyes, K. (in print). Design principles for developing and analyzing formative assessments to support mechanistic reasoning. In R. W. Bachtiar, M. Haskel-Ittah, & N. Graulich (Eds.), *Mechanistic Reasoning in Science Education*.

Cooper, M. M., Kouyoumdjian, H., & Underwood, S. M. (2016). Investigating students' reasoning about acid−base reactions. *Journal of Chemical Education*, *93*(10), 1703-1712. https://doi.org/10.1021/acs.jchemed.6b00417

Cooper, M. M., Posey, L. A., & Underwood, S. M. (2017). Core ideas and topics: Building up or drilling down? *Journal of Chemical Education*, *94*(5), 541-548. https://doi.org/10.1021/acs.jchemed.6b00900

Crandell, O. M., Kouyoumdjian, H., Underwood, S. M., & Cooper, M. M. (2019). Reasoning about reactions in organic chemistry: Starting it in general chemistry. *Journal of Chemical Education*, *96*(2), 213-226. https://doi.org/10.1021/acs.jchemed.8b00784

Crandell, O. M., Lockhart, M. A., & Cooper, M. M. (2020). Arrows on the page are not a good gauge: Evidence for the importance of causal mechanistic explanations about nucleophilic substitution in organic chemistry. *Journal of Chemical Education*, *97*(2), 313-327. https://doi.org/10.1021/acs.jchemed.9b00815

Crandell, O. M., & Pazicni, S. (2023). Leveraging cognitive resources to investigate the impact of molecular orientation on students' activation of symmetry resources. *Chemistry Education Research and Practice*, *24*(1), 353-368. https://doi.org/10.1039/D2RP00164K

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355-376. https://doi.org/10.1007/s11229-006-9097-x

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Clarendon Press.

Crowder, C. J., & Raker, J. R. (2024). Patterns in explanations of organic chemistry reaction mechanisms: A text analysis by level of explanation sophistication. *Journal of Chemical Education*, *101*(12), 5203-5220. https://doi.org/10.1021/acs.jchemed.4c01042

Dai, H., Liu, Z., Liao, W., Huang, X., Wu, Z., Zhao, L., … & Zhu, D. (2023). *AugGPT: Leveraging ChatGPT for text data augmentation* [preprint arXiv:2302.13007]. arXiv. https://doi.org/10.48550/arXiv.2302.13007

Darden, L. (2002). Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward/backward chaining. *Philosophy of Science*, *69*(S3), S354-S365. https://doi.org/10.1086/341858

Darden, L., & Craver, C. F. (2002). Strategies in the interfield discovery of the mechanism of protein synthesis. *Studies in History and Philosophy of Biological and Biomedical Sciences, 33*(1), 1-28. https://doi.org/10.1016/S1369-8486(01)00021-8

Deng, J. M., & Flynn, A. B. (2021). Reasoning, granularity, and comparisons in students' arguments on two organic chemistry items. *Chemistry Education Research and Practice*, *22*(3), 749-771. https://doi.org/10.1039/D0RP00320D

Deunk, M. I., Doolaard, S., Smale-Jacobse, A. E., & Bosker, R. J. (2015). *Differentiation within and across classrooms: A systematic review of studies into the cognitive effects of differentiation practices*. RUG/GION.

Deunk, M. I., Smale-Jacobse, A. E., de Boer, H., Doolaard, S., & Bosker, R. J. (2018). Effective differentiation practices: A systematic review and meta-analysis of studies on the cognitive effects of differentiation practices in primary education. *Educational Research Review*, *24*, 31-54. https://doi.org/10.1016/j.edurev.2018.02.002

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding* [preprint arXiv:1810.04805]. arXiv. https://doi.org/10.48550/arXiv.1810.04805

DiSessa, A. A., Sherin, B. L., & Levin, M. (2016). Knowledge analysis. In M. Levin & N. J. Brown (Eds.), *Knowledge and interaction: A synthetic agenda for the learning sciences* (pp. 30-71). Routledge.

Dood, A. J., Dood, J. C., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2020a). Analyzing explanations of substitution reactions using lexical analysis and logistic regression techniques. *Chemistry Education Research and Practice*, *21*(1), 267-286. https://doi.org/10.1039/C9RP00148D

Dood, A. J., Dood, J. C., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2020b). Using the research literature to develop an adaptive intervention to improve student explanations of an S_N1 reaction mechanism. *Journal of Chemical Education*, *97*(10), 3551-3562. https://doi.org/10.1021/acs.jchemed.0c00569

Dood, A. J., Fields, K. B., Cruz-Ramírez de Arellano, D., & Raker, J. R. (2019). Development and evaluation of a Lewis acid–base tutorial for use in postsecondary organic chemistry courses. *Canadian Journal of Chemistry*, *97*(10), 711-721. https://doi.org/10.1139/cjc-2018-0479

Dood, A. J., Fields, K. B., & Raker, J. R. (2018). Using lexical analysis to predict Lewis acid–base model use in response to an acid–base proton-transfer reaction. *Journal of Chemical Education*, *95*(8), 1267-1275. https://doi.org/10.1021/acs.jchemed.8b00177

Dood, A. J., & Watts, F. M. (2022). Mechanistic reasoning in organic chemistry: A scoping review of how students describe and explain mechanisms in the chemistry education research literature. *Journal of Chemical Education*, *99*(8), 2864-2876. https://doi.org/10.1021/acs.jchemed.2c00313

Dood, A. J., Watts, F. M., Connor, M. C., & Shultz, G. V. (2024). Automated text analysis of organic chemistry students' written hypotheses. *Journal of Chemical Education*, *101*(3), 807-818. https://doi.org/10.1021/acs.jchemed.3c00757

Dood, A. J., Winograd, B. A., Finkenstaedt-Quinn, S. A., Gere, A. R., & Shultz, G. V. (2022). PeerBERT: Automated characterization of peer review comments across courses. In *Proceedings of the 12th International Learning Analytics and Knowledge Conference* (LAK22) (pp. 492-499). Association for Computing Machinery. https://doi.org/10.1145/3506860.350689

Duncan, R. G. (2007). The role of domain-specific knowledge in generative reasoning about complicated multileveled phenomena. *Cognition and Instruction*, *25*(4), 271-336. https://doi.org/10.1080/07370000701632355

Durant, J. (1994). What is scientific literacy? *European Review*, *2*(1), 83-89. https://doi.org/10.1017/S1062798700000922

Esselman, B. J., Hill, N. J., DeGlopper, K. S., Ellison, A. J., Stowe, R. L., Schwarz, C. E., & Ellias, N. J. (2023). Authenticity-driven design of a high-enrollment organic laboratory course. *Journal of Chemical Education*, *100*(12), 4674-4685. https://doi.org/10.1021/acs.jchemed.3c00845

Fang, L., Lee, G.-G., & Zhai, X. (2023). *Using GPT-4 to augment unbalanced data for automatic scoring* [preprint arXiv:2310.18365]. arXiv. https://doi.org/10.48550/arXiv.2310.18365

Ferguson, R., & Bodner, G. M. (2008). Making sense of the arrow-pushing formalism among chemistry majors enrolled in organic chemistry. *Chemistry Education Research and Practice*, *9*(2), 102-113. https://doi.org/10.1039/B806225K

Flynn, A. B., & Featherstone, R. B. (2017). Language of mechanisms: Exam analysis reveals students' strengths, strategies, and errors when using the electron-pushing formalism (curved arrows) in new reactions. *Chemistry Education Research and Practice*, *18*(1), 64-77. https://doi.org/10.1039/C6RP00126B

Franovic, C. G.-C., Noyes, K., Stoltzfus, J. R., Schwarz, C. V., Long, T. M., & Cooper, M. M. (2023). Undergraduate chemistry and biology students' use of causal mechanistic reasoning to explain and predict preferential protein–ligand binding activity. *Journal of Chemical Education*, *100*(5), 1716-1727. https://doi.org/10.1021/acs.jchemed.2c00737

Frost, S. J. H., Yik, B. J., Dood, A. J., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2023). Evaluating electrophile and nucleophile understanding: A large-scale study of learners' explanations of reaction mechanisms. *Chemistry Education Research and Practice*, *24*(2), 706-722. https://doi.org/10.1039/D2RP00327A

Fuhrmann, T., Rosenbaum, L., Wagh, A., Eloy, A., Wolf, J., Blikstein, P., & Wilkerson, M. (2025). Right but wrong: How students' mechanistic reasoning and conceptual understandings shift when designing agent-based models using data. *Science Education*, *109*(1), 3-26. https://doi.org/10.1002/sce.21890

Fussell, R. K., Stump, E. M., & Holmes, N. G. (2024). Method to assess the trustworthiness of machine coding at scale. *Physical Review Physics Education Research*, *20*(1), Article 010113. https://doi.org/10.1103/PhysRevPhysEducRes.20.010113

Gao, S., Outlaw, T. C., Liang-Lin, J. G., Feng, A., Shimomura, R., Roizen, J. L., & Cox, C. T. (2024). Analysis of resources applied to rationalize elimination mechanisms. *Chemistry Education Research and Practice*, *25*(1), 62-78. https://doi.org/10.1039/D3RP00031A

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software*, *65*(4), 1-34. https://doi.org/10.18637/jss.v065.i04

Genolini, C., & Falissard, B. (2010). Kml: k-means for longitudinal data. *Computational Statistics*, *25*(2), 317-328. https://doi.org/10.1007/s00180-009-0178-4

Genolini, C., & Falissard, B. (2011). Kml: A package to cluster longitudinal data. *Computer Methods and*

Programs in Biomedicine, *104*(3), e112-e121. https://doi.org/10.1016/j.cmpb.2011.05.008

Glennan, S., & Illari, P. M. (2017). Varieties of mechanisms. In S. Glennan & P. M. Illari (Eds.), *Handbook of mechanisms and mechanical philosophy* (pp. 91-103). Routledge.

Gombert, S., Di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., … & Drachsler, H. (2023). Coding energy knowledge in constructed responses with explainable NLP models. *Journal of Computer Assisted Learning*, *39*(3), 767-786. https://doi.org/10.1111/jcal.12767

Goodwin, W. M. (2003). Explanation in organic chemistry. *Annals of the New York Academy of Sciences*, *988*(1), 141-153. https://doi.org/10.1111/j.1749-6632.2003.tb06093.x

Goodwin, W. M. (2008). Structural formulas and explanation in organic chemistry. *Foundations of Chemistry*, *10*(2), 117-127. https://doi.org/10.1007/s10698-007-9033-2

Graulich, N. (2025). The tip of the iceberg in organic chemistry – revisited. *Chemistry Education Research and Practice,* Advance Article. https://doi.org/10.1039/D4RP00345D

Graulich, N., & Caspari, I. (2021). Designing a scaffold for mechanistic reasoning in organic chemistry. *Chemistry Teacher International*, *3*(1), 19-30. https://doi.org/10.1515/cti-2020-0001

Graulich, N., Hedtrich, S., & Harzenetter, R. (2019). Explicit versus implicit similarity: Exploring relational conceptual understanding in organic chemistry. *Chemistry Education Research and Practice*, *20*(4), 924-936. https://doi.org/10.1039/C9RP00054B

Graulich, N., & Schween, M. (2018). Concept-oriented task design: Making purposeful case comparisons in organic chemistry. *Journal of Chemical Education*, *95*(3), 376-383. https://doi.org/10.1021/acs.jchemed.7b00672

Grotzer, T. A. (2003). Learning to understand the forms of causality implicit in scientifically accepted explanations. *Studies in Science Education*, *39*(1), 1-74. https://doi.org/10.1080/03057260308560195

Grove, N. P., Cooper, M. M., & Cox, E. L. (2012). Does mechanistic thinking improve student success in organic chemistry? *Journal of Chemical Education*, *89*(7), 850-853. https://doi.org/10.1021/ed200394d

Grove, N. P., Cooper, M. M., & Rush, K. M. (2012). Decorating with arrows: Toward the development of representational competence in organic chemistry. *Journal of Chemical Education*, *89*(7), 844-849. https://doi.org/10.1021/ed2003934

Haas, D. B., Watts, F. M., Dood, A. J., & Shultz, G. V. (2024). Analysis of organic chemistry students' developing reasoning elicited by a scaffolded case comparison activity. *Chemistry Education Research and Practice*, *25*(3), 742-759. https://doi.org/10.1039/D4RP00021H

Hammer, D., Elby, A., Scherr, R. E., & Redish, E. F. (2005). Resources, framing, and transfer. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 89-120). Information Age Publishing.

Haskel-Ittah, M. (2023). Explanatory black boxes and mechanistic reasoning. *Journal of Research in Science Teaching*, *60*(4), 915-933. https://doi.org/10.1002/tea.21817

Haskel-Ittah, M. (2024). Causal thinking in science education and the challenges it holds. In P. M. Illari & F. Russo (Eds.), *Handbook of causality and causal methods* (pp. 661-670). Routledge.

Haskel-Ittah, M., & Duncan, R. G. (2021). Undergraduate students reasoning about genetic mechanisms. In J. A. E. de Vries & Y. Hod (Eds.), *Proceedings of the 15th International Conference of the Learning Sciences – ICLS 2021* (pp. 107-114). International Society of the Learning Sciences.

Haskel-Ittah, M., Duncan, R. G., Vázquez-Ben, L., & Yarden, A. (2020). Reasoning about genetic mechanisms: Affordances and constraints for learning. *Journal of Research in Science Teaching*, *57*(3), 342-367. https://doi.org/10.1002/tea.21595

Haudek, K. C., Moscarella, R. A., Weston, M., Merrill, J. E., & Urban-Lurain, M. (2015). Construction of rubrics to evaluate content in students' scientific explanation using computerized text analysis. *Proceedings of the National Association of Research in Science Teaching Annual Conference*, Chicago, IL.

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J. E., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid–base chemistry in introductory biology. *CBE—Lifes Science Education*, *11*(3), 283-293. https://doi.org/10.1187/cbe.11-08-0084

Hedtrich, S., & Graulich, N. (2018a). Translating numbers into feedback: Providing students with automatically generated feedback. *American Journal of Educational Research*, *6*(2), 108-116. https://doi.org/10.12691/education-6-2-3

Hedtrich, S., & Graulich, N. (2018b). Using software tools to provide students in large classes with individualized formative feedback. *Journal of Chemical Education*, *95*(12), 2263-2267. https://doi.org/10.1021/acs.jchemed.8b00173

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength natural language processing in Python* [Computer software]. https://spacy.io/

Houchlei, S. K., Bloch, R. R., & Cooper, M. M. (2021). Mechanisms, models, and explanations: Analyzing the mechanistic paths students take to reach a product for familiar and unfamiliar organic reactions. *Journal of Chemical Education*, *98*(9), 2751-2764. https://doi.org/10.1021/acs.jchemed.1c00099

Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*(1), 119-135. https://doi.org/10.1007/s13194-011-0038-2

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes, and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 624-635). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445923

Karademir, O., Borgards, L., Di Mitri, D., Strauß, S., Kubsch, M., Brobeil, M., … & Drachsler, H. (2024). Following the impact chain of the LA cockpit: An intervention study investigating a teacher dashboard's effect on student learning. *Journal of Learning Analytics*, *11*(2), 215-228. https://doi.org/10.18608/jla.2024.8399

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., … & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, Article 100074. https://doi.org/10.1016/j.caeai.2022.100074

Kieser, F., Tschisgale, P., Rauh, S., Bai, X., Maus, H., Petersen, S., … & Wulff, P. (2024). David vs. Goliath: Comparing conventional machine learning and a large language model for assessing students' concept use in a physics problem. *Frontiers in Artificial Intelligence*, *7*, Article 1408817. https://doi.org/10.3389/frai.2024.1408817

Knoop-van Campen, C., & Molenaar, I. (2020). How teachers integrate dashboards into their feedback practices. *Frontline Learning Research*, *8*(4), 37-51. https://doi.org/10.14786/flr.v8i4.641

Koedinger, K. R., Carvalho, P. F., Liu, R., & McLaughlin, E. A. (2023). An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences*, *120*(13), Article e2221311120. https://doi.org/doi:10.1073/pnas.2221311120

Kohn, K. P., Underwood, S. M., & Cooper, M. M. (2018a). Connecting structure–property and structure–function relationships across the disciplines of chemistry and biology: Exploring student perceptions. *CBE—Life Sciences Education*, *17*(2), Article ar33. https://doi.org/10.1187/cbe.18-01-0004

Kohn, K. P., Underwood, S. M., & Cooper, M. M. (2018b). Energy connections and misconnections across chemistry and biology. *CBE—Life Sciences Education*, *17*(1), Article ar3. https://doi.org/10.1187/cbe.17-08-0169

Kraft, A., Strickland, A. M., & Bhattacharyya, G. (2010). Reasonable reasoning: Multi-variate problem-solving in organic chemistry. *Chemistry Education Research and Practice*, *11*(4), 281-292. https://doi.org/10.1039/C0RP90003F

Kranz, D., Martin, P. P., Schween, M., & Graulich, N. (under review). Should we scaffold it? Analysing students' learning gains to evaluate the effect of task format and scaffolding. *Chemistry Education Research and Practice*.

Kranz, D., Schween, M., & Graulich, N. (2023). Patterns of reasoning: Exploring the interplay of students' work with a scaffold and their conceptual knowledge in organic chemistry. *Chemistry Education Research and Practice*, *24*(2), 453-477. https://doi.org/10.1039/D2RP00132B

Krist, C., Schwarz, C. V., & Reiser, B. J. (2019). Identifying essential epistemic heuristics for guiding mechanistic reasoning in science learning. *Journal of the Learning Sciences*, *28*(2), 160-205. https://doi.org/10.1080/10508406.2018.1510404

Kubsch, M., Krist, C., & Rosenberg, J. M. (2023). Distributing epistemic functions and tasks—A framework for augmenting human analytic power with machine learning in science education research. *Journal of Research in Science Teaching*, *60*(2), 423-447. https://doi.org/10.1002/tea.21803

Küchemann, S., Klein, P., Becker, S., Kumari, N., & Kuhn, J. (2020). Classification of students' conceptual understanding in STEM education using their visual attention distributions: A comparison of three machine-learning approaches. In *Proceedings of the 12th International Conference on Computer Supported Education 2020* (pp. 36-46). SciTePress. https://doi.org/10.5220/0009359400360046

Lieber, L. S., & Graulich, N. (2020). Thinking in alternatives—A task design for challenging students' problem-solving approaches in organic chemistry. *Journal of Chemical Education*, *97*(10), 3731-3738. https://doi.org/10.1021/acs.jchemed.0c00248

Lieber, L. S., & Graulich, N. (2022). Investigating students' argumentation when judging the plausibility of alternative reaction pathways in organic chemistry. *Chemistry Education Research and Practice*, *23*(1), 38-53. https://doi.org/10.1039/D1RP00145K

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022a). Closing the gap of organic chemistry students' performance with an adaptive scaffold for argumentation patterns. *Chemistry Education Research and Practice*, *23*(4), 811-828. https://doi.org/10.1039/D2RP00016D

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022b). Students' individual needs matter: A training to adaptively address students' argumentation skills in organic chemistry. *Journal of Chemical Education*, *99*(7), 2754-2761. https://doi.org/10.1021/acs.jchemed.2c00213

Lo Piano, S. (2020). Ethical principles in machine learning and artificial intelligence: Cases from the field and possible ways forward. *Humanities and Social Sciences Communications*, *7*(1), Article 9. https://doi.org/10.1057/s41599-020-0501-9

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4765-4774). Curran Associates, Inc. https://doi.org/10.48550/arXiv.1705.07874

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms *Philosophy of Science*, *67*(1), 1-25. https://doi.org/10.1086/392759

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B., & Krajcik, J. (2021). Using machine learning to score multi-dimensional assessments of chemistry and physics. *Journal of Science Education and Technology*, *30*(2), 239-254. https://doi.org/10.1007/s10956-020-09895-9

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia Medica*, *23*(2), 143-149. https://doi.org/10.11613/bm.2013.018

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, *2*(11), 205-206. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for dimension reduction* [preprint arXiv:1802.03426]. arXiv. https://doi.org/10.48550/arXiv.1802.03426

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software*, *3*(29), 861-862. https://doi.org/10.21105/joss.00861

McNeill, K. L., & Krajcik, J. (2011). *Supporting grade 5-8 students in constructing explanations in science: The claim, evidence, and reasoning framework for talk and writing*. Pearson Allyn & Bacon.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences*, *15*(2), 153-191. https://doi.org/10.1207/s15327809jls1502_1

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*(10), 955-966. https://doi.org/10.1037/0003-066X.30.10.955

Messick, S. (1979). Test validity and the ethics of assessment. *ETS Research Report Series*, *1979*(1), i-43. https://doi.org/10.1002/j.2333-8504.1979.tb01178.x

Møller, A. G., Dalsgaard, J. A., Pera, A., & Aiello, L. M. (2023). *The parrot dilemma: Human-labeled vs. LLM-augmented data in classification tasks* [preprint arXiv:2304.13861]. arXiv. https://doi.org/10.48550/arXiv.2304.13861

Moon, A., Stanford, C., Cole, R., & Towns, M. (2016). The nature of students' chemical reasoning employed in scientific argumentation in physical chemistry. *Chemistry Education Research and Practice*, *17*(2), 353-364. https://doi.org/10.1039/C5RP00207A

Moreira, P., Marzabal, A., & Talanquer, V. (2019). Using a mechanistic framework to characterise chemistry students' reasoning in written explanations. *Chemistry Education Research and Practice*, *20*(1), 120-131. https://doi.org/10.1039/C8RP00159F

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academic Press.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, *21*(1), 183-196. https://doi.org/10.1007/s10956-011-9300-9

Nelson, L. K. (2020). Computational Grounded Theory: A methodological framework. *Sociological Methods & Research*, *49*(1), 3-42. https://doi.org/10.1177/0049124117729703

Nieswandt, M. (2007). Student affect and conceptual understanding in learning chemistry. *Journal of Research in Science Teaching*, *44*(7), 908-937. https://doi.org/10.1002/tea.20169

Noyes, K., Carlson, C. G.-C., Stoltzfus, J. R., Schwarz, C. V., Long, T. M., & Cooper, M. M. (2022). A deep look into designing a task and coding scheme through the lens of causal mechanistic reasoning *Journal of Chemical Education*, *99*(2), 874-885. https://doi.org/10.1021/acs.jchemed.1c00959

Noyes, K., & Cooper, M. M. (2019). Investigating student understanding of London dispersion forces: A longitudinal study. *Journal of Chemical Education*, *96*(9), 1821-1832. https://doi.org/10.1021/acs.jchemed.9b00455

Noyes, K., McKay, R. L., Neumann, M., Haudek, K. C., & Cooper, M. M. (2020). Developing computer resources to automate analysis of students' explanations of London dispersion forces. *Journal of Chemical Education*, *97*(11), 3923-3936. https://doi.org/10.1021/acs.jchemed.0c00445

Pabuccu, A., & Erduran, S. (2017). Beyond rote learning in organic chemistry: The infusion and impact of argumentation in tertiary education. *International Journal of Science Education*, *39*(9), 1154-1172. https://doi.org/10.1080/09500693.2017.1319988

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 33* (pp. 8024-8035). Curran Associates, Inc. https://dl.acm.org/doi/abs/10.5555/3454287.3455008

Perkins, D. N., & Grotzer, T. A. (2005). Dimensions of causal understanding: The role of complex causal models in students' understanding of science. *Studies in Science Education*, *41*(1), 117-165. https://doi.org/10.1080/03057260508560216

Pölloth, B., Diekemper, D., Bosch, C., & Schwarzer, S. (2024). Reply to the 'Comment on "What resources do high school students activate to link energetic and structural changes in chemical reactions? A qualitative study"' by K. S. Taber, Chem. Educ. Res. Pract., 2024, 25. *Chemistry Education Research and Practice*, *25*(3), 958-965. https://doi.org/10.1039/D4RP00031E

Pölloth, B., Diekemper, D., & Schwarzer, S. (2023). What resources do high school students activate to link energetic and structural changes in chemical reactions? A qualitative study. *Chemistry Education Research and Practice*, *24*(4), 1153-1173. https://doi.org/10.1039/D3RP00068K

Prevost, L. B., Haudek, K. C., Merrill, J. E., & Urban-Lurain, M. (2012). Examining student constructed explanations of thermodynamics using lexical analysis. *Proceedings of the 42nd IEEE Frontiers in Education Conference Proceedings*, Seattle, WA.

Raker, J. R., Yik, B. J., & Dood, A. J. (2022). Development of a generalizable framework for machine learning-based evaluation of written explanations of reaction mechanisms from the postsecondary organic chemistry curriculum. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (pp. 304-319). The Royal Society of Chemistry. https://doi.org/10.1039/9781839167782-00304

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. SAGE.

Reid, D. P., & Drysdale, T. D. (2024). Student-facing learning analytics dashboard for remote lab practical work. *IEEE Transactions on Learning Technologies*, *17*, 1037-1050. https://doi.org/10.1109/TLT.2024.3354128

Rodriguez, J.-M. G., Stricker, A. R., & Becker, N. M. (2020). Students' interpretation and use of graphical representations: Insights afforded by modeling the varied population schema as a coordination class. *Chemistry Education Research and Practice*, *21*(2), 536-560. https://doi.org/10.1039/C9RP00249A

Rosenberg, J. M., & Krist, C. (2021). Combining machine learning and qualitative methods to elaborate students' ideas about the generality of their model-based explanations. *Journal of Science Education and Technology*, *30*(2), 255-267. https://doi.org/10.1007/s10956-020-09862-4

Rotich, F., Ward, L., Beck, C., & Popova, M. (2024). Attention is currency: How surface features of Lewis structures influence organic chemistry student reasoning about stability. *Chemistry Education Research and Practice*, *25*(4), 1071-1089. https://doi.org/10.1039/D4RP00030G

Rupp, A. A. (2018). Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied Measurement in Education*, *31*(3), 191-214. https://doi.org/10.1080/08957347.2018.1464448

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education*, *92*(3), 499-525. https://doi.org/10.1002/sce.20264

Sbeglia, G. C., & Nehm, R. H. (2024). Building conceptual and methodological bridges between SSE's diversity, equity, and inclusion statement and educational actions in evolutionary biology. *Evolution*, *78*(5), 809-820. https://doi.org/10.1093/evolut/qpae026

Scharlott, L. J., Rippey, D. W., Rosa, V., & Becker, N. M. (2024). Progression toward causal mechanistic reasoning through phenomenon-based learning in introductory chemistry. *Journal of Chemical Education*, *101*(3), 777-788. https://doi.org/10.1021/acs.jchemed.3c00517

Sevian, H., & Talanquer, V. (2014). Rethinking chemistry: A learning progression on chemical

thinking. *Chemistry Education Research and Practice*, *15*(1), 10-23. https://doi.org/10.1039/C3RP00111C

Sherin, B. (2013). A computational study of commonsense science: An exploration in the automated analysis of clinical interview data. *Journal of the Learning Sciences*, *22*(4), 600-638. https://doi.org/10.1080/10508406.2013.836654

Shiroda, M., Franovic, C. G.-C., de Lima, J., Noyes, K., Babi, D., Beltran-Flores, E., … & Stoltzfus, J. R. (2024). Examining and supporting mechanistic explanations across chemistry and biology courses. *CBE—Life Sciences Education*, *23*(3), Article ar38. https://doi.org/10.1187/cbe.23-08-0157

Smale-Jacobse, A. E., Meijer, A., Helms-Lorenz, M., & Maulana, R. (2019). Differentiated instruction in secondary education: A systematic review of research evidence. *Frontiers in Psychology*, *10,* Article 2366. https://doi.org/10.3389/fpsyg.2019.02366

Southard, K. M., Espindola, M. R., Zaepfel, S. D., & Bolger, M. S. (2017). Generative mechanistic explanation building in undergraduate molecular and cellular biology. *International Journal of Science Education*, *39*(13), 1795-1829. https://doi.org/10.1080/09500693.2017.1353713

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360-407. https://www.jstor.org/stable/747612

Talanquer, V. (2018). Exploring mechanistic reasoning in chemistry. In J. Yeo, T. W. Teo, & K.-S. Tang (Eds.), *Science education research and practice in Asia-Pacific and beyond* (pp. 39-52). Springer Singapore. https://doi.org/10.1007/978-981-10-5149-4_3

Talanquer, V. (2022). The complexity of reasoning about and with chemical representations. *JACS Au*, *2*(12), 2658-2669. https://doi.org/10.1021/jacsau.2c00498

Tschisgale, P., Wulff, P., & Kubsch, M. (2023). Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review Physics Education Research*, *19*(2), Article 020123. https://doi.org/10.1103/PhysRevPhysEducRes.19.020123

van Mil, M. H. W., Boerwinkel, D. J., & Waarlo, A. J. (2013). Modelling molecular mechanisms: A framework of scientific reasoning to construct molecular-level explanations for cellular behaviour. *Science & Education*, *22*(1), 93-118. https://doi.org/10.1007/s11191-011-9379-7

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450-469. https://doi.org/10.1093/pan/mpq025

Vo, K., Sarkar, M., White, P. J., & Yuriev, E. (2024). Development of problem-solving skills supported by metacognitive scaffolding: Insights from students' written work. *Chemistry Education Research and Practice*, *25*(4), 1197-1209. https://doi.org/10.1039/D3RP00284E

Wang, C., Liu, X., Wang, L., Sun, Y., & Zhang, H. (2021). Automated scoring of Chinese grades 7-9 students' competence in interpreting and arguing from evidence. *Journal of Science Education and Technology*, *30*(2), 269-282. https://doi.org/10.1007/s10956-020-09859-z

Watts, F. M., Dood, A. J., & Shultz, G. V. (in print). Using unsupervised machine learning to explore organic chemistry students' mechanistic reasoning in Writing-to-Learn assignments. In R. W. Bachtiar, M. Haskel-Ittah, & N. Graulich (Eds.), *Mechanistic Reasoning in Science Education*.

Watts, F. M., Dood, A. J., & Shultz, G. V. (2022). Developing machine learning models for automated analysis of organic chemistry students' written descriptions of organic reaction mechanisms. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (pp. 285-303). The Royal Society of Chemistry. https://doi.org/10.1039/9781839167782-00285

Watts, F. M., Schmidt-McCormack, J. A., Wilhelm, C. A., Karlin, A., Sattar, A., Thompson, B. C., … & Shultz, G. V. (2020). What students write about when students write about mechanisms: Analysis of features present in students' written descriptions of an organic reaction mechanism. *Chemistry Education Research and Practice*, *21*(4), 1148-1172. https://doi.org/10.1039/C9RP00185A

Watts, F. M., Zaimi, I., Kranz, D., Graulich, N., & Shultz, G. V. (2021). Investigating students' reasoning over time for case comparisons of acyl transfer reaction mechanisms. *Chemistry Education Research and Practice*, *22*(2), 364-381. https://doi.org/10.1039/D0RP00298D

Weinrich, M. L., & Sevian, H. (2017). Capturing students' abstraction while solving organic reaction mechanism problems across a semester. *Chemistry Education Research and Practice*, *18*(1), 169-190. https://doi.org/10.1039/C6RP00120C

Weinrich, M. L., & Talanquer, V. (2016). Mapping students' modes of reasoning when thinking about chemical reactions used to make a desired

product. *Chemistry Education Research and Practice*, *17*(2), 394-406. https://doi.org/10.1039/C5RP00208G

Wulff, P., Buschhüter, D., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2022). Bridging the gap between qualitative and quantitative assessment in science education research with machine learning—A case for pretrained language models-based clustering. *Journal of Science Education and Technology*, *31*(4), 490-513. https://doi.org/10.1007/s10956-022-09969-w

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2023). Utilizing a pretrained language model (BERT) to classify preservice physics teachers' written reflections. *International Journal of Artificial Intelligence in Education*, *33*(3), 439-466. https://doi.org/10.1007/s40593-022-00290-6

Wulff, P., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2023). Enhancing writing analytics in science education research with machine learning and natural language processing—Formative assessment of science and non-science preservice teachers' written reflections. *Frontiers in Education*, *7*, Article 1061461. https://doi.org/10.3389/feduc.2022.1061461

Yan, F., & Talanquer, V. (2015). Students' ideas about how and why chemical reactions happen: Mapping the conceptual landscape. *International Journal of Science Education*, *37*(18), 3066-3092. https://doi.org/10.1080/09500693.2015.1121414

Yik, B. J., Dood, A. J., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2021). Development of a machine learning-based tool to evaluate correct Lewis acid–base model use in written responses to open-ended formative assessment items. *Chemistry Education Research and Practice*, *22*(4), 866-885. https://doi.org/10.1039/D1RP00111F

Yik, B. J., Dood, A. J., Frost, S. J. H., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2023). Generalized rubric for level of explanation sophistication for nucleophiles in organic chemistry reaction mechanisms. *Chemistry Education Research and Practice*, *24*(1), 263-282. https://doi.org/10.1039/D2RP00184E

Zaimi, I., Dood, A. J., & Shultz, G. V. (2024). The evolution of an assignment: How a Writing-to-Learn assignment's design shapes organic chemistry students' elaborations on reaction mechanisms. *Chemistry Education Research and Practice*, *25*(1), 327-342. https://doi.org/10.1039/D3RP00197K

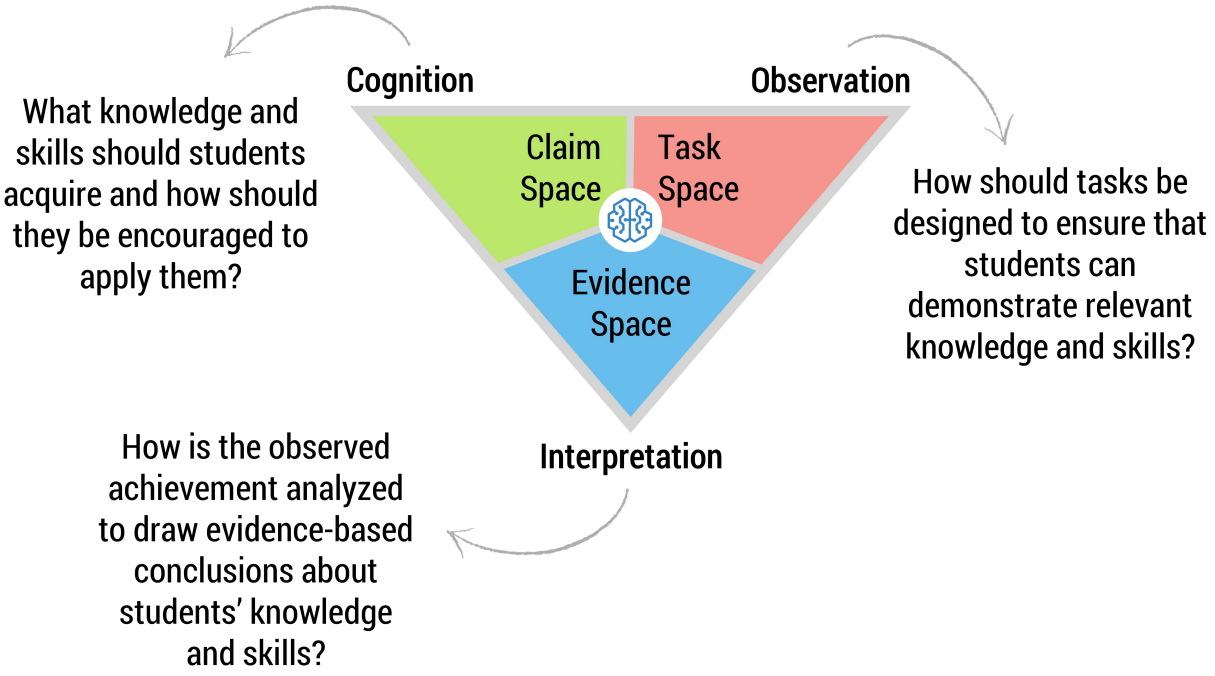
Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement*, *76*(2), 280-303. https://doi.org/10.1177/0013164415590022

Zhai, X. (2021). Practices and theories: How can machine learning assist in innovative assessment practices in science education. *Journal of Science Education and Technology*, *30*(2), 139-149. https://doi.org/10.1007/s10956-021-09901-8

Zhai, X., Krajcik, J., & Pellegrino, J. W. (2021). On the validity of machine learning-based next generation science assessments: A validity inferential network. *Journal of Science Education and Technology*, *30*(2), 298-312. https://doi.org/10.1007/s10956-020-09879-9

2 | When a machine detects student reasoning: A review of machine learning-based formative assessment of mechanistic reasoning

Chapter 2 presents a review article. It is reproduced from Martin and Graulich (2023) with permission from the Royal Society of Chemistry. No changes have been made to the original article.



Martin, P. P., & Graulich, N. (2023). When a machine detects student reasoning: A review of machine learning-based formative assessment of mechanistic reasoning. *Chemistry Education Research and Practice*, *24*(2), 407-427. https://doi.org/10.1039/D2RP00287F

Chemistry Education Research and Practice

REVIEW ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Educ. Res. Pract.*, 2023, **24**, 407

When a machine detects student reasoning: a review of machine learning-based formative assessment of mechanistic reasoning

Paul P. Martin and Nicole Graulich *

Received 21st October 2022,
Accepted 28th January 2023

DOI: 10.1039/d2rp00287f

rsc.li/cerp

In chemistry, reasoning about the underlying mechanisms of observed phenomena lies at the core of scientific practices. The process of uncovering, analyzing, and interpreting mechanisms for explanations and predictions requires a specific kind of reasoning: mechanistic reasoning. Several frameworks have already been developed that capture the aspects of mechanistic reasoning to support its formative assessment. However, evaluating mechanistic reasoning in students' open responses is a time- and resource-intense, complex, and challenging task when performed by hand. Emerging technologies like machine learning (ML) can automate and advance the formative assessment of mechanistic reasoning. Due to its usefulness, ML has already been applied to assess mechanistic reasoning in several research projects. This review focuses on 20 studies dealing with ML in chemistry education research capturing mechanistic reasoning. We developed a six-category framework based on the evidence-centered design (ECD) approach to evaluate these studies in terms of *pedagogical purpose*, *rubric design*, *construct assessment*, *validation approaches*, *prompt structure*, and *sample heterogeneity*. Contemporary effective practices of ML-based formative assessment of mechanistic reasoning in chemistry education are emphasized to guide future projects by these practices and to overcome challenges. Ultimately, we conclude that ML has advanced replicating, automating, and scaling human scoring, while it has not yet transformed the quality of evidence drawn from formative assessments.

Introduction

Across all sciences, mechanistic reasoning is a critical thinking skill (Machamer *et al.*, 2000; Glennan, 2002; Russ *et al.*, 2008; Bolger *et al.*, 2012; Illari and Williamson, 2012; Southard *et al.*, 2016; van Mil *et al.*, 2016). However, students experience significant challenges in generating mechanistic explanations in science subjects (*cf.*, Bachtiar *et al.*, 2022). Research in chemistry education has as well documented that these challenges include reasoning about chemical mechanisms in organic (*cf.*, Graulich, 2015; Dood and Watts, 2022a, 2022b) or general chemistry (*cf.*, Talanquer, 2009; Sevian and Talanquer, 2014) since many cognitive operations must co-occur. For example, when solving mechanistic problems, students must reason about submicroscopic processes, think of multiple influences, and weigh different variables (Machamer *et al.*, 2000; Russ *et al.*, 2008). Due to the importance of all those skills for a profound chemical understanding, a major goal of chemistry education is supporting students in constructing causal, mechanistic explanations (Cooper, 2015). Consequently, teaching needs to provide opportunities for students to reason about *how* and *why* mechanisms occur to emphasize the importance of a process-oriented understanding (Cooper *et al.*, 2016).

However, students will only construct such explanations and develop the desired, cognitively demanding reasoning skills if these skills are regularly assessed (Stowe and Cooper, 2017; Stowe *et al.*, 2021; DeGlopper *et al.*, 2022). Traditional chemistry assessment tends to emphasize rote learning strategies, such as recall, application of simple algorithms, and pattern recognition (Stowe and Cooper, 2017). For this reason, students may rely on memorization in chemistry classes (Grove and Lowery Bretz, 2012). Accordingly, instructors need tools for formative assessment that capture the quality of mechanistic reasoning. Such tools can be assessment items that allow students to explain *how* and *why* mechanisms occur (Cooper *et al.*, 2016) so that instructors can regularly assess and foster the desired reasoning skills.

Developing high-quality science assessments that elicit an extensive set of competencies as well as content knowledge and give adequate feedback is challenging in daily teaching practices (Songer and Ruiz-Primo, 2012; Pellegrino, 2013; Pellegrino *et al.*, 2016). Creating valid and reliable tools for drawing evidence-based inferences about mechanistic reasoning and individual

Justus-Liebig-University Giessen, Institute of Chemistry Education, Heinrich-Buff-Ring 17, 35392 Giessen, Germany.
E-mail: Nicole.Graulich@didaktik.chemie.uni-giessen.de

This journal is © The Royal Society of Chemistry 2023

View Article Online
Review Article Chemistry Education Research and Practice

support needs is particularly challenging. So far, closed-response questions, for example, single- or multiple-choice items, have often been the standard type of assessment in large classrooms. These items can be evaluated quickly, but the forced choice comes with validity threats (Birenbaum and Tatsuoka, 1987; Kuechler and Simkin, 2010; Lee *et al.*, 2011). Students may just guess the correct answer, preventing future learning progression. Additionally, closed-response questions cannot assess students' normative scientific and naïve ideas as accurately as human- or computer-scored open-ended items (Beggrow *et al.*, 2014).

Hence, science education research (*e.g.*, Haudek *et al.*, 2019) calls for the application of open-ended assessments such as constructed responses, essays, simulations, educational games, and interdisciplinary assessments since these formats may promote detailed explanations of chemical phenomena. Until now, resource constraints like workload, costs, and time for human evaluation have often prevented teachers and faculty members from using such formative assessments regularly in large-enrollment courses. Machine learning (ML) can be used to increase the application of formative open-ended assessments (*cf.*, Zhai *et al.*, 2020a, 2020c) because it offers new approaches for capturing student understanding, facilitating immediate decision-making as well as action-taking (Zhai, 2021). Notably, ML eases human effort in scoring, improves approaches to evidence collection, and has great potential in tapping manifold constructs (Zhai, 2019), which enables more reasoning-focused assessment and teaching (for more information on ML terms, see the glossary in Appendix).

The primary goal of this review is to illustrate the advancements and shortcomings that ML has brought to the formative assessment of mechanistic reasoning. For that purpose, we discuss six interdependent categories: *pedagogical purpose of ML application*, *rubric design*, *construct assessment*, *validation approaches*, *prompt structure*, and *sample heterogeneity*. We emphasize the importance of these categories for designing and implementing formative assessments and provide an in-depth analysis of how mechanistic reasoning is captured in ML-based chemistry assessments. Three questions guide the objectives of this review: First, what contemporary effective practices can be deduced from current research projects? Second, how did the selected research projects already advance the field of ML-based formative chemistry assessment capturing mechanistic reasoning? Third, what are the shortcomings in applying ML to formative chemistry assessment capturing mechanistic reasoning? After analyzing the selected studies, we suggest implications for the future implementation of ML in formative assessments to extend assessment functionality.

Theoretical considerations

Evidence-centered design approach

The discussion of ML-based formative assessment is, in this review, anchored in the evidence-centered design (ECD) approach to educational assessment (*e.g.*, Mislevy *et al.*, 2003a, 2003b; Mislevy, 2006; Mislevy and Haertel, 2007; Rupp *et al.*, 2012; Riconscente *et al.*, 2015; Pellegrino *et al.*, 2016; Kubsch *et al.*, 2022b). ECD offers a conceptual framework for assessment practices in which evidence is gathered and interpreted in coherence with the intended purpose of the assessment (Mislevy *et al.*, 2003a). Following this, ECD considers assessments as a theory-guided process of evidentiary reasoning aiming at drawing coherent conclusions from the things students say, write, or make to their competencies (Mislevy *et al.*, 2003a; Mislevy and Haertel, 2007; Mislevy, 2016; Pellegrino *et al.*, 2016). Drawing such evidentiary conclusions helps bring the expectations about students' domain-specific competencies in line with the observations needed to evidence them (Glaser *et al.*, 1987; Mislevy *et al.*, 2003a).

Influenced by these considerations, three highly interrelated *spaces* are highlighted in the ECD (*cf.*, Fig. 1): the *claim*, *evidence*, and *task space* (Pellegrino *et al.*, 2016; Kubsch *et al.*, 2022b). The *claim space* addresses the objectives of an assessment, *i.e.*, the competencies that should be evaluated through the assessment (Messick, 1994). Consequently, before designing an assessment, one needs to analyze the targeted cognitive domain to make a precise claim about students' competencies within this domain (Rupp *et al.*, 2012; Mislevy, 2016; Pellegrino *et al.*, 2016; Kubsch *et al.*, 2022b). Making a claim includes unpacking the fine-grained, domain-specific characteristics, analyzing their connections, specifying how they contribute to the acquisition of essential competencies, and arranging all

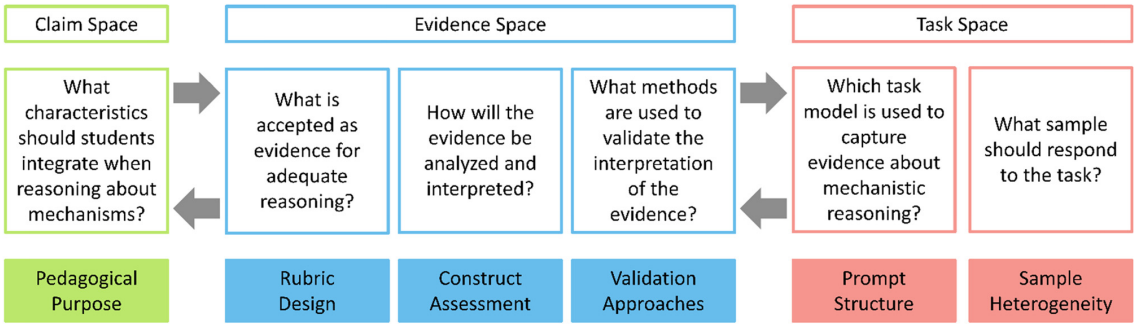


Fig. 1 Simplified representation of the evidence-centered design (ECD) approach, including the *claim*, *evidence*, and *task space*, the guiding questions central to each *space*, as well as the six machine learning-related categories derived from ECD.

 This journal is © The Royal Society of Chemistry 2023

View Article Online
Chemistry Education Research and Practice Review Article

competencies in an order with increasing complexity (Pellegrino *et al.*, 2016). In the context of mechanistic reasoning, several frameworks have been developed that characterize the specificities of this complex type of reasoning (Machamer *et al.*, 2000; Russ *et al.*, 2008; Kraft *et al.*, 2010; Sevian and Talanquer, 2014; Cooper *et al.*, 2016; Caspari *et al.*, 2018a; Krist *et al.*, 2019; Dood *et al.*, 2020a; Raker *et al.*, 2023; Yik *et al.*, 2023). Although all these frameworks focus on the dynamic interplay of entities and activities a scalar level below the phenomenon, they concentrate on different conceptualizations and elaborateness of mechanistic reasoning (*cf.*, Dood and Watts, 2022a). Most frameworks differentiate between forms of reasoning that capture, among others, descriptive, causal, and mechanistic reasoning (Machamer *et al.*, 2000; Russ *et al.*, 2008; Cooper *et al.*, 2016; Caspari *et al.*, 2018a). Descriptive reasoning entails identifying the entities of a mechanism as well as their explicit properties without necessarily outlining the cause for the underlying processes. Such reasoning is more focused on describing the system as a whole (Sevian and Talanquer, 2014; Dood *et al.*, 2020a; Raker *et al.*, 2023; Yik *et al.*, 2023). Causal reasoning involves understanding cause-and-effect relationships between different variables and *why* they contribute to the outcome of a chemical phenomenon. Causal reasoning is, therefore, concerned with understanding principles or factors that are driving a reaction, rather than the specific steps or processes involved (diSessa, 1993; Carey, 1995; Russ *et al.*, 2008; Cooper *et al.*, 2016). Mechanistic reasoning focuses on understanding *how* a chemical reaction proceeds at the molecular level, including the mechanistic steps that convert the reactants into the products. So, mechanistic reasoning aims to explain electronically, how underlying processes occur (Machamer *et al.*, 2000; Russ *et al.*, 2008; Cooper *et al.*, 2016; Caspari *et al.*, 2018a, 2018b; Krist *et al.*, 2019).

After reflecting about the *claim space*, evidence statements that outline the performance accepted as evidence must be composed (Mislevy, 2016; Pellegrino *et al.*, 2016; Kubsch *et al.*, 2022b). Formulating evidence statements involves specifying the desired competencies in detail and setting a framework for the interpretation of this evidence (Kubsch *et al.*, 2022b). Eliciting evidence of mechanistic reasoning is a challenging, iterative procedure that requires multiple refinements of the task model and a repeated evaluation of students' responses (Cooper *et al.*, 2016; Noyes *et al.*, 2022). In particular, a coding scheme should be developed based on expected and actual answers to outline how a high level of reasoning sophistication could look like (Cooper, 2015). Here, evidence rules (Mislevy *et al.*, 2003a; Rupp *et al.*, 2012) define how observable aspects, for instance, written phrases in students' work products about a mechanism's plausibility, contribute to students' competence in mechanistic reasoning. Measurement models, in turn, specify the method for making diagnostic inferences about the examined construct (Rupp *et al.*, 2012), *e.g.*, how phrases in students' written accounts are coded as evidence of mechanistic reasoning.

Coherently defining evidence rules and measurement models helps develop a construct-centered task model that characterizes the form of tasks needed to collect evidence about students' competencies (Pellegrino *et al.*, 2016). Rather than representing a single task, task models characterize the central features that a set of potential tasks must possess as well as the elements that could be varied across different assessment contexts (Mislevy *et al.*, 2003a; Mislevy and Haertel, 2007; Kubsch *et al.*, 2022b). Besides defining these contexts, a task model for capturing mechanistic reasoning must find a compromise between initiating as much sophisticated reasoning as possible without providing too much information on the structure and content of a high-quality response (Noyes *et al.*, 2022). Structuring prompts adequately and specifying the degree of scaffolding is, thus, key in developing task models (Caspari and Graulich, 2019; Graulich and Caspari, 2020; Noyes *et al.*, 2022), for instance, that guide students in judging the plausibility of a reaction mechanism (Lieber and Graulich, 2020, 2022; Lieber *et al.*, 2022a, 2022b; Watts *et al.*, 2022). Consistently connecting *claim*, *evidence*, and *task space* ultimately enables the design and use of formative assessments to be better aligned with the construct of interest.

Compared to previous methods, ML has great potential in automatically evaluating mechanistic reasoning since these algorithms can *learn* from previous experience without explicit programming (Samuel, 1959). For detecting mechanistic reasoning with supervised techniques, human raters have to consistently label evidence of mechanistic reasoning in students' responses according to a literature-based or inductively derived coding scheme (Raker *et al.*, 2023). Afterward, the selected algorithm can be trained with the labeled data, so that it can decipher the underlying patterns. With this approach, it is not necessary to explicitly program an algorithm that, for example, distinguishes between a chain of phenomenological links and mechanistic reasoning. The level of causality, *i.e.*, different types of reasoning, can simply be determined by training the algorithm with coherently labeled data and letting the algorithm *learn* from these patterns (Dood *et al.*, 2018, 2020a; Noyes *et al.*, 2020). In doing so, even subtleties in mechanistic reasoning can be identified if heterogenous and coherently labeled training data is available (Watts *et al.*, 2023).

ML-related categories derived from ECD

Aligned with the *claim*, *evidence*, and *task space*, six categories can be defined which are present in many studies – not only in those which deal with the formative assessment of mechanistic reasoning (*cf.*, Fig. 1): *pedagogical purpose of ML application*, *rubric design*, *construct assessment*, *validation approaches*, *prompt structure*, and *sample heterogeneity*. The *pedagogical purpose of ML application* articulates the pedagogical benefits of using ML in science assessment. This category can be located in the *claim space* because, first, it must be asked for which pedagogical purpose an assessment is to be used. The categories *rubric design*, *construct assessment*, and *validation approaches* are consistent with the *evidence space* since they correspond to the evidence rules and measurement models outlined above. More specifically, *rubric design* refers to the definition of the fine-grained characteristics accepted as evidence. *Construct assessment* describes the algorithmic model that translates the input data into diagnostic details about students' competencies, which means that this category builds the bridge between the technological implementation and

This journal is © The Royal Society of Chemistry 2023

View Article Online
Review Article Chemistry Education Research and Practice

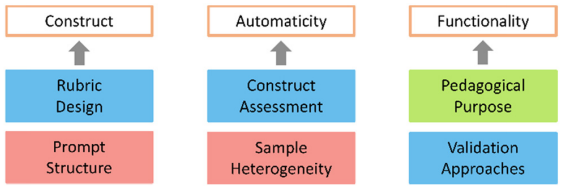


Fig. 2 Mapping the six categories of the machine learning-adapted evidence-centered design to the three dimensions of Zhai *et al.*'s (2020a) framework.

diagnostic granularity of ML-based formative assessment. Moreover, *validation approaches* characterize the method used for validating the reliability of an algorithm. Last, *prompt structure* and *sample heterogeneity* are assigned to the *task space. Prompt structure* specifies the tasks that elicit the desired evidence about students' competencies. *Sample heterogeneity* refers to the sample responding to this task; here, a heterogenous sample is a prerequisite for building equitable ML models. Considering the six categories in the development of formative assessments helps translate cognitive theories into assessments that yield evidence concerning the construct of interest (Pellegrino *et al.*, 2016). The ECD approach adapted for ML-based formative assessment (called ML-adapted ECD hereafter) will serve as a theoretical underpinning for the evaluation of the selected studies in this review.

Another framework conceptualizing the implementation of ML in science assessment in general, developed by Zhai *et al.* (2020a), considers the dimensions of *construct*, *automaticity*, and *functionality*. Some cross-references between the ML-adapted ECD and Zhai *et al.*'s (2020a) framework can be made (*cf.*, Fig. 2). Zhai *et al.*'s (2020a) dimension *construct* is covered by our categories *rubric design* and *prompt structure.* For *automaticity*, Zhai *et al.* (2020a) introduced, among others, the variables *machine extracts attributes* and *model generalization. Machine extracts attributes* is reflected in our category *construct assessment* because both deal with the fundamental methodology of an ML algorithm. Zhai *et al.*'s (2020a) *model generalization* analyzes whether an ML algorithm trained with one dataset can be applied to another dataset. Building upon this distinction, we used the category *sample heterogeneity*. For *functionality*, Zhai *et al.* (2020a) established, among others, the variables *score use* and *measurement models*. The variable *score use* evaluates whether the ML output is embedded in a learning activity, which we highlight in *pedagogical purpose of ML application*. Last, Zhai *et al.*'s (2020a) variable *measurement models* corresponds to our category *validation approaches* since both refer to the method used for the validation of an algorithm.

Multi-hierarchical ML workflow for capturing students' mechanistic reasoning

Based on the ML-adapted ECD and the six categories identified in the *claim*, *evidence*, and *task space* (*cf.*, Fig. 1), we further divided each category into three hierarchical levels to describe the extent to which the selected studies applied the ECD principles (*cf.*, Fig. 3). These hierarchical levels may help put ML-adapted ECD into practice and characterize the advancements and shortcomings that ML has brought to the formative assessment of mechanistic reasoning. Besides, it may serve as a tool to evaluate ongoing research processes. We illustrate these categories and levels first with research approaches in other domains that can also be applied to chemistry assessment to show the generalizability and topicality of the ML-adapted ECD principles in science assessment.

Claim space

The *claim space* comprises the category *pedagogical purpose of ML application* that covers primarily how to integrate ML into learning environments. The pedagogical purpose sets the basis for applying ML in educational assessments since it specifies the objectives of the respective assessment.

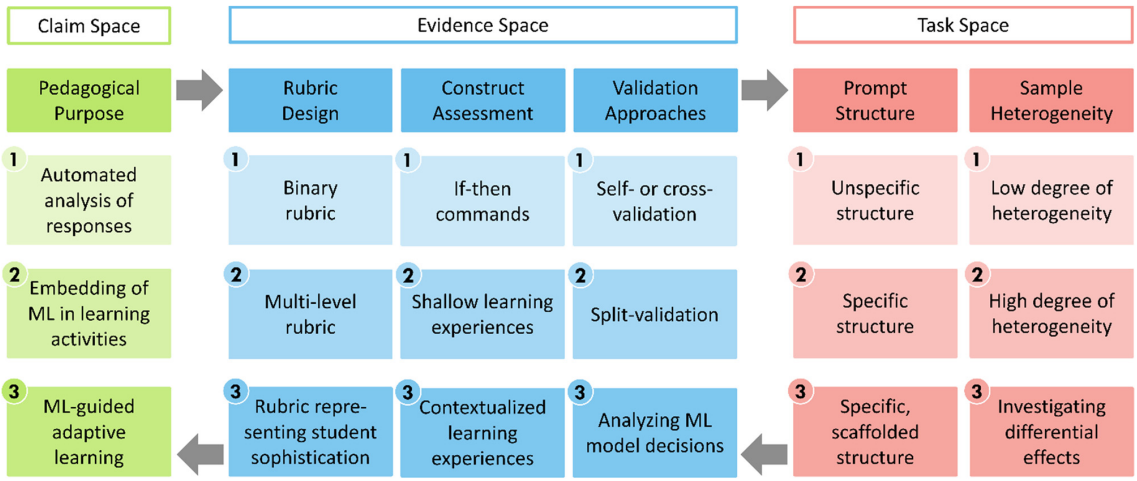


Fig. 3 Representation of the workflow for capturing complex constructs with machine learning (ML) based on the evidence-centered design (ECD) approach. Three hierarchical levels for the degree of implementing the ECD principles when applying ML are proposed in each category.

 This journal is © The Royal Society of Chemistry 2023

View Article Online
Chemistry Education Research and Practice Review Article

Pedagogical purpose of ML application. A key component of applying ML in science education is verifying the purpose of the designed assessment, which leads to the pedagogical benefits of ML. Applying ML for the *automated analysis of responses* (level 1) helps evaluate students' proficiency in real-time. This ML application can be considered the first level since the *embedding of ML in learning activities* (level 2) is linked to higher achievement in science assessment (*cf.*, Zhai *et al.*, 2020c). Easily accessible embeddings eventually provide an opportunity for *ML-guided adaptive learning* (level 3) through personalized feedback or tailored exercises (Kerr, 2016).

To exemplify what we mean by a well-planned *pedagogical purpose of ML application* in terms of adaptive learning, we delineate the studies of Zhu *et al.* (2017), Mao *et al.* (2018), and Lee *et al.* (2021). The authors investigated students' scientific argumentation skills with a focus on the linkage between claim and evidence, the limitations of the given data, and the factors that cause uncertainty. By doing so, they developed an ML-based instructional system that provided automatically individualized real-time feedback for students so that they had the chance to revise their arguments accordingly. Zhu *et al.* (2017) found that most students revised their arguments about factors that affect climate change after getting adapted feedback. Consequently, students who revised their arguments had significantly higher final scores than those who did not adjust their arguments.

Lee *et al.* (2021) built upon the work of Zhu *et al.* (2017) and Mao *et al.* (2018) by applying ML in three simulation-based scientific argumentation tasks. In addition to argumentation feedback, Lee *et al.* (2021) provided ML-enabled simulation feedback to improve students' interactions with simulations about groundwater systems. Students who received adapted simulation feedback were more likely to re-run the simulation, leading to significantly higher scores in some tasks. This result shows that ML algorithms can provide effective individualized feedback not only for written arguments but also for simulation interactions. So, ML can be used to ease the implementation of adaptive learning.

Evidence space

The *evidence space* contains the three categories *rubric design*, *construct assessment*, and *validation approaches* dealing with the characteristics accepted as evidence for a construct, the way these characteristics are analyzed, and the methods used for the validation of the analysis. The *evidence space* connects the *claim* and *task space* which ensures a valid evaluation of the investigated construct.

Rubric design. In any type of assessment, validity is a prerequisite. Notably, a computer algorithm can only capture reasoning validly if humans can. So, valid rubrics need to be designed by students' responses or the literature. Analytic rubrics measure the presence or absence of multiple binary conceptual components. Here, every analytic bin outlines a single idea, and every response must be coded for each of the non-mutually exclusive bins. For a more comprehensive evaluation of student reasoning, holistic rubrics can be applied since they measure the characteristics of students' responses by assigning a single, mutually exclusive score. Analytic and holistic rubrics are interrelated. On the one hand, a holistic rubric can be derived by analyzing the combinations of the analytic bins (Noyes *et al.*, 2022). On the other hand, holistic rubrics can be deconstructed into analytic bins (Jescovitch *et al.*, 2019b). Integrating both types eventually has the advantage that students' responses can be evaluated at an overarching and a fine-grained level.

In their study on scientific argumentation, Jescovitch *et al.* (2021) and Wang *et al.* (2021) compared a multi-level, holistic coding approach with a dichotomous, analytic coding approach. Both found that ML models trained with analytic bins performed slightly better than those trained with a holistic rubric, especially for items that introduced a complex scenario. They inferred that analytic rubrics might have the potential to identify and unpack additional complexity in students' responses, which can be explained by the reduced human coding effort, resulting in an increase in human–human inter-rater reliability and an improvement of the model performance (Jescovitch *et al.*, 2019a, 2021). However, there is no evidence that one coding approach outperforms the other in every context for human and machine coding (Zhai *et al.*, 2020c). For this reason, more important than the rubric type is that rubrics allow for consistent classifications of data, both within and between human coders (Raker *et al.*, 2023).

For the application of ML, fine-grained rubrics seem to offer deeper insights into student reasoning. Therefore, we classify *binary rubrics* as level 1, *multi-level rubrics* as level 2, and *multi-level rubrics representing different levels of reasoning sophistication* as level 3 in the ML-adapted ECD. Rubrics classified as level 2, in general, differentiate between *incorrect*, *partially correct*, and *correct* responses, whereas rubrics classified as level 3 distinguish between different levels of reasoning sophistication. In all cases, a valid rubric and an iterative process of its application and refinement are necessary to elicit the construct of interest (Allen and Tanner, 2006).

Construct assessment. Applying ML in science assessment offers a great opportunity to automatically evaluate complex constructs and detect hidden patterns. Before designing an assessment, it is necessary to decide whether supervised or unsupervised techniques should be applied to analyze the dataset (*cf.*, Appendix). The magnitude of available data, the explored phenomena, and the generated hypotheses guide the ML model selection (Mjolsness and Decoste, 2001). In general, no algorithm outperforms any other in every context (Zhai *et al.*, 2020c).

For example, lexical analysis software can automatically extract domain-specific categories that are iteratively refined by human experts afterward (Haudek *et al.*, 2011). Defining meaningful categories can, consequently, be a major task in applying lexical analysis (Urban-Lurain *et al.*, 2010). Categories should offer enough information to detect different sophistication levels; however, infrequently used categories have less predictive power. With statistical methods, it is, thus, possible to identify meaningful categories and collapse non-discriminatory ones. Sometimes, categories are formed based on human- or computer-set if-then commands, which come with limitations. On the one hand, case-specific custom-libraries need to be defined to decide which keywords are included in the if-then

This journal is © The Royal Society of Chemistry 2023

View Article Online
Review Article Chemistry Education Research and Practice

commands and how fine-grained the categories should be, which is labor- and time-intense. On the other hand, if-then commands do not allow for flexible adjustments of the categories over time. So, we define *construct assessment based on if–then commands* as the first level in this category.

Rather than just running human- or computer-defined commands, ML models can also gain experience on their own (Bishop, 2006). Once the algorithm has *learned* from the training data, it can classify or predict the characteristics of new data (*cf.*, Appendix). For such ML algorithms, features can, for instance, be extracted with *n*-grams. These *n*-grams are sequences of *n* words that allow for identifying repetitive patterns, *i.e.*, *n* consecutive words in students' responses. Besides word *n*-grams, ML algorithms can also process character *n*-grams, response length, syntactic dependencies, and semantic role labels (*e.g.*, Mao *et al.*, 2018). These methods require less human effort, especially when a reliable scoring rubric has already been applied to many student responses (Haudek *et al.*, 2011). However, the mentioned natural language processing techniques are built on the simplified assumption that the word order is irrelevant to the meaning of a sentence (Wulff *et al.*, 2022b), which complicates the detection of implicit semantic embeddings. So, traditional ML models are only sensitive to key conceptual components, which is why we define *construct assessment based on shallow learning experiences* as the second level of the ML-adapted ECD. Here, the term experience refers to the algorithm's process of collecting information from pre-processed data to make automated decisions.

Other techniques like Bidirectional Encoder Representations from Transformers (BERT), a natural language processing technique for calculating embedding vectors, may assess the relationships between words more comprehensively since it *knows* contextualized embeddings (Wulff *et al.*, 2022a). In other words, BERT is capable of grasping the implicit meaning of individual words, *understands* dynamic contextualized relationships of a word within any other word in a sentence, and recognizes filler words. Since BERT processes all words of a sentence simultaneously, it can detect both conceptual and semantic features, which is why we consider *construct assessment based on contextualized learning experiences* as the third level in this category.

Validation approaches. The primary metric for the accuracy of an ML model is the agreement between human and machine scores (Williamson *et al.*, 2012). For a high machine-human score agreement, a high human–human inter-rater reliability in the coding process is necessary. Since supervised models are built on human codes, cases in which humans did not reach a consensus may also induce scoring problems for the algorithm. Hence, it seems helpful to feed only data where consensus is reached in the algorithm.

Three methods can be applied to calculate the machine-human score agreement: self-validation, split-validation, and cross-validation (*cf.*, Appendix). Nehm *et al.* (2012) found that their machine-human score agreement slightly decreased when validating their algorithms with a new testing set compared to the agreement reached when using the same dataset for model training and validation. In their systematic review, Zhai *et al.* (2020c) could underpin this result by finding that cross-validation yield better machine-human score agreements than self- and split-validation. However, using self-validation brings equity and generalizability constraints since the same data is used for the training and evaluation of the algorithm (Zhai *et al.*, 2020a, 2020c). For this reason, we classify *self- and cross-validation approaches* as the first and *split-validation approaches* as the second level in the ML-adapted ECD. For educational purposes, the transparency of ML-based decisions should, additionally, not be underestimated (Cheuk, 2021). Besides calculating the level of agreement, it is helpful to understand why the respective model decisions were made by comparing human- and computer-set clusters, interpreting misclassifications, or conducting interviews. *Analyzing ML model decisions* is, therefore, the third level in this category.

Irrespective of the validation approach, various metrics are used to measure the magnitude of machine-human score agreement (*cf.*, Zhai *et al.*, 2020c). For a more consistent indication of machine-human score agreement, Zhai *et al.* (2021) call for the introduction of standardized rules, *e.g.*, to report a confusion matrix, accuracy, Cohen's κ, and the relevant variance.

Task space

The *task space* consists of two categories *prompt structure* and *sample heterogeneity* defining the task model used for capturing evidence of mechanistic reasoning as well as the sample responding to that task. The considerations of the *task space* help practically implement science assessments.

Prompt structure. When developing a prompt, a precise wording should clearly communicate the concepts that sophisticated reasoning needs to address. Comparing the so-initiated students' ideas with an expected answer helps adjust the prompt iteratively. When modifying a prompt, one has to keep in mind that small changes in the wording can drastically change the way students respond (Noyes *et al.*, 2022). A prompt has to offer enough information for students to work with a task. If a prompt is too vague, students may not be engaged to reason in a sophisticated way. In contrast, a too specific prompt may give students detailed information on what the answer requires. This may lead to a repetition of key terms, phrases, and ideas denoted in the question stem. Preparing a well-elaborated prompt is, following this, crucial to ensure that the assessment captures students' understanding, not the repetition of buzzwords.

During the prompt refinement process, one has to consider not only the wording but also the structure of a prompt in terms of scaffolding (*e.g.*, Kang *et al.*, 2014). Scaffolding, for example, combined with contrasting cases (Graulich and Schween, 2018), helps focus students' attention on productive concepts (Wood *et al.*, 1976) and activate resources that might otherwise be overlooked (Hammer, 2000). When adjusting the scaffolding, one has to consider the resources that students have already applied in their reasoning, the parts of the task model that have activated those resources, and the modifications that can be implemented to activate further appropriate resources (Noyes *et al.*, 2022). In ML-based science assessment, scaffolding can also be realized by providing multiple response boxes in

 This journal is © The Royal Society of Chemistry 2023

View Article Online
Chemistry Education Research and Practice Review Article

which students can enter their answers (Urban-Lurain *et al.*, 2010). This creation of separate text boxes has two major benefits. On the one hand, multiple boxes provide students with an easily accessible form of scaffolding since these boxes may help students structure their explanations as desired. On the other hand, multiple response boxes facilitate automated text analysis as shown by Urban-Lurain *et al.* (2010). In their study, a single text box was provided to investigate student thinking on acid–base chemistry. They realized that the students used similar terms in this box but with entirely different ideas in mind. Due to the similar wording, ML models could neither elicit the context to which the technical terms referred nor students' cognitive concepts. The introduction of text boxes for different prompts helped them to overcome this limitation.

Designing a prompt to capture mechanistic reasoning is, considering all these aspects, a complex process requiring a detailed analysis of the phenomenon under investigation and of the competencies that students should acquire. Simply asking students to explain *how* and *why* phenomena occur, without providing comprehensive support, is often not sufficient to initiate sophisticated reasoning (Cooper, 2015). Scaffolded prompts can better guide students in activating their resources as the expectations of what a good answer should look like are more clearly outlined and assessable afterward. Based on the known effectiveness of scaffolded prompts, we define the levels *unspecific* (level 1), *specific* (level 2), and *specific*, *scaffolded* (level 3) in the category *prompt structur*e. We understand by *unspecific prompts* questions that can be asked regardless of the subject domain, such as "*Explain your reasoning*". *Specific prompts* provide better orientation for students on how to solve a task. Due to their potential instructional support on how to structure a response, *specific*, *scaffolded* prompts comprise the third level in this category.

Sample heterogeneity. Because language background influences how mechanistic reasoning is expressed (Deng *et al.*, 2022), the training of an ML model ideally requires a heterogeneous data sample. Responses from students with different language backgrounds should be fed into the algorithm to accurately assess various reasoning types. Following this, the heterogeneity of answers is key to developing ML models that can be applied in different learning environments, independent of institution, course, curriculum, and question prompt.

Some studies on ML in science education already emphasized the importance of creating equitable ML models. Ha *et al.* (2011) investigated whether an ML model developed using biology students' written explanations of evolutionary change from one institution can score responses from another institution. They found that their predictive model could accurately score responses at both universities for most key concepts; however, the model performance slightly decreased for a few concepts at the second university. The partly lower accuracy of the ML model was explained with unique language patterns, low concept frequencies, misspellings, and informal vocabulary. Liu *et al.* (2016) analyzed differential effects concerning gender, native language, and use of a computer for homework and found no generalizable differences between human and machine scoring across all items for any of these variables. Similar to Ha *et al.* (2011), Liu *et al.* (2016) traced minor differences in human and machine scoring back to misspellings and linguistic diversity. These findings led Ha and Nehm (2016) to examine the influence of misspelled words on machine scoring. Although students who learned English as an additional language produced significantly more misspellings, this did not meaningfully impact computer scoring efficiency. For a valid examination of these differential effects, a heterogeneous data sample is required (Cheuk, 2021).

Putting it all together, considerable heterogeneity of responses is needed to investigate the generalizability of an ML model, which means that data must cover all sophistication levels. If specific words, terms, or phrases are missing from the training data, the algorithm cannot predict them in the testing data. This prerequisite is reflected in the ML-adapted ECD (*cf.*, Fig. 3). A *low degree of heterogeneity* (level 1) may produce a biased ML model, since students who meet specific external characteristics may be favored, whereas students who do not meet these characteristics may be disadvantaged. By *low degree of heterogeneity*, we understand certain external characteristics that tend to be identical across the investigated sample such as demographic background, instructor, and curricula. A *high degree of heterogeneity* (level 2) enlarges the training vocabulary of the ML model and increases equity by considering different demographic backgrounds, instructors, and curricula. *Investigating differential effects* (level 3) helps finally uncover and prevent bias.

Selection of studies

This review investigates the state-of-the-art in the area of ML-detected mechanistic reasoning. We aim to provide an in-depth analysis of the prospects and obstacles that ML techniques have brought to the evaluation of this special type of reasoning, rather than to give a broad overview of applying ML in chemistry or science education. Please, see Zhai *et al.* (2020a, 2020c) for a review or a framework, respectively, on applying ML in science assessment, Deeva *et al.* (2021) for a review on automated feedback systems, and Gerard *et al.* (2015) for a *meta*-analysis on the effectiveness of automated, adaptive guidance.

We defined a set of keywords to look for studies that contributed to the automated analysis of mechanistic reasoning in chemistry. After an iterative process of refinement, the search term *chemistry AND* ("*machine learning*" *OR* "*lexical analysis*" *OR* "*automated guidance*" *OR* "*automated feedback*" *OR* "*automated text analysis*") was established. We then used the three databases Google Scholar, ProQuest, and ERIC to find suitable articles. Due to the immense amount of returned results (44 600 in total), we modified where the search terms must occur. We specified that the keywords must occur in the title when using Google Scholar, in the abstract when using ProQuest, and anywhere in the article when using ERIC. Eventually, the search process resulted in 996 articles (333 Google Scholar, 640 ProQuest, 23 ERIC). First, we screened the title of these papers to sort all articles related to the field of chemistry but not to chemistry education. After that, we looked closely at the

View Article Online
Review Article Chemistry Education Research and Practice

abstracts to examine whether the articles contributed to our area of interest. Additionally, we defined some including and excluding criteria that the reviewed articles must comply with.

The included studies had to meet the following criteria:

• Topic of research: research studies must focus on ML's application in chemistry education research. The participants in the studies did not have to be chemistry students exclusively.

• Goal of research: research studies must contribute to the analysis of mechanistic reasoning. In this review, mechanistic reasoning generally refers to reasoning about the underlying mechanisms of observed phenomena, *i.e.*, general or organic chemistry phenomena. For a study to be included, it was not sufficient that chemistry content or reasoning about macroscopic phenomena was addressed. Included studies must capture reasoning about underlying mechanisms, *i.e.*, *how* and *why* phenomena occur.

• Time range of publications: since the first paper meeting the defined selection criteria was published in 2009, we considered all articles since 2009.

The following criteria led to the exclusion of a study:

• Studies concentrating on science assessments with predominantly non-chemical content are excluded (*e.g.*, Liu *et al.*, 2016; Zhai *et al.*, 2022a).

• Studies analyzing inquiry or modeling skills in drawing activities are excluded (*e.g.*, Rafferty *et al.*, 2013, 2014; Gerard *et al.*, 2016; Zhai *et al.*, 2022b).

• Studies focusing on the automated analysis of inquiry skills in designing and conducting experiments are excluded (*e.g.*, Gobert *et al.*, 2013, 2015; Sao Pedro *et al.*, 2013).

• Studies using ML to evaluate student-written interpretations of chemical representations are excluded (*e.g.*, Prevost *et al.*, 2014).

• Studies investigating the automated analysis of student-constructed graphs are excluded (*e.g.*, Vitale *et al.*, 2015).

• Studies using written responses to explore methodological issues of ML, *e.g.*, the effect of construct complexity on scoring accuracy (Haudek and Zhai, 2021), are excluded.

• Studies discussing the teacher role in ML application or their pedagogical content knowledge assessed by ML are excluded (*e.g.*, Zhai *et al.*, 2020b).

We screened the reference section of each selected study to find other appropriate papers. The final set of articles contains 20 studies (*cf.*, Table 1). To be noted, Urban-Lurain *et al.* (2009) summarized parts of Haudek *et al.*'s (2009) study. Prevost *et al.* (2012a) document the use of an automatically evaluated item, which is even presented in a more extensive study by Prevost *et al.* (2012b). Dood *et al.* (2018) report the development of a

Table 1 Overview of the selected research articles

Author	Software	Method	Education level	Domain	Science practice	Item	Sample sizes
Haudek *et al.* (2009), Urban-Lurain *et al.* (2009)	SPSS text analytics for surveys	Category-based lexical analysis	Under graduate	Thermodynamics, acid–base chemistry	Explanation	CR	158/153/382
Haudek *et al.* (2012)	SPSS text analytics for surveys	Category-based lexical analysis	Under graduate	Acid–base chemistry	Explanation	CR	1172/323
Prevost *et al.* (2012a, 2012b)	SPSS text analytics for surveys/SPSS modeler text analytics	Category-based lexical analysis	Under graduate	Thermodynamics	Explanation	CR	168/329/386
Liu *et al.* (2014)	c-Rater	Natural language processing	Middle school	Thermodynamics	Explanation	CR	412/362/321/356
Donnelly *et al.* (2015)	c-Rater	Natural language processing	Middle school	Thermodynamics	Inquiry	CR, SI	346
Haudek *et al.* (2015)	SPSS modeler text analytics	Category-based lexical analysis	Under graduate	Acid–base chemistry	Explanation	CR	336
Vitale *et al.* (2016)	c-Rater-ML	Support vector regressor	Middle school	Climate change	Inquiry	CR, FD, SI	283
Tansomboon *et al.* (2017)	c-Rater-ML	Support vector regressor	Middle school	Thermodynamics	Inquiry	CR, SI	482
Dood *et al.* (2018, 2019)	SPSS modeler text analytics	Category-based lexical analysis	Under graduate	Acid–base chemistry	Explanation	CR	752/1058
Haudek *et al.* (2019)	R/RStudio	Ensemble algorithm	Middle school	Structures and properties	Argumentation	CR	246/775/763
Dood *et al.* (2020a, 2020b)	SPSS modeler text analytics/python	Category-based lexical analysis	Post-secondary	Nucleophilic substitution	Explanation	CR	1041
Noyes *et al.* (2020)	R/RStudio	Ensemble algorithm	Under graduate	Intermolecular forces	Explanation	CR	1730
Maestrales *et al.* (2021)	AACR web portal	Ensemble algorithm	High school	Structures and properties	Explanation	CR	26 800
Winograd *et al.* (2021b)	Python	BERT/convolutional neural network	Under graduate	Chemical equilibrium	Communication	WTL	297
Yik *et al.* (2021)	R/RStudio	Support vector machine	Post-secondary	Acid–base chemistry	Explanation	CR	852
Watts *et al.* (2023)	Python	Convolutional neural network	Under graduate	Racemization, acid hydrolysis, Wittig reaction	Explanation	WTL	771

Note: Haudek *et al.* (2009) and Urban-Lurain *et al.* (2009), Prevost *et al.* (2012a, 2012b), Dood *et al.* (2018, 2019), and Dood *et al.* (2020a, 2020b) refer to the same predictive model, which is why they are represented within a single row. BERT = Bidirectional Encoder Representations from Transformers, CR = constructed response, SI = simulation, FD = flow diagram, WTL = writing-to-learn.

 This journal is © The Royal Society of Chemistry 2023

View Article Online
Chemistry Education Research and Practice Review Article

predictive model for Lewis acid–base concept use, whereas Dood *et al.* (2019) focus on the pedagogical benefits of their model. Similarly, the pedagogical use of Dood *et al.*'s (2020a) computerized scoring model is reported in the study by Dood *et al.* (2020b). In this review, we refer up to now, if possible, to Haudek *et al.* (2009), Prevost *et al.* (2012b), Dood *et al.* (2018), and Dood *et al.* (2020a).

Discussion of the selected studies on ML-based formative assessment of mechanistic reasoning

Based on the defined levels, we categorized the 20 selected articles. First, author PPM assigned each article in each of the six categories to one of three hierarchical levels. After discussing the categorization with author NG, we refined it until a total consensus was reached (*cf.*, Table 2). Based on this categorization, we illustrate some contemporary effective practices of ML-based formative assessment of mechanistic reasoning, *i.e.*, studies that were assigned to the third level of the ML-adapted ECD in the respective category (*cf.*, Fig. 3). After that, we deduce trends as well as advancements and shortcomings that ML has brought to the formative assessment of mechanistic reasoning.

Pedagogical purpose of ML-detected reasoning

The pedagogical purpose of most studies was focused on automating and scaling the analysis of constructed responses in large classes. Thirteen of the reviewed studies (*cf.*, Table 2) set *automating human scoring* (level 1) as their pedagogical purpose. Here, these studies used the ML output to scan the number of responses in each rubric category, which helped predict students' competence. However, these studies concentrated less on investigating the pedagogical benefits of ML, but more on the technical and validity features of their assessment.

Dood *et al.* (2020a, 2020b) developed an adaptive online tutorial to improve reasoning about nucleophilic substitution reactions. Before working with this tutorial, students' explanations of a unimolecular nucleophilic substitution were automatically scored. Depending on the determined level of explanation sophistication, students were assigned one of two adapted tutorials that addressed leaving group departure, carbocation stability, nucleophilic addition, and acid–base proton transfer. Dood *et al.* (2020a, 2020b) found that completing the adapted tutorial significantly improved students' reasoning skills. Their work illustrates how ML can be used to implement aspects of *adaptive learning* (level 3).

Also marked as contemporary effective practices in the category *pedagogical purpose*, Donnelly *et al.* (2015), Vitale *et al.* (2016), and Tansomboon *et al.* (2017) used ML techniques to investigate the effectiveness of automated, adaptive guidance in readily accessible online curricula units on thermodynamics and climate change. Donnelly *et al.* (2015) assigned middle school students to either a revisit condition, where they were prompted to review a dynamic visualization, or a critique condition, where they had to criticize a sample explanation. With automated scoring, they found that combining adapted guidance with either the revisit or critique condition helped students significantly in acquiring knowledge and revising their explanations. Here, both conditions were more effective for low-performing students indicating the importance of adapted guidance. Building upon this work, Tansomboon *et al.* (2017) found that transparent automated guidance, that explicitly clarified how the ML output was generated, supported low-performing students in revising an explanation better than typical automated guidance. Similarly, Vitale *et al.* (2016) found that content guidance, that specified missing ideas in students' responses, resulted in instant learning gains within a simulation, but knowledge integration guidance, that delivered tailored hints, was better suited to enable knowledge transfer. Together, these studies indicate that ML output can serve as a basis for investigating further qualitative and quantitative research questions in science education.

Table 2 Levels of the selected studies according to our framework. The left-aligned, orange dots indicate level 1, the centered, blue dots indicate level 2, and the right-aligned, green dots indicate level 3

Study	Pedagogical purpose	Rubric design	Construct assessment	Validation approaches	Prompt structure	Sample heterogeneity
Haudek *et al.* (2009), Urban-Lurain *et al.* (2009)	●	●	●	●	●	●
Haudek *et al.* (2012)	●	●	●	●	●	●
Prevost *et al.* (2012a, 2012b)	●	●	●	●	●	●
Liu *et al.* (2014)	●	●	●	●	●	●
Donnelly *et al.* (2015)	●	●	●	●	●	●
Haudek *et al.* (2015)	●	●	●	●	●	●
Vitale *et al.* (2016)	●	●	●	●	●	●
Tansomboon *et al.* (2017)	●	●	●	●	●	●
Dood *et al.* (2018, 2019)	●	●	●	●	●	●
Haudek *et al.* (2019)	●	●	●	●	●	●
Dood *et al.* (2020a, 2020b)	●	●	●	●	●	●
Noyes *et al.* (2020)	●	●	●	●	●	●
Maestrales *et al.* (2021)	●	●	●	●	●	●
Winograd *et al.* (2021b)	●	●	●	●	●	●
Yik *et al.* (2021)	●	●	●	●	●	●
Watts *et al.* (2023)	●	●	●	●	●	●

This journal is © The Royal Society of Chemistry 2023

View Article Online
Review Article Chemistry Education Research and Practice

According to these contemporary effective practices, the next steps in ML-based assessment are to explore the benefits of ML models in daily teaching practices, shifting ML-influenced research from replicating and validating human scoring to guiding students in adaptively acquiring reasoning skills. Comprehensively discussing the *pedagogical purpose of ML application* before designing and using an assessment helps ensure that we take advantage of ML's comprehensive pedagogical benefits.

Rubric design in ML-detected reasoning

When looking at the selected studies in the category *rubric design*, it is apparent that most of them (*cf.*, Table 2) already used *multi-level rubrics representing different levels of reasoning sophistication* (level 3). Only three studies used *binary rubrics* (level 1) to evaluate whether a student applied the Lewis acid–base concept, while three others designed *multi-level rubrics* (level 2) that distinguished between *incorrect*, *partially correct*, and *correct* responses, also with a focus on acid–base reactions.

To elicit three-dimensional learning (National Research Council, 2012), Maestrales *et al.* (2021) created a holistic, three-level rubric classifying responses as *incorrect*, *correct*, and *multi-dimensional correct*. By listing the explanatory components expected in each dimension, Maestrales *et al.* (2021) did not only show what an expert answer should look like but also created a valid foundation for human scoring. Liu *et al.* (2014), Donnelly *et al.* (2015), Vitale *et al.* (2016), and Tansomboon *et al.* (2017) developed holistic, multi-level knowledge integration rubrics to assess student reasoning in computer-guided inquiry learning (Linn *et al.*, 2014). By distinguishing between *off-task responses*, *scientifically non-normative ideas*, *unelaborated links*, *full links*, and *complex links*, these rubrics valued students' skills in generating ideas, applying interdisciplinary concepts, and gathering evidence (Linn and Eylon, 2011). Similarly, Dood *et al.* (2020a) and Noyes *et al.* (2020) generated holistic, three-level rubrics to categorize student thinking about substitution reactions and London dispersion forces, respectively. Dood *et al.* (2020a) distinguished between responses that described *what* happens in a reaction, *why* the reaction occurs at a surface level, and *why* the reaction occurs at a deeper level. Noyes *et al.* (2020) differentiated between *non-electrostatic*, *electrostatic causal*, and *causal mechanistic* explanations.

Just like Dood *et al.* (2020a) and Noyes *et al.* (2020), Watts *et al.* (2023) analyzed fine-grained reasoning features, that were in alignment with Russ *et al.*'s (2008) framework for mechanistic reasoning, to automatically analyze the *how* and *why* of mechanisms. In contrast to the studies mentioned above, Watts *et al.* (2023) applied an analytic coding approach. With convolutional neural networks they identified the presence of different features of mechanistic reasoning, *e.g.*, *identifying activities*. Ultimately, their convolutional neural network provided a means to distinguish between *explicit* and *implicit* as well as *static* and *dynamic* explanations. Winograd *et al.* (2021b) and Watts *et al.* (2023) share the fine-grained analytic coding approach; however, Winograd *et al.*'s (2021b) approach differs in the way that four dimensions of cognitive operations (Grimberg and Hand, 2009), namely *consequences*, *cause and effect*, *deduction*, and *argumentation*, were evaluated. Winograd *et al.* (2021b) argue that causal reasoning includes *consequences*, *deduction*, and *cause and effect*; subsequently, the identification of these three operations may help detect instances of causal reasoning.

Haudek *et al.* (2019) combined analytic and holistic coding approaches when assessing argumentation skills of middle school students. For human coding, they applied a dichotomous analytic rubric that outlined, for instance, the presence of one claim, one piece of evidence, and three pieces of reasoning. For computer coding, the analytic components were combined into a holistic score indicating whether students formed *incomplete*, *partially complete*, or *complete* arguments about sugar dissolving in water.

Prevost *et al.* (2012b) and Haudek *et al.* (2015) chose a completely different approach to design rubrics for reaction spontaneity and acid–base behavior of functional groups, respectively. Based on human- and computer-set categories, both applied *k*-means clustering, an unsupervised ML technique, to group explanations into mutually exclusive clusters. After clustering, they investigated the key aspects of each cluster by analyzing sample responses to inductively develop rubrics. Their process of inductive rubric development has two benefits: as an unsupervised technique, the application of *k*-means clustering does not require human coding, which is less time- and cost-consuming. Additionally, rubrics can be built exploratively based on student responses, allowing for more extensive analyses of students' mechanistic reasoning and for adjusting clusters easily over time.

Overall, most studies in our sample constructed ML algorithms that could accurately evaluate different levels of sophistication, illustrating already a high standard of granularity in *rubric design*. The high accuracy of the ML-based evaluation in these studies indicates that ML has the potential to adopt fine-grained rubrics for detailed analyses. A prerequisite, however, is that the defined levels are well-distinguishable for humans so that the algorithm can consistently decipher the underlying patterns. In the future, it might be necessary to assess not only the sophistication level but also the correctness of applied concepts for summative purposes as well as students' mixed ideas. Applying unsupervised ML more extensively helps additionally detect patterns across all responses so that rubrics can be adjusted inductively.

Construct assessment in ML-detected reasoning

To evaluate mechanistic reasoning, different methods have been used so far (*cf.*, Table 2). In the earlier years of ML application in chemistry assessment, studies used feature extraction *based on human- and computer-defined if-then commands* (level 1), leveraging both supervised and unsupervised techniques. In particular, these studies set up domain-specific categories, *e.g.*, *accept protons* or *sharing electrons* (Dood *et al.*, 2018), to classify the use of essential keywords. The technological shortcomings of these studies can be explained by the limited possibilities at that time. The authors had to resort to basic ML techniques since applying complex algorithms was much more complex than today. By contrast, six more recent studies let algorithms *learn* from *shallow experiences* (level 2).

 This journal is © The Royal Society of Chemistry 2023

View Article Online
Chemistry Education Research and Practice Review Article

Different from these research projects, two studies considered *contextualized features* (level 3) of written responses more comprehensively. Winograd *et al.* (2021b) applied pre-trained BERT models to characterize the syntactic dependencies of conceptual components according to different cognitive operations (Grimberg and Hand, 2009) while students reasoned about ocean acidification. Their method can be seen as a starting point for fostering complex reasoning in essays, providing feedback for learners and instructors about applied cognitive operations, and uncovering unproductive learning processes. With a similar goal, Watts *et al.* (2023) applied convolutional neural networks, a cutting-edge deep learning algorithm, to evaluate how students reason about entities, properties of entities, and activities.

Collectively, the studies show that selecting a method for natural language processing as well as a ML technique is a critical step in designing any assessment. The construct assessment method must be chosen so that detailed diagnostic information can be derived from the ML output, potentially enabling individualized feedback and adapted guidance. Due to the enormous technological progress, cutting-edge algorithms can now be employed without programming from scratch. Compared to the use of ML in earlier years, this opened up the possibility for an extensive evaluation of conceptual and semantic features in text responses.

Validation approaches in ML-detected reasoning

In our sample (*cf.*, Table 2), six studies relied on *cross-validation* or did not report their validation approach (level 1), while eight used additionally *split-validation* (level 2). Six other studies even *analyzed ML model decisions* (level 3) by discussing misclassifications, comparing human- and computer-set clusters, as well as conducting face-to-face interviews.

Maestrales *et al.* (2021) found that responses including more advanced vocabulary like *density* or *velocity* were more likely to be misclassified. The algorithm's difficulties in dealing with advanced vocabulary could be explained by the underrepresentation of these technical terms in the training set. Since students did not frequently apply specific technical terms, the algorithm could not score them correctly. Yik *et al.* (2021) also investigated the shortcomings of their generalized model for evaluating the Lewis acid–base model use. They examined three reasons for false positive and two reasons for false negative predictions, *i.e.*, cases in which the algorithm was wrong. For instance, responses that included the term *electrons* without referring to the Lewis acid–base theory led to false positives. Yik *et al.*'s (2021) identification of misclassifications can inform future research to anticipate the limitations of ML models and to adjust rubric, coding process, and choice of algorithm accordingly.

Some misclassifications can be avoided by using clustering techniques to inductively group students' responses into mutually exclusive clusters. As outlined in the section *Rubric design in ML-detected reasoning*, Prevost *et al.* (2012b) and Haudek *et al.* (2015) used such an approach, namely *k*-means clustering, to exploratively group students' responses based on their conceptual ideas. In this context, they compared the clusters they expected before ML analysis with those identified by the ML algorithm. This human interpretation in conjunction with ML-based analysis finally helped them derive clusters that comprehensively reflected student reasoning.

Haudek *et al.* (2012), in turn, investigated whether machine-scored written responses reflected students' mental models by conducting interviews. They figured out that there was no significant difference in the number of acid–base-related categories, *e.g.*, *strong base*, *raise pH*, and *ionization*, identified by an ML algorithm and used in oral interviews. With these categories, Haudek *et al.* (2012), subsequently, classified responses as *incorrect*, *partially correct*, and *correct* and found that *partially correct* responses were the most difficult to score accurately. The lack of accuracy was explained by the vagueness of what *partially correct* answers contained, as human raters interpreted the *partially correct* level of student sophistication differently; thus, leading to validity bias. To be noted, Donnelly *et al.* (2015) conducted interviews just like Haudek *et al.* (2012); however, the purpose of these interviews was not to analyze ML model decisions, which is why this study was not assigned to the third level of the ML-adapted ECD in *validation approaches*.

Together, the studies indicate that the validity features of ML-based formative assessment seem to be well-researched, yet. In almost every study, human scoring could be replicated with high accuracy, which led to major advancements in automating and scaling formative assessments. Moreover, several sources for misclassifications could as well be identified. Future research can build upon these insights to increase the machine-human score agreement of their assessment.

Prompt structure in ML-detected reasoning

When the possibility of applying ML in chemistry assessment emerged, many methodological considerations were initially necessary to design ML-based assessments. This has led to a research gap in integrating technological findings with aspects of mechanistic reasoning in chemistry education research. In some of the earlier studies in our sample (*cf.*, Table 2) rather *unspecific prompts* (level 1) were developed, for example, "Explain your answer for the above question" (Haudek *et al.*, 2009, p. 4; Haudek *et al.*, 2012, p. 284; Haudek *et al.*, 2015, p. 7), "How can you tell?" (Prevost *et al.*, 2012a, p. 3; Prevost *et al.*, 2012b, p. 2), "Explain your reasoning for your choice above", and "Why do you think this is so?" (Prevost *et al.*, 2012b, p. 4, 5). In their limitation section, Prevost *et al.* (2012b) realized that specific prompts probably would have motivated students more to apply their chemical content knowledge extensively. Reaching the next level, five other studies developed more *specific prompts* (level 2) to elicit mechanistic reasoning.

Dood *et al.* (2018, 2020a), Noyes *et al.* (2020), Yik *et al.* (2021), and Watts *et al.* (2023) developed their research instrument based on the findings of Cooper *et al.* (2016): in Dood *et al.* (2018, 2020a) and Yik *et al.* (2021), students first had to describe *what* happens on a molecular level for a specific acid–base reaction or a unimolecular nucleophilic substitution, respectively. After that, students had to explain *why* the reaction occurs. Similarly, Noyes *et al.* (2020) investigated learners' explanations of London dispersion forces by asking *what* happens with the electrons during

View Article Online
Review Article Chemistry Education Research and Practice

the attraction process and *why* the atoms attract. With a focus on the *how* and *why*, Watts *et al.* (2023) elicited students' content knowledge of reaction mechanisms in a writing-to-learn activity by prompting them to describe and explain racemization and acid hydrolysis of a thalidomide molecule as well as a base-free Wittig reaction. In another writing-to-learn activity, Winograd *et al.* (2021b) asked students to respond to a fake social media prompt that addressed chemical equilibrium and Le Châtelier's principle in the authentic context of ocean acidification. By scaffolding the problem-solving process with structured questions (level 3), these studies used appropriate prompts for an ML analysis of student reasoning.

Taken together, mechanistic reasoning, with or without ML techniques, can only be assessed, if prompts elicit the desired, reasoning skills. Recent studies in chemistry education research have shown the effectiveness of scaffolded prompts (Cooper *et al.*, 2016; Caspari and Graulich, 2019; Graulich and Caspari, 2020; Watts *et al.*, 2021; Kranz *et al.*, 2023). According to these findings, ML-related studies should pay attention to the structure and wording of prompts to collect evidence of mechanistic reasoning.

Sample heterogeneity in ML-detected reasoning

In many ML-related research projects focusing on mechanistic reasoning, the homogeneity of the sample population is mentioned as a limitation (*cf.*, Table 2). In thirteen studies, data is collected at a single institution (level 1), while six others collected data at different institutions (level 2). However, only Noyes *et al.* (2020) gathered data from three institutions to *investigate differential effects* (level 3), that is, how different institutional backgrounds, curricula, and student ethnicity influence the accuracy of ML models. They built an initial model fed with responses from a single student population and a combined model fed with responses from four diverse student populations. Cross-validation showed that the combined model performed just like the initial model. However, split-validation revealed that the combined model performed slightly better than the initial model in predicting the level of causal mechanistic reasoning for responses from different institutions. Consequently, the combined model could capture mechanistic reasoning about London dispersion forces more equitably in other contexts, documenting the need for collecting responses from students with diverse backgrounds.

Overall, the reviewed studies highlight that there is a risk of producing biased ML models. When the data sample is too homogenous or when bias is introduced in the way prompts are structured, rubrics are designed, and responses are coded, ML models will replicate this bias. The development of automated scoring models goes, according to Cheuk (2021), hand in hand with questions of equity. Future studies should consider how bias is introduced into their ML model, how the parameters of an algorithm must be set to reduce bias, and which differential effects occur.

Implications and future work

Based on the ML-adapted ECD, we reviewed the selected studies in each of the six categories. However, due to the rapid ML progresses, the question arises of what needs to be done to further advance the application of ML in science assessment. We expanded the ML-adapted ECD by describing a fourth level that points out future implications (*cf.*, Fig. 4).

Pedagogical purpose of ML application

The pedagogical benefits of ML's application have not yet been fully explored. For instance, in a learning environment using Just-in-Time Teaching (Novak *et al.*, 1999) guided by ML, an instructor can longitudinally assign online homework questions to students. After working on their homework, students can receive automated feedback and tailored exercises. This concept can also be realized with an Application Program Interface, which can automatically assign tailored learning assistance in terms of follow-up tutorials over a semester to each student. Including students' experiences, interests, and interdisciplinary and extracurricular skills in the evaluation may be possible as well. This technology-based approach to adaptive learning helps students self-assess their improvements and deficits, whereas instructors can monitor students' proficiency – just before the next class session starts. With an overview of students' competencies, teachers can then tailor their lessons instantly.

Beyond research focusing on delivering automated guidance directly to students, there is a need for research on teacher dashboards. Showing ML output to provide instructors with insights into the ways students respond can be useful, as it helps streamline the evaluation process. Furthermore, dashboards allow

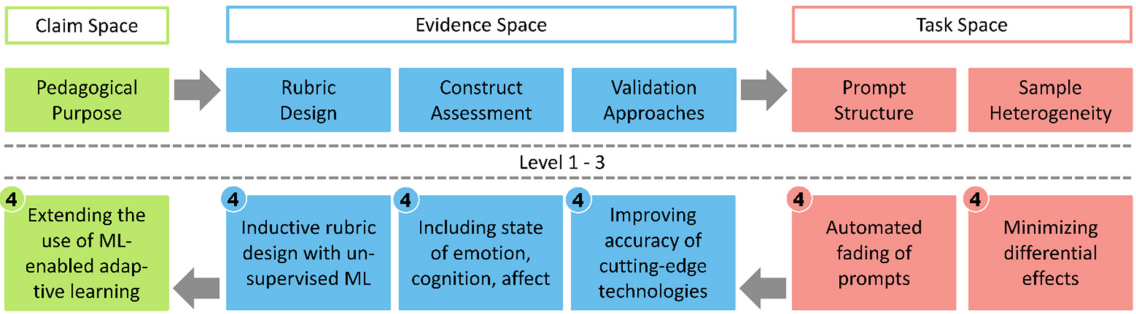


Fig. 4 Next level for implementing evidence-centered design in machine learning-based science assessment.

 This journal is © The Royal Society of Chemistry 2023

View Article Online
Chemistry Education Research and Practice Review Article

instructors to tailor direct guidance and provide individualized feedback to students (Urban-Lurain *et al.*, 2013).

Rubric design

To facilitate the rubric design for ML-based formative assessment, Raker *et al.* (2023) aggregated previously applied frameworks for evaluating mechanistic reasoning by defining four hierarchical levels of explanation sophistication: *absent*, d*escriptive*, *foundational*, and *complex*. They argue that their framework allows to predict an overall level of sophistication as well as to evaluate different reaction aspects, *e.g.*, electrophiles, nucleophiles, carbocation stability, leaving group stability, proton transfer, and solvent effects. To show the applicability of their framework, Raker *et al.* (2023) derived a rubric for the evaluation of explanation sophistication for electrophiles, independent of the reaction context. The operationalization of the proposed framework for rubric design in ML-based assessment may allow for a fine-grained evaluation of student responses.

However, most rubrics have been built deductively, which means that researchers set frameworks first and analyze data according to this framework afterward. Unsupervised pattern recognition hold, though, great potential in designing rubrics inductively. By applying a three-step process called *computational grounded theory* (Nelson, 2020), unsupervised ML can be integrated with domain-specific human expertise to raise the reliability and validity of rubrics: in a first step, unsupervised algorithms can detect patterns in data exploratively to gain breadth and quantity. After that, humans can interpret these emerging patterns to add depth and quality. In the end, algorithms can check if the qualitative human coding represents the structure of the data and if it can be applied reliably. Unsupervised ML and humans, thus, form a symbiosis in data interpretation.

For example, Rosenberg and Krist (2021) applied *computational grounded theory* to develop a nuanced rubric for evaluating the epistemic characteristics of the generality of model-based explanations. They combined a two-step clustering approach with qualitative, interpretative coding to derive a fine-grained rubric inductively from students' responses. The combination of unsupervised ML and qualitative coding finally revealed the complexity of students' conceptions. In the future, this approach may help develop computerized instructional systems that cover the whole range of student sophistication, *e.g.*, in terms of mechanistic reasoning.

Construct assessment

In many studies, constructs have been assessed by grouping responses into multiple categories. When running human- or computer-set if-then commands, this category-based approach comes with limitations. Categories formed on human- or computer-set if-then commands may not necessarily capture the heterogeneity of students' ideas since they only identify the presence of chemical concepts but not their sophistication. To mitigate this constraint, categories need to be generated to which only responses with a high level of sophistication are assigned (*e.g.*, Haudek *et al.*, 2012; Dood *et al.*, 2018). Furthermore, the covariance of different categories should be examined using web diagrams (*e.g.*, Haudek *et al.*, 2012; Prevost *et al.*, 2012b). Web diagrams do not only visualize the number of responses in each category, but they also represent connections between categories. However, one has to keep in mind that a wide range of responses may be assigned to a single category. So, the examination of conceptual categories may only reveal that specific concepts are integrated into a response, but not to what extent the concepts are elaborated, how plausible their connections are, and how complex an explanation is.

Moreover, *n*-grams and term frequency weighting allow only for identifying and counting *n* adjacent words so that a response is just represented by a sequence of numbers, called a *bag of words*. Given that the meaning of a word significantly depends on the context, semantic, logical, and technical relationships between phrases may only be analyzed superficially with this technique. *n*-Grams and related approaches cannot elicit contextual embeddings, cross-references, and implicit statements. In the future, cutting-edge technologies like BERT may better reveal contextual embeddings in written responses.

Furthermore, the definition of constructs often relies on written features when capturing mechanistic reasoning, resulting in constructs that are solely based on students' learning products (Mislevy, 2016; Zhai *et al.*, 2020a). However, ML algorithms can include other types of data in their predictive model, *e.g.*, log data (*e.g.*, Zhu *et al.*, 2017; Lee *et al.*, 2021). Gathering process data with high-tech devices or sensory information may also facilitate the definition of new constructs. For example, Ghali *et al.* (2016) developed the educational game LewiSpace to teach drawing Lewis structures. They collected data about emotional and cognitive states when predicting whether a learner will succeed or fail at a level of the game. Primarily, they used three types of sensors, such as electroencephalography, eye tracking, and facial expression recognition with an optical camera, to predict learners' success rates. Integrating knowledge from neuroscience, psychology, and computer science like behavioral, psychological, and brain wave data (*e.g.*, Wehbe *et al.*, 2014; Mason and Just, 2016) may finally help include state of emotion, cognition, and affect into science assessment (Kubsch *et al.*, 2022a).

Validation approaches

In most cases, algorithms have been trained to detect categories or *n*-grams, but this approach is a bit outdated. Recent applications of natural language processing use BERT, either for evaluating a writing-to-learn activity (Winograd *et al.*, 2021b) or peer review (Winograd *et al.*, 2021a; Dood *et al.*, 2022). BERT is pretrained with billions of sentences, which means that fewer input data, but more computation costs are needed. Notably, BERT helps detect the structure of arguments in longer texts, which makes it useful for analyzing essays with domain-general content. In science assessment, ML is, however, mostly applied for the evaluation of short, content-rich explanations. So, future research needs to check whether such cutting-edge techniques can also be used to accurately evaluate short, content-rich scientific explanations. Maybe, such technologies will produce better outcome metrics than the *n*-gram approach; Dood *et al.* (2022) and Winograd *et al.* (2021b, 2021b) in chemistry education research as well as Wulff *et al.* (2022a, 2022b) in physics education research have laid the foundation for future research in this area.

This journal is © The Royal Society of Chemistry 2023

View Article Online
Review Article Chemistry Education Research and Practice

Prompt structure

The construction of sound, logical, and coherent explanations is needed to reason appropriately about underlying processes, which requires well-designed prompts. Many prompts that elicit not only *what* happens but also *why* reactions occur have already been developed in ML-based formative assessments (Dood *et al.*, 2018, 2020a; Noyes *et al.*, 2020; Yik *et al.*, 2021; Watts *et al.*, 2023). These scaffolded prompts serve as an instructional guide when organizing explanations about mechanistic processes; however, any scaffold must be understood as a temporary support. Once a student has internalized the intended explanatory pattern, the instructional support should be faded (McNeill *et al.*, 2006; Noroozi *et al.*, 2018). To advance ML's application in the category *prompt structure*, automated text analysis could be used to immediately evaluate which prompt type is most suitable for students. By offering adapted and faded prompts afterward, students can get instructional support based on their current level of sophistication.

Sample heterogeneity

If the data used to train an ML algorithm is not representative of a diverse student population, the algorithm may be biased toward certain groups or individuals. Building upon the work of Noyes *et al.* (2020) it might, thus, be necessary to investigate the differential effects of an algorithm further, *i.e.*, how gender, age, cultural and linguistic background, socioeconomic status, educational level and resources, learning environment and challenges, prior knowledge, individual experiences, or motivation impact the conclusions drawn from automated analysis. With the help of such investigations, several steps could be taken to minimize the observed differential effects. First, gathering a heterogeneous dataset for the training of an algorithm helps ensure that the model is not biased toward a particular group of students. Subsequently, counterfactual analyses can provide insight into how particular algorithmic decisions are made, which helps identify and mitigate any biases. Influenced by the findings of counterfactual analyses, bias-sensitive algorithms can be programmed, which can be validated using metrics accounting for fairness. However, despite all technological opportunities, domain experts remain irreplaceable in revising and adjusting the algorithm.

Conclusions

The primary goal of this review was to illustrate the advancements and shortcomings that ML has brought to the formative assessment of mechanistic reasoning by providing thorough detail of the contemporary effective practices in this area. To answer our first guiding question, various contemporary effective practices have been identified as indicated by the level 3 classification in Table 2. However, the small number of articles shows that much research may still be needed to further advance ML's application in chemistry assessment, especially to capture complex reasoning.

Synthesizing the reviewed studies to answer the other two guiding questions revealed that applying ML advanced the analysis of mechanistic reasoning in some categories of the ML-adapted ECD. The tenets of the *evidence space* are comprehensively addressed by current research studies. Due to the design of multi-level rubrics capturing different sophistication levels, evidence rules that determine how to analyze features in written responses are characterized in detail. Additionally, the technological progress of recent years has facilitated the construction of measurement models that assess conceptual and semantic features as shown by Winograd *et al.* (2021b) and Watts *et al.* (2023). Besides, validation approaches are extensively reported in the reviewed studies. Future research should, following this, use fine-grained rubrics in conjunction with cutting-edge technologies to continue these contemporary effective practices. About the *task space*, the analysis turns out to be more differentiated as both, advancements and shortcomings, can be identified. Developing scaffolded prompts helped elicit evidence of mechanistic reasoning as shown by many studies. However, only Noyes *et al.* (2020) investigated differential effects in their sample, which leads us to the conclusion that a low sample heterogeneity is a major limitation in many studies. Further shortcomings can be identified in the *claim space* since in more than half of the selected studies the pedagogical purpose of ML application currently does not go beyond automating the scoring of students' mechanistic reasoning. That is, many studies did not envisage beforehand how to embed ML analysis optimally in a learning environment. To this point, research needs to be done that investigates students' and instructors' expectations of ML output in automated science assessment.

The identified advancements and shortcomings of our analysis are in alignment with other reviews and perspectives (Zhai *et al.*, 2020a, 2020c; Kubsch *et al.*, 2023). In their dimension *construct*, Zhai *et al.* (2020a) found that 81% of the reviewed studies tapped complex constructs indicating the great potential of ML in assessing such constructs. Since mechanistic reasoning is a complex construct *per se* (*cf.*, Bachtiar *et al.*, 2022), our review confirms Zhai *et al.*'s (2020a) finding. However, in many studies of our sample, there is no evidence that the investigated construct could solely be analyzed by ML. Instead, the primary intention of most reviewed studies was to replicate human scoring and deliver conventional science assessments faster. Hence, ML's potential in deepening human evaluation, analyzing more nuanced cognitive processes, and dealing with *big data* is not fully utilized, yet (Zhai *et al.*, 2020c).

For the *automaticity* of an assessment, we have, just like Zhai *et al.* (2020a), found that most studies advanced the application of straightforward supervised techniques to automate, replicate, and scale human scoring of students' responses from valid and reliable assessment instruments, while only a few employed unsupervised techniques to reveal hidden patterns. This finding indicates that the *automaticity* of scoring complex constructs with supervised techniques is already well-researched. However, as shown by the studies we reviewed, many formative assessments are only automated for a rather homogeneous sample because most studies collected data at a single institution.

Furthermore, many studies in our sample provide limited improvement to assessment *functionality*. Although technological advances have made it increasingly possible to capture

 This journal is © The Royal Society of Chemistry 2023

View Article Online
Chemistry Education Research and Practice Review Article

conceptual as well as semantic features in text responses, the pedagogical purpose of applying ML is usually not envisioned *a priori*. This seems to be a common issue in current ML studies, as Zhai *et al.* (2020a, 2020c) found as well that not even half of the reviewed studies reported how to embed their assessment in a learning environment. In the future, more studies are needed to investigate the opportunities of ML for individually and adaptively supporting students' reasoning. Integrating high-quality learning assistance into high-functional ML-based instructional settings ultimately allows for adaptively supporting students in acquiring scientific competence.

In sum, ML techniques hold great potential in supporting humans in data analysis. Subsequently, additional data sources can be combined to extend the validity of conclusions (Kubsch *et al.*, 2021). To date, research has mostly focused on assessment automation, which only changes the quantity of research efforts, but not its quality (Nelson, 2020; Kubsch *et al.*, 2023). Hence, we get only more, but the same insights as in traditional assessments. Focusing on automation leaves many ML methods unexplored, which is why the quality of evidence from data does not change. In the future, it is, for instance, necessary to integrate computational techniques with interpretative human coding to augmentatively raise the validity of both and gain new insights into cognitive processes (Sherin, 2013; Nelson, 2020; Nelson *et al.*, 2021; Kubsch *et al.*, 2023).

Along with other researchers (Zhai *et al.*, 2020a, 2020c; Kubsch *et al.*, 2023), we suggest exploring the transformative character of ML in science education beyond automating formative assessment. This includes moving away from perceiving humans as consumers of algorithmic output that takes certain decision-making processes away since such output rarely answers research questions related to the qualitative analysis of cognitive operations (Kubsch *et al.*, 2023). In other words, ML algorithms do not replace human qualitative interpretation; thoughtful human interpretation is more crucial than ever.

Author contributions

Paul P. Martin: conceptualization, investigation, methodology, formal analysis, visualization, writing – original draft. Nicole Graulich: supervision, conceptualization, investigation, methodology, visualization, writing – review and editing.

Conflicts of interest

There are no conflicts to declare.

Appendix: extended glossary for ML-related terms

Definition of ML

ML is a field that intersects computer science, data science, statistics, and artificial intelligence (AI), and assumes that algorithms can *learn* from previous experience without explicit programming (Samuel, 1959). Therefore, ML can be defined as "a computer program [that] is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell, 1997, p. 2). The key of this definition is that ML "is concerned with the question of how to construct computer programs that automatically improve with experience" (Mitchell, 1997, p. XV). In this context, the term experience refers to some kind of information, such as labels in data.

The procedure of using ML

Applying ML mainly contains two stages: learning from existing experience (training phase) and predicting new inputs (testing phase). In the training phase, a set of features is initially extracted from a dataset using pre-processing methods. Then, an algorithmic model can be generated based on the examined features of the existing data. Finally, this algorithmic model can be deployed to analyze the testing data.

Supervised ML

Supervised ML has the goal of mimicking the human ability to extract features from datasets. In doing so, the algorithm must be fed with already labeled data. After replicating the known input-output relationship, the algorithm deciphers the underlying patterns. Finally, labels in unknown data can be predicted.

Unsupervised ML

Unsupervised ML techniques operate with unlabeled data, which means these techniques are deployed to reveal hidden patterns. All these hidden patterns can afterward be utilized to establish new labels. Since there is no need for human coding at all, unsupervised approaches save time, resources, costs, and effort.

Semi-supervised ML

A hybrid form of both supervised and unsupervised ML is semi-supervised ML, deployed for data consisting of labeled and unlabeled cases. After building a model using labeled data, the semi-supervised algorithm is utilized to forecast unlabeled data. The model can then be adjusted and re-trained with larger datasets and less coding effort.

Self-validation

Self-validation uses the same data for constructing and evaluating the ML model. Since the quality of the model is not measured for new data, the approach of split-validation is often preferred.

Split-validation

When applying split-validation, the dataset must be randomly divided into a training and a testing set to analyze the accuracy of predictions even for unknown inputs.

Cross-validation

Cross-validation carries the idea of split-validation. For k-fold cross-validation, data is first partitioned into k randomly selected subsets with the same number of responses. Whereas $(k-1)$ folds are then used for model training, the remaining fold is withheld for model validation. The codes of the remaining fold are predicted by the algorithm, which is built on the other $(k-1)$ folds. This method is repeated until each of the k validation sets has been coded by the on $(k-1)$ folds built model. Finally, an average agreement between human and computer codes can be calculated.

Reference to AI

The application of ML and AI increases not only in the field of science education but also in medicine, mobility, media, marketing, fraud detection, and face recognition; however, there is hardly a distinction between the two terms (Kühl *et al.*, 2020). This fact can be explained by the various relationships between the two technologies. ML algorithms benefit from training to solve real-world problems automatically through computational resources (Mohri *et al.*, 2012). ML represents, thus, a set of methods for learning from patterns in data (Kühl *et al.*, 2020). AI is, in turn, an application of several techniques, among others ML, which aim to perform intelligent and cognitive tasks while imitating human behavior. In other words, AI refers to machines that simulate human reasoning steps (Newell and Simon, 1961; Bellmann, 1978; Haugeland, 1989), leading to the criterion that AI must accomplish human tasks at least as well as humans do (Rich *et al.*, 2009).

Acknowledgements

This publication is part of the first author's doctoral thesis (Dr. rer. nat.) at the Faculty of Biology and Chemistry, Justus-Liebig-University Giessen, Germany. We thank Marcus Kubsch for a thoughtful discussion of the manuscript. Furthermore, we thank all members of the Graulich group for fruitful discussions. Finally, Paul P. Martin thanks Amber Dood, Field Watts, and Brandon Yik for the helpful meetings on machine learning in chemistry education research.

References

Allen D. and Tanner K., (2006), Rubrics: Tools for Making Learning Goals and Evaluation Criteria Explicit for Both Teachers and Learners, *CBE Life Sci. Educ.*, **5**, 197–203.

Bachtiar R. W., Meulenbroeks R. F. G. and van Joolingen W. R., (2022), Mechanistic reasoning in science education: a literature review, *EURASIA J. Math. Sci. Tech. Ed.*, **18**, em2178.

Beggrow E. P., Ha M., Nehm R. H., Pearl D. and Boone W. J., (2014), Assessing Scientific Practices Using Machine-Learning Methods: How Closely Do They Match Clinical Interview Performance? *J. Sci. Educ. Technol.*, **23**, 160–182.

Bellmann R., (1978), *An Introduction to Artificial Intelligence: Can Computers Think?* Boyd and Fraser.

Birenbaum M. and Tatsuoka K. K., (1987), Open-Ended Versus Multiple-Choice Response Formats – It Does Make a Difference for Diagnostic Purposes, *Appl. Psychol. Meas.*, **11**, 385–395.

Bishop C. M., (2006), *Pattern Recognition and Machine Learning*, New York: Springer.

Bolger M. S., Kobiela M., Weinberg P. J. and Lehrer R., (2012), Children's Mechanistic Reasoning, *Cogn. Instr.*, **30**, 170–206.

Carey S., (1995), *Causal cognition: A multidisciplinary debate*, New York, NY, US: Clarendon Press/Oxford University Press, pp. 268–308.

Caspari I. and Graulich N., (2019), Scaffolding the structure of organic chemistry students' multivariate comparative mechanistic reasoning, *Int. J. Physc. Chem. Ed.*, **11**, 31–43.

Caspari I., Kranz D. and Graulich N., (2018a), Resolving the complexity of organic chemistry students' reasoning through the lens of a mechanistic framework, *Chem. Educ. Res. Pract.*, **19**, 1117–1141.

Caspari I., Weinrich M., Sevian H. and Graulich N., (2018b), This mechanistic step is "*productive*": organic chemistry students' backward-oriented reasoning, *Chem. Educ. Res. Pract.*, **19**, 42–59.

Cheuk T., (2021), Can AI be racist? Color-evasiveness in the application of machine learning to science assessments, *Sci. Educ.*, **105**, 825–836.

Cooper M. M., (2015), Why Ask Why? *J. Chem. Educ.*, **92**, 1273–1279.

Cooper M. M., Kouyoumdjian H. and Underwood S. M., (2016), Investigating Students' Reasoning about Acid–Base Reactions, *J. Chem. Educ.*, **93**, 1703–1712.

Deeva G., Bogdanova D., Serral E., Snoeck M. and De Weerdt J., (2021), A review of automated feedback systems for learners: Classification framework, challenges and opportunities, *Comput. Educ.*, **162**, 104094.

DeGlopper K. S., Schwarz C. E., Ellias N. J. and Stowe R. L., (2022), Impact of Assessment Emphasis on Organic Chemistry Students' Explanations for an Alkene Addition Reaction, *J. Chem. Educ.*, **99**, 1368–1382.

Deng J. M., Rahmani M. and Flynn A. B., (2022), The role of language in students' justifications of chemical phenomena, *Int. J. Sci. Educ.*, **44**, 2131–2151.

diSessa A. A., (1993), Toward an Epistemology of Physics, *Cogn. Instr.*, **10**, 105–225.

Donnelly D. F., Vitale J. M. and Linn M. C., (2015), Automated Guidance for Thermodynamics Essays: Critiquing Versus Revisiting, *J. Sci. Educ. Technol.*, **24**, 861–874.

Dood A. J. and Watts F. M., (2022a), Mechanistic Reasoning in Organic Chemistry: A Scoping Review of How Students Describe and Explain Mechanisms in the Chemistry Education Research Literature, *J. Chem. Educ.*, **99**, 2864–2876.

Dood A. J. and Watts F. M., (2022b), Students' Strategies, Struggles, and Successes with Mechanism Problem Solving in Organic Chemistry: A Scoping Review of the Research Literature, *J. Chem. Educ.*, **100**, 53–68.

Dood A. J., Fields K. B. and Raker J. R., (2018), Using Lexical Analysis to Predict Lewis Acid-Base Model Use in Response to an Acid-Base Proton-Transfer Reaction, *J. Chem. Educ.*, **95**, 1267–1275.

 This journal is © The Royal Society of Chemistry 2023

Dood A. J., Fields K. B., Cruz-Ramírez de Arellano D. and Raker J. R., (2019), Development and evaluation of a Lewis acid-base tutorial for use in postsecondary organic chemistry courses, *Can. J. Chem.*, **97**, 711–721.

Dood A. J., Dood J. C., Cruz-Ramírez de Arellano D., Fields K. B. and Raker J. R., (2020a), Analyzing explanations of substitution reactions using lexical analysis and logistic regression techniques, *Chem. Educ. Res. Pract.*, **21**, 267–286.

Dood A. J., Dood J. C., Cruz-Ramírez de Arellano D., Fields K. B. and Raker J. R., (2020b), Using the Research Literature to Develop an Adaptive Intervention to Improve Student Explanations of an S_N1 Reaction Mechanism, *J. Chem. Educ.*, **97**, 3551–3562.

Dood A. J., Winograd B. A., Finkenstaedt-Quinn S. A., Gere A. R. and Shultz G. V., (2022), PeerBERT: Automated Characterization of Peer Review Comments Across Courses, in *Proceedings of the LAK22: 12th International Learning Analytics and Knowledge Conference*, New York, NY, pp. 492–499.

Gerard L. F., Matuk C., McElhaney K. and Linn M. C., (2015), Automated, adaptive guidance for K-12 education, *Educ. Reas. Rev.*, **15**, 41–58.

Gerard L. F., McElhaney K. W., Rafferty A. N., Ryoo K., Liu O. L. and Linn M. C., (2016), Automated Guidance for Student Inquiry, *J. Educ. Psychol.*, **108**, 60–81.

Ghali R., Ouellet S. and Frasson C., (2016), LewiSpace: an Exploratory Study with a Machine Learning Model in an Educational Game, *J. Educ. Train. Stud.*, **4**, 192–201.

Glaser R., Lesgold A. and Lajoie S., (1987), Toward a Cognitive Theory for the Measurement of Achievement, in Ronning R. R., Glover J. A., Conoley J. C. and Witt J. C. (ed.), *The Influence of Cognitive Psychology on Testing and Measurement*, Lawrence Erlbaum, pp. 41–85.

Glennan S., (2002), Rethinking Mechanistic Explanation, *Philos. Sci.*, **69**, S342–S353.

Gobert J. D., Sao Pedro M., Raziuddin J. and Baker R. S., (2013), From Log Files to Assessment Metrics: Measuring Students' Science Inquiry Skills Using Educational Data Mining, *J. Learn. Sci.*, **22**, 521–563.

Gobert J. D., Baker R. and Wixon M. B., (2015), Operationalizing and Detecting Disengagement Within Online Science Microworlds, *Educ. Psychol.*, **50**, 43–57.

Graulich N., (2015), The tip of the iceberg in organic chemistry classes: how do students deal with the invisible? *Chem. Educ. Res. Pract.*, **16**, 9–21.

Graulich N. and Caspari I., (2020), Designing a scaffold for mechanistic reasoning in organic chemistry, *Chem. Teach. Int.*, **3**, 19–30.

Graulich N. and Schween M., (2018), Concept-Oriented Task Design: Making Purposeful Case Comparisons in Organic Chemistry, *J. Chem. Educ.*, **95**, 376–383.

Grimberg B. I. and Hand B., (2009), Cognitive Pathways: Analysis of students' written texts for science understanding, *Int. J. Sci. Educ.*, **31**, 503–521.

Grove N. P. and Lowery Bretz S., (2012), A continuum of learning: from rote memorization to meaningful learning in organic chemistry, *Chem. Educ. Res. Pract.*, **13**, 201–208.

Ha M. and Nehm R., (2016), The Impact of Misspelled Words on Automated Computer Scoring: A Case Study of Scientific Explanations, *J. Sci. Educ. Technol.*, **25**, 358–374.

Ha M., Nehm R. H., Urban-Lurain M. and Merrill J. E., (2011), Applying Computerized-Scoring Models of Written Biological Explanations across Courses and Colleges: Prospects and Limitations, *CBE Life Sci. Educ.*, **10**, 379–393.

Hammer D., (2000), Student resources for learning introductory physics, *Am. J. Phys.*, **68**, S52–S59.

Haudek K. C. and Zhai X., (2021), Exploring the Effect of Assessment Construct Complexity on Machine Learning Scoring of Argumentation, Presented in part at the National Association of Research in Science Teaching Annual Conference, Virtual.

Haudek K. C., Moscarella R. A., Urban-Lurain M., Merrill J. E., Sweeder R. D. and Richmond G., (2009), Using lexical analysis software to understand student knowledge transfer between chemistry and biology, Presented in part at the National Association of Research in Science Teaching Annual Conference, Garden Grove, CA.

Haudek K. C., Kaplan J. J., Knight J., Long T. M., Merrill J. E., Munn A., Nehm R. H., Smith M. and Urban-Lurain M., (2011), Harnessing Technology to Improve Formative Assessment of Student Conceptions in STEM: Forging a National Network, *CBE Life Sci. Educ.*, **10**, 149–155.

Haudek K. C., Prevost L. B., Moscarella R. A., Merrill J. E. and Urban-Lurain M., (2012), What Are They Thinking? Automated Analysis of Student Writing about Acid-Base Chemistry in Introductory Biology, *CBE Life Sci. Educ.*, **11**, 283–293.

Haudek K. C., Moscarella R. A., Weston M., Merrill J. E. and Urban-Lurain M., (2015), Construction of rubrics to evaluate content in students' scientific explanation using computerized text analysis, Presented in part at the National Association of Research in Science Teaching Annual Conference, Chicago, IL.

Haudek K. C., Wilson C. D., Stuhlsatz M. A. M., Donovan B., Bracey Z. B., Gardner A., Osborne J. F. and Cheuk T., (2019), Using automated analysis to assess middle school students' competence with scientific argumentation, Presented in part at the National Conference on Measurement in Education (NCME), Annual Conference, Toronto, ON.

Haugeland J., (1989), *Artificial Intelligence: The Very Idea*, MIT Press.

Illari P. M. and Williamson J., (2012), What is a mechanism? Thinking about mechanisms across the sciences, *Eur. J. Philos. Sci.*, **2**, 119–135.

Jescovitch L. N., Doherty J. H., Scott E. E., Cerchiara J. A., Wenderoth M. P., Urban-Lurain M., Merrill J. E. and Haudek K. C., (2019a), Challenges in Developing Computerized Scoring Models for Principle-Based Reasoning in a Physiology Context, Presented in part at the National Association of Research in Science Teaching Annual Conference, Baltimore, MD.

Jescovitch L. N., Scott E. E., Cerchiara J. A., Doherty J. H., Wenderoth M. P., Merrill J. E., Urban-Lurain M. and Haudek K. C., (2019b), Deconstruction of Holistic Rubrics into Analytic Bins for Large-Scale Assessments of Students'

Reasoning of Complex Science Concepts, *Pract. Assess. Res. Eval.*, **24**, 1–13, DOI: **10.7275/9h7f-mp76**.

Jescovitch L. N., Scott E. E., Cerchiara J. A., Merrill J. E., Urban-Lurain M., Doherty J. H. and Haudek K. C., (2021), Comparison of Machine Learning Performance Using Analytic and Holistic Coding Approaches Across Constructed Response Assessments Aligned to a Science Learning Progression, *J. Sci. Educ. Technol.*, **30**, 150–167.

Kang H., Thompson J. and Windschitl M., (2014), Creating Opportunities for Students to Show What They Know: The Role of Scaffolding in Assessment Tasks, *Sci. Educ.*, **98**, 674–704.

Kerr P., (2016), Adaptive learning, *ELT J.*, **70**, 88–93.

Kraft A., Strickland A. M. and Bhattacharyya G., (2010), Reasonable reasoning: multi-variate problem-solving in organic chemistry, *Chem. Educ. Res. Pract.*, **11**, 281–292.

Kranz D., Schween M. and Graulich N., (2023), Patterns of reasoning – exploring the interplay of students' work with a scaffold and their conceptual knowledge in organic chemistry, *Chem. Educ. Res. Pract.*, DOI: **10.1039/d2rp00132b**.

Krist C., Schwarz C. V. and Reiser B. J., (2019), Identifying Essential Epistemic Heuristics for Guiding Mechanistic Reasoning in Science Learning, *J. Learn. Sci.*, **28**, 160–205.

Kubsch M., Rosenberg J. M. and Krist C., (2021), Beyond Supervision: Human/Machine Distributed Learning in Learning Sciences Research, in *Proceedings of the 15th International Conference of the Learning Sciences-ICLS 2021*, Bochum, Germany, pp. 897–898.

Kubsch M., Caballero D. and Uribe P., (2022a), Once More with Feeling: Emotions in Multimodal Learning Analytics, in Giannakos M., Spikol D., Di Mitri D., Sharma K., Ochoa X. and Hammad R. (ed.), *The Multimodal Learning Analytics Handbook*, Cham: Springer International Publishing, pp. 261–285.

Kubsch M., Czinczel B., Lossjew J., Wyrwich T., Bednorz D., Bernholt S., Fiedler D., Strauß S., Cress U., Drachsler H., Neumann K. and Rummel N., (2022b), Toward learning progression analytics—Developing learning environments for the automated analysis of learning using evidence centered design, *Front. Educ.*, **7**, 1–15, DOI: **10.3389/feduc.2022.981910**.

Kubsch M., Krist C. and Rosenberg J. M., (2023), Distributing epistemic functions and tasks – A framework for augmenting human analytic power with machine learning in science education research, *J. Res. Sci. Teach.*, **60**, 423–447.

Kuechler L. W. and Simkin M. G., (2010), Why Is Performance on Multiple-Choice Tests and Constructed-Response Tests Not More Closely Related? Theory and an Empirical Test, *Dec. Sci. J. Innov. Educ.*, **8**, 55–73.

Kühl N., Goutier M., Hirt R. and Satzger G., (2020), Machine Learning in Artificial Intelligence: Towards a Common Understanding, *arXiv*, preprint, arXiv:2004.04686, DOI: **10.48550/arXiv.2004.04686**.

Lee H.-S., Liu O. L. and Linn M. C., (2011), Validating Measurement of Knowledge Integration in Science Using Multiple-Choice and Explanation Items, *Appl. Meas. Educ.*, **24**, 115–136.

Lee H.-S., Gweon G.-H., Lord T., Paessel N., Pallant A. and Pryputniewicz S., (2021), Machine Learning-Enabled Automated Feedback: Supporting Students' Revision of Scientific Arguments Based on Data Drawn from Simulation, *J. Sci. Educ. Technol.*, **30**, 168–192.

Lieber L. S. and Graulich N., (2020), Thinking in Alternatives—A Task Design for Challenging Students' Problem-Solving Approaches in Organic Chemistry, *J. Chem. Educ.*, **97**, 3731–3738.

Lieber L. S. and Graulich N., (2022), Investigating students' argumentation when judging the plausibility of alternative reaction pathways in organic chemistry, *Chem. Educ. Res. Pract.*, **23**, 38–53.

Lieber L. S., Ibraj K., Caspari-Gnann I. and Graulich N., (2022a), Closing the gap of organic chemistry students' performance with an adaptive scaffold for argumentation patterns, *Chem. Educ. Res. Pract.*, **23**, 811–828.

Lieber L. S., Ibraj K., Caspari-Gnann I. and Graulich N., (2022b), Students' Individual Needs Matter: A Training to Adaptively Address Students' Argumentation Skills in Organic Chemistry, *J. Chem. Educ.*, **99**, 2754–2761.

Linn M. C. and Eylon B.-S., (2011), *Science learning and instruction: Taking advantage of technology to promote knowledge integration*, New York, NY: Routledge.

Linn M. C., Gerard L. F., Ryoo K., McElhaney K., Liu O. L. and Rafferty A. N., (2014), Education technology. Computer-guided inquiry to improve science learning, *Science*, **344**, 155–156.

Liu O. L., Brew C., Blackmore J., Gerard L., Madhok J. and Linn M. C., (2014), Automated Scoring of Constructed-Response Science Items: Prospects and Obstacles, *Educ. Meas*, **33**, 19–28.

Liu O. L., Rios J. A., Heilman M., Gerard L. and Linn M. C., (2016), Validation of Automated Scoring of Science Assessments, *J. Res. Sci. Teach.*, **53**, 215–233.

Machamer P., Darden L. and Craver C. F., (2000), Thinking About Mechanisms, *Philos. Sci.*, **67**, 1–25.

Maestrales S., Zhai X., Touitou I., Baker Q., Schneider B. and Krajcik J., (2021), Using Machine Learning to Score Multi-Dimensional Assessments of Chemistry and Physics, *J. Sci. Educ. Technol.*, **30**, 239–254.

Mao L., Liu O. L., Roohr K., Belur V., Mulholland M., Lee H.-S. and Pallant A., (2018), Validation of Automated Scoring for a Formative Assessment that Employs Scientific Argumentation, *Educ. Assess.*, **23**, 121–138.

Mason R. A. and Just M. A., (2016), Neural Representations of Physics Concepts, *Psychol. Sci.*, **27**, 904–913.

McNeill K. L., Lizotte D. J., Krajcik J. and Marx R. W., (2006), Supporting Students' Construction of Scientific Explanations by Fading Scaffolds in Instructional Materials, *J. Learn. Sci.*, **15**, 153–191.

Messick S., (1994), The Interplay of Evidence and Consequences in the Validation of Performance Assessments, *Educ. Res.*, **23**, 13–23.

Mislevy R. J., (2006), Cognitive psychology and educational assessment, in Brennan R. L. (ed.), *Educational measurement*, Phoenix: Greenwood Press, vol. 4, pp. 257–305.

 This journal is © The Royal Society of Chemistry 2023

Mislevy R. J., (2016), How Developments in Psychology and Technology Challenge Validity Argumentation, *J. Educ. Meas.*, **53**, 265–292.

Mislevy R. J. and Haertel G. D., (2007), Implications of Evidence-Centered Design for Educational Testing, *Educ. Meas*, **25**, 6–20.

Mislevy R. J., Almond R. G. and Lukas J. F., (2003a), A Brief Introduction to Evidence-Centered Design, *ETS Res. Rep. Ser.*, **2003**, i–29.

Mislevy R. J., Steinberg L. S. and Almond R. G., (2003b), Focus Article: On the Structure of Educational Assessments, *Meas. Interdiscip. Sci Res. Per.*, **1**, 3–62.

Mitchell T. M., (1997), *Machine Learning*, New York, NY: McGraw Hill.

Mjolsness E. and Decoste D., (2001), Machine Learning for Science: State of the Art and Future Prospects, *Science*, **293**, 2051–2055.

Mohri M., Rostamizadeh A. and Talwalkar A., (2012), *Foundation of Machine Learning*, Cambridge, MA London, England: The MIT Press.

National Research Council, (2012), *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*, National Academic Press.

Nehm R. H., Ha M. and Mayfield E., (2012), Transforming Biology Assessment with Machine Learning: Automated Scoring of Written Evolutionary Explanations, *J. Sci. Educ. Technol.*, **21**, 183–196.

Nelson L. K., (2020), Computational Grounded Theory: A Methodological Framework, *Sociol. Methods Res.*, **49**, 3–42.

Nelson L. K., Burk D., Knudsen M. and McCall L., (2021), The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods, *Sociol. Methods Res.*, **50**, 202–237.

Newell A. and Simon H. A., (1961), GPS, A Program that Simulates Human Thought, in Billing H. (ed.), *Lernende Automaten*, München: Oldenbourg, pp. 109–124.

Noroozi O., Kirschner P. A., Biemanns H. J. A. and Mulder M., (2018), Promoting Argumentation Competence: Extending from First- to Second-Order Scaffolding Through Adaptive Fading, *Educ. Psychol. Rev.*, **30**, 153–176.

Novak G. M., Gavrin A., Patterson E. and Christian W., (1999), *Just-In-Time Teaching: Blending Active Learning with Web Technology*, Upper Saddle River NJ: Prentice Hall.

Noyes K., McKay R. L., Neumann M., Haudek K. C. and Cooper M. M., (2020), Developing Computer Resources to Automate Analysis of Students' Explanations of London Dispersion Forces, *J. Chem. Educ.*, **97**, 3923–3936.

Noyes K., Carlson C. G., Stoltzfus J. R., Schwarz C. V., Long T. M. and Cooper M. M., (2022), A Deep Look into Designing a Task and Coding Scheme through the Lens of Causal Mechanistic Reasoning, *J. Chem. Educ.*, **99**, 874–885.

Pellegrino J. W., (2013), Proficiency in Science: Assessment Challenges and Opportunities, *Science*, **340**, 320–323.

Pellegrino J., DiBello L. and Goldman S., (2016), A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments, *Educ. Psychol.*, **51**, 59–81.

Prevost L. B., Haudek K. C., Merrill J. E. and Urban-Lurain M., (2012a), Deciphering student ideas on thermodynamics using computerized lexical analysis of student writing, Presented in part at the ASEE Annual Conference & Exposition, San Antonio, TX.

Prevost L. B., Haudek K. C., Merrill J. E. and Urban-Lurain M., (2012b), Examining student constructed explanations of thermodynamics using lexical analysis, Presented in part at the 2012 IEEE Frontiers in Education Conference, Seattle, WA.

Prevost L. B., Haudek K. C., Cooper M. M. and Urban-Lurain M., (2014), Computerized Lexical Analysis of Students' Written Interpretations of Chemical Representations, Presented in part at the National Association of Research in Science Teaching Annual Conference, Pittsburgh, PA.

Rafferty A. N., Gerard L. F., McElhaney K. W. and Linn M. C., (2013), Automating Guidance for Students' Chemistry Drawings, Presented in part at the Artificial Intelligence in Education Conference, Memphis, TN.

Rafferty A. N., Gerard L. F., McElhaney K. and Linn M. C., (2014), Promoting Student Learning through Automated Formative Guidance on Chemistry Drawings, in *Proceedings of the International Society of the Learning Sciences*, Boulder, CO, pp. 386–393.

Raker J. R., Yik B. J. and Dood A. J., (2023), Development of a Generalizable Framework for Machine Learning-Based Evaluation of Written Explanations of Reaction Mechanisms from the Postsecondary Organic Chemistry Curriculum, in Graulich N. and Shultz G. V. (ed.), *Student Reasoning in Organic Chemistry*, The Royal Society of Chemistry, pp. 304–319.

Rich E., Knight K. and Nair S. B., (2009), *Artificial Intelligence*, McGraw-Hill.

Riconscente M. M., Mislevy R. J. and Corrigan S., (2015), Evidence-Centered Design, in Lane S., Raymond M. R. and Haladyna T. M. (ed.), *Handbook of Test Development*, New York, NY: Taylor & Francis/Routledge, vol. 2, pp. 40–63.

Rosenberg J. M. and Krist C., (2021), Combining Machine Learning and Qualitative Methods to Elaborate Students' Ideas About the Generality of their Model-Based Explanations, *J. Sci. Educ. Technol.*, **30**, 255–267.

Rupp A. A., Levy R., Dicerbo K. E., Sweet S. J., Crawford A. V., Caliço T., Benson M., Fay D., Kunze K. L., Mislevy R. J. and Behrens J. T., (2012), Putting ECD into Practice: The Interplay of Theory and Data in Evidence Models within a Digital Learning Environment, *J. Educ. Data Min.*, **4**, 49–110.

Russ R. S., Scherr R. E., Hammer D. and Mikeska J., (2008), Recognizing Mechanistic Reasoning in Student Scientific Inquiry: A Framework for Discourse Analysis Developed From Philosophy of Science, *Sci. Educ.*, **92**, 499–525.

Samuel A. L., (1959), Some Studies in Machine Learning Using the Game of Checkers, *IBM J. Res. Dev.*, **3**, 211–229.

Sao Pedro M. A., de Baker R. S. J., Gobert J. D., Montalvo O. and Nakama A., (2013), Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill, *User Model. User-Adapt. Interact.*, **23**, 1–39, DOI: 10.1007/s11257-011-9101-0.

Sevian H. and Talanquer V., (2014), Rethinking chemistry: a learning progression on chemical thinking, *Chem. Educ. Res. Pract.*, **15**, 10–23.

Sherin B., (2013), A Computational Study of Commonsense Science: An Exploration in the Automated Analysis of Clinical Interview Data, *J. Learn. Sci.*, **22**, 600–638.

Songer N. B. and Ruiz-Primo M. A., (2012), Assessment and Science Education: Our Essential New Priority? *J. Res. Sci. Teach.*, **49**, 683–690.

Southard K., Wince T., Meddleton S. and Bolger M. S., (2016), Features of Knowledge Building in Biology: Understanding Undergraduate Students' Ideas about Molecular Mechanisms, *CBE Life Sci. Educ.*, **15**, ar7.

Stowe R. L. and Cooper M. M., (2017), Practicing What We Preach: Assessing ''Critical Thinking'' in Organic Chemistry, *J. Chem. Educ.*, **94**, 1852–1859.

Stowe R. L., Scharlott L. J., Ralph V. R., Becker N. M. and Cooper M. M., (2021), You Are What You Assess: The Case for Emphasizing Chemistry on Chemistry Assessments, *J. Chem. Educ.*, **98**, 2490–2495.

Talanquer V., (2009), On Cognitive Constraints and Learning Progressions: The case of ''structure of matter'', *Int. J. Sci. Educ.*, **31**, 2123–2136.

Tansomboon C., Gerard L. F., Vitale J. M. and Linn M. C., (2017), Designing Automated Guidance to Promote Productive Revision of Science Explanations, *Int. J. Artif. Intell. Educ.*, **27**, 729–757.

Urban-Lurain M., Moscarella R. A., Haudek K. C., Giese E., Sibley D. F. and Merrill J. E., (2009), Beyond Multiple Choice Exams: Using Computerized Lexical Analysis to Understand Students' Conceptual Reasoning in STEM Disciplines, Presented in part at the 2009 IEEE Frontiers in Education Conference, San Antonio, TX.

Urban-Lurain M., Moscarella R. A., Haudek K. C., Giese E., Merrill J. E. and Sibley D., (2010), Insight into Student Thinking in STEM: Lessons Learned from Lexical Analysis of Student Writing, Presented in part at the National Association of Research in Science Teaching Annual Conference, Philadelphia, PA.

Urban-Lurain M., Prevost L., Haudek K. C., Henry E. N., Berry M. and Merrill J. E., (2013), Using Computerized Lexical Analysis of Student Writing to Support Just-in-Time Teaching in Large Enrollment STEM Courses, Presented in part at the 2013 IEEE Frontiers in Education Conference, Oklahoma City, OK.

van Mil M. H. W., Postma P. A., Boerwinkel D. J., Klaasen K. and Waarlo A. J., (2016), Molecular Mechanistic Reasoning: Toward Bridging the Gap Between the Molecular and Cellular Levels in Life Science Education, *Sci. Educ.*, **100**, 517–585.

Vitale J. M., Lai K. and Linn M. C., (2015), Taking advantage of automated assessment of student-constructed graphs in science, *J. Res. Sci. Teach.*, **52**, 1426–1450.

Vitale J. M., McBride E. and Linn M. C., (2016), Distinguishing complex ideas about climate change: knowledge integration vs. specific guidance, *Int. J. Sci. Educ.*, **38**, 1548–1569.

Wang C., Liu X., Wang L., Sun Y. and Zhang H., (2021), Automated Scoring of Chinese Grades 7–9 Students' Competence in Interpreting and Arguing from Evidence, *J. Sci. Educ. Technol.*, **30**, 269–282.

Watts F. M., Zaimi I., Kranz D., Graulich N. and Shultz G. V., (2021), Investigating students' reasoning over time for case comparisons of acyl transfer reaction mechanisms, *Chem. Educ. Res. Pract.*, **22**, 364–381.

Watts F. M., Dood A. J. and Shultz G. V., (2023), Developing machine learning models for automated analysis of organic chemistry students' written descriptions of organic reaction mechanisms, in Graulich N. and Shultz G. V. (ed.), *Student Reasoning in Organic Chemistry*, The Royal Society of Chemistry, pp. 285–303.

Watts F. M., Park G. Y., Petterson M. N. and Shultz G. V., (2022), Considering alternative reaction mechanisms: students' use of multiple representations to reason about mechanisms for a writing-to-learn assignment, *Chem. Educ. Res. Pract.*, **23**, 486–507.

Wehbe L., Murphy B., Talukdar P., Fyshe A., Ramdas A. and Mitchell T. M., (2014), Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses, *PLoS One*, **9**, e112575.

Williamson D. M., Xi X. and Breyer F. J., (2012), A Framework for Evaluation and Use of Automated Scoring, *Educ. Meas*, **31**, 2–13.

Winograd B. A., Dood A. J., Finkenstaedt-Quinn S. A., Gere A. R. and Shultz G. V., (2021a), Automating Characterization of Peer Review Comments in Chemistry Courses, in *Proceedings of the 14th Computer-Supported Collaborative Learning (CSCL)*, Bochum, Germany, pp. 11–18.

Winograd B. A., Dood A. J., Moon A., Moeller R., Shultz G. V. and Gere A. R., (2021b), Detecting High Orders of Cognitive Complexity in Students' Reasoning in Argumentative Writing About Ocean Acidification, in *Proceedings of the LAK21: 11th International Learning Analytics and Knowledge Conference*, New York, NY, pp. 586–591.

Wood D., Bruner J. S. and Ross G., (1976), The role of tutoring in problem solving, *J. Child Psychol. Psychiatry*, **17**, 89–100.

Wulff P., Buschhüter D., Westphal A., Mientus L., Nowak A. and Borowski A., (2022a), Bridging the Gap Between Qualitative and Quantitative Assessment in Science Education Research with Machine Learning—A Case for Pretrained Language Models-Based Clustering, *J. Sci. Educ. Technol.*, **31**, 490–513.

Wulff P., Mientus L., Nowak A. and Borowski A., (2022b), Utilizing a Pretrained Language Model (BERT) to Classify Preservice Physics Teachers' Written Reflections, *Int. J. Artif. Intell. Educ.*, 1–28, DOI: **10.1007/s40593-022-00290-6**.

Yik B. J., Dood A. J., Cruz-Ramírez de Arellano D., Fields K. B. and Raker J. R., (2021), Development of a machine learning-based tool to evaluate correct Lewis acid-base model use in written responses to open-ended formative assessment items, *Chem. Educ. Res. Pract.*, **22**, 866–885.

Yik B. J., Dood A. J., Frost S. J. H., Cruz-Ramírez de Arellano D., Fields K. B. and Raker J. R., (2023), Generalized rubric for level of explanation sophistication for nucleophiles in

 This journal is © The Royal Society of Chemistry 2023

View Article Online
Chemistry Education Research and Practice Review Article

organic chemistry reaction mechanisms, *Chem. Educ. Res. Pract.*, **24**, 263–282.

Zhai X., (2019), Call for Papers: Applying Machine Learning in Science Assessment: Opportunity and Challenge, *J. Sci. Educ. Technol.*, 1–3, DOI: **10.13140/RG.2.2.10914.07365**.

Zhai X., (2021), Practices and Theories: How Can Machine Learning Assist in Innovative Assessment Practices in Science Education, *J. Sci. Educ. Technol.*, **30**, 139–149.

Zhai X., Haudek K. C., Shi L., Nehm R. H. and Urban-Lurain M., (2020a), From substitution to redefinition: a framework of machine learning-based science assessment, *J. Res. Sci. Teach.*, **57**, 1430–1459.

Zhai X., Haudek K. C., Stuhlsatz M. A. M. and Wilson C. D., (2020b), Evaluation of Construct-Irrelevant Variance Yielded by Machine and Human Scoring of a Science Teacher PCK Constructed Response Assessment, *Stud. Educ. Eval.*, **67**, 100916.

Zhai X., Yin Y., Pellegrino J. W., Haudek K. C. and Shi L., (2020c), Applying machine learning in science assessment: a systematic review, *Stud. Sci. Educ.*, **56**, 111–151.

Zhai X., Haudek K. C. and Ma W., (2022a), Assessing Argumentation Using Machine Learning and Cognitive Diagnostic Modeling, *Res. Sci. Educ.*, 1–20, DOI: **10.1007/s11165-022-10062-w**.

Zhai X., He P. and Krajcik J., (2022b), Applying machine learning to automatically assess scientific models, *J. Res. Sci. Teach.*, 1–30, DOI: **10.1002/tea.21773**.

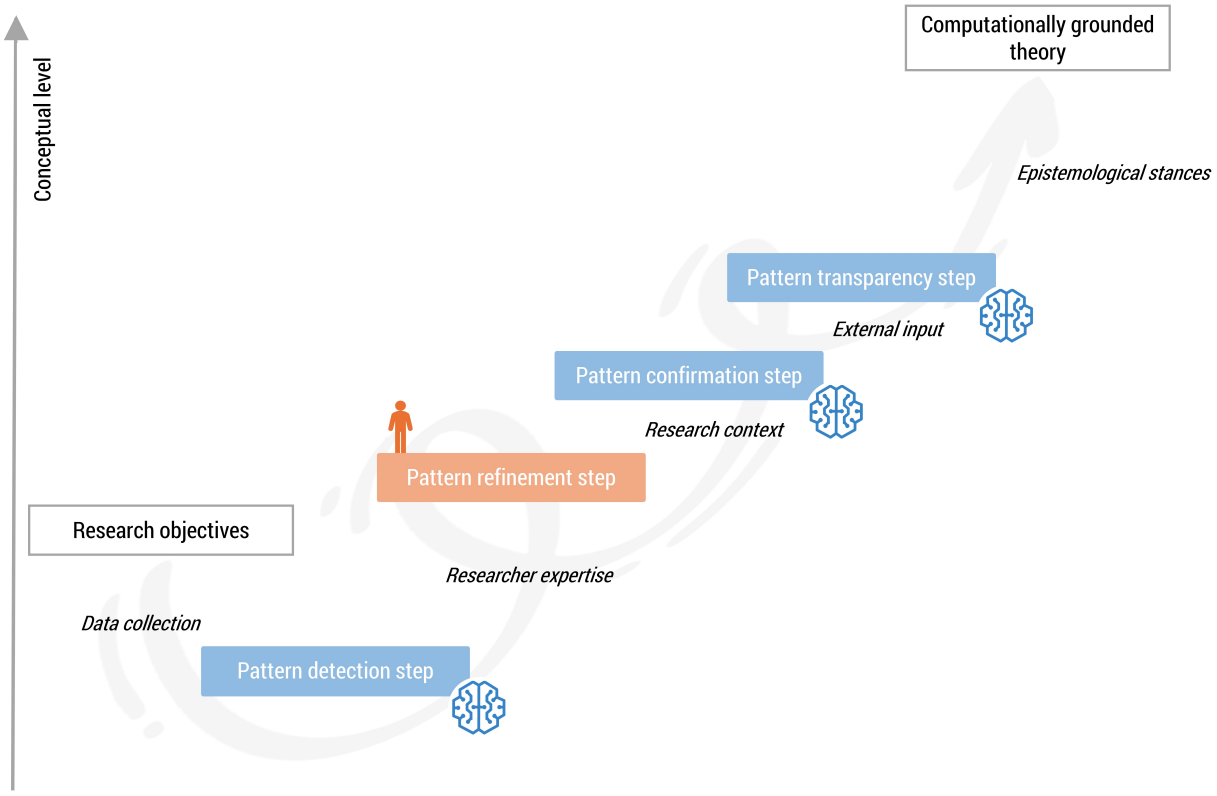
Zhai X., Shi L. and Nehm R. H., (2021), A Meta-Analysis of Machine Learning-Based Science Assessments: Factors Impacting Machine-Human Score Agreements, *J. Sci. Educ. Technol.*, **30**, 361–379.

Zhu M., Lee H.-S., Wang T., Liu O. L., Belur V. and Pallant A., (2017), Investigating the impact of automated feedback on students' scientific argumentation, *Int. J. Sci. Educ.*, **39**, 1648–1668.

This journal is © The Royal Society of Chemistry 2023

3 | Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry

Chapter 3 presents an original research paper. It is reproduced from Martin, Kranz, Wulff, and Graulich (2024) under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0) (https://creativecommons.org/licenses/by-nc-nd/4.0/). No changes have been made to the original article.



Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2024). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*, *61*(8), 1757-1792. https://doi.org/10.1002/tea.21903

Received: 2 May 2023 | Revised: 24 August 2023 | Accepted: 26 August 2023

DOI: 10.1002/tea.21903

RESEARCH ARTICLE

JRST | WILEY

Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry

Paul P. Martin[1] | **David Kranz**[1] | **Peter Wulff**[2] | **Nicole Graulich**[1]

[1]Institute of Chemistry Education, Justus-Liebig-University, Giessen, Germany

[2]Physics Education Research, University of Education, Heidelberg, Germany

Correspondence
Nicole Graulich, Institute of Chemistry Education, Justus-Liebig-University, Giessen, Heinrich-Buff-Ring 17, 35392 Giessen, Germany.
Email: nicole.graulich@didaktik.chemie.uni-giessen.de

Abstract

Constructing arguments is essential in science subjects like chemistry. For example, students in organic chemistry should learn to argue about the plausibility of competing chemical reactions by including various sources of evidence and justifying the derived information with reasoning. While doing so, students face significant challenges in coherently structuring their arguments and integrating chemical concepts. For this reason, a reliable assessment of students' argumentation is critical. However, as arguments are usually presented in open-ended tasks, scoring assessments manually is resource-consuming and conceptually difficult. To augment human diagnostic capabilities, artificial intelligence techniques such as machine learning or natural language processing offer novel possibilities for an in-depth analysis of students' argumentation. In this study, we extensively evaluated students' written arguments about the plausibility of competing chemical reactions based on a methodological approach called *computational grounded theory*. By using an unsupervised clustering technique, we sought to evaluate students' argumentation patterns in detail, providing new insights into the *modes of reasoning* and

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2023 The Authors. *Journal of Research in Science Teaching* published by Wiley Periodicals LLC on behalf of National Association for Research in Science Teaching.

levels of granularity applied in students' written accounts. Based on this analysis, we developed a holistic 20-category rubric by combining the data-driven clusters with a theory-driven framework to automate the analysis of the identified argumentation patterns. Pre-trained large language models in conjunction with deep neural networks provided *almost perfect* machine-human score agreement and well-interpretable results, which underpins the potential of the applied state-of-the-art deep learning techniques in analyzing students' argument complexity. The findings demonstrate an approach to combining human and computer-based analysis in uncovering written argumentation.

KEYWORDS

argumentation competence, computational grounded theory, machine learning, natural language processing, organic chemistry learning

1 | INTRODUCTION

In science education, argumentation skills are essential components of critical thinking and problem-solving (e.g., Driver et al., 2000; Duschl & Osborne, 2002; Erduran et al., 2004; Faize et al., 2018; Jiménez-Aleixandre & Erduran, 2007). In particular, the ability to construct, justify, and defend an argument based on evidence and reasoning is key to evaluating scientific hypotheses (Toulmin, 2003). Additionally, a coherent, evidence-based argumentation allows one to reflect on the quality of arguments, clearly clarify or critique ideas, propose solutions and alternatives for complex problems, and convince others to accept valid conclusions (Kuhn, 2010; Kuhn & Udell, 2003). Hence, science education should support students in developing strong argumentation skills as these skills help students become more critical and analytical thinkers, facilitating engagement in meaningful discussions about scientific ideas (Erduran, 2019). By allowing students to practice how to construct arguments, students can increase the complexity of their written argumentation and acquire conceptual knowledge (Cetin, 2014; Lieber et al., 2022a, b), which also encourages them to discuss opposing positions (Chin & Osborne, 2010; Kuhn & Udell, 2003).

Consistent argumentation involves several steps, which include identifying the claim being made, gathering, evaluating, and weighing evidence to support the claim, justifying the linkage between claim and evidence with reasoning, and critiquing constructed arguments (e.g., Driver et al., 2000; Manz, 2016; McNeill et al., 2006; McNeill & Krajcik, 2011; Osborne et al., 2004; Osborne & Patterson, 2011). However, students often experience challenges in constructing well-justified arguments (Driver et al., 2000; McNeill & Krajcik, 2011). For instance, students integrate non-normative ideas in their argumentation (Sampson et al., 2011; Walker et al., 2019), misinterpret relevant data needed to justify statements (Chinn & Malhotra, 2002), mix up empirical evidence and personal interpretation (Berland & Reiser, 2009; Jeong et al., 2007), or neglect sources of uncertainty (Kanari & Millar, 2004; Lee et al., 2014). Students

experience these challenges in science subjects because they may not have had enough chances to continuously explore, communicate, interpret, and criticize data in a guided fashion (Driver et al., 2000; Jiménez-Aleixandre et al., 2000; Newton et al., 1999).

To assess and foster students' argumentation, it seems helpful to integrate open-ended questions in argumentation-focused, computer-based formative assessments (cf., Kuo & Wu, 2013). Notably, a formative assessment that aims to capture the strength of an argument needs to be well-developed (e.g., Liu, 2020). For example, students should be prompted to construct various evidence-based arguments as well as counterarguments. However, evaluating students' responses to open-ended questions is resource-consuming when performed by hand (Nehm & Haertig, 2012; Ullmann, 2019). Emerging computer-based techniques like machine learning (ML) can be used to efficiently process complex language data (Goldberg, 2017; Jurafsky & Martin, 2023). Over the last couple of years, this computer-based approach to automatically analyzing students' responses has been increasingly applied to advance the formative assessment of open-ended items (cf., Martin & Graulich, 2023; Zhai, Haudek, et al., 2020; Zhai, Yin, et al., 2020). ML can improve teaching and learning by affording data-driven insights into students' reasoning, automating the evaluation of formative assessments, tailoring guidance, feedback, or exercises to students' needs, and easing longitudinal, large-scale data collection.

In this article, we analyze students' scientific argumentation about the plausibility of competing chemical reactions by leveraging deep learning, an advanced ML type, following a methodological approach called *computational grounded theory*. Precisely, we used pre-trained large language models built on deep neural networks because they offer novel capabilities in systematically analyzing natural language. With these techniques, we aimed to provide an in-depth analysis of a diverse set of characteristics present in students' argumentation, like the *modes of reasoning* (Sevian & Talanquer, 2014) or the *levels of granularity* (Bodé et al., 2019; Deng & Flynn, 2021). The objectives of this study are, accordingly, guided by pedagogical and methodological considerations. The pedagogical considerations correspond to the pedagogical feature introduced in Zhai, Yin, et al.'s (2020) analytical triangle for applying ML in science assessment, while the methodological considerations cover the technical and validity features proposed in this triangle. To be noted, the term *deep learning* refers in this article to an advanced ML type, not to a science learning or instructional approach as conceptualized in educational psychology (Chin & Brown, 2000; Miller & Krajcik, 2019).

2 | METHODOLOGICAL CONSIDERATIONS

2.1 | ML and NLP basics

In general, ML is a subfield of artificial intelligence focusing on the development of algorithms that learn from data and make decisions with minimal human intervention (Bishop, 2006; Mitchell, 1997; Mohri et al., 2012). Following this, ML comprises algorithms that learn automatically from data without being explicitly programmed (Samuel, 1959). ML techniques can be roughly classified as supervised or unsupervised algorithms: Supervised algorithms are trained on human-labeled data so that the output to be predicted, the so-called ground truth, is already provided in the training set. Based on the input–output mapping, the algorithm detects underlying patterns to predict labels in new data. In turn, unsupervised algorithms do not require labeled training data; instead, they extract patterns independently. During the last 15 years, the application of ML has significantly increased in science education research (cf., Deeva et al., 2021; Gerard et al., 2015; Martin & Graulich, 2023; Zhai, Haudek, et al., 2020; Zhai, Yin, et al., 2020).

For analyzing human language, text data needs to be preprocessed first with suitable natural language processing (NLP) techniques. NLP transforms language data systematically so that computers can read, analyze, and produce human language. In recent years, numerous techniques have been developed that enable computers to engage with human language. As a cutting-edge approach, large language models emerged in the field of NLP. Large language models refer to a specific type of advanced ML algorithm that is pre-trained on enormous amounts of unlabeled text data to capture the semantics of the human language (Manning, 2022). For example, the large language model Bidirectional Encoder Representations from Transformers (BERT) has been developed to solve NLP problems like text classification (Devlin et al., 2018). Researchers can fine-tune this model for their domain-specific purposes, for instance, to automatically classify student-written responses, which is called a downstream task (Ruder, 2019). Practitioners in the field of artificial intelligence often use BERT for several reasons. First, BERT outperformed other NLP techniques in analyzing human language (Devlin et al., 2018) because it can process words in relation to each other, even if they are far apart in a sentence (Taher Pilehvar & Camacho-Collados, 2020). For this reason, BERT captures the context-dependent meaning of individual words as well as dynamic relationships between words (Mikolov et al., 2013; Taher Pilehvar & Camacho-Collados, 2020). Moreover, less data might be required to train an ML model when using BERT as an NLP technique because the pre-training of BERT on massive text corpora enhances model generalizability and accuracy. Accordingly, BERT is relatively fast to train, for example, to evaluate students' responses in science assessment (Dood et al., 2022; Gombert et al., 2022; Winograd, Dood, Finkenstaedt-Quinn, et al., 2021; Winograd, Dood, Moon, et al., 2021; Wulff, Mientus, et al., 2022). Furthermore, compared to other large language models such as GPT-4, BERT has the advantage of being exclusively executable on private devices without outsourcing the research data to third parties such as private companies.

Since 2018, BERT has formed the template for various large language models with specific purposes and advantages such as *BERT base uncased*, *BERT large uncased* (Devlin et al., 2018), *RoBERTa base*, *RoBERTa large* (Liu et al., 2019), or *SciBERT scivocab uncased* (Beltagy et al., 2019). These models work all with the same architecture (cf., Vaswani et al., 2017) but are trained on different text corpora. BERT is trained on a large corpus of general text provided, among others, through English Wikipedia (Devlin et al., 2018). RoBERTa (Robustly Optimized BERT Pretraining Approach) addresses some of the limitations of BERT by training the model longer on more text data and using dynamic masking, which helps the algorithm focus on the most relevant words in a sentence (Liu et al., 2019). In contrast, SciBERT (Scientific BERT) is fine-tuned on a dataset of scientific publications, making it particularly useful for scientific text classification (Beltagy et al., 2019). Moreover, *base* models include significantly fewer features than *large* models. For instance, *BERT base uncased* comprises 110 million parameters and 768 hidden layers, while *BERT large uncased* comprises 340 million parameters and 1024 hidden layers (Devlin et al., 2018). Generally, *base* models seem to be a good trade-off between performance and resource requirements, while *large* models tend to perform better on a wide range of tasks, but at the cost of increased computational resources and longer training times.

2.2 | Literature overview on capturing students' argumentation with ML

ML and NLP offer new possibilities for assessing scientific argumentation because corresponding techniques analyze language data automatically, enabling the use of open-ended tasks to capture the whole range of students' proficiency (Harris et al., 2019). Some researchers

have already investigated students' scientific argumentation skills according to a pre-selected argumentation model while focusing on different aspects. These aspects can be categorized as establishing design principles (Cheuk, 2021; Cheuk et al., 2019), automating human scoring (Haudek et al., 2019; Wang et al., 2021), evaluating the validity of ML-based scores (Haudek & Zhai, 2021; Kaldaras & Haudek, 2022; Wilson et al., 2024), creating fine-grained, individualized, diagnostic skill profiles (Zhai et al., 2023), and delivering automated personalized feedback (Huang et al., 2011; Lee et al., 2019, 2021; Mao et al., 2018; Zhu et al., 2017).

Collectively, these studies demonstrate that students' argumentation skills have been analyzed with different research foci. However, most of the studies have in common that they relied on supervised techniques to automate the scoring of students' written arguments based on predefined categories, while unsupervised ML has not been applied to exploratively identify patterns within the data. Leveraging unsupervised ML to reveal patterns across arguments that might not have been apparent to a human analyst may, nevertheless, provide a better understanding of the types of arguments students generate and identify areas where they need additional support. Applying unsupervised ML should not prevent humans from analyzing arguments according to predefined theory-rich categories but offers the potential to integrate data-driven classifications into research-informed theoretical frameworks. This, in turn, may extend human diagnostic capabilities and facilitate the fine-grained automated evaluation of students' arguments through supervised ML.

3 | STUDY OBJECTIVES

In this study, we applied an approach to analyzing students' written arguments that integrated theory-driven human interpretation with data-driven deep learning techniques (Carlsen & Ralund, 2022; Nelson, 2020). With this approach, we envisioned a more fine-grained evaluation of complex argumentation as we expected to identify intricate argumentation patterns. Following this, the study objectives are pedagogical and methodological in nature. From a pedagogical point of view, we combined unsupervised techniques with human interpretation to explore novel characteristics that might appear in students' argumentation about chemical reactions, such as the *modes of reasoning* (Sevian & Talanquer, 2014) or the *levels of granularity* (Bodé et al., 2019; Deng & Flynn, 2021). From a methodological point of view, we aligned the data-driven results with theory-driven considerations to establish a transparent and interpretable deep learning architecture capable of accurately assessing students' argumentation across different chemical reactions.

Taken together, we aimed to provide a more fine-grained analysis of additional patterns in students' written argumentation by applying deep learning methodology. Such an in-depth analysis may be more appropriate for the long-term goal of supporting students longitudinally in ML-based instructional settings as feedback and exercises can be better adapted to students' learning needs (e.g., Donnelly et al., 2015; Dood et al., 2020a, b; Lim et al., 2023; Sailer et al., 2023; Tansomboon et al., 2017; Vitale et al., 2016; Watts et al., 2023a). We hypothesized that deep learning has the potential to uncover additional facets in students' argumentation as it can evaluate complex text data while representing intricate relations (cf., Hernández-Blanco et al., 2019).

3.1 | Research questions

The study is guided by four research questions (RQs), which reflect the four steps of *computational grounded theory* (Figure 2) and can be divided according to the pedagogical or methodological focus.

Pedagogical focus:

1. What argumentation patterns can be uncovered by integrating unsupervised deep learning with human interpretation?
2. What *modes of reasoning* and *levels of granularity* do students apply when judging the plausibility of competing chemical reactions?

Methodological focus:

3. To what extent can a deep learning model evaluate argument complexity in students' argumentation about the plausibility of competing chemical reactions?
4. To what extent do the features that the deep learning model used to classify the data correspond to the human scoring guidelines?

4 | METHODS

4.1 | Setting of data collection

The data was collected at a private, research-intensive, liberal arts university in the Northeastern United States in April and May 2021. Sixty-four students from an Organic Chemistry II course participated voluntarily while receiving extra credit for completing the exercises. The age of the participants ranged from 18 to 22 years; 34 students identified themselves as females, 29 as males, and one as non-binary. Students majored in various fields, including biochemistry, chemistry, biology, and chemical engineering, among others.

Students wrote on average 27 words (SD = 12.7, max = 138, min = 5) per argument. In total, 1108 arguments were collected. Further details on the setting and participants can be found in Lieber et al. (2022a).

4.2 | Research instrument

Successfully building arguments is influenced by students' conceptual understanding and their epistemic knowledge of scientific argumentation (Lieber et al., 2022a, b; Lieber & Graulich, 2020, 2022; Sandoval & Millwood, 2005). Scaffolding the process of building arguments can, consequently, be a valuable instructional approach to support students' argumentation skills as it significantly improves argument complexity (Luo et al., 2020). Using a scaffold can be helpful because it slows down the decision-making process and guides students in paying attention to implicit features that they might otherwise overlook (Caspari & Graulich, 2019; Graulich & Caspari, 2020; Kang et al., 2014; Kranz et al., 2023; Watts et al., 2021). In particular, adaptively scaffolding students' argumentation skills, that is, tailoring support to a student's individual needs, can close existing performance gaps in backing claims with data and integrating essential concepts in an argument (Lieber et al., 2022a, b).

With the goal of supporting students' scientific argumentation skills and their conceptual organic chemistry knowledge, Lieber et al. (2022a, b) designed an adaptive instructional setting where students were prompted to judge the plausibility of competing chemical reactions. In organic chemistry, reflecting on competing reactions is important as reactants can potentially

undergo multiple reaction pathways to form more or less plausible products (cf., Illari & Williamson, 2012). The plausibility of these different reaction pathways depends on several factors, including the energetic level of the products, the rate of the reaction, as well as the properties of the respective molecules depending on the reaction conditions and the nature of the reagents. Arguing about alternative reaction pathways requires the integration of multiple chemical concepts (Lieber et al., 2022a; Lieber & Graulich, 2022). Chemical ideas and principles that are often considered separately must, thus, be connected to weigh them against each other (Lieber & Graulich, 2022; Watts et al., 2022).

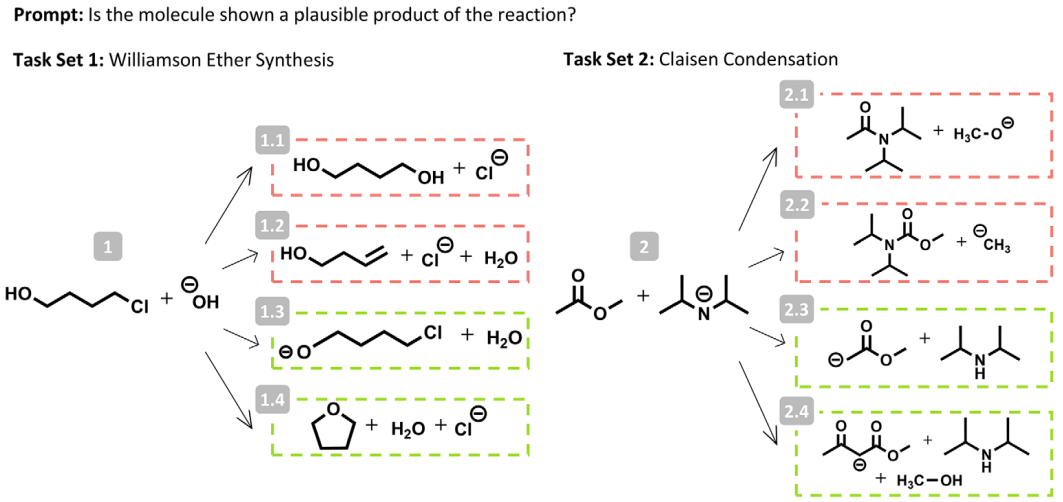
To scaffold the argumentation about competing chemical reactions, Lieber et al. (2022a, b) utilized a simplified version of Toulmin's argumentation pattern: the frequently applied three-component CER model based on the terms *claim*, *evidence (data)*, and *reasoning (warrant)* (McNeill et al., 2006). A claim acts as a statement, which is why it needs causal support (McNeill et al., 2006; McNeill & Krajcik, 2011). Causal support can be provided based on an explanation or scientific data, called evidence (McNeill et al., 2006; McNeill & Krajcik, 2011). Generally, the evidence uses information from various resources to support the claim (McNeill & Krajcik, 2011). Eventually, reasoning must be applied to provide a cause-and-effect relationship between claim and evidence (McNeill & Krajcik, 2011).

Lieber et al. (2022a, b) developed a two-part training centered on constructing CER-aligned arguments on competing chemical reactions in organic chemistry. Specifically, students judged the plausibility of four alternative reaction products for intramolecular Williamson ether synthesis and Claisen condensation (Figure 1). When judging the plausibility, students were prompted to make a claim about whether the displayed product is (im-)plausible, construct evidence to provide chemical concepts of why the respective product may be (im-)plausible, and link the claim and evidence with reasoning by including electronic, steric, and energetic effects (Figure 1). For example, students argued that 4-chlorobutan-1-ol and hydroxide ions (Figure 1, reactants 1) react in a nucleophilic substitution reaction, in which the hydroxide ion acts as a nucleophile attacking the electrophilic carbon of the alkyl chloride, to form a diol (Figure 1, product 1.1) (Lieber & Graulich, 2022). In fact, the reaction results after a proton transfer (Figure 1, product 1.3) and a subsequent intramolecular S_N2 reaction in the formation of tetrahydrofuran (Figure 1, product 1.4). With this task design, Lieber et al. (2022a, b) showed that adaptive scaffolding significantly improved students' performance in the respective area of support. Given this positive impact on students' organic chemistry learning, we built upon this research instrument, which means that the nature of students' argumentation about the plausibility of competing reactions guided the pedagogical objectives of this study.

The described scaffolds were implemented online via *Qualtrics*. The two task sets (Figure 1) were presented on 2 days with a 3-week interval. The four alternative reaction products were shown one after the other.

4.3 | Computational grounded theory

The analysis of students' argumentation skills in this study is anchored in a three-step methodological framework called *computational grounded theory* (Nelson, 2020). *Computational grounded theory* is based on the traditional grounded theory approach, which is a method for developing theories about social phenomena through the close analysis of data (Glaser & Strauss, 1999). Particularly, grounded theory enables themes to emerge inductively from data, rather than being imposed on it, intending to generate data-driven understandings of not



Sample solution for product 1.1

Claim: The product is not plausible.

Evidence	Reasoning
The hydroxide ion is a strong base.	Due to the small atom size and, thus, low polarizability of the oxygen atom, the negative charge is localized allowing for deprotonation of the hydroxyl group.
In this reaction, the hydroxide ion acts as a base, not as a nucleophile.	Water and the hydroxyl group have very similar pKa values, which results in an equilibrium of alkoxide and hydroxyl.
Acid-base reactions occur faster than substitution reactions.	Acid-base reactions have no need for a change of geometry and protons transport little mass, which lowers the activation energy.

FIGURE 1 Students evaluated the plausibility of four alternative reaction products for the reaction of 4-chlorobutan-1-ol with hydroxide ions (task set 1) and for the reaction of methyl acetate with diisopropylamide (task set 2). The plausible reaction products are highlighted in green. Three sample arguments for the implausibility of product 1.1 are shown; further sample solutions can be found in Lieber et al. (2022a).

observable processes (Charmaz, 2014). However, the nature of grounded theory requires subjective assumptions about how to collect and interpret data, which may result in biased decisions (Saldana, 2015), low reproducibility of the generated hypotheses (Biernacki, 2012), and limited applicability for large, unstructured datasets (Bail, 2014).

To mitigate these constraints, *computational grounded theory* combines qualitative research and computational techniques to systematically analyze *big data* for inductive theory-generating research (Kubsch et al., 2021, 2023; Nelson, 2020; Rosenberg & Krist, 2021). This allows for identifying patterns in unstructured data, inductively generating theories about the underlying phenomena being studied, and deductively verifying the generated theories with quantitative tests (Nelson, 2020). For instance, Rosenberg and Krist (2021) applied *computational grounded theory* with ML methods to design a fine-grained construct map for evaluating students' considerations of generality in science. They combined a two-step clustering approach with interpretative coding to inductively reveal the complexity of students' cognitive conceptions, which helped them uncover the whole range of the epistemic characteristics of students' model-based explanations.

From a wider perspective, *computational grounded theory* proposes a framework that clarifies which tasks should be performed by domain experts and which by algorithms (Kubsch

et al., 2023; Nelson, 2020). Humans set the pedagogical purpose of an assessment first, while algorithms assign codes to data in a supervised or unsupervised manner (Kubsch et al., 2023; Nelson et al., 2021). In the case of unsupervised learning, humans need to interpret the assigned codes to make sense of the categorization (Nelson, 2020). Finally, combining human and algorithmic work helps draw evidence-based inferences about phenomena as data- and theory-driven considerations are complementarily integrated (Kubsch et al., 2023).

4.3.1 | Pattern detection step

For the practical application, *computational grounded theory* proposes a three-step procedure to gain breadth and depth in data analysis (Figure 2) (Kubsch et al., 2021; Nelson, 2020). In the first step, the pattern detection step (Figure 2), computational techniques are applied to transform complex, content-rich text into interpretable categories or networks (Nelson, 2020). Especially, unsupervised ML algorithms can be used in this step to reveal reproducible patterns that researchers may not have previously considered.

To detect patterns in complex language data, we first employed large language models to transform students' written arguments into so-called contextualized embeddings, which are high-dimensional numerical representations of words. Transforming human language into contextualized embeddings enhances the performance of ML methods (Zehner et al., 2016). The large language models *BERT base uncased*, *BERT large uncased* (Devlin et al., 2018), *RoBERTa base*, *RoBERTa large* (Liu et al., 2019), and *SciBERT scivocab uncased* (Beltagy et al., 2019) were used to calculate these contextualized embeddings. We compared different large language models because each model has unique strengths depending on the task to be performed and the given dataset. Eventually, we continued our analysis with *BERT base uncased* because this large language model extracted the fewest number of noise points in the clustering approach, which was desirable to classify most data.

Afterward, we used Uniform Manifold Approximation and Projection (UMAP) to reduce the dimensionality of the contextualized embeddings since it effectively reduces high-dimensional data (McInnes et al., 2018). We reduced the dimensionality of the 768-dimensional vectors to computationally ease clustering and to visualize the clusters in a two-dimensional space, given that most information in complex data is typically stored in only few dimensions (Brunton & Kutz, 2019; Zehner et al., 2016). UMAP includes several hyperparameters; in our analysis, dimensionality was reduced to 10 and the number of neighbors was set to 15 because this configuration yielded well-interpretable results in prior studies in science education (Wulff, Buschhüter, et al., 2022) and beyond (Grootendorst, 2020).

Ultimately, the unsupervised ML technique Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) was applied to extract compact clusters (Campello et al., 2013, 2015; McInnes et al., 2017). HDBSCAN can identify clusters of arbitrary shape, automatically determine the number of clusters, and handle noisy language data, which is typical for students' reasoning in science subjects (e.g., Wulff et al., 2023; Wulff, Buschhüter, et al., 2022a). While clustering, HDBSCAN identifies so-called noise points, which correspond to non-normative or mixed ideas as well as to concepts that do not exceed the threshold for the minimum cluster size. Excluding noise from analysis allows for a more discriminating analysis of students' ambiguous argumentation and provides a more thorough understanding of the data characteristics. Because of these benefits, we applied HDBSCAN to characterize the degree of elaborateness, interconnectedness, and specificity of students' arguments. The minimal cluster

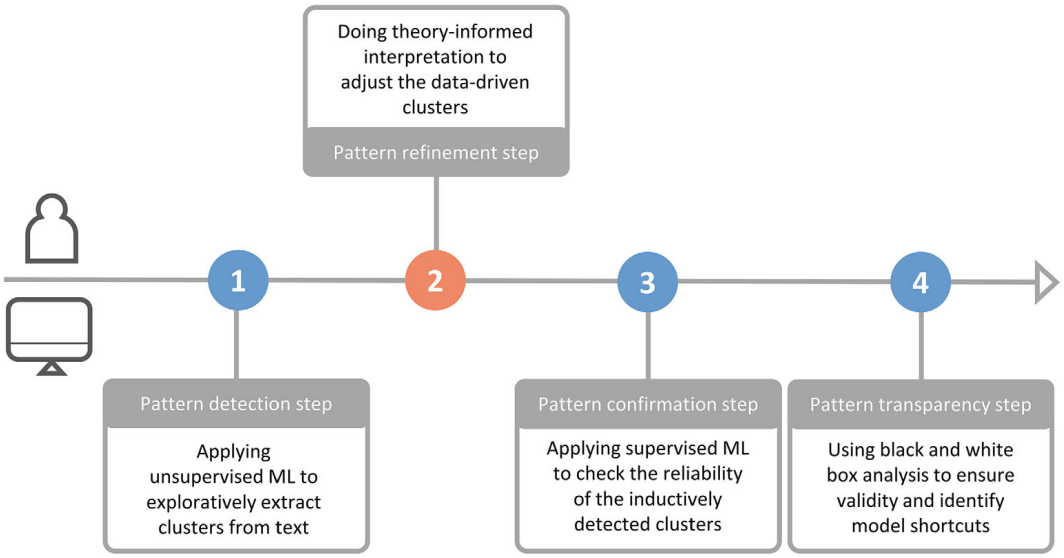


FIGURE 2 Revisited four-step computational grounded theory framework.

size was set to 12, so clusters were only extracted if they comprised at least 12 arguments. The number 12 was chosen since a minimal cluster size of 12 corresponds to 1% of our dataset.

4.3.2 | Pattern refinement step

In the subsequent pattern refinement step (Figure 2), human experts interpret the obtained data-driven patterns holistically through research-informed thematic content analysis to confirm the plausibility of the computationally detected categories (Nelson, 2020). In particular, domain experts interpret the data-driven analysis, check the usefulness of the derived categories, and modify these categories if necessary.

4.3.3 | Pattern confirmation step

Eventually, computational techniques such as supervised ML can be deployed in the pattern confirmation step (Figure 2) to verify the reliability and generalizability of the inductively derived patterns (Nelson, 2020). Leveraging deep learning techniques seems helpful in this step since they can handle complex data and achieve high accuracy (Mathew et al., 2021).

Deep learning is a subfield of ML that involves the training of deep neural networks with multiple layers of interconnected neurons, inspired by the structure of the human brain. The term *deep* refers to the use of multiple layers in the network, which allow for learning complex relationships. The training of deep learning models involves feeding large amounts of data into the network and adjusting the model parameters to improve output. As an advantage, such architectures can automatically represent noisy text data and do not require preprocessing tasks like removing non-informative words or transforming words into their base form (Angelov, 2020; Goodfellow et al., 2016).

In our project, we used the deep learning framework PyTorch (Paszke et al., 2019) in Python. When feeding data in the deep neural network, we split our dataset in a 65:15:20 ratio as a training, validation, and test set. We used the training set for the training of the neural network, the validation set for determining the best configuration of hyperparameters, and the test set for checking the accuracy of the model. As hyperparameters, we calibrated the number of epochs, the batch size, and the learning rate. Epochs refer to the number of cycles the model is trained on the training set. During each epoch, the model's parameters are updated. Typically, the more epochs the model is trained on, the more it improves its performance on the training data up to a certain point of saturation. If the model is trained for too many epochs, it may start to overfit the training data, which means that it will perform poorly on new data. To assess overfitting, we checked the performance of our model on a new dataset: the test set. Additionally, we monitored the training loss, which is a measure of change for the discrepancy between the true score and predicted output. As another hyperparameter, the batch size specifies the number of concurrent training samples in one pass through the network. The learning rate, in turn, controls the step size at which an optimizer updates the network parameters to adapt to a certain context.

In contrast to other studies using deep learning in chemistry education (e.g., Watts et al., 2023b; Winograd, Dood, Moon, et al., 2021), we applied our neural network for holistic scoring that assigns every argument to one of 20 mutually exclusive categories. First, we applied *BERT base uncased* to identify the most performant hyperparameter configuration based on the validation set. We tested 100 different epochs between 1 and 100, seven different learning rates between 1e-7 and 1e-4, and five different batch sizes between 2 and 32, resulting in a total of 3500 configurations. We then identified the best combinations of the learning rate and batch size with varying epochs and applied these to *BERT large uncased*, *RoBERTa base*, *RoBERTa large*, and *SciBERT scivocab uncased*. We continued to vary the epochs during model training as we determined that they had the greatest impact on the model performance.

4.3.4 | Pattern transparency step

We expanded *computational grounded theory* in our analysis by a fourth step, the pattern transparency step (Figure 2). During this step, methods for explaining the internal and external workings of an ML algorithm such as black and white box analysis or external validation can be applied. Black box analysis helps understand how an algorithm operates solely based on its observable input–output behavior. As the internal workings of ML models can be difficult to interpret, black box analysis is a time-efficient method for determining which features are driving a model's output. For implementing black box analysis, we modified four input sentences with a low level of argument complexity by systematically adding or removing words or phrases according to predefined criteria of our scoring rubric so that these sentences just reached the next level of the rubric (see sections 5.2 and 5.4). Subsequently, we checked whether the changed input led to the desired algorithmic classification.

However, black box analysis does not investigate the internal workings of an algorithm, which has led to the development of complex techniques for white boxing an algorithm. White box analysis aims to interpret the decision-making process of ML models by identifying the feature importance of individual words via post-hoc explanations (cf., Gombert et al., 2022). For white box analysis, we applied a technique called SHapley Additive exPlanations (SHAP) to calculate the importance scores of individual words for each rubric category (Lundberg &

Lee, 2017). This technique extracts keywords associated with positive or negative classifications in a rubric category and identifies critical model shortcuts (cf., Geirhos et al., 2020).

In sum, the pattern transparency step is essential for ensuring that an algorithm makes unbiased decisions. Transparently explaining the results of an ML algorithm is especially helpful when its decisions have a significant impact on human lives since white boxing an algorithm can build trust in ML methodology (e.g., Caruana & Nori, 2022).

5 | RESULTS AND DISCUSSION

5.1 | RQ1: Argumentation patterns uncovered by integrating unsupervised deep learning with human interpretation

To answer RQ1, we used HDBSCAN to extract 22 distinct clusters, an additional noise cluster, their 10 most representative words, and sample arguments (Table 1). By qualitatively evaluating the most representative words as well as sample responses, we derived descriptions of each topic. Figure 3 presents the two-dimensional embedding space of the extracted clusters. Clusters that are similar in content also appear close to each other in the two-dimensional embedding space: For example, arguments in clusters 2 and 3 apply acid–base theories on an electronic level, with minor differences in the interconnectedness of applied concepts. Arguments in clusters 4 and 5 correlate the thermodynamic state of the reagents to implicit structural properties, such as bond energies or resonance. In clusters 8, 9, 10, 11, and 12, arguments focus on the energetic aspects of a reaction in various degrees of elaborateness, in contrast to structural features addressed in many other clusters. Arguments in clusters 14, 15, and 16 apply acid–base theories on a rather phenomenological level, e.g., by citing pKa-values, or connect acid–base chemistry to thermodynamics or kinetics, e.g., by quantifying chemical equilibrium or estimating reaction rate. Arguments in clusters 17, 20, and 21 use steric hindrance in conjunction with hybridization or leaving group quality, while arguments in clusters 18 and 19 connect nucleophilicity and electrophilicity to basicity or electronic effects.

Clustering students' written arguments uncovered various argumentation patterns (Table 1) and revealed the degree of interconnectedness, elaborateness, and specificity of the applied concepts. The clusters comprised varying chemical topics and different levels of sophistication, indicating that the clustering approach captured the content *and* complexity of students' reasoning. Since students were explicitly prompted to include different chemical concepts in their argumentation, it is not surprising that these two aspects are also reflected in the clusters. However, clustering was exclusively data-driven, meaning that it groups data points based on the distance in the embedding space. While clustering helped us identify patterns across many responses, this approach did not provide a classification of the identified clusters influenced by theoretical frameworks, instructional goals, or learning objectives. Including research-informed classifications of students' argumentation grounded on subject matter expertise may, nonetheless, extend the validity of the analysis and ease the subsequent process of designing effective personalized instruction (Carlsen & Ralund, 2022; Kubsch et al., 2023; Nelson, 2020; Sherin, 2013). Hence, we combined the clustering approach with research-informed considerations to ensure that instructional decisions are data- *and* theory-driven.

TABLE 1 Overview of the extracted clusters.

ID	Cluster size	Description	Top words	Student example
−1	255	Redescribing the displayed reaction without referring to chemical concepts	Reaction, carbon, group, forms, S_N2, attacks, molecule, stabilizes, occurs, product	"Because the hydroxide ion can attack the carbon in an S_N2 fashion."
0	51	Estimating entropic changes based on the number of reactants and products	Entropy, reactant, product, favorable, increases, molecule, reaction, changes, unfavorable, forms	"This will increase the entropy of the system. There are more molecules in the products than in the reactants, which is a favorable transformation for entropy."
1	122	Estimating leaving group quality based on base strength, electronegativity, or polarizability	Leaves, good, group, poor, solvent, octet, electronegative, base, weak, stabilizes	"Chloride is a good leaving group. Chloride is polarizable and electronegative, making it well able to stabilize a negative charge effectively."
2	13	Invoking the Lewis acid–base theory	Acid, donates, base, proton, electron, acts, Lewis, accepts, carbon, pair	"The hydroxide anion will function as a Lewis base because it will donate an electron pair to the alpha carbon which forms the bond between the alpha carbon and hydroxide."
3	19	Applying electronic effects to infer acid strength	Carbon, electron, reacts, molecule, near, attacks, withdrawing, pair, forms, unstable	"The hydrogen atom, that is removed from the first molecule, is the most acidic one since the inductive effects from the oxygens in the molecule withdraw electron density, lowering its pKa."
4	15	Identifying resonance stabilization in carbanions	Carbanion, stabilizes, resonance, carbocation, results, product, forms, ketone, enolate, unstable	"The carbanion produced is stable because resonance forms are present that stabilize it."
5	21	Referring to bond enthalpies	Unfavorable, bond, wants, carbon, group, takes, molecule, satisfies, forms, stabilizes	"The reaction is energetically unfavorable. The breaking of the C-O bond is more expensive than the energy released from forming a C–N bond."
6	16	Identifying electrophiles based on electron density	Electrophile, density, carbon, electron, atom, withdrawing, attacks, takes, electronegative, makes	"Chlorine is much more electronegative than carbon; thus, it withdraws electron density and makes carbon an excellent electrophile."
7	12	Identifying acids and bases	Weak, acid, base, water, strong, proton, donates, molecule, atom, small	"The amine ion is a strong base that can deprotonate the ester. The amine ion acts as a base and the ester acts as an acid."

(Continues)

TABLE 1 (Continued)

ID	Cluster size	Description	Top words	Student example
8	78	Applying electronegativity to estimate the stability of the products	Electronegative, negative, charge, stabilizes, atom, carbon, better, effects, unstable, higher	"The negative charge is on a more electronegative atom. Oxygen is more electronegative than nitrogen since it is closer to completing its octet and so pulls electrons toward its nucleus. Products that have the negative charge on the more electronegative atom are more stable and so are more likely to be products."
9	48	Locating resonance stabilization	Resonance, stabilizes, bond, negative, charge, delocalized, electron, ring, pushes, pair	"The negative charge in the product is stabilized. There is resonance, so the electrons can form a double bond and then shift to the oxygen atom."
10	15	Focusing on charges as a surface-level indicator of stability	Negative, charge, positive, partial, makes, bond, reforms, prefers, carbon, likely	"The product is negatively charged. Molecules are most stable when limited charges are present."
11	98	Attributing energy levels to molecules	Bond, product, enthalpy, energy, stabilizes, reactant, favorable, reaction, forms, breaks	"The products are not very stable since they are not the lowest energy of this reaction."
12	15	Identifying resonance stabilization	Aldol, enolate, resonance, stabilizes, intermediate, energy, low, product, reactant, able	"The product participates in resonance, allowing it to be stabilized."
13	13	Weighing nucleophilic and basic properties by considering steric effects	Amine, bulky, alkoxy, acts, amide, nucleophile, approaches, easy, molecule, tertiary	"The amine is more likely to act as a base than as a nucleophile. The amine's bulkiness makes it difficult to approach the molecule, but it is much easier to remove a proton."
14	14	Citing pKa-values for identifying the strongest acid	pKa, low, favorable, higher, product, small, reversible, reaction, large, water	"The product has a higher pKa. With a higher pKa, the reaction is more likely to happen."
15	54	Describing proton transfer or comparing its reaction rate	Deprotonates, base, alcohol, strong, acid, proton, group, pKa, water, anion	"Acid–base reactions occur faster. Thus, this would not be a nucleophilic attack."

TABLE 1 (Continued)

ID	Cluster size	Description	Top words	Student example
16	22	Quantifying chemical equilibrium based on pKa-values	Ester, pKa, acid, base, amine, conjugate, low, deprotonates, weak, equilibrium	"Water, the conjugate acid of hydroxide, and alcohols have a very similar pKa, so there will be an equilibrium between the deprotonation of the alcohol by hydroxide and the protonation of the alkoxide by water."
17	37	Identifying leaving groups	Alkyl, primary, eliminates, group, substitution, reaction, better, favorable, S_N2, leaves	"Primary alkyl halide is more likely to undergo S_N2 since chloride is a good leaving group."
18	106	Identifying electrophiles and nucleophiles based on charges	Nucleophile, electrophile, attacks, electron, acts, good, lone, pair, charge, positive	"Hydroxide ion is a nucleophile. It bears a negative charge."
19	48	Identifying nucleophiles or weighing nucleophilic and basic properties	S_N2, nucleophile, reaction, E2, E1, primary, base, attacks, good, strong	"LDA is acting like a strong base, not like a nucleophile in an S_N2 reaction."
20	22	Connecting steric hindrance to hybridization	Steric, hindrance, S_N2, reaction, primary, interference, bulky, prevents, reaches, hybridization	"S_N2 reactions can occur. The carbon center attached to the chlorine is primary and sp3 hybridized, so there is no steric hindrance or orbital interference."
21	14	Connecting steric hindrance to the number of substituents	Hindrance, steric, substituent, bulky, attacks, effects, tertiary, likely, primary, spatial	"Chlorine is primary. There is very little steric hindrance."

5.2 | RQ2: Analysis of the *modes of reasoning* and *levels of granularity* applied in students' argumentation

As proposed by *computational grounded theory*, we looked for a research-informed framework to establish links between the identified clusters so that the ML output can be used to guide students' learning progression. Looking for a theoretical framework was led by two requirements: On the one hand, the theory-driven considerations needed to differentiate students' argumentation skills and their chemical concept knowledge in as much detail as possible. On the other hand, these considerations must allow for using ML in an instructional setting to accompany students' learning progression over time. A theoretical framework consisting of the *modes of reasoning* (Sevian & Talanquer, 2014) and the *levels of granularity* (Bodé et al., 2019; Darden, 2002; Deng et al., 2023; Deng & Flynn, 2021; Luisi, 2002; Machamer et al., 2000; Southard et al., 2016; Talanquer, 2018; van Mil et al., 2013) best reflected the cluster analysis findings.

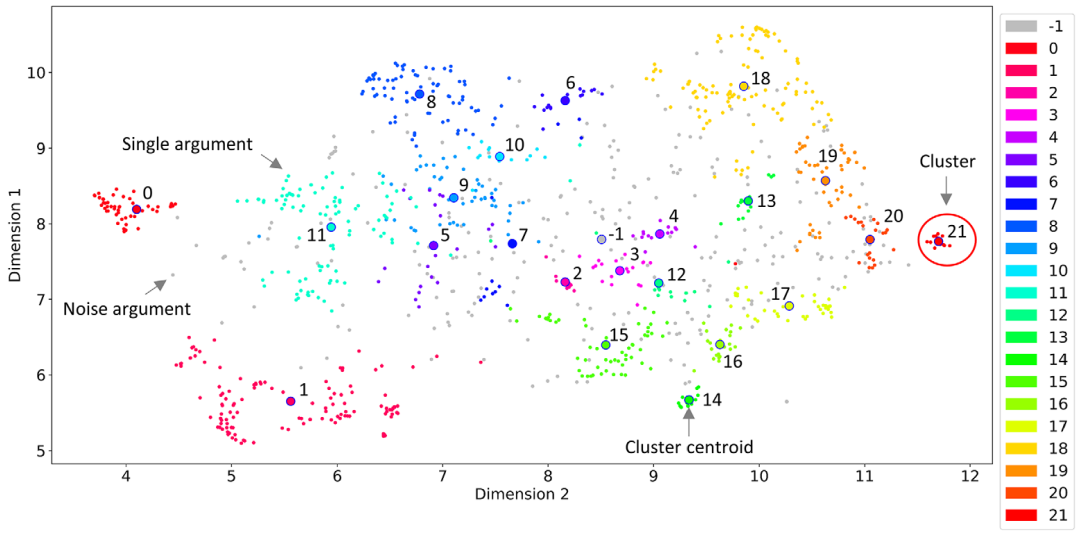


FIGURE 3 Two-dimensional embedding space of the extracted clusters. Each dot in the visualization corresponds to a single argument and different clusters are indicated by different colors. Noisy arguments, which are arguments that are too vague or include multiple topics, are represented by gray dots. Larger dots visualize cluster centroids. Dimensions 1 and 2 indicate some sort of principal components.

To evaluate the elaborateness and interconnectedness of applied concepts in students' scientific argumentation, Sevian and Talanquer (2014) proposed four *modes of reasoning*, namely, *descriptive*, *relational*, *linear causal*, and *multicomponent causal* (Table 2). These *modes of reasoning* discuss the complexity of novice or expert reasoning regarding their abilities to connect various concepts, provide sound justifications, and develop sophisticated explanations (Russ et al., 2008; Sevian & Talanquer, 2014). A major commitment underpinning these *modes of reasoning* is that students' understanding of science subjects cannot only be examined by assessing their content knowledge but also by evaluating how they integrate new information into their existing cognitive network (Sevian & Talanquer, 2014). So far, the *modes of reasoning* have been applied in various research contexts to analyze student reasoning in chemistry (e.g., Bodé et al., 2019; Carle et al., 2021; Deng et al., 2022, 2023; Deng & Flynn, 2021; Moon et al., 2019; Moreira et al., 2019; Sevian & Talanquer, 2014; Weinrich & Talanquer, 2016).

Arguments about scientific phenomena cannot only be generated in different *modes of reasoning* but also at different *levels of granularity* (Bodé et al., 2019; Deng et al., 2023; Deng & Flynn, 2021; Soo, 2019). The concept of *granularity* is characterized by the level of specificity to which an argument refers, assuming that different tasks require different grain sizes to explain the underlying processes. In general, the *level of granularity* varies depending on the context at hand but often ranges from global to submicroscopic perspectives (Darden, 2002). In our study, we adopt four *levels of granularity*, namely, *structural*, *energetic*, *phenomenological*, and *electronic* (Table 3), that were already applied in Deng and Flynn (2021) and Deng et al. (2023). These four levels are in alignment with Krist et al.'s (2019) *epistemic heuristics* which can be applied to characterize students' mechanistic reasoning across scientific content areas. The *epistemic heuristics* include identifying all factors of the phenomenon under investigation, considering the scalar level below, decomposing the factors at this lower scalar level, and linking those factors back to the target phenomenon. The target phenomenon of Krist et al. (2019) corresponds to *phenomenological* descriptions, which are characterized by argumentation about

TABLE 2 *Modes of reasoning* to characterize the elaborateness and interconnectedness of students' argumentation.

Mode of reasoning	Description	Student example arguing about reaction 1.3 (Figure 1)
Descriptive	Redescribing a phenomenon by recognizing explicit properties of single salient components without providing any connections or cause-and-effect relationships between these components	"This reaction is plausible. The alcohol acts as an acid and hydroxide ions act as a base." (Cluster 7)
Relational	Discussing the relationships between different components as well as their explicit and implicit properties in a correlative fashion or providing surface-level justifications of the established relations constrained by a reduction of variables	"This reaction is plausible because the pKa value of water is like the pKa value of an alcohol. So, the acidic alcohol is easily deprotonated by the basic hydroxide ions to form alkoxide." (Cluster 14)
Linear causal	Identifying cause-and-effect interactions between explicit and implicit properties of most components or explaining mechanisms based on the static interplay of different concepts	"This reaction is plausible. Alkoxide and hydroxyl exist in equilibrium. Hydroxide ions act as a base and deprotonate the acidic alcohol. The two involved OH groups have very similar pKa values and are, thus, expected to interchange protons easily." (Cluster 16)
Multi-component causal	Perceiving phenomena as the dynamic interplay of several interrelated concepts, explaining systematically how different properties affect the reactivity of the components, and weighing how different concepts relate to each other	*Not observed in the data*

Note: Descriptions of the *modes of reasoning* are adapted from Sevian and Talanquer (2014), student examples are chosen from the data analyzed in this study. Cluster IDs can be found in Table 1.

explicit or measurable properties of the given structures. The lower scalar level arises, in turn, from the interplay of the *energetic*, *structural*, and *electronic* properties of the involved components.

Based on these theoretical considerations, we qualitatively analyzed students' arguments again. A *mode of reasoning* and *level of granularity* was assigned to every argument, which is in the following illustrated by the student examples listed in Table 2. These examples address the concepts of acidity and basicity while arguing about reaction 1.3 (Figure 1).

In the *descriptive* example, the student claimed that the reaction is plausible; however, it is only described that the alcohol molecule can act as an acid and the hydroxide ion as a base, lacking any cause-and-effect relationship. In the *relational* argument, the student compared the pKa values of water and alcohol molecules, which are essential chemical properties, to support the claim. Highlighting the similarity in pKa values demonstrates a stronger understanding of the acid–base concept, resulting in an increased elaborateness of the argument. In the *linear causal* example, the student explains, based on pKa values, that alkoxide and hydroxyl exist in equilibrium. This detailed explanation of why in a chemical equilibrium a fast exchange of

TABLE 3 *Levels of granularity* proposed by Deng and Flynn (2021) to classify the grain size of students' argumentation.

Level of granularity	Description
Structural	Focusing on chemical compounds as well as their properties, which are derived from structural features such as the connectivity of atoms, atom size, the presence of charges, and salient centers
Energetic	Focusing on thermodynamic and kinetic changes during chemical reactions, which are whether a reaction is reactant- or product-favored, whether any entropic effects impact the reaction process, and at which rate a reaction occurs
Phenomenological	Reasoning about measurable properties of components such as pKa values, the direction of chemical equilibrium, or reaction rate without including electron distribution
Electronic	Reasoning about the distribution and movement of electrons in chemical compounds by including electronic effects like resonance, hyperconjugation, or inductive effects to explain how electron density impacts the properties of a molecule

protons can occur strengthens the causal link in the argument and provides a comprehensive understanding of the factors contributing to the plausibility of acid–base reactions. The *modes of reasoning* helped us classify the degree of causality in each argument.

Despite different levels of causality, the three arguments all address structural properties related to acid–base reaction 1.3 (Figure 1) by relying on the identification of acidic and basic functional groups and the influence of their pKa values on the reaction outcome. In contrast, other students argued about acid–base reaction 1.3 (Figure 1) from a kinetic point of view by simply claiming that acid–base reactions occur faster than any other reaction type. Such an argumentation does not include any structural properties; as a result, we distinguished, inspired by the *levels of granularity*, between a *structural* and *energetic* argumentation.

In addition, none of the three examples above included electronic properties such as electron density or orbital interactions, which is why we classified these examples as *phenomenological*. Nonetheless, other students integrated electronic properties in their argumentation about acid–base reaction 1.3 (Figure 1). For example, one student argued that hydroxide ions are strong bases due to their small atom size and low polarizability, leading to a localization of the negative charge. This argument highlights electronic properties related to the concept of basicity, showing the necessity to differentiate between *phenomenological* and *electronic* arguments as proposed in the *levels of granularity*.

With this qualitative analysis in mind, we established a two-dimensional holistic scoring rubric by adapting the *modes of reasoning* and *levels of granularity* (Figure 4). This scoring rubric is nominally scaled. The number of clusters detected with HDBSCAN guided the process of rubric design as this number served as a reference for estimating how many categories could be clearly distinguished in the scoring rubric. For the *levels of granularity*, we found by interpreting the detected clusters that classifying every argument into two of the four proposed levels described the responses in our sample best. Accordingly, every argument has one of four codes *structural-phenomenological*, *structural-electronic*, *energetic-phenomenological*, or *energetic-electronic*. Additionally, the *structural level of granularity* was divided into three sublevels because HDBSCAN detected topics with a *structural* focus for three different concepts.

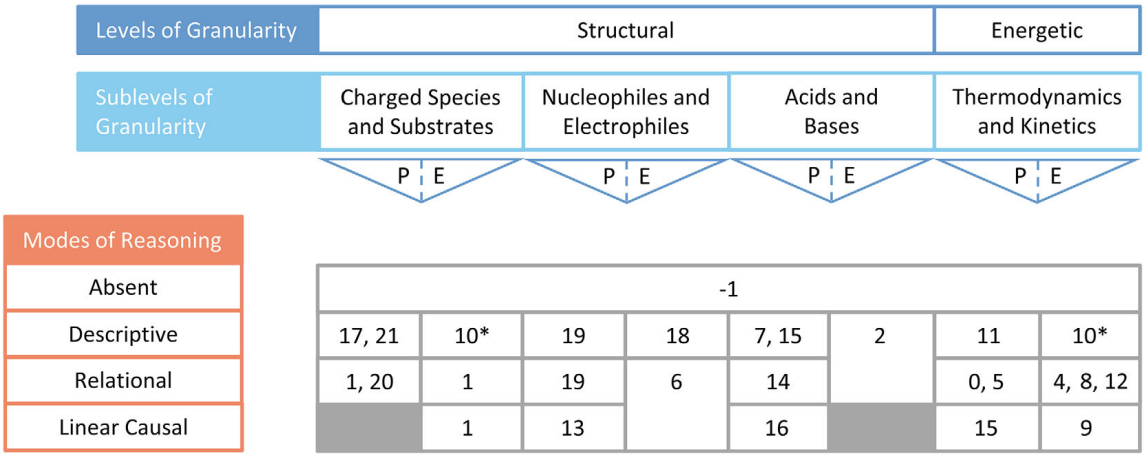


FIGURE 4 Classification of the clusters identified through HDBSCAN according to the two-dimensional scoring rubric. For instance, cluster −1 mostly contains *absent* responses. Numbers indicate the cluster ID listed in Table 1. P stands for *phenomenological*, E for *electronic*. The asterisk shows that clusters belong together.

Here, we introduced the *structural sublevels* of *charged species and substrates*, *nucleophiles and electrophiles*, and *acids and bases*. This distinction helped us consider the chemical entities students included in their argumentation. To clarify what we mean by *energetic*, we renamed this level to *thermodynamics and kinetics*. For the *modes of reasoning*, we introduced the mode *absent* for off-task responses. Since instances of *multicomponent causal reasoning* were not prompted, we did not include this *mode* in our coding scheme.

After including research-informed considerations, we examined if the theory- and data-driven analysis of students' argumentation matched by assigning every cluster to the *mode of reasoning* and *level of granularity* that was most frequently applied across all student arguments in that cluster (Figure 4). In doing so, we found that the clusters partially represented the theoretical framework—and vice versa. Consequently, we iteratively modified the classification of HDBSCAN by dividing some large clusters into further subclusters and aggregating clusters that contained similar ideas (Table 4). For example, cluster 19 could be divided into two further subclusters since this cluster incorporated two different *modes of reasoning* with enough sample responses to precisely differentiate between these two *modes*. By contrast, clusters 17 and 21 were aggregated because both referred to a *descriptive*, *phenomenological* argumentation about *charged species and substrates*. Aligning the results of the data-driven cluster analysis with theoretical considerations contributed to establishing well-distinguishable coding categories (cf., Carlsen & Ralund, 2022). However, because of ambiguous wording in students' argumentation as well as the small number of arguments in clusters 2, 6, and 10, it was not useful for subsequent supervised automation to further divide these clusters as conceptualized in the theoretical framework (Figure 4). Besides, *structural-phenomenological* reasoning with a focus on *charged species and substrates* and a *linear causal mode of reasoning* was not present in our dataset at all, which is why we excluded this category from our analysis, leading to a total of 21 categories (Figure 4).

HDBSCAN facilitated rubric design as it detected patterns across many arguments, which enhanced the validity of our scoring rubric as it is grounded in data. Furthermore, HDBSCAN identified distinct topics students reasoned about. In particular, this approach assisted us in differentiating between well-interpretable and vague categories. Integrating research-informed

TABLE 4 Final modifications of the clusters identified through HDBSCAN guided by the *modes of reasoning* and *levels of granularity*.

Modification	Justification
Dividing a cluster into subclusters	
Dividing cluster 1 into three subclusters	Responses in cluster 1 explain why a leaving group is good or bad by referring to base strength, electronegativity, or polarizability. The high number of responses indicates the heterogeneity and coarse level of data differentiation. The cluster was divided into three subclusters which attribute leaving group quality to (i) low base strength, (ii) high electronegativity, and (iii) high polarizability to better differentiate the data.
Dividing cluster 15 into two subclusters	Responses in cluster 15 justify the occurrence of proton transfer by identifying acids and bases or consider that acid–base reactions occur faster than substitution reactions. Despite the similar wording, these two ideas are different in quality when judging the plausibility of chemical reactions. Hence, the cluster was divided into two subclusters to distinguish between describing proton transfer and considering kinetics.
Dividing cluster 19 into two subclusters	Responses in cluster 19 identify nucleophiles and electrophiles or weigh nucleophilic properties against basic properties. Since combining nucleophilic and basic properties is considered a higher *mode of reasoning*, these arguments were separated from rather *descriptive* ones by dividing this cluster into two subclusters.
Aggregating various clusters	
Aggregating clusters 0 and 5	Responses in cluster 0 reason about entropic changes by referring to the number of reactants and products, while responses in cluster 5 reason about enthalpic changes by referring to bond energies. Although these clusters are distinct and easily interpretable, they address, from a technical point of view, a system's thermodynamic changes. Due to the technical comparability, the clusters were unified.
Aggregating clusters 4, 8, and 12	Responses in cluster 4 identify resonance stabilization in carbanions, while responses in cluster 12 identify resonance stabilization in general, which is why they were aggregated. Similarly, responses in cluster 8 apply electronegativity, which is another electronic property, to estimate the energetic level of molecules. So, these clusters address the thermodynamic stability of molecules on an electronic level, leading to their combination.
Aggregating clusters 17 and 21	Clusters 17 and 21 are *phenomenological* and *descriptive* in nature. Responses in both clusters only recognize and describe salient properties of the displayed molecules, which are properties that are explicitly shown by Lewis structures. Consequently, these clusters were aggregated.

considerations finally showed where clusters needed to be further differentiated or, in contrast, be aggregated (cf., Carlsen & Ralund, 2022). Taken together, refining the detected clusters influenced by subject matter expertise uncovered new argumentation patterns and laid the foundation for the subsequent process of pattern confirmation.

TABLE 5 Metrics to measure score agreements.

Accuracy	Cohen's κ	Weighted F_1 score	Macro F_1 score
Percentage of agreement between two raters	Agreement between two raters beyond chance agreement	Agreement metrics for imbalanced datasets accounting for the number and type of prediction errors as the harmonic mean of precision and recall	
		Weighted mean of all per-class F_1 scores	Arithmetic, i.e., unweighted mean of all per-class F_1 scores

Based on the new scoring rubric, a student research assistant and the first author scored the entire dataset holistically by assigning one of 20 mutually exclusive scores to answer RQ2. Here, we needed to exclude cluster 3 from analysis (Figure 4) since it contained only a few arguments, which could also be classified into one of the other clusters. Across all rubric categories, we achieved *almost perfect* inter-rater reliability (cf., Landis & Koch, 1977) as shown by an accuracy of 84%, Cohen's κ of 0.83, a weighted F_1 score of 0.85, and a macro F_1 score of 0.78 (for a definition of these metrics, see Table 5). After that, codes were discussed until full agreement was reached. The distribution of the arguments across the 20 categories is shown in Table 6. Along with other researchers, we found that students at the same educational level applied different *modes of reasoning* to solve the same exercise (Bodé et al., 2019; Deng & Flynn, 2021; Moreira et al., 2019; Weinrich & Talanquer, 2016). The students in this study constructed mostly *descriptive* or *relational* arguments, which is also a common observation in other studies (cf., Carle et al., 2021; Deng et al., 2022; Deng & Flynn, 2021). Moreover, when arguing about Williamson ether synthesis, students applied predominantly a *phenomenological level of granularity*.

5.3 | RQ3: Level of accuracy of the deep learning model

After human scoring, we checked if a deep neural network can accurately classify students' arguments according to the holistic, 20-category rubric. For this reason, we compared the accuracy of different large language models on the validation set. *BERT base uncased*, *RoBERTa base, RoBERTa large*, and *SciBERT scivocab uncased* reached a maximum macro F_1 score of 0.88 across all hyperparameter configurations, while *BERT large uncased* performed best with a maximum macro F_1 score of 0.89. *BERT large uncased* achieved the highest macro F_1 score with 38 epochs, a batch size of 8, and a learning rate of 1e-5. We then applied this best hyperparameter configuration to measure the accuracy of the model on the unseen test set.

Across all 20 categories, our multi-class model achieved *almost perfect* machine-human score agreement as indicated by an accuracy of 87%, Cohen's κ of 0.86, a weighted F_1 score of 0.86, and a macro F_1 score of 0.80. Machine-human score agreements were slightly higher than human-human score agreements. To be noted, we used the initial human scores, which may contain errors as a rater may have lacked a complete understanding of the rubric categories, to calculate human-human score agreements while training the model on consensus scores. As presented in Table 7, the macro F_1 scores per category ranged from 0.50 to 1.00, showing that the applied deep learning model identified some categories more accurately than others. Predictably, the algorithm struggled to detect categories with fewer sample arguments because ML is based on an inductive learning procedure that is dependent on diverse datasets.

TABLE 6 Number of arguments per rubric category and task.

	Structural						Energetic	
	Charged species and substrates		Nucleophiles and electrophiles		Acids and bases		Thermodynamics and kinetics	
	P	E	P	E	P	E	P	E
Intramolecular Williamson ether synthesis								
Absent	6							
Descriptive	78	24*	27	13	63	35	37	24*
Relational	23	20	48	36	34		40	19
Linear causal	0	21	7		11	0	14	7
Claisen condensation								
Absent	17							
Descriptive	49	46*	20	18	27	35	10	46*
Relational	15	12	16	38	24		42	104
Linear causal	0	15	15		9	0	8	25

Note: The asterisk shows that clusters belong together.
Abbreviations: E, *electronic*; P, *phenomenological*.

TABLE 7 F_1 scores listed per rubric category.

	Structural						Energetic	
	Charged species and substrates		Nucleophiles and electrophiles		Acids and bases		Thermodynamics and kinetics	
	P	E	P	E	P	E	P	E
Absent	0.57							
Descriptive	0.89	0.92*	0.96	0.86	0.91	0.88	0.56	0.92*
Relational	0.60	0.57	0.81	0.95	1.00		0.94	0.98
Linear causal	n.e.	0.67	0.57		1.00	n.e.	0.50	0.82

Note: The asterisk shows that clusters belong together.
Abbreviations: E, *electronic*; P, *phenomenological*; n.e., non-existent.

To subsequently perform a baseline comparison, we also tried to automate the holistic scoring with a support vector machine. A support vector machine is a more traditional, so-called shallow supervised ML algorithm that seeks to find the best possible boundary between different classes in a dataset. We chose a support vector machine since it performed well in a classification task on students' use of the Lewis acid–base theory across different assessment items (Yik et al., 2021). However, by adopting the *modes of reasoning* and *levels of granularity*, the support vector machine achieved only *fair* machine-human score agreement with an accuracy of 31%, Cohen's κ of 0.25, a weighted F_1 score of 0.25, and a macro F_1 score of 0.21. This demonstrates that applying a deep neural network was necessary to solve the holistic scoring problem, which is in line with other findings demonstrating that deep neural networks perform better than more traditional ML models in educational contexts

(e.g., Gombert et al., 2022; Wulff, Mientus, et al., 2022). However, the accuracy of ML methods strongly depends on the complexity of the task at hand. For instance, Küchemann et al. (2020) found that more traditional ML models outperformed deep neural networks for simpler scoring tasks.

In sum, we built a deep learning model with *almost perfect* accuracy capable of automatically evaluating the nature of students' arguments across different chemical reactions and concepts. This model can be used in further formative assessments to analyze students' argumentation irrespective of the displayed chemical reaction.

5.4 | RQ4: Consensus between human and deep learning scoring guidelines

To answer RQ4, we validated the accuracy of our deep learning architecture in three different ways. First, we used *ChatGPT*, a large language model that can analyze and produce human-like language, to generate 20 fictitious student arguments (cf., Dai et al., 2023), one per category, based on the cluster descriptions (Table 1). We then tested whether our deep learning architecture can classify these *ChatGPT*-generated arguments into the correct category, which helped us estimate if the 20 categories are distinct from each other and if the cluster descriptions are valid. Our deep learning architecture classified 17 fictitious student arguments into the correct category. This implies that the cluster descriptions are comprehensible, distinguishable, and valid for generating additional fictitious arguments. In addition, this approach provides further support for the validity of the clustering method used to define the categories.

For black box analysis, we selected a *phenomenological*, *descriptive* argument from each of the four rubric sublevels *charged species and substrates*, *nucleophiles and electrophiles*, *acids and bases*, and *thermodynamics and kinetics*. Afterward, we added word groups listed in our scoring guidelines to these *phenomenological*, *descriptive* arguments so that they reached the next category of the scoring rubric in the respective sublevel. For example, we added the phrase "as they are a weak base" to a *phenomenological*, *descriptive* argument in the sublevel *charged species and substrates* and checked whether this additional phrase was sufficient for the algorithm to classify the argument as *phenomenological*, *relational*. With this approach, we simulated if students' learning progression can be accompanied by the deep learning architecture over time. The algorithm correctly classified 18 out of 20 arguments, which highlights the potential of the algorithm in monitoring students' learning progression even over a longer period.

Furthermore, the model explainer SHAP, a technique for white boxing an algorithm (Lundberg & Lee, 2017), was used to calculate the importance scores of individual words for each rubric category, that is, how individual words impact algorithmic output. In doing so, we formed the arithmetic mean of the importance scores across all arguments. For clarity, we present only the output for four exemplary rubric categories herein (Figure 5). Positive scores have a positive impact on the model output in the respective category and increase, thus, the probability of a positive classification, while negative scores have a negative impact on the model output in the respective category and decrease, thus, the probability of a positive classification. In general, our algorithm utilized similar keywords than human raters in categories with higher F_1 scores (Figure 5a, b). Words like *resonance*, *electron*, *octet*, and *dipole* relate, for instance, the thermodynamic stability of the reagents to electron distribution, which is why these words correspond to the human scoring guidelines in the category *thermodynamics and kinetics, electronic*, *relational* (Figure 5a). Furthermore, the words *electron*, *electronegative*, *inductive*,

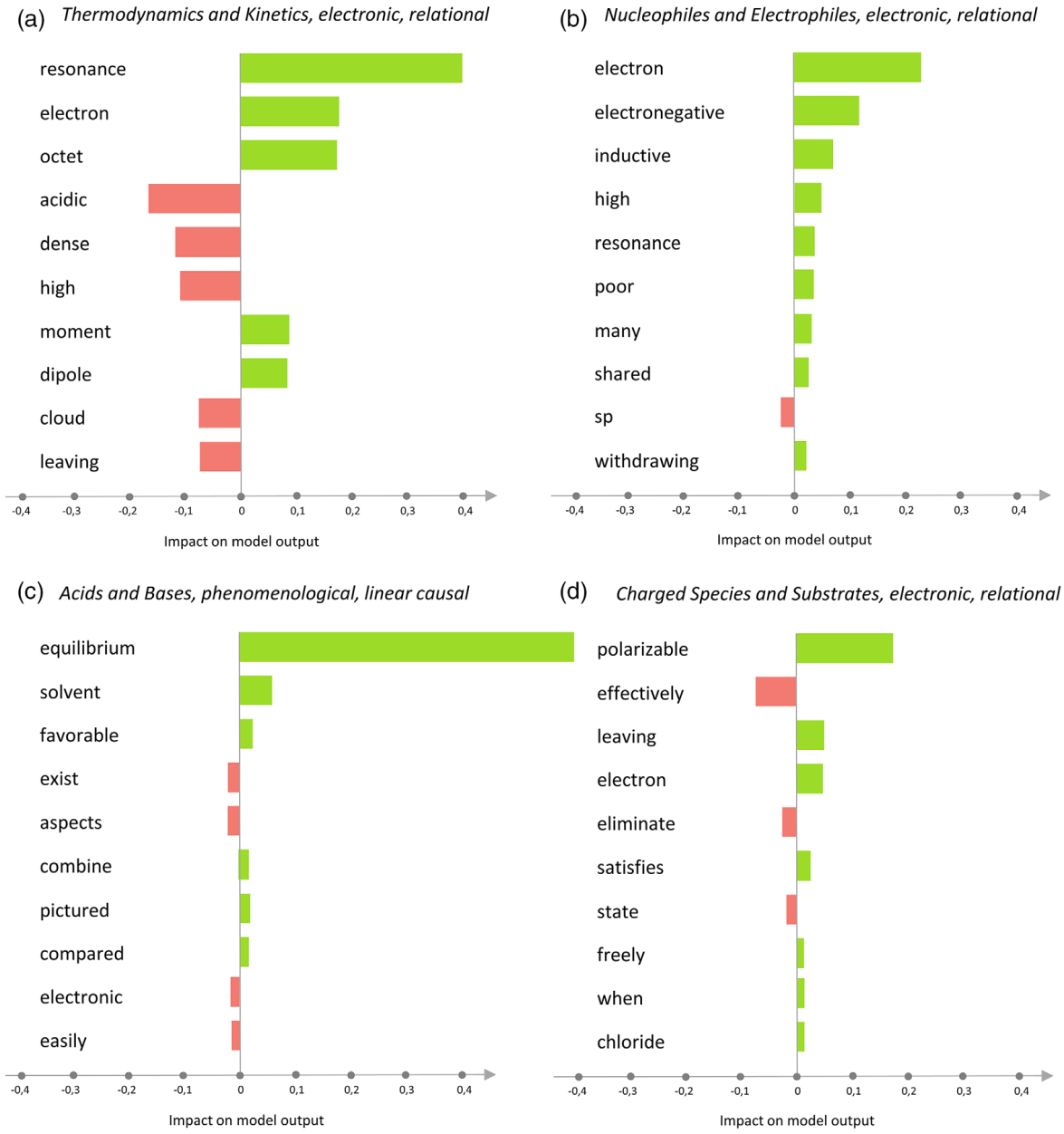


FIGURE 5 Importance scores of the 10 words that impact algorithmic output most significantly for four sample categories.

resonance, and *withdrawing* can be used to relate nucleophilic and electrophilic properties of reactive centers to their electron density (Figure 5b). In one category, we identified that the algorithmic decision relied almost exclusively on a single word, which is a critical model shortcut (Figure 5c). Although the word *equilibrium* is key to argumentation about *acids and bases*, the algorithm neglected other words like *pKa*, *acidic*, or *basic* in this category, which also mirrors human shortcuts in the scoring procedure. In the category *charged species and substrate*, *electronic*, *phenomenological*, a category with a lower F_1 score, the algorithm relied partly on expectable words like *polarizable*, *leaving*, and *electron*, but the impact of the other words on the model output was not explainable based on human scoring guidelines (Figure 5d).

Collectively, the implemented validation techniques demonstrated that the deep learning model adopts human scoring guidelines to a great extent. In particular, the deep learning architecture achieved an accuracy of 85% when using *ChatGPT* for external validation and an accuracy of 90% when conducting black box analysis. These metrics are in alignment with the overall model accuracy of 87%. Unsurprisingly, arguments from categories with lower F_1 scores (Table 7) were more difficult to classify. Furthermore, white box analysis determined the most significant words associated with positive or negative classifications in each category, which helped estimate if the words consulted by the deep learning model correspond to human scoring guidelines. Expectedly, the words used to classify arguments in categories with higher F_1 scores were comparable to keywords human raters looked for, while discrepancies arose in categories with lower F_1 scores. Overall, these validation methods provided additional performance metrics, helped us gain a better understanding of the model generalizability, and may increase confidence in decisions derived from the algorithm reported herein.

6 | LIMITATIONS

From a technological point of view, large language models such as BERT are trained on uncleaned language data, among others, from the Internet or Wikipedia. For this reason, large language models might acquire stereotyped biases (e.g., Caliskan et al., 2017), which can lead to the perpetuation of discrimination against racially marginalized groups in science assessment (Cheuk, 2021; Cheuk et al., 2019), such as second language learners (Deng et al., 2022; Deng & Flynn, 2023). Reducing bias can include a more controlled training process or applying subject-sensitive, situation-specific standards for examining bias along with creating countermeasures against discrimination (Grimm, Steegh, Çolakoğlu, et al., 2023; Grimm, Steegh, Kubsch, & Neumann, 2023).

Moreover, we did not investigate the difficulty of detecting the clusters without deep learning. We only showed via post-hoc explanations that the features consulted by the deep learning model correspond to human scoring guidelines. Specifically, the deep learning model relied frequently on specific words in responses (Figure 5, Wulff, Mientus, et al., 2022) or similarities between sentences as encoded in the embedding space (Figure 3, Wulff, Mientus, et al., 2022). However, we did not explore a priori if the clusters were easy to detect by humans. The high level of human-human inter-rater reliability indicates, nonetheless, that human analysts can reliably adopt the developed scoring rubric.

Besides, students' written arguments were assigned to one of 20 categories depending on the applied *mode of reasoning* and *level of granularity*. However, automating the analysis of some categories was not possible at all due to the lack of sample arguments. For example, we aggregated the categories of *relational- and linear causal-electronic reasoning* in the sublevel *nucleophiles and electrophiles* or the categories of *descriptive- and relational-electronic reasoning* in the sublevel *acids and bases*. In the future, further student-written arguments need to be collected to expand the scoring rubric to precisely differentiate between all conceivable categories. Here, it may be particularly helpful to gather data at different institutions, from courses with varying instructors and curricula, and from students with diverse demographic backgrounds. Future research can, therefore, increase the validity and reliability of the deep learning architecture with a larger training set.

A limitation related to the absence of arguments in some rubric categories is a low sample heterogeneity in our study since our model is trained on a sample that belongs to a rather

homogeneous demographic group. In other words, the data used to train our deep learning model originated from a single set of students enrolled at the same institution. Hence, we could not determine to what extent the model is generalizable to other institutions or cohorts. Nevertheless, prior research documented that pre-trained large language models applied for analyzing students' reasoning in science subjects could be generalized across contexts, such as institutions and disciplines (Wulff et al., 2023). In detail, a fine-tuned BERT model developed for classifying pre-service physics teachers' written reflections could be reliably utilized across universities as well as for analyzing written reflections of non-physics students (Wulff et al., 2023). This suggests that BERT models as used in this study show strong generalizability across various contexts, enabling researchers to share their models and adapt these models to specific student populations. In the future, we aim to further improve model generalizability by fine-tuning the model to new contexts. Based on more heterogeneous data, the potential bias of our deep learning architecture could also be reduced (cf., Noyes et al., 2020), which may build more confidence in the results produced by our model.

7 | CONCLUSIONS AND IMPLICATIONS

We employed ML at an early stage in the assessment development process to create and refine a holistic scoring rubric for evaluating students' argument complexity when judging the plausibility of competing chemical reactions. Our approach combined data-driven deep learning methods with theory-driven human interpretation, which aligns with emerging recommendations for using ML in conjunction with research-informed theories to get a more comprehensive understanding of complex phenomena (Carlsen & Ralund, 2022; Kubsch et al., 2023; Nelson, 2020; Rosenberg and Krist, 2021; Tschisgale et al., 2023).

In contrast to other projects analyzing students' scientific argumentation with ML, we applied a clustering approach as the first step in our analysis to extend the level of detail of our classification. Here, we used HDBSCAN in an exploratory way to detect common patterns in students' arguments, which gave us an idea of the categories suitable for subsequent supervised automation. Cluster analysis supported us in formulating evidence statements that specify the performance accepted as evidence, which is an integral part of rubric development as postulated by the evidence-centered design approach to educational assessment (Kubsch et al., 2022; Mislevy, 2016; Mislevy et al., 2003; Pellegrino et al., 2016). Although ML is increasingly prevalent in educational assessment, such an unsupervised approach is not frequently applied, yet (Zhai, Yin, et al., 2020).

Cluster analysis informed the process of rubric development, but we did not adopt the findings of the cluster analysis without theory-driven reflection. So, cluster analysis uncovered to some extent how students judged the plausibility of competing chemical reactions but only the theoretical framework enabled the systematic classification of students' argumentation. Combining a data- and theory-driven approach ensured that the rubric categories did not only represent statistical values of similarity evident in the data but also pre-existing, empirical theories and human expertise. A similar methodological approach was used by Anderson et al. (2020), Prevost et al. (2012), Sherin (2013), Wulff, Buschhüter, et al. (2022), Wulff et al. (2023), and Zehner et al. (2016) to justify the validity of their qualitative coding, but not with a focus on rubric development. By contrast, Jescovitch, Doherty, et al. (2019), Jescovitch, Scott, et al. (2019), Jescovitch et al. (2021), Kaldaras et al. (2022), and Mao et al. (2018) focused on rubric development for ML-based scoring, but without the use of unsupervised techniques. In this

article, instead, we provide just like Haudek et al. (2015) and Rosenberg and Krist (2021) an example of how to incorporate unsupervised ML techniques into human qualitative interpretation for rubric development.

After the application of unsupervised techniques, we used large language models, which made significant advancements in addressing a broad spectrum of NLP issues (cf., Koroteev, 2021), as most notably experienced in conversational chatbots like *ChatGPT*. In conjunction with a deep neural network, we could automate the evaluation of students' arguments according to a 20-category holistic scoring rubric. As far as we know, the possibility of applying deep learning to evaluate argumentative patterns in students' responses according to a complex multi-class rubric was not reported before. Other projects in chemistry education research chose either an analytic scoring approach with a single or multiple binary classifiers (e.g., Dood et al., 2018; Watts et al., 2023b; Winograd, Dood, Moon, et al., 2021; Yik et al., 2021) or applied a holistic rubric with three to five levels (e.g., Donnelly et al., 2015; Dood et al., 2020a; Haudek et al., 2009, 2012, 2019; Liu et al., 2014; Maestrales et al., 2021; Noyes et al., 2020; Tansomboon et al., 2017; Vitale et al., 2016). With *computational grounded theory* as a methodological framework, we demonstrated that deep learning techniques can be used to evaluate students' arguments according to an abstract, holistic, multi-class rubric with *almost perfect* machine-human score agreement. Therefore, we assume that by strategically integrating theory-driven considerations into data-driven classifications it is possible to develop a scoring rubric that can be more easily incorporated into supervised analysis.

The diagnosis of students' argumentation skills can, in the future, be used to design instructional settings that realize aspects of adaptive learning (cf., Plass & Pawar, 2020). Here, our analysis enables providing personalized feedback and tailored exercises accounting for the *modes of reasoning* and *levels of granularity* applied in students' argumentation. This ML-assisted approach to adaptive learning seems promising as it increased students' level of explanation sophistication in another chemistry context (Dood et al., 2020a, b).

Along with realizing aspects of adaptive learning, the established deep learning architecture can be used to monitor students' learning progression over time (cf., Swiecki et al., 2022). Specifically, tracking students over a more extended period may help evaluate how argumentation skills change, how students move from one argumentation cluster to another, how and when clusters emerge, and what competencies need to be trained so that students reason in high-level clusters (Zhai et al., 2023). Moreover, this longitudinal data collection may enable measuring the effects of adaptive learning on students' argumentation skills and help determine optimal conditions for adaptive learning including how to fade support. Ultimately, this could allow for new approaches to Just-in-Time Teaching (Novak et al., 1999) realized with adaptively faded scaffolds (McNeill et al., 2006; Noroozi et al., 2018).

Furthermore, students responded in our instructional setting to the same prompt for each of the proposed reactions, that is, they repeatedly judged whether the displayed molecule is a plausible product. In the future, the prompt could be varied to generalize our deep learning model not only across different chemical reactions but also across different prompt types. Here, it may be possible to instantly assess which prompt type is most appropriate for a student and to adapt the prompt afterward. This may allow for providing automatically adapted prompts with varying degrees of scaffolding and guidance.

ACKNOWLEDGMENTS

This publication is part of the first author's doctoral thesis (Dr. rer. nat.) at the Faculty of Biology and Chemistry, Justus-Liebig-University Giessen, Germany. We thank Leonie Lieber, Ira

Caspari-Gnann, and Krenare Ibraj for their groundbreaking research on students' argumentation in chemistry as well as for providing access to their data. We thank Felix Blödtner for scoring students' arguments and all members of the Graulich group for fruitful discussions. Finally, Paul P. Martin thanks Brandon Yik, Amber Dood, and Field Watts for sharing their innovative thoughts on the application of ML in science education. There are no conflicts of interest to declare. Open Access funding enabled and organized by Projekt DEAL.

ORCID

Paul P. Martin https://orcid.org/0000-0001-8648-4250
David Kranz https://orcid.org/0000-0002-2054-6882
Peter Wulff https://orcid.org/0000-0002-5471-7977
Nicole Graulich https://orcid.org/0000-0002-0444-8609

REFERENCES

Anderson, D., Rowley, B., Stegenga, S., Irvin, P. S., & Rosenberg, J. M. (2020). Evaluating content-related validity evidence using a text-based machine learning procedure. *Educational Measurement: Issues and Practice*, *39*(4), 53–64.

Angelov, D. (2020). Top2Vec: Distributed representations of topics. *arXiv Preprint*, arXiv:2008.09470. https://doi.org/10.48550/arXiv.2008.09470

Bail, C. A. (2014). The cultural environment: Measuring culture with big data. *Theory and Society*, *43*(3/4), 465–482.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv Preprint*, arXiv:1903.10676. https://doi.org/10.48550/arXiv.1903.10676

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26–55.

Biernacki, R. (2012). *Reinventing evidence in social inquiry: Decoding facts and variables*. Palgrave Macmillan.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bodé, N. E., Deng, J. M., & Flynn, A. B. (2019). Getting past the rules and to the WHY: Causal mechanistic arguments when judging the plausibility of organic reaction mechanisms. *Journal of Chemical Education*, *96*(6), 1068–1082.

Brunton, S. L., & Kutz, J. N. (2019). *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in knowledge discovery and data mining: PAKDD 2013* (pp. 160–172). Springer.

Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, *10*(1), 1–51.

Carle, M. S., el Issa, R. J., Pilote, N., & Flynn, A. B. (2021). Ten essential delocalization learning outcomes: How well are they achieved? *ChemRxiv Preprint*, 1–28. https://doi.org/10.26434/chemrxiv.13322771.v1

Carlsen, H. B., & Ralund, S. (2022). Computational grounded theory revisited: From computer-led to computer-assisted text analysis. *Big Data & Society*, *9*(1), 20539517221080146.

Caruana, R., & Nori, H. (2022). Why data scientists prefer glassbox machine learning: Algorithms, differential privacy, editing and bias mitigation, In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 4776-4777). Association for Computing Machinery.

Caspari, I., & Graulich, N. (2019). Scaffolding the structure of organic chemistry students' multivariate comparative mechanistic reasoning. *International Journal of Physics and Chemistry Education*, *11*(2), 31–43.

Cetin, P. S. (2014). Explicit argumentation instruction to facilitate conceptual understanding and argumentation skills. *Research in Science & Technological Education*, *32*(1), 1–20.

Charmaz, K. (2014). *Constructing grounded theory*. Sage Publications.

Cheuk, T. (2021). Can AI be racist? Color-evasiveness in the application of machine learning to science assessments. *Science Education*, *105*(5), 825–836.

Cheuk, T., Osborne, J., Cunningham, K. R., Haudek, K., Santiago, M. M., Urban-Lurain, M., Merrill, J., Wilson, C. D., Stuhlsatz, M. A. M., Donovan, B., Bracey, Z. B., & Gardner, A. (2019). Towards an equitable design framework of developing argumentation in science items and rubrics for machine learning. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Baltimore, MD.

Chin, C., & Brown, D. E. (2000). Learning in science: A comparison of deep and surface approaches. *Journal of Research in Science Teaching*, *37*(2), 109–138.

Chin, C., & Osborne, J. (2010). Supporting argumentation through students' questions: Case studies in science classrooms. *Journal of the Learning Sciences*, *19*(2), 230–284.

Chinn, C., & Malhotra, B. (2002). Children's responses to anomalous scientific data: How is conceptual change impeded? *Journal of Educational Psychology*, *94*(2), 327–343.

Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., Liu, N., Li, S., Zhu, D., Cai, H., Sun, L., Li, Q., Shen, D., Liu, T., & Li, X. (2023). AugGPT: Leveraging ChatGPT for text data augmentation. *arXiv Preprint*, arXiv:2302.13007. https://doi.org/10.48550/arXiv.2302.13007

Darden, L. (2002). Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward/backward chaining. *Philosophy of Science*, *69*(S3), S354–S365.

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & de Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, *162*(104094), 1–43.

Deng, J. M., Carle, M. S., & Flynn, A. B. (2023). Students' reasoning in chemistry arguments and designing resources using constructive alignment. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (pp. 74–89). The Royal Society of Chemistry.

Deng, J. M., & Flynn, A. B. (2021). Reasoning, granularity, and comparisons in students' arguments on two organic chemistry items. *Chemistry Education Research and Practice*, *22*(3), 749–771.

Deng, J. M., & Flynn, A. B. (2023). "I am working 24/7, but I can't translate that to you": The barriers, strategies, and needed supports reported by chemistry trainees from english-as-an-additional language backgrounds. *Journal of Chemical Education*, *100*(4), 1523–1536.

Deng, J. M., Rahmani, M., & Flynn, A. B. (2022). The role of language in students' justifications of chemical phenomena. *International Journal of Science Education*, *44*(13), 2131–2151.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint*, arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Donnelly, D. F., Vitale, J. M., & Linn, M. C. (2015). Automated guidance for thermodynamics essays: Critiquing versus revisiting. *Journal of Science Education and Technology*, *24*(6), 861–874.

Dood, A. J., Dood, J. C., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2020a). Analyzing explanations of substitution reactions using lexical analysis and logistic regression techniques. *Chemistry Education Research and Practice*, *21*(1), 267–286.

Dood, A. J., Dood, J. C., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2020b). Using the research literature to develop an adaptive intervention to improve student explanations of an S_N1 reaction mechanism. *Journal of Chemical Education*, *97*(10), 3551–3562.

Dood, A. J., Fields, K. B., & Raker, J. R. (2018). Using lexical analysis to predict Lewis acid-base model use in response to an acid-base proton-transfer reaction. *Journal of Chemical Education*, *95*(8), 1267–1275.

Dood, A. J., Winograd, B. A., Finkenstaedt-Quinn, S. A., Gere, A. R., & Shultz, G. V. (2022). PeerBERT: Automated characterization of peer review comments across courses. In *LAK22: 12th International learning analytics and knowledge conference* (pp. 492–499). Association for Computing Machinery.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, *84*(3), 287–312.

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education*, *38*(1), 39–72.

Erduran, S. (2019). Argumentation in chemistry education: An overview. In S. Erduran (Ed.), *Argumentation in chemistry education: Research, policy and practice* (pp. 1–10). The Royal Society of Chemistry.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education*, *88*(6), 915–933.

Faize, F. A., Husain, W., & Nisar, F. (2018). A critical review of scientific argumentation in science education. *EURASIA Journal of Mathematics, Science and Technology Education*, *14*(1), 475–483.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673.

Gerard, L. F., Matuk, C., McElhaney, K., & Linn, M. C. (2015). Automated, adaptive guidance for K-12 education. *Educational Research Review*, *15*, 41–58.

Glaser, B., & Strauss, A. (1999). *Discovery of grounded theory: Strategies for qualitative research* (Vol. 1). Routledge.

Goldberg, Y. (2017). *Neural network methods for natural language processing*. Morgan & Claypool Publishers.

Gombert, S., di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., Grimm, A., Bohm, I., Neumann, K., & Drachsler, H. (2022). Coding energy knowledge in constructed responses with explainable NLP models. *Journal of Computer Assisted Learning*, *39*(3), 767–786.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Graulich, N., & Caspari, I. (2020). Designing a scaffold for mechanistic reasoning in organic chemistry. *Chemistry Teacher International*, *3*(1), 19–30.

Grimm, A., Steegh, A., Çolakoğlu, J., Kubsch, M., & Neumann, K. (2023a). Positioning responsible learning analytics in the context of STEM identities of under-served students. *Frontiers in Education*, *7*(1082748), 1–12.

Grimm, A., Steegh, A., Kubsch, M., & Neumann, K. (2023b). Learning analytics in physics education: Equity-focused decision-making lacks guidance! *Journal of Learning Analytics*, *10*(1), 71–84.

Grootendorst, M. (2020). Topic Modeling with BERT: Leveraging BERT and TF-IDF to create easily interpretable topics. https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6

Harris, C., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice*, *38*(2), 53–67.

Haudek, K. C., Moscarella, R. A., Urban-Lurain, M., Merrill, J. E., Sweeder, R. D., & Richmond, G. (2009). Using lexical analysis software to understand student knowledge transfer between chemistry and biology. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

Haudek, K. C., Moscarella, R. A., Weston, M., Merrill, J. E., & Urban-Lurain, M. (2015). Construction of rubrics to evaluate content in students' scientific explanation using computerized text analysis. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Chicago, IL.

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J. E., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. *CBE: Life Sciences Education*, *11*(3), 283–293.

Haudek, K. C., Wilson, C. D., Stuhlsatz, M. A. M., Donovan, B., Bracey, Z. B., Gardner, A., Osborne, J. F., & Cheuk, T. (2019). Using automated analysis to assess middle school students' competence with scientific argumentation. Paper presented at the annual meeting of the National Conference on Measurement in Education (NCME), Toronto, ON.

Haudek, K. C., & Zhai, X. (2021). Exploring the effect of assessment construct complexity on machine learning scoring of argumentation. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Virtual.

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A systematic review of deep learning approaches to educational data mining. *Complexity*, *2019*(1306039), 1–22.

Huang, C.-J., Wang, Y.-W., Huang, T.-H., Chen, Y.-C., Chen, H.-M., & Chang, S.-C. (2011). Performance evaluation of an online argumentation learning assistance agent. *Computers & Education*, *57*(1), 1270–1280.

Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal of Philosophy of Science*, *2*(1), 119–135.

Jeong, H., Songer, N. B., & Lee, S.-Y. (2007). Evidentiary competence: Sixth graders' understanding for gathering and interpreting evidence in scientific investigations. *Research in Science Education*, *37*(1), 75–97.

Jescovitch, L. N., Doherty, J. H., Scott, E. E., Cerchiara, J. A., Wenderoth, M. P., Urban-Lurain, M., Merrill, J. E., & Haudek, K. C. (2019a). Challenges in developing computerized scoring models for principle-

based reasoning in a physiology context. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Baltimore, MD.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Doherty, J. H., Wenderoth, M. P., Merrill, J. E., Urban-Lurain, M., & Haudek, K. C. (2019b). Deconstruction of holistic rubrics into analytic bins for large-scale assessments of students' reasoning of complex science concepts. *Practical Assessment, Research & Evaluation*, *24*(7), 1–13.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J. E., Urban-Lurain, M., Doherty, J. H., & Haudek, K. C. (2021). Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression. *Journal of Science Education and Technology*, *30*(2), 150–167.

Jiménez-Aleixandre, M. P., Bugallo Rodríguez, A., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": Argument in high school genetics. *Science Education*, *84*(6), 757–792.

Jiménez-Aleixandre, M. P., & Erduran, S. (2007). Argumentation in science education: An overview. In S. Erduran & M. P. Jiménez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from classroom-based research* (pp. 3–27). Springer.

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing: An Introduction to natural language processing, computational linguistics, and speech recognition* (Vol. 3). Prentice Hall.

Kaldaras, L., & Haudek, K. C. (2022). Validation of automated scoring for learning progression-aligned next generation science standards performance assessments. *Frontiers in Education*, *7*(968289), 1–22.

Kaldaras, L., Yoshida, N. R., & Haudek, K. C. (2022). Rubric development for AI-enabled scoring of three-dimensional constructed-response assessment aligned to NGSS learning progression. *Frontiers in Education*, *7*(983055), 1–15.

Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching*, *41*(7), 748–769.

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: The role of scaffolding in assessment tasks. *Science Education*, *98*(4), 674–704.

Koroteev, M. (2021). BERT: A review of applications in natural language processing and understanding. *arXiv Preprint*, arXiv:2103.11943. https://doi.org/10.48550/arXiv.2103.11943

Kranz, D., Schween, M., & Graulich, N. (2023). Patterns of reasoning: Exploring the interplay of students' work with a scaffold and their conceptual knowledge in organic chemistry. *Chemistry Education Research and Practice*, *24*(2), 453–477.

Krist, C., Schwarz, C. V., & Reiser, B. J. (2019). Identifying essential epistemic heuristics for guiding mechanistic reasoning in science learning. *Journal of the Learning Sciences*, *28*(2), 160–205.

Kubsch, M., Czinczel, B., Lossjew, J., Wyrwich, T., Bednorz, D., Bernholt, S., Fiedler, D., Strauß, S., Cress, U., Drachsler, H., Neumann, K., & Rummel, N. (2022). Toward learning progression analytics: Developing learning environments for the automated analysis of learning using evidence centered design. *Frontiers in Education*, *7*(981910), 1–15.

Kubsch, M., Krist, C., & Rosenberg, J. M. (2023). Distributing epistemic functions and tasks: A framework for augmenting human analytic power with machine learning in science education research. *Journal of Research in Science Teaching*, *60*(2), 423–447.

Kubsch, M., Rosenberg, J. M., & Krist, C. (2021). Beyond supervision: Human/machine distributed learning in learning sciences research. In E. de Vries, Y. Hod, & J. Ahn (Eds.), *Proceedings of the 15th international conference of the learning sciences: ICLS 2021* (pp. 897–898). International Society of the Learning Sciences.

Küchemann, S., Klein, P., Becker, S., Kumari, N., & Kuhn, J. (2020). Classification of students' conceptual understanding in STEM education using their visual attention distributions: A comparison of three machine-learning approaches. In *Proceedings of the 12th international conference on computer supported education 2020* (pp. 36–46). SciTePress.

Kuhn, D. (2010). Teaching and learning science as argument. *Science Education*, *94*(5), 810–824.

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development*, *74*(5), 1245–1260.

Kuo, C.-Y., & Wu, H.-K. (2013). Toward an integrated model for designing assessment systems: An analysis of the current status of computer-based assessments in science. *Computers & Education*, *68*, 388–403.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lee, H.-S., Gweon, G.-H., Lord, T., Paessel, N., Pallant, A., & Pryputniewicz, S. (2021). Machine learning-enabled automated feedback: Supporting students' revision of scientific arguments based on data drawn from simulation. *Journal of Science Education and Technology*, *30*(2), 168–192.

Lee, H.-S., Liu, O. L., Pallant, A., Roohr, K. C., Pryputniewicz, S., & Buck, Z. E. (2014). Assessment of uncertainty-infused scientific argumentation. *Journal of Research in Science Teaching*, *51*(5), 581–605.

Lee, H.-S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education*, *103*(3), 590–622.

Lieber, L. S., & Graulich, N. (2020). Thinking in alternatives: A task design for challenging students' problem-solving approaches in organic chemistry. *Journal of Chemical Education*, *97*(10), 3731–3738.

Lieber, L. S., & Graulich, N. (2022). Investigating students' argumentation when judging the plausibility of alternative reaction pathways in organic chemistry. *Chemistry Education Research and Practice*, *23*(1), 38–53.

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022a). Closing the gap of organic chemistry students' performance with an adaptive scaffold for argumentation patterns. *Chemistry Education Research and Practice*, *23*(4), 811–828.

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022b). Students' individual needs matter: A training to adaptively address students' argumentation skills in organic chemistry. *Journal of Chemical Education*, *99*(7), 2754–2761.

Lim, L., Bannert, M., van der Graaf, J., Singh, S., Fan, Y., Surendrannair, S., Rakovic, M., Molenaar, I., Moore, J., & Gašević, D. (2023). Effects of real-time analytics-based personalized scaffolds on students' self-regulated learning. *Computers in Human Behavior*, *139*(107547), 1–18.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, *33*(2), 19–28.

Liu, X. (2020). *Using and developing measurement instruments in science education: A Rasch modeling approach* (Vol. 2). Information Age Publishing.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv Preprint*, arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

Luisi, P. L. (2002). Emergence in chemistry: Chemistry as the embodiment of emergence. *Foundations of Chemistry*, *4*(3), 183–200.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc.

Luo, X., Wei, B., Shi, M., & Xiao, X. (2020). Exploring the impact of the reasoning flow scaffold (RFS) on students' scientific argumentation: Based on the structure of observed learning outcomes (SOLO) taxonomy. *Chemistry Education Research and Practice*, *21*(4), 1083–1094.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25.

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B., & Krajcik, J. (2021). Using machine learning to score multi-dimensional assessments of chemistry and physics. *Journal of Science Education and Technology*, *30*(2), 239–254.

Manning, C. D. (2022). Human language understanding & reasoning. *Daedalus*, *151*(2), 127–138.

Manz, E. (2016). Examining evidence construction as the transformation of the material world into community knowledge. *Journal of Research in Science Teaching*, *53*(7), 1113–1140.

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H.-S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, *23*(2), 121–138.

Martin, P. P., & Graulich, N. (2023). When a machine detects student reasoning: A review of machine learning-based formative assessment of mechanistic reasoning. *Chemistry Education Research and Practice*, *24*(2), 407–427.

Mathew, A., Amudha, P., & Sivakumari, S. (2021). Deep learning techniques: An overview. In A. E. Hassanien, R. Bhatnagar, & A. Darwish (Eds.), *Advanced machine learning technologies and applications: Proceedings of AMLTA 2020* (pp. 599–608). Springer.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density-based clustering. *Journal of Open Source Software*, *2*(11), 205–206.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv Preprint*, arXiv:1802.03426. https://doi.org/10.48550/arXiv.1802.03426

McNeill, K. L., & Krajcik, J. (2011). *Supporting grade 5–8 students in constructing explanations in science: The claim, evidence, and reasoning framework for talk and writing*. Pearson Allyn & Bacon.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences*, *15*(2), 153–191.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv Preprint*, arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781

Miller, E. C., & Krajcik, J. S. (2019). Promoting deep learning through project-based learning: A design problem. *Disciplinary and Interdisciplinary Science Education Research*, *1*(1), 1–10.

Mislevy, R. J. (2016). How developments in psychology and technology challenge validity argumentation. *Journal of Educational Measurement*, *53*(3), 265–292.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), 1–29.

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundation of machine learning*. MIT Press.

Moon, A., Moeller, R., Gere, A. R., & Shultz, G. V. (2019). Application and testing of a framework for characterizing the quality of scientific reasoning in chemistry students' writing on ocean acidification. *Chemistry Education Research and Practice*, *20*(3), 484–494.

Moreira, P., Marzabal, A., & Talanquer, V. (2019). Using a mechanistic framework to characterise chemistry students' reasoning in written explanations. *Chemistry Education Research and Practice*, *20*(1), 120–131.

Nehm, R. H., & Haertig, H. (2012). Human vs. computer diagnosis of students' natural selection knowledge: Testing the efficacy of text analytic software. *Journal of Science Education and Technology*, *21*(1), 56–73.

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, *49*(1), 3–42.

Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2021). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, *50*(1), 202–237.

Newton, P., Driver, R., & Osborne, J. (1999). The place of argumentation in the pedagogy of school science. *International Journal of Science Education*, *21*(5), 553–576.

Noroozi, O., Kirschner, P. A., Biemanns, H. J. A., & Mulder, M. (2018). Promoting argumentation competence: Extending from first- to second-order scaffolding through adaptive fading. *Educational Psychology Review*, *30*(1), 153–176.

Novak, G. M., Gavrin, A., Patterson, E., & Christian, W. (1999). *Just-In-time teaching: Blending active learning with web technology*. Prentice Hall.

Noyes, K., McKay, R. L., Neumann, M., Haudek, K. C., & Cooper, M. M. (2020). Developing computer resources to automate analysis of students' explanations of London dispersion forces. *Journal of Chemical Education*, *97*(11), 3923–3936.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching*, *41*(10), 994–1020.

Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education*, *95*(4), 627–638.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., & Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. Paper presented at the 33rd Conference on Neural Information Processing Systems, Vancouver, CAN.

Pellegrino, J., DiBello, L., & Goldman, S. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, *51*(1), 1–23.

Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education*, *52*(3), 275–300.

Prevost, L. B., Haudek, K. C., Merrill, J. E., & Urban-Lurain, M. (2012). Examining student constructed explanations of thermodynamics using lexical analysis. *Paper presented at the IEEE Frontiers in Education Conference*.

Rosenberg, J. M., & Krist, C. (2021). Combining machine learning and qualitative methods to elaborate students' ideas about the generality of their model-based explanations. *Journal of Science Education and Technology*, *30*(2), 255–267.

Ruder, S. (2019). *Neural transfer learning for natural language processing*. National University of Ireland.

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education*, *92*(3), 499–525.

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction*, *83*(101620), 1–10.

Saldana, J. (2015). *The coding manual for qualitative researchers* (Vol. 3). Sage Publications.

Sampson, V., Grooms, J., & Walker, J. P. (2011). Argument-driven inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments: An exploratory study. *Science Education*, *95*(2), 217–257.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 211–229.

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction*, *23*(1), 23–55.

Sevian, H., & Talanquer, V. (2014). Rethinking chemistry: A learning progression on chemical thinking. *Chemistry Education Research and Practice*, *15*(1), 10–23.

Sherin, B. (2013). A computational study of commonsense science: An exploration in the automated analysis of clinical interview data. *Journal of the Learning Sciences*, *22*(4), 600–638.

Soo, K. W. (2019). *The role of granularity in causal learning*. University of Pittsburgh.

Southard, K., Wince, T., Meddleton, S., & Bolger, M. S. (2016). Features of knowledge building in biology: Understanding undergraduate students' ideas about molecular mechanisms. *CBE: Life Sciences Education*, *15*(1), ar7.1–ar7.16.

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*(100075), 1–10.

Taher Pilehvar, M., & Camacho-Collados, J. (2020). *Embeddings in natural language processing: Theory and advances in vector representations of meaning*. Morgan & Claypool Publishers.

Talanquer, V. (2018). Progressions in reasoning about structure-property relationships. *Chemistry Education Research and Practice*, *19*(4), 998–1009.

Tansomboon, C., Gerard, L. F., Vitale, J. M., & Linn, M. C. (2017). Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education*, *27*(4), 729–757.

Toulmin, S. E. (2003). *The uses of argument (Rev. Ed.)*. Cambridge University Press.

Tschisgale, P., Wulff, P., & Kubsch, M. (2023). Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review Physics Education Research*, *19*(2), 020123-1-020123-24.

Ullmann, T. D. (2019). Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education*, *29*(2), 217–257.

van Mil, M. H. W., Boerwinkel, D. J., & Waarlo, A. J. (2013). Modelling molecular mechanisms: A framework of scientific reasoning to construct molecular-level explanations for cellular behaviour. *Science & Education*, *22*(1), 93–118.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30*. Curran Associates, Inc.

Vitale, J. M., McBride, E., & Linn, M. C. (2016). Distinguishing complex ideas about climate change: Knowledge integration vs. specific guidance. *International Journal of Science Education*, *38*(9), 1548–1569.

Walker, J. P., Van Duzor, A. G., & Lower, M. A. (2019). Facilitating argumentation in the laboratory: The challenges of claim change and justification by theory. *Journal of Chemical Education*, *96*(3), 435–444.

Wang, C., Liu, X., Wang, L., Sun, Y., & Zhang, H. (2021). Automated scoring of Chinese grades 7-9 students' competence in interpreting and arguing from evidence. *Journal of Science Education and Technology*, *30*(2), 269–282.

Watts, F. M., Dood, A. J., & Shultz, G. V. (2023a). Automated, content-focused feedback for a writing-to-learn assignment in an undergraduate organic chemistry course. In *LAK23: 13th international learning analytics and knowledge conference* (pp. 531–537). Association for Computing Machinery.

Watts, F. M., Dood, A. J., & Shultz, G. V. (2023b). Developing machine learning models for automated analysis of organic chemistry students' written descriptions of organic reaction mechanisms. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (pp. 285–303). The Royal Society of Chemistry.

Watts, F. M., Park, G. Y., Petterson, M. N., & Shultz, G. V. (2022). Considering alternative reaction mechanisms: Students' use of multiple representations to reason about mechanisms for a writing-to-learn assignment. *Chemistry Education Research and Practice*, *23*(2), 486–507.

Watts, F. M., Zaimi, I., Kranz, D., Graulich, N., & Shultz, G. V. (2021). Investigating students' reasoning over time for case comparisons of acyl transfer reaction mechanisms. *Chemistry Education Research and Practice*, *22*(2), 364–381.

Weinrich, M. L., & Talanquer, V. (2016). Mapping students' modes of reasoning when thinking about chemical reactions used to make a desired product. *Chemistry Education Research and Practice*, *17*(2), 394–406.

Wilson, C. D., Haudek, K. C., Osborne, J. F., Buck Bracey, Z. E., Cheuk, T., Donovan, B. M., Stuhlsatz, M. A. M., Santiago, M. M., & Zhai, X. (2024). Using automated analysis to assess middle school students' competence with scientific argumentation. *Journal of Research in Science Teaching*, *61*(1), 38–69.

Winograd, B. A., Dood, A. J., Finkenstaedt-Quinn, S. A., Gere, A. R., & Shultz, G. V. (2021a). Automating characterization of peer review comments in chemistry courses. In C. E. Hmelo-Silver, B. de Wever, & J. Oshima (Eds.), *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning: CSCL 2021* (pp. 11–18). International Society of the Learning Sciences.

Winograd, B. A., Dood, A. J., Moon, A., Moeller, R., Shultz, G. V., & Gere, A. R. (2021b). Detecting high orders of cognitive complexity in students' reasoning in argumentative writing about ocean acidification. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 586–591). Association for Computing Machinery.

Wulff, P., Buschhüter, D., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2022). Bridging the gap between qualitative and quantitative assessment in science education research with machine learning — A case for pretrained language models-based clustering. *Journal of Science Education and Technology*, *31*(4), 490–513.

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2022). Utilizing a pretrained language model (BERT) to classify preservice physics teachers' written reflections. *International Journal of Artificial Intelligence in Education*, *33*(3), 439–466.

Wulff, P., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2023). Enhancing writing analytics in science education research with machine learning and natural language processing: Formative assessment of science and non-science preservice teachers' written reflections. *Frontiers in Education*, *7*(1061461), 1–18.

Yik, B. J., Dood, A. J., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2021). Development of a machine learning-based tool to evaluate correct Lewis acid-base model use in written responses to open-ended formative assessment items. *Chemistry Education Research and Practice*, *22*(4), 866–885.

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement*, *76*(2), 280–303.

Zhai, X., Haudek, K. C., & Ma, W. (2023). Assessing argumentation using machine learning and cognitive diagnostic modeling. *Research in Science Education*, *53*(2), 405–424.

Zhai, X., Haudek, K. C., Shi, L., Nehm, R. H., & Urban-Lurain, M. (2020a). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching*, *57*(9), 1430–1459.

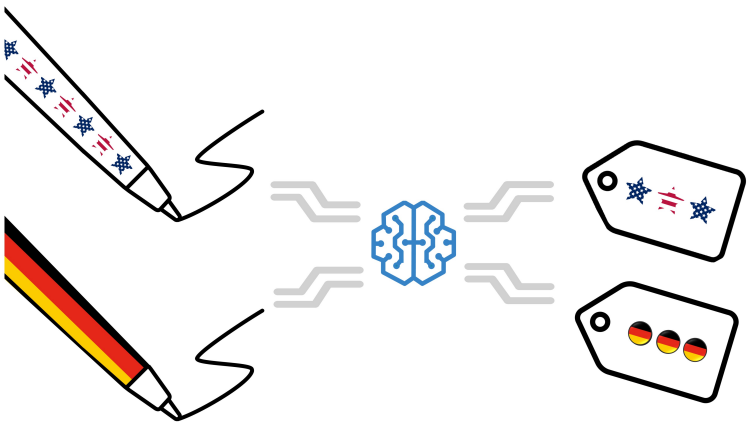
Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020b). Applying machine learning in science assessment: A systematic review. *Studies in Science Education*, *56*(1), 111–151.

Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education*, *39*(12), 1648–1668.

How to cite this article: Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2024). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*, *61*(8), 1757–1792. https://doi.org/10.1002/tea.21903

4 | Beyond language barriers: Allowing multiple languages in postsecondary chemistry classes through multilingual machine learning

Chapter 4 presents an original research paper. It is reproduced from Martin and Graulich (2024) under the terms of a Creative Commons Attribution 4.0 International License (CC BY 4.0) (https://creativecommons.org/licenses/by/4.0/). No changes have been made to the original article.



Martin, P. P., & Graulich, N. (2024). Beyond language barriers: Allowing multiple languages in postsecondary chemistry classes through multilingual machine learning. *Journal of Science Education and Technology*, *33*(3), 333-348. https://doi.org/10.1007/s10956-023-10087-4

Journal of Science Education and Technology (2024) 33:333–348
https://doi.org/10.1007/s10956-023-10087-4

Beyond Language Barriers: Allowing Multiple Languages in Postsecondary Chemistry Classes Through Multilingual Machine Learning

Paul P. Martin[1] · **Nicole Graulich**[1]

Accepted: 16 November 2023 / Published online: 14 February 2024
© The Author(s) 2024

Abstract
Students who learn the language of instruction as an additional language represent a heterogeneous group with varying linguistic and cultural backgrounds, contributing to classroom diversity. Because of the manifold challenges these students encounter while learning the language of instruction, additional barriers arise for them when engaging in chemistry classes. Adapting teaching practices to the language skills of these students, for instance, in formative assessments, is essential to promote equity and inclusivity in chemistry learning. For this reason, novel educational practices are needed to meet each student's unique set of language capabilities, irrespective of course size. In this study, we propose and validate several approaches to allow undergraduate chemistry students who are not yet fluent in the language of instruction to complete a formative assessment in their preferred language. A technically easy-to-implement option for instructors is to use translation tools to translate students' reasoning in any language into the instructor's language. Besides, instructors could also establish multilingual machine learning models capable of automatically analyzing students' reasoning regardless of the applied language. Herein, we evaluated both opportunities by comparing the reliability of three translation tools and determining the degree to which multilingual machine learning models can simultaneously assess written arguments in different languages. The findings illustrate opportunities to apply machine learning for analyzing students' reasoning in multiple languages, demonstrating the potential of such techniques in ensuring equal access for learners of the language of instruction.

Keywords Equity · Language inclusivity · Formative assessment · Machine learning · Natural language processing

Introduction

Linguistic diversity is increasing in science subjects like chemistry, so the language of instruction does not necessarily correspond to a student's first language. Students who learn the language of instruction as an additional language face unique challenges in language acquisition, science learning, and the intersection of both (del Rosario Basterra et al., 2011; Lee & Fradd, 1998). Therefore, further barriers arise for them to actively participate in chemistry classes (Deng & Flynn, 2023).

So far, much research has focused on students who learn English as an additional language (Eng+). These Eng+ students represent a heterogeneous group with diverse linguistic backgrounds, different prior opportunities in learning English, and varying experiences in communicating in another language (Deng & Flynn, 2023; Deng et al., 2022; Flores & Smith, 2013). Although these Eng+ students may possess strong content knowledge, they may lack the fluency to fully express their knowledge in English (Deng & Flynn, 2023; Deng et al., 2022; Flores & Smith, 2013; Lyon et al., 2012; Swanson et al., 2014). This language barrier can hinder effective communication, impede the exchange of ideas, and potentially exclude valuable contributions from Eng+ students (Deng & Flynn, 2023; Deng et al., 2022). Additionally, Eng+ students may experience linguistic insecurity, which involves feelings of anxiety about their English language usage (Deng & Flynn, 2023).

In educational settings such as chemistry classes, students taught in a different language than their first language face

✉ Nicole Graulich
Nicole.Graulich@didaktik.chemie.uni-giessen.de

Paul P. Martin
Paul.Martin@didaktik.chemie.uni-giessen.de

[1] Institute of Chemistry Education, Justus-Liebig-University, Heinrich-Buff-Ring 17, 35392 Giessen, Germany

various challenges in understanding, applying, and communicating domain-specific scientific concepts as well as in constructing evidence-based arguments (Curtis & Millar, 1988; Deng et al., 2022; Lee & Fradd, 1998). For instance, Eng+ students struggled to grasp the intended meaning of terms (Solano-Flores & Trumbull, 2003), especially when the scientific and everyday meanings of these terms diverge (Lee & Orgill, 2022). Moreover, undergraduate Eng+ students in general chemistry encountered both extrinsic and intrinsic challenges in engaging in lectures, comprehending laboratory procedures, and expressing their content knowledge (Lee et al., 2020). Additionally, postsecondary Eng+ students generated the fewest causal arguments in chemistry in a recent study (Deng et al., 2022). Beyond that, linguistic or cultural references in science assessments may not be accessible to Eng+ students (Luykx et al., 2007; Solano-Flores & Nelson-Barber, 2001; Solano-Flores & Trumbull, 2003).

Collectively, this indicates that students' performance in science assessments is linked to their language proficiency, irrespective of their capacity to comprehend and apply scientific principles (Afitska & Heaton, 2019; Curtis & Millar, 1988; Deng et al., 2022; Maerten-Rivera et al., 2010; Noble et al., 2014; Solano-Flores & Trumbull, 2003; Turkan & Liu, 2012). Language learners may face additional challenges because of unique cognitive operations associated with learning an additional language (Valdés & Figueroa, 1994). For this reason, Eng+ students might solve some assessment items better in their first language and others better in English (Solano-Flores & Trumbull, 2003). Assessment items represent, thus, a unique set of linguistic challenges depending on the language in which they are administered (Lee & Orgill, 2022; Solano-Flores & Trumbull, 2003). This may establish inequities when assessing students who learn the language of instruction as an additional language in chemistry and beyond.

So, educators need to ensure equitable means for these language learners (Lee, 2005; Wolf et al., 2008). On the one hand, educators could help them overcome their language barrier by continuously providing opportunities for language development (Amano et al., 2021). On the other hand, educators could design equitable instructional settings, among others, formative assessments, that allow for a more diverse educational landscape.

Proposing Approaches for Allowing Multiple Languages in Chemistry Classes

In their study, Deng et al. (2022) found that most but not all students preferred the language of their chemistry instruction for communicating on chemistry assessments, regardless of their first language. Some language learners might, therefore, benefit from communicating scientific ideas in the language of instruction, whereas others might enhance their reasoning about chemical phenomena by using their first language. Consequently, a time- and resource-efficient approach to increasing equity for the latter students could be providing them with the opportunity to complete exercises and formative assessments in their first language (cf., Afitska & Heaton, 2019; Buxton et al., 2014; Lee, 2005). However, since an instructor is not expected to be fluent in the respective language, a translation tool such as Google Translate, DeepL Translate, or ChatGPT could be used to translate back and forth between languages. Due to recent advancements in natural language processing (NLP) and machine learning (ML), these tools made significant progress in translating human language. In addition, learners of the language of instruction already utilize such tools to a great extent (Almusharraf & Bailey, 2023), which is why we quantitatively compared their performance. By leveraging these translation tools, educators could enable equal access to educational resources for language learners, which promotes inclusivity.

Nevertheless, allowing formative assessments in multiple languages is also possible by other means. Besides translation tools, educators could also design ML-based instructional settings, which have the potential to evaluate students' reasoning in every language automatically. While developing such ML models is resource-intensive, analyzing students' responses is fully automated once the model has been constructed. However, there are fundamental concerns that ML algorithms take a majority-driven language as the benchmark against which all other data is evaluated (Cheuk, 2021; Li et al., 2023). Consequently, a majority-focused linguistic practice may be overvalued (Ha et al., 2011; Liu et al., 2016), even if scientifically non-normative ideas are expressed (Nehm et al., 2012). In contrast, responses including uncommon vocabulary are more likely to be misclassified (Ha et al., 2011; Liu et al., 2016; Maestrales et al., 2021), which may penalize learners of the language of instruction (Cheuk, 2021). So, ML-based science assessments may favor a limited range of linguistic practices deemed suitable for scientific discourse, prioritizing academic language over students' informal everyday vocabulary (Cheuk, 2021). Therefore, establishing ML-based systems can unconsciously deepen structural forms of historically grown inequalities (Cheuk, 2021; Grimm, Steegh, Çolakoğlu et al., 2023; Grimm, Steegh, Kubsch et al., 2023; Li et al., 2023).

Following this, it is essential to identify and reduce the potential biases of ML algorithms. In doing so, Wilson et al. (2023) compared human, machine analytic, and machine holistic scoring regarding their scoring severity as well as their scoring gap between Eng+ and non-Eng+ students. They found that Eng+ students received on average lower human, machine holistic, and machine analytic scores than

non-Eng+ students. Moreover, machine holistic scores were on average lower than human and machine analytic scores across both linguistic groups; however, the performance gap between Eng+ and non-Eng+ students was greater for machine analytic scoring than for human and machine holistic scoring, specifically for the most difficult items. This finding indicates that machine analytic scoring can be biased when analyzing Eng+ students' responses, while machine holistic scoring may increase equity in evaluating Eng+ students' written arguments.

A multilingual approach to developing ML models for science assessments may also acknowledge students' linguistic diversity since such models can automatically assess students' reasoning across languages. For generating corresponding multilingual training datasets, instructors can either collect multilingual data or translate monolingual data into another language with a suitable translation tool so that students can express themselves beyond language barriers.

Utilizing deep learning techniques such as large language models and deep neural networks seems especially helpful when evaluating multilingual student responses since these techniques can handle complex data and achieve high accuracy. By applying these techniques, instructors have several opportunities: They can apply either a multilingual large language model for analyzing student responses across languages or monolingual models for each language separately. For this reason, the accuracy of different multi- and monolingual large language models is key to providing reliable assessments.

Methodological Considerations

Basics to Artificial Intelligence, Machine Learning, and Deep Learning

Due to recent technological progress, algorithms of artificial intelligence (AI) have taken over activities that have been, up to now, exclusively associated with human abilities. In this way, the term AI broadly describes software that automatically performs cognitive activities such as software-based planning, problem-solving, and decision-making (Bellmann, 1978; Haugeland, 1989). In education, AI can transform teaching and learning as evident in various articles (Kubsch et al., 2023; Martin & Graulich, 2023; Zhai, Haudek et al., 2020). For example, intelligent tutoring systems can automatically evaluate unique student challenges, provide immediate feedback, and deliver tailored exercises (Deeva et al., 2021). Accordingly, these systems can reduce the workload for educators while expediting assessment procedures (Urban-Lurain et al., 2013).

ML is, in turn, a subarea of AI that deals primarily with developing algorithms that learn from data to make predictions (Bishop, 2006; Mitchell, 1997; Mohri et al., 2012). Supervised ML algorithms are trained on human-labeled data so that the desired output, known as ground truth, is already included in the training data. The algorithm can then discern the underlying patterns and predict labels for new data. Over the past 15 years, supervised ML has been increasingly applied in science education research (e.g., Deeva et al., 2021; Gerard et al., 2015; Zhai, Yin et al., 2020), for example, to detect student reasoning in formative chemistry assessments (Martin & Graulich, 2023). In this way, supervised ML has contributed to significant advancements in automatically analyzing chemistry students' reasoning (e.g., Dood et al., 2018, 2020; Haudek et al., 2019; Maestrales et al., 2021; Noyes et al., 2020; Tansomboon et al., 2017; Vitale et al., 2016; Wilson et al., 2023; Yik et al., 2021, 2023).

Furthermore, deep learning is a specific ML technique that concentrates on training deep neural networks containing multiple layers of interconnected artificial neurons, inspired by the function of the human brain (Goodfellow et al., 2016). During network training, large amounts of data are fed into the network to adjust the parameters of the neuron connections. This technique has enabled significant advancements in various fields and remains an active research area (Mathew et al., 2021). Particularly, deep learning has gained prominence in educational studies due to its ability to accurately analyze unstructured data, as stored in images (e.g., Lee et al., 2023; Zhai et al., 2022) or written language (e.g., Dood et al., 2022; Gombert et al., 2023; Martin et al., 2023; Tschisgale et al., 2023; Watts et al., 2023; Winograd, Dood, Finkenstaedt-Quinn et al., 2021; Winograd, Dood, Moon et al., 2021; Wulff et al., 2023).

Basics to Natural Language Processing and Large Language Models

To efficiently analyze human language with ML, text data needs to be preprocessed by applying suitable NLP methods. NLP enables computers to analyze human language; so far, numerous techniques are available for facilitating human–computer interaction. Over the last couple of years, large language models emerged as a state-of-the-art technique in NLP. These models can perform various language-related tasks such as question-answering or text generation (Radford et al., 2019). Large language models have been trained on massive amounts of text data to capture complex language patterns and produce contextually relevant responses.

One such cutting-edge large language model is Bidirectional Encoder Representations from Transformers (BERT). BERT is pretrained on large corpora of unlabeled text data, such as Wikipedia, to understand semantic relationships between different words in a sentence, even if they are far apart (Devlin et al., 2018). Compared to formerly applied NLP methods, which often ignore word order (Angelov, 2020; Jurafsky & Martin, 2023), BERT can capture syntactic nuances of language, implicit meanings of phrases, and

the context in which words are used (Mikolov et al., 2013; Taher Pilehvar & Camacho-Collados, 2020). Hence, BERT has achieved state-of-the-art performance on a wide range of NLP benchmarks (Devlin et al., 2018) and has become a template for many subsequent large language models. Instructors can fine-tune BERT models for domain-specific purposes such as text classification in science assessment, which is called a downstream task (Ruder, 2019).

Research Questions

In this article, we propose and validate several ML-based approaches designed to allow students who have not yet mastered the language of instruction to complete formative assessments in their preferred language. However, as instructors are not expected to be fluent in the respective language, questions of reliability arise when analyzing students' reasoning across multiple languages. We approach the following research questions (RQs) by simulating the language transition between English and German as both authors are fluent in these languages. While doing so, we focused on undergraduate students' argumentation in organic chemistry about the plausibility of competing reactions. The research questions examined herein can inform instructors how to acknowledge multiple languages in the classroom.

1. Which translation tool produces German translations of English-written scientific arguments that deep learning architectures can analyze most accurately?
2. Which large language model achieves the highest level of reliability when analyzing German-written scientific arguments?
3. To what extent does the reliability of mono- and multilingual ML approaches to analyzing English- and German-written scientific arguments differ?
4. To what extent does an augmentation of the model's training data by combining translations of different translation tools impact the ML model's reliability?

Research Context

Setting of Data Collection

The data used in this second data analysis was gathered at a private research-intensive, liberal arts university in the North-eastern United States during April and May 2021. Sixty-four undergraduate organic chemistry II students voluntarily participated in the study, receiving extra credit for completing the exercises. The age of the participants ranged from 18 to 22 years. Among the participants, 34 identified as female, 29 as male, and one as non-binary. The students majored in various subjects, including biochemistry, chemistry, biology, and chemical engineering.

The tasks on building arguments about competing reactions in organic chemistry were implemented online using Qualtrics. The two task sets (Fig. 1) were presented on two different days, with a 3-week gap between them. The four alternative reaction products were sequentially displayed to the students. Students could build more than one argument per alternative reaction product. The original English-written arguments contain on average 27 words ($SD = 12.7$, $max = 138$, $min = 5$); the German translations comprise, in turn, on average 24 words ($SD = 11.6$, $max = 127$, $min = 4$). A total of 1108 arguments were collected. For

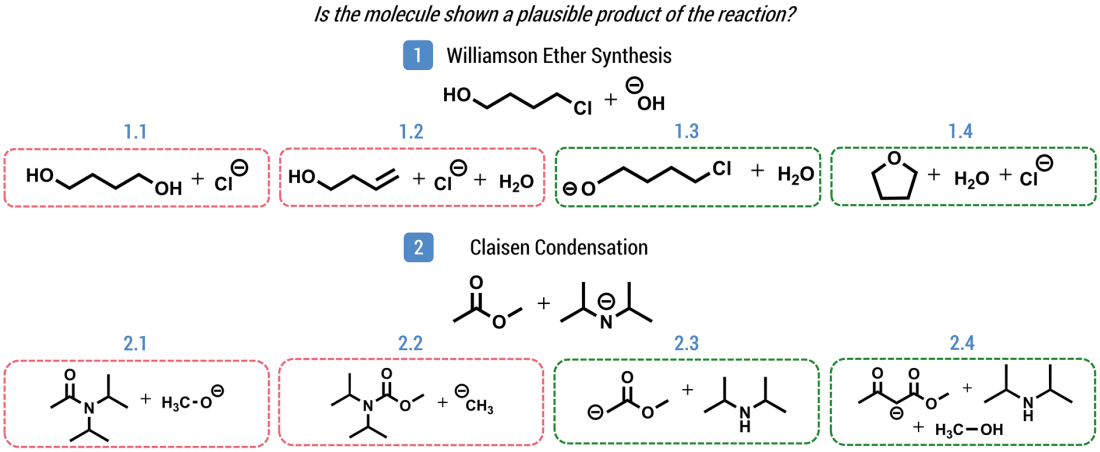


Fig. 1 Students judged the plausibility of four alternative reaction products for an intramolecular Williamson ether synthesis between 4-chlorobutan-1-ol and hydroxide (task 1) and a Claisen condensation between methyl acetate and diisopropylamide (task 2). Plausible products (1.3, 1.4, 2.3, 2.4) are highlighted in green, implausible products (1.1, 1.2, 2.1, 2.2) are highlighted in red

more comprehensive information about the study setting, see Lieber et al. (2022a).

Research Instrument

To advance students' argumentation in organic chemistry, Lieber et al. (2022a, b) developed an adaptive instructional setting, where students were prompted to judge the plausibility of competing chemical reactions for intramolecular Williamson ether synthesis and Claisen condensation (Fig. 1). In organic chemistry, competing reactions arise when reactants have the potential to undergo various reaction pathways, resulting in more or less plausible reaction products. Engaging in discussions about these alternative reaction products requires integrating multiple chemical concepts, which must be weighed to build evidence-based arguments as well as counterarguments (Lieber & Graulich 2020, 2022; Lieber et al., 2022a, b; Watts et al., 2022). In Lieber et al.'s (2022a, b) setting, students were prompted to make a claim about whether the displayed reaction product is plausible, provide evidence by using chemical concepts to support their claim, and establish a logical connection between claim and evidence by considering electronic, steric, or energetic effects (Fig. 2). Through this task design, Lieber et al. (2022a, b) demonstrated that adaptive scaffolding significantly enhanced students' argumentation.

Scoring Rubric

We developed a two-dimensional holistic rubric (Fig. 3) for evaluating students' arguments with ML drawing from the *levels of granularity* (Bodé et al., 2019; Deng & Flynn, 2021; Deng et al., 2023; Soo, 2019) and the *modes of reasoning* (Sevian & Talanquer, 2014). The levels of granularity refer to the grain size at which phenomena are explained. In general, diverse tasks demand varying levels of granularity to adequately reason about the underlying processes (Darden, 2002). In our study, we applied four levels of granularity, namely, *structural*, *energetic*, *phenomenological*, and *electronic*, which have been used in studies by Deng and Flynn (2021) and Deng et al. (2023). Besides, Sevian and Talanquer's (2014) modes of reasoning, namely, *descriptive*, *relational*, *linear causal*, and *multicomponent causal*, comprise the second dimension of our rubric. These modes of reasoning characterize the sophistication level demonstrated by students in terms of their ability to establish connections between concepts and to provide well-justified explanations for why phenomena occur (Russ et al., 2008; Sevian & Talanquer, 2014). The modes of reasoning imply that evaluating students' understanding of scientific principles should not be limited to assessing their content knowledge alone; it should also involve examining how they integrate new information into their existing cognitive network (Sevian & Talanquer, 2014). In sum, our rubric helped

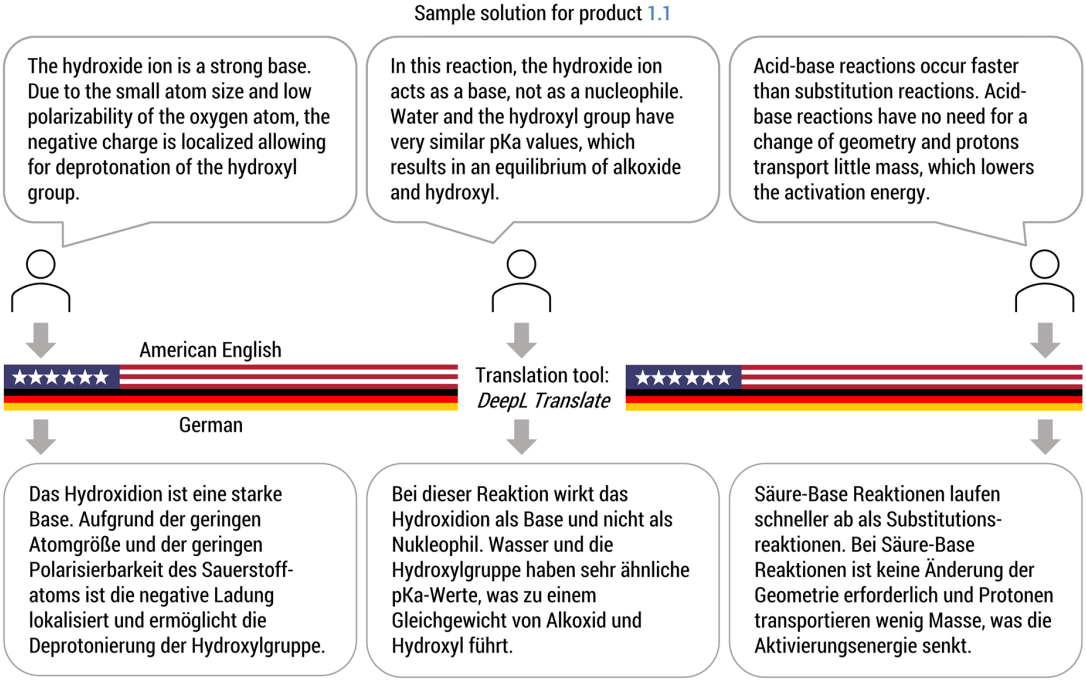


Fig. 2 Sample solutions and their translations for the implausibility of product 1.1 (Fig. 1). For more detailed sample solutions, see Lieber et al. (2022a)

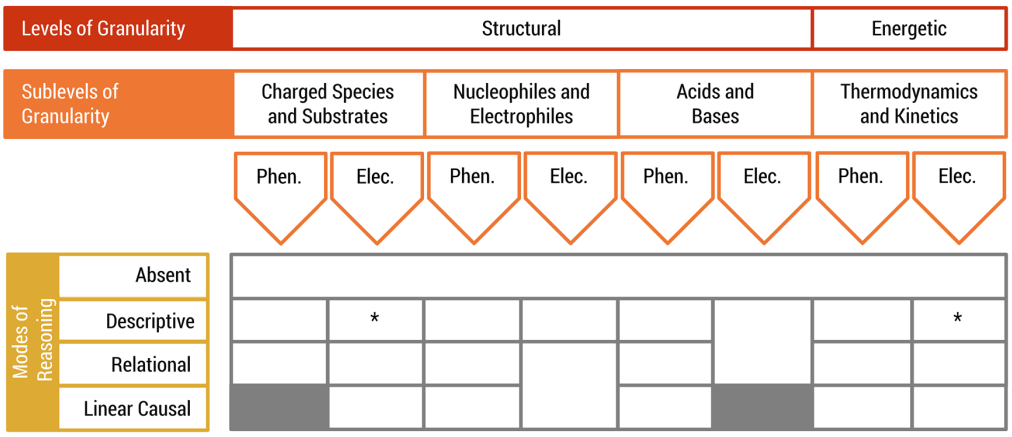


Fig. 3 Two-dimensional holistic rubric for classifying students' chemistry-related arguments into one of twenty categories. *Note*: The asterisk shows that categories belong together. Phen. = *phenomenological*, Elec. = *electronic*

us examine the concepts and relationships, the level of causality, as well as the grain size that students addressed in their argumentation. A more in-depth description of the rubric development process is published in Martin et al. (2023).

Methods

We used the PyTorch deep learning framework (Paszke et al., 2019) implemented in Python to examine the RQs. We split our data into a training, validation, and test set with a ratio of 65:15:20. The training set was used to train a deep neural network, the validation set was employed to determine the optimal hyperparameter configuration, and the test set was utilized to check the model accuracy based on four metrics (cf., Table 1). We adjusted the number of epochs, the learning rate, and the batch size as hyperparameters. Epochs indicate the number of cycles the model is trained on. Generally, training the model on more epochs enhances the model performance on the training data, but excessive training can lead to poor performance on new data. As we determined that the number of epochs greatly impacts model performance (Martin et al., 2023), we varied this number between 1 and 100. The learning rate, in turn, refers to the rate at which an optimizer updates the parameters of the model during training. Consequently, the learning rate impacts how quickly the model adapts to a specific context. Here, we tested learning rates of 1e−6, 5e−6, 1e−5, 5e−5, and 1e−4. Last, the batch size controls how many concurrent training examples are processed together in one network pass. We tried a batch size of 2, 4, 8, 16, and 32. In sum, a total of 2500 hyperparameter configurations were tested for each model.

Translation Tools

To identify the translation tool producing translations that ML techniques can analyze most accurately, we tested how reliable our deep learning architecture can classify translations generated by Google Translate (Google LLC, 2006), DeepL Translate (DeepL SE, 2017), and ChatGPT (Open-AI, 2022). In other words, human raters did not score the level of accuracy of translations generated by different tools; instead, we investigated how accurate various large language models evaluate the German translations of students' English-written scientific argumentation. Therefore, the German translations of the English-written arguments were, for each tool separately, used to train and test our deep learning architecture. When using the translation tools, we did not adjust the output; the translations were fed unmodified into the deep learning architecture. For the ChatGPT analysis,

Table 1 Metrics to measure machine-human score agreements

Accuracy	Cohen's κ	Weighted F_1-score	Macro F_1-score
Percentage of agreement between two raters	Agreement between two raters beyond chance agreement	Agreement metrics for imbalanced datasets accounting for the number and type of prediction errors as the harmonic mean of precision and recall	
		Weighted mean of all per-class F_1-scores	Arithmetic, i.e., unweighted mean of all per-class F_1-scores

the free accounts of the first author and a research assistant were used to access the GPT-3.5 Feb 13 version (Brown et al., 2020; OpenAI, 2023). In ChatGPT, arguments were translated in a single chat using the prompt "Please, translate the following sentence in German."

We employed the monolingual large language model *GottBert-cased* as well as the multilingual model *xlm-RoBERTa-base* to compare the accuracy of the different translation tools. We used both mono- and multilingual large language models for the analysis to validate the generalizability and robustness of our findings. As hyperparameters, we varied the number of epochs, the learning rate, and the batch size as described above.

Large Language Models for Analyzing German-Written Arguments

We leveraged three German-specific, monolingual large language models *BERT-base-German-cased* (Chan et al., 2019), *dbmdz/BERT-base-German-cased* (MDZ Digital Library team, 2020), and *GottBERT-cased* (Scheible et al., 2020) as well as three multilingual models *BERT-base-multilingual-cased* (Devlin et al., 2018), *xlm-RoBERTa-base* (Conneau et al., 2019), and *xlm-RoBERTa-large* (Conneau et al., 2019) to identify the best performing one for analyzing German-written arguments. We used *cased* large language models, which are models that retain the distinction between uppercase and lowercase letters, to preserve the case information provided in the German language. Since students' arguments were originally written in English, we utilized DeepL Translate to gather German translations. Again, hyperparameters were varied as described above.

Mono- and Multilingual ML Approaches

If mono- and multilingual large language models would show similar accuracy in analyzing students' argumentation, educators could use a single multilingual model for analyzing students' reasoning across languages. Hence, we compared the accuracy of the English-specific deep learning architecture reported by Martin et al. (2023) with the best-performing German-specific architecture created for answering RQ1 and RQ2, and multilingual architectures that simultaneously analyzed students' argumentation in both languages. We used three multilingual large language models *BERT-base-multilingual-cased* (Devlin et al., 2018), *xlm-RoBERTa-base* (Conneau et al., 2019), and *xlm-RoBERTa-large* (Conneau et al., 2019) with varying epochs, learning rates, and batch sizes to identify the best-performing one. To be noted, the monolingual models are built on 1108 student-written arguments, while the multilingual ones are built on twice the amount of data, which are 1108 English-written and 1108 German-written arguments. For the multilingual approach, we ensured that an original English-written argument and its translation are either both included in the test set or not included in the test set (see the "Text Data Augmentation" section).

Text Data Augmentation

In general, the performance of ML and NLP largely depends on the quantity and quality of the training data, so training a generalizable model becomes challenging with limited data. As we hypothesized that the English-specific model can evaluate students' argumentation more accurately than the German-specific one, we looked for time-efficient ways to increase the scoring accuracy of the latter. Therefore, we tripled our German-written dataset by combining the translations of Google Translate, DeepL Translate, and ChatGPT. Since the output of the three translation tools slightly differs, combining the translations might increase model accuracy.

This process of creating additional data for ML model training is called data augmentation, which is a technique for increasing the size and diversity of the data by applying various modifications. Techniques for text data augmentation involve adding synonyms, inserting or replacing words, changing word order, altering sentence structures, or applying other linguistic transformations. Text data augmentation often helps expand the sample size, increase data heterogeneity, and boost model performance (e.g., Bayer et al., 2022; Feng et al., 2021; Shorten et al., 2021; Wei & Zou, 2019).

However, data augmentation involves the risk of overfitting the data, which means that the model becomes overly specialized. As a result, the model performs exceptionally well on the training data but fails to make accurate classifications on new data. Following that, incorporating the translations of the same sentence in both the training and test set would distort model performance as it is not measured how the model classifies new data. Accordingly, we paid attention that the translations of the same sentence are either all in the test set or not in the test set. For evaluating performance changes, we used the large language model *GottBERT-cased* while adjusting the hyperparameters as mentioned above.

Results and Discussion

RQ 1: Comparing the Performance of Translation Tools for English to German Translations

To determine the translation tool that produces translations most accurately analyzable by our deep learning

architecture, we used the validation set to identify the optimal hyperparameter configuration. Here, the machine-human score agreements varied tremendously depending on the number of epochs, the learning rate, and the batch size. So, we compared the performance of the deep learning architecture across all hyperparameter configurations, rather than for a predetermined set of those. After that, we evaluated the performance of the deep learning architecture analyzing German translations of English-written scientific arguments based on the test set to identify the best translation tool in our context. We hypothesized that ChatGPT would translate English-written arguments best into German due to the recent advancements of the GPT models in analyzing natural language.

Our deep learning architecture performed similarly across the three translation tools, with only minor variations in machine-human score agreements (Table 2). As indicated by a Cohen's κ value above 0.80 (Landis & Koch, 1977), the architecture achieved *almost perfect* machine-human score agreements when using Google Translate or DeepL Translate as a translation tool. This trend was further supported by accuracy, with Google Translate and DeepL Translate exhibiting higher values than ChatGPT. From a qualitative point of view, Google Translate and DeepL Translate handled complex sentence structures and idiomatic expressions best, resulting in more contextually appropriate translations. By contrast, ChatGPT achieved slightly lower levels of accuracy and Cohen's κ because it occasionally added or paraphrased sentences (cf., Bang et al., 2023), which led to modifying some phrases relevant to coding. Our initial hypothesis that ChatGPT would produce the most accurate translations was, thus, not confirmed. Nonetheless, especially macro F_1-score, which provides a more balanced performance evaluation across all twenty categories, indicates that ChatGPT is also a good translation tool with nearly comparable performance metrics.

Despite these valuable insights, it should be emphasized that the performance of such tools can vary depending on the specific context of the text and the language pair. In a domain-general study, Jiao et al. (2023) also compared the translation capabilities of Google Translate, DeepL Translate, and ChatGPT showing that Google Translate performed best while ChatGPT performed the least effectively. Particularly, ChatGPT's translation quality was slightly dependent on the applied prompt, with the most accurate translations obtained by asking: "Please provide the [target language] translation for these sentences." Furthermore, ChatGPT achieved almost comparable performance to other tools on high-resource language pairs, i.e., language pairs where millions of sentences are available as databases, such as German and English (Jiao et al., 2023). However, the performance gap between ChatGPT and other tools widened when dealing with low-resource language pairs like Romanian and English and when translating between languages of different language families like Chinese and English, which was confirmed by Bang et al. (2023) based on further language pairs. Moreover, the translation quality depended on the domain, where ChatGPT performed worst when translating data from biomedical abstracts or an online forum but outperformed Google Translate and DeepL Translate when translating common voice data (Jiao et al., 2023). Interestingly, the performance gap between ChatGPT and other tools got smaller when using GPT-4 instead of GPT-3.5 (Jiao et al., 2023) or when prompting ChatGPT to postedit its translations (Bang et al., 2023).

Taken together, our analysis shows that deep learning architectures can reliably classify unmodified German translations of English-written scientific arguments. This high level of machine-human score agreements can be explained, among others, by the improved performance of translation tools (Conneau & Lample, 2019), which, in turn, can be attributed to different reasons. On the one hand, improved attention mechanisms enhanced translation quality (Conneau & Lample, 2019; Vaswani et al., 2017) since these mechanisms allow translation tools to capture contextual information. On the other hand, the availability of large amounts of multilingual training data increased translation accuracy since more training data helps large language models learn more nuances of different languages. Nonetheless, comparing different translation tools is important to identify the

Table 2 Comparison of the performance of Google Translate, DeepL Translate, and ChatGPT for English to German translations

Tool	Epochs	Learning rate	Batch size	Accuracy	Cohen's κ	Weighted F_1-score	Macro F_1-score
Monolingual model: *GottBERT-cased*							
Google Translate	62	5e−6	2	0.8239	0.8114	0.8229	0.7877
DeepL Translate	16	5e−5	32	0.8194	0.8067	0.8228	0.7845
ChatGPT	45	5e−5	32	0.8062	0.7924	0.8062	0.7880
Multilingual model: *xlm-RoBERTa-base*							
Google Translate	91	5e−6	32	0.8106	0.7966	0.8050	0.7740
DeepL Translate	97	1e−5	2	0.8150	0.8015	0.8140	0.7759
ChatGPT	72	1e−5	2	0.8062	0.7920	0.8055	0.7753

most suitable one for domain-specific translation needs. Specifically, when translating students' chemistry-related arguments from English to German, our deep learning architecture analyzed translations from Google Translate and DeepL Translate most accurately. These insights are valuable for optimal translation solutions in chemistry classes.

RQ 2: Evaluating the Performance of Large Language Models in Analyzing German Translations of English-Written Scientific Arguments

Comparable to the applied translation tools, different large language models excel in different areas. Accordingly, we compared the machine-human score agreements of six mono- or multilingual large language models when processing German translations of English-written scientific arguments to identify the most reliable one. Again, we used the validation set to identify the best configuration of hyperparameters for each model. Because of the tremendous variation of the model performance depending on the hyperparameter settings, we compared the large language models based on their highest machine-human score agreements. Subsequently, we assessed their performance by using the test set.

Generally, monolingual large language models are developed to analyze text in a single language, which means that they have mostly a more fine-grained understanding of that language. In contrast, multilingual models can analyze text simultaneously in various languages, but often with lower accuracy in a single language. We are aware that we did not leverage the full potential of the multilingual large language models in the context of this RQ since we solely used them to analyze German-written arguments, i.e., arguments written in a single language. Nevertheless, comparing the accuracy of mono- and multilingual large language models in analyzing a single language helps determine their performance differences. We hypothesized that the German-specific, monolingual models would significantly outperform the multilingual ones.

The performance of each model can be found in Table 3. The monolingual large language model *GottBERT-cased* achieved the highest machine-human score agreements across all four metrics when analyzing the German translations of our sample. Surprisingly, the multilingual large language models *xlm-RoBERTa-base* and *xlm-RoBERTa-large* performed nearly as well as *GottBERT-cased*, highlighting the potential of multilingual models in analyzing students' scientific argumentation. Moreover, the large language model *dbmdz/BERT-base-German-cased* demonstrated slightly lower performance in accuracy, Cohen's κ, and weighted F_1-score than the three aforementioned models. Conversely, this model outperformed *xlm-RoBERTa-base* in terms of macro F_1-score, highlighting its capability to evaluate all categories equally. Finally, *BERT-base-German-cased* ranked as the least effective monolingual model across all metrics, while *BERT-base-multilingual-cased* performed the least among the multilingual models.

The observed variations in performance can be ascribed to the differences in model sizes, particularly influenced by the volume of data available in the German language (Table 4). Specifically, *GottBERT-cased* demonstrated the best performance because it is trained on the largest dataset in the German language. Furthermore, the multilingual large language models *xlm-RoBERTa-large* and *xlm-RoBERTa-base*, trained on the second highest amount of German language data, performed second best and showed, thus, improved performance compared to the other monolingual large language models. In detail, *xlm-RoBERTa-large* slightly outperformed *xlm-RoBERTa-base* due to an increased number of layers and hidden states in model architecture (Table 4). Ultimately, the decline in performance from *dbmdz/BERT-base-German-cased* to *BERT-base-German-cased* to *BERT-base-multilingual-cased* regarding accuracy, Cohen's κ, and weighted F_1-score also aligns with the amount of German language data these models are trained on.

In sum, a thorough model comparison allowed us to make informed decisions about the most suitable large language model for analyzing German translations of English-written

Table 3 Comparison of the performance of different large language models in analyzing German translations of English-written scientific arguments

Large language model	Epochs	Learning rate	Batch size	Accuracy	Cohen's κ	Weighted F_1-score	Macro F_1-score
Monolingual models: German-specific							
BERT-base-German-cased	52	5e−6	4	0.7885	0.7736	0.7884	0.7428
dbmdz/BERT-base-German-cased	40	1e−5	2	0.8018	0.7876	0.8022	0.7748
GottBERT-cased	16	5e−5	32	0.8194	0.8067	0.8228	0.7845
Multilingual models							
BERT-base-multilingual-cased	46	1e−5	16	0.7841	0.7686	0.7844	0.7462
xlm-RoBERTa-base	97	1e−5	2	0.8150	0.8015	0.8140	0.7689
xlm-RoBERTa-large	97	1e−5	2	0.8194	0.8061	0.8170	0.7766

Table 4 Details on model sizes: *#lgs* = number of included languages, *size: eng* = data size in the English language, *size: ger* = data size in the German language, L = number of layers, H = number of hidden states, V = vocabulary size, P = total number of parameters

Large language model	#lgs	Size: eng (in GB)	Size: ger (in GB)	L	H	V	P
Monolingual model: English-specific							
BERT-large-uncased	1	16.0	0	24	1024	30k	336M
Monolingual models: German-specific							
BERT-base-German-cased	1	0	12.0	12	768	31k	110M
dbmdz/BERT-base-German-cased	1	0	16.0	12	768	31k	111M
GottBERT-cased	1	0	145.0	12	768	52k	127M
Multilingual models							
BERT-base-multilingual-cased	104	15.5	2.1	12	768	110k	119M
xlm-RoBERTa-base	100	323.4	71.6	12	768	250k	279M
xlm-RoBERTa-large	100	323.4	71.6	24	1024	250k	561M

scientific arguments. Particularly, we found in non-conformity with our initial hypothesis that the monolingual large language models did not significantly outperform the multilingual ones as *xlm-RoBERTa-base* and *xlm-RoBERTa-large* also achieved high levels of machine-human score agreements. This finding highlights the promising abilities of both mono- and multilingual large language models in automating the analysis of students' scientific argumentation.

RQ 3: Comparing the Performance of Mono- and Multilingual ML Approaches in Analyzing Scientific Argumentation

As evidenced from the previous findings, multilingual large language models have the potential to accurately analyze scientific arguments in different languages. Following this, we compared the accuracy of the English-specific model reported by Martin et al. (2023), the best-performing German-specific model identified in the prior sections, as well as multilingual models that simultaneously analyzed English- and German-written arguments. The analysis steps are similar to the previous two sections. We hypothesized that the accuracy of the multilingual models would correspond approximately to the accuracy of the German-specific counterpart.

First, the English-specific model performed much better than the German-specific and multilingual ones (Table 5). On the one hand, this result was expected as the BERT models were initially developed to analyze the English language (Devlin et al., 2018). However, the details on the model sizes do not make explaining this result straightforward. For example, based on the size of the training data and vocabulary (Table 4), one could hypothesize that *GottBERT-cased*, *xlm-RoBERTa-base*, or *xlm-RoBERTa-large* perform best, whereas the total number of parameters might suggest that *xlm-RoBERTa-large* performs at least as well as *BERT-large-uncased*. We hypothesize that the remarkable performance of *BERT-large-uncased* is a consequence of different interrelated factors: In addition to the original intention of the BERT models to analyze the English language, the high number of layers, hidden states, and parameters (Table 4) likely contributes to improved performance. Furthermore, the limited accuracy of the translation tools may have led to minor modifications in certain phrases relevant to coding, potentially complicating the classification of the German translations by the deep learning architecture. Consequently, the accuracy of the English-specific approach to analyzing written argumentation is more reliable than the German-specific and multilingual ones.

Besides, the German-specific model *GottBERT-cased* slightly outperformed its multilingual counterparts, which is not surprising given that *GottBERT-cased* is trained on the most extensive dataset of the German language

Table 5 Comparison of the performance of mono- and multilingual ML approaches in analyzing scientific argumentation

Large language model	Epochs	Learning rate	Batch size	Accuracy	Cohen's κ	Weighted F_1-score	Macro F_1-score
Monolingual model: English-specific							
BERT-large-uncased	38	1e−5	8	0.8711	0.8610	0.8640	0.7979
Monolingual model: German-specific							
GottBERT-cased	62	5e−6	2	0.8239	0.8114	0.8229	0.7877
Multilingual models							
BERT-base-multilingual-cased	36	1e−5	8	0.8150	0.8016	0.8153	0.7772
xlm-RoBERTa-base	53	1e−5	4	0.8018	0.7866	0.7921	0.7874
xlm-RoBERTa-large	15	5e−6	4	0.8084	0.7942	0.8024	0.7685

(Table 4). In essence, since the multilingual models also analyze German language data but are trained on a smaller sample of such data, a slight decrease in performance could be anticipated. Nonetheless, *BERT-base-multilingual-cased* also achieved high machine-human score agreements in terms of accuracy, Cohen's κ, and weighted F_1-score and performed—in alignment with our hypothesis—about as well as the German-specific model (Table 5). Additionally, the macro F_1-score of *xlm-RoBERTa-base* is nearly equivalent to that of *GottBERT-cased*.

Together, the high reliability of the multilingual large language models observed again has various reasons. On the one hand, multilingual large language models are pretrained on large text corpora of different languages, extending their accuracy and generalizability across languages. On the other hand, we used twice the amount of training data to fine-tune the multilingual models since the English- and German-written arguments were combined for model training. Here, our chemistry-related arguments contain many domain-specific terms that share common linguistic features across English and German (Fig. 4). Because of the alignment of some of these terms, the multilingual models pretrained on both English and German text corpora analyzed students' scientific arguments almost as well as the monolingual models solely pretrained on English or German corpora. In other words, the analysis of the multilingual arguments benefitted from the larger training corpora of the multilingual models due to a linguistic overlap in chemistry-related terminology between both languages.

The remarkable level of accuracy shows that multilingual large language models can help bridge language gaps and improve accessibility to formative chemistry assessments for learners of the language of instruction. Especially, instructors can employ multilingual models in their courses to offer and reliably score assessments in more than one language. Students can then choose the language in which they want to complete an assessment, which may foster a sense of inclusivity and create avenues for a more equitable and enriched educational landscape. Additionally, the automated text analysis of the large language models potentially enables implementing adaptive learning, including personalized guidance, feedback, and learner support (Plass & Pawar, 2020), or Just-in-Time Teaching (Novak et al., 1999) in multiple languages. As these models continue to advance, they will further ease communicating across different languages and transform chemistry learning in multilingual contexts.

RQ 4: Investigating the Impact of Training Data Augmentation on ML Model Performance

As identified in the previous section, the accuracy of the German-specific model was almost 5% less than the accuracy of the English-specific one (Table 5). Therefore, we looked for easily implementable ways to increase the accuracy of the German-specific model. In doing so, we augmented our training data by combining the translations of the three applied tools to not only improve model performance but to also ensure that the model's accuracy is upheld across linguistic contexts. This technique for text data augmentation is inspired by Sennrich et al. (2015), who translated text into a different language and subsequently translated it back to the original language. Due to randomness in the translation process, their augmented text differed from the original text while conceptual consistency was preserved. We hypothesized that text data augmentation would contribute to a significant increase in machine-human score agreements.

Combining the translations of Google Translate, DeepL Translate, and ChatGPT slightly improved model performance from accuracy = 0.8239, Cohen's κ = 0.8114, weighted F_1-score = 0.8229, macro F_1-score = 0.7877 to accuracy = 0.8267, Cohen's κ = 0.8145, weighted F_1-score = 0.8267, and macro F_1-score = 0.7954. Hence, combining translations of different tools can be a first step in ensuring that an ML model maintains its accuracy across linguistic contexts, which confirmed our initial hypothesis to some extent. Nevertheless, the performance of the English-specific model was still much better than that of the augmented German-specific one (Table 5). Future work can, for example, use ChatGPT to rephrase text, generating multiple conceptually similar but semantically different samples (Dai et al., 2023).

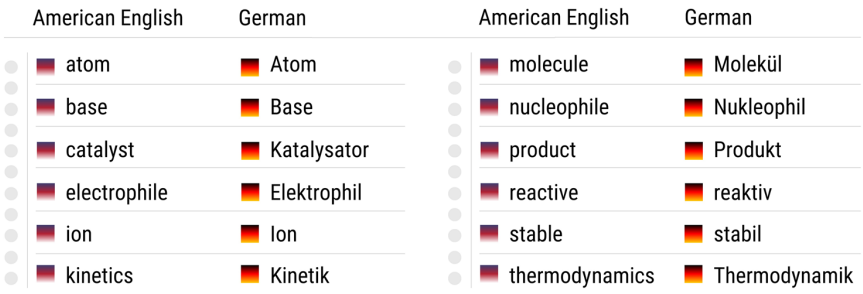


Fig. 4 Examples of chemistry-related terms that share linguistic similarities in English and German

Conclusions and Implications

In this study, we analyzed how one can reliably evaluate students' reasoning across multiple languages. We found that, when translating students' reasoning from English to German, Google Translate and DeepL Translate were most applicable. Furthermore, multilingual large language models showed great potential in automatically analyzing students' reasoning across languages. Specifically, we determined that both the monolingual model *GottBERT-cased* and the multilingual models *xlm-RoBERTa-base* and *xlm-RoBERTa-large* achieved *almost perfect* machine-human score agreements when analyzing the German translations of the English-written arguments. This finding was confirmed by the high agreement metrics of the multilingual models when simultaneously analyzing English- and German-written arguments. Therefore, instructors could use these multilingual models to allow learners to complete an assessment in their preferred language, creating chances for a more equitable and enriched educational setting for all learners. In the future, multilingual models may play a major role in facilitating communication across various languages, empowering students to engage in learning opportunities in their preferred language.

However, enough sample data in a certain language is a prerequisite for model training. Translating back and forth between languages is a way to gather data for automatically analyzing students' reasoning regardless of the language. We observed that combining translations of different tools slightly improved model accuracy, but a significant performance gap between the English-specific and German-specific models was still prevalent.

This analysis reflects the current level of technology; however, translation tools and large language models are continually improving with the ongoing advancements in ML. Following that, future research could investigate how the reliability of translation tools and large language models improves with technological progress. In addition, future research could also determine to what extent multilingual large language models can assess students' reasoning when switching between languages within a sentence, which is called translanguaging. Translanguaging allows individuals to move flexibly between languages while using different languages for different purposes. In educational settings, translanguaging has gained attention as it values students' multilingual abilities to support their learning, rather than suppressing or restricting their language use (Grapin et al., 2023; Jakobsson et al., 2021; Ryu, 2019). By employing multilingual models, students can alter between languages within a response, which may encourage them to use their full linguistic repertoire when communicating their ideas. Appreciating students' multilingual language proficiency may help account for the fluidity and interconnectedness of different languages in science contexts and encourage students to reason beyond language barriers.

Limitations

We composed a German dataset by translating English-written scientific arguments into German. Translating back and forth between languages does, however, not necessarily carry the broader social or cultural background of students' reasoning. So, we may have not fully captured the language characteristics representative of German students, which means that the data used for model training may not adequately mirror a diverse population of German students. Additionally, the original English-written arguments came from a rather homogeneous demographic group, which may constrain the capability of our model to generate precise classifications across diverse multilingual populations and institutions. To address this constraint, additional data from multilingual students needs to be collected. By using more training data, we may also mitigate some biases that we have not identified yet (cf., Noyes et al., 2020), leading to increased confidence in the outcomes of our model. In addition, an analysis of the internal workings of our algorithm, as performed by Martin et al. (2023) for the original English-written arguments, could also be applied to the German-specific and multilingual deep learning architectures reported in this article.

Moreover, only two reactions, a Williamson ether synthesis and a Claisen condensation, were evaluated in this study. The results of the analysis could possibly differ if the students were prompted to judge the plausibility of alternative reaction products for additional mechanisms. Beyond that, the arguments were constructed within a traditional organic chemistry curriculum. Curriculums such as "Mechanisms before Reactions" (Flynn & Ogilvie, 2015) or "Organic Chemistry, Life, the Universe and Everything" (Cooper et al., 2019) place more emphasis on explanations and argumentation, which may affect the dataset, the model training process, and yield different machine-human score agreements.

Besides, we only investigated the language transition between English and German, i.e., between Western languages that share certain linguistic characteristics and historical roots. For instance, both languages utilize the same alphabet and have been influenced by Latin, French, and other Roman languages over time; so, English and German show similarities in vocabulary, syntax, and grammatical structures. Our findings may probably not be generalizable across non-Western languages. Accordingly, future research should investigate how translating between other languages impacts the machine-human score agreements of different mono- and multilingual approaches.

Acknowledgements This publication is part of the first author's doctoral thesis (Dr. rer. nat.) at the Faculty of Biology and Chemistry, Justus-Liebig-University Giessen, Germany. We thank Peter Wulff and David Kranz for their help in implementing the machine learning analysis. Moreover, we thank Leonie Lieber, Ira Caspari-Gnann, and Krenare Ibraj for their pioneering research on students' argumentation in organic chemistry as well as for sharing their research data. Finally, we thank Felix Blödtner for evaluating and translating students' arguments and all members of the Graulich group for fruitful discussions.

Author Contribution PPM: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, and writing original draft. NG: conceptualization, investigation, project administration, resources, supervision, writing, review, and editing.

Funding Open Access funding enabled and organized by Projekt DEAL. This work is financially supported by the "Verband der Chemischen Industrie" (German Chemical Industry Association).

Data Availability The data used in this study is available from the corresponding author upon reasonable request.

Declarations

Ethical Approval All data collection procedures received Institutional Review Board approval for human subject research (STUDY00001480).

Consent to Participate Informed consent was obtained from all participants.

Conflict of Interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Afitska, O., & Heaton, T. J. (2019). Mitigating the effect of language in the assessment of science: A study of English-language learners in primary classrooms in the United Kingdom. *Science Education, 103*(6), 1396–1422. https://doi.org/10.1002/sce.21545

Almusharraf, A., & Bailey, D. (2023). Machine translation in language acquisition: A study on EFL students' perceptions and practices in Saudi Arabia and South Korea. *Journal of Computer Assisted Learning, 39*(6), 1988–2003. https://doi.org/10.1111/jcal.12857

Amano, T., Rios Rojas, C., Boum Ii, Y., Calvo, M., & Misra, B. B. (2021). Ten tips for overcoming language barriers in science. *Nature Human Behaviour, 5*(9), 1119–1122. https://doi.org/10.1038/s41562-021-01137-1

Angelov, D. (2020). Top2Vec: Distributed representations of topics. *arXiv preprint*. arXiv:2008.09470. https://doi.org/10.48550/arXiv.2008.09470

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., & Chung, W. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint.* arXiv:2302.04023. https://doi.org/10.48550/arXiv.2302.04023

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2022). A survey on data augmentation for text classification. *ACM Computing Surveys, 55*(7), 1–39. https://doi.org/10.1145/3544558

Bellmann, R. (1978). *An introduction to artificial intelligence: Can computers think?* Boyd and Fraser.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bodé, N. E., Deng, J. M., & Flynn, A. B. (2019). Getting past the rules and to the WHY: Causal mechanistic arguments when judging the plausibility of organic reaction mechanisms. *Journal of Chemical Education, 96*(6), 1068–1082. https://doi.org/10.1021/acs.jchemed.8b00719

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., et al. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (33rd ed., pp. 1877–1901). Curran Associates Inc.

Buxton, C., Allexsaht-Snider, M., Aghasaleh, R., Kayumova, S., Kim, S., Choi, Y.-J., & Cohen, A. (2014). Potential benefits of bilingual constructed response science assessments for understanding bilingual learners' emergent use of language of scientific investigation practices. *Double Helix, 2*(1), 1–21. https://doi.org/10.37514/DBH-J.2014.2.1.05

Chan, B., Möller, T., Pietsch, M., & Soni, T. (2019). *German BERT*. Hugging Face. Retrieved September 21, 2023, from https://huggingface.co/bert-base-german-cased

Cheuk, T. (2021). Can AI be racist? Color-evasiveness in the application of machine learning to science assessments. *Science Education, 105*(5), 825–836. https://doi.org/10.1002/sce.21671

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (32nd ed., pp. 7057–7067). Curran Associates Inc.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint*. arXiv:1911.02116. https://doi.org/10.48550/arXiv.1911.02116

Cooper, M. M., Stowe, R. L., Crandell, O. M., & Klymkowsky, M. W. (2019). Organic chemistry, life, the universe and everything (OCLUE): A transformed organic chemistry curriculum. *Journal of Chemical Education, 97*(4), 1858–1872. https://doi.org/10.1021/acs.jchemed.9b00401

Curtis, S., & Millar, R. (1988). Language and conceptual understanding in science: A comparison of English and Asian language speaking children. *Research in Science & Technological Education, 6*(1), 61–77. https://doi.org/10.1080/0263514880060106

Dai, H., Liu, Z., Liao, W., Huang, X., Wu, Z., Zhao, L., Liu, W., Liu, N., Li, S., & Zhu, D. (2023). AugGPT: Leveraging ChatGPT for text data augmentation. *arXiv preprint.* arXiv:2302.13007. https://doi.org/10.48550/arXiv.2302.13007

Darden, L. (2002). Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward/backward chaining. *Philosophy of Science, 69*(S3), S354–S365. https://doi.org/10.1086/341858

DeepL SE. (2017). *DeepL Translate* [Computer program].

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities.

Computers & Education, 162(104094), 1–43. https://doi.org/10.1016/j.compedu.2020.104094

del Rosario Basterra, M., Trumbull, E., & Solano-Flores, G. (2011). *Cultural validity in assessment: Addressing linguistic and cultural diversity*. Routledge.

Deng, J. M., & Flynn, A. B. (2021). Reasoning, granularity, and comparisons in students' arguments on two organic chemistry items. *Chemistry Education Research and Practice, 22*(3), 749–771. https://doi.org/10.1039/D0RP00320D

Deng, J. M., & Flynn, A. B. (2023). "I am working 24/7, but I can't translate that to you": The barriers, strategies, and needed supports reported by chemistry trainees from English-as-an-additional language backgrounds. *Journal of Chemical Education, 100*(4), 1523–1536. https://doi.org/10.1021/acs.jchemed.2c01063

Deng, J. M., Carle, M. S., & Flynn, A. B. (2023). Students' reasoning in chemistry arguments and designing resources using constructive alignment. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (1st ed., pp. 74–89). The Royal Society of Chemistry.

Deng, J. M., Rahmani, M., & Flynn, A. B. (2022). The role of language in students' justifications of chemical phenomena. *International Journal of Science Education, 44*(13), 2131–2151. https://doi.org/10.1080/09500693.2022.2114299

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*. arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Dood, A. J., Dood, J. C., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2020). Analyzing explanations of substitution reactions using lexical analysis and logistic regression techniques. *Chemistry Education Research and Practice, 21*(1), 267–286. https://doi.org/10.1039/C9RP00148D

Dood, A. J., Fields, K. B., & Raker, J. R. (2018). Using lexical analysis to predict Lewis acid-base model use in response to an acid-base proton-transfer reaction. *Journal of Chemical Education, 95*(8), 1267–1275. https://doi.org/10.1021/acs.jchemed.8b00177

Dood, A. J., Winograd, B. A., Finkenstaedt-Quinn, S. A., Gere, A. R., & Shultz, G. V. (2022). PeerBERT: Automated characterization of peer review comments across courses. *LAK22: 12th International Learning Analytics and Knowledge Conference* (12th ed., pp. 492–499). Association for Computing Machinery.

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A survey of data augmentation approaches for NLP. *arXiv preprint*. arXiv:2105.03075. https://doi.org/10.48550/arXiv.2105.03075

Flores, A., & Smith, K. C. (2013). Spanish-speaking English language learners' experiences in high school chemistry education. *Journal of Chemical Education, 90*(2), 152–158. https://doi.org/10.1021/ed300413j

Flynn, A. B., & Ogilvie, W. W. (2015). Mechanisms before reactions: A mechanistic approach to the organic chemistry curriculum based on patterns of electron flow. *Journal of Chemical Education, 92*(5), 803–810. https://doi.org/10.1021/ed500284d

Gerard, L. F., Matuk, C., McElhaney, K., & Linn, M. C. (2015). Automated, adaptive guidance for K-12 education. *Educational Research Review, 15*, 41–58. https://doi.org/10.1016/j.edurev.2015.04.001

Gombert, S., di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., Grimm, A., Bohm, I., Neumann, K., & Drachsler, H. (2023). Coding energy knowledge in constructed responses with explainable NLP models. *Journal of Computer Assisted Learning, 39*(3), 767–786. https://doi.org/10.1111/jcal.12767

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Google LLC. (2006). *Google Translate* [Computer program].

Grapin, S. E., Pierson, A., González-Howard, M., Ryu, M., Fine, C., & Vogel, S. (2023). Science education with multilingual learners: Equity as access and equity as transformation. *Science Education, 107*(4), 999–1032. https://doi.org/10.1002/sce.21791

Grimm, A., Steegh, A., Çolakoğlu, J., Kubsch, M., & Neumann, K. (2023). Positioning responsible learning analytics in the context of STEM identities of under-served students. *Frontiers in Education, 7*(1082748), 1–12. https://doi.org/10.3389/feduc.2022.1082748

Grimm, A., Steegh, A., Kubsch, M., & Neumann, K. (2023). Learning analytics in physics education: Equity-Focused decision-making lacks guidance! *Journal of Learning Analytics, 10*(1), 71–84. https://doi.org/10.18608/jla.2023.7793

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: Prospects and limitations. *CBE - Life Sciences Education, 10*(4), 379–393. https://doi.org/10.1187/cbe.11-08-0081

Haudek, K. C., Wilson, C. D., Stuhlsatz, M. A. M., Donovan, B., Bracey, Z. B., Gardner, A., Osborne, J. F., & Cheuk, T. (2019). Using automated analysis to assess middle school students' competence with scientific argumentation. Paper presented at the National Conference on Measurement in Education (NCME), Annual Conference, Toronto, ON.

Haugeland, J. (1989). *Artificial intelligence: The very idea*. MIT Press.

Jakobsson, A., Larsson, P. N., & Karlsson, A. (2021). *Translanguaging in science education*. Springer.

Jiao, W., Wang, W., Huang, J.-T., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint*. arXiv:2301.08745. https://doi.org/10.48550/arXiv.2301.08745

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Vol. 3). Prentice Hall.

Kubsch, M., Krist, C., & Rosenberg, J. M. (2023). Distributing epistemic functions and tasks—A framework for augmenting human analytic power with machine learning in science education research. *Journal of Research in Science Teaching, 60*(2), 423–447. https://doi.org/10.1002/tea.21803

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lee, O. (2005). Science education with English language learners: Synthesis and research agenda. *Review of Educational Research, 75*(4), 491–530. https://doi.org/10.3102/00346543075004491

Lee, O., & Fradd, S. H. (1998). Science for all, including students from non-English-language backgrounds. *Educational Researcher, 27*(4), 12–21. https://doi.org/10.3102/0013189x027004012

Lee, E. N., & Orgill, M. (2022). Toward equitable assessment of English language learners in general chemistry: Identifying supportive features in assessment items. *Journal of Chemical Education, 99*(1), 35–48. https://doi.org/10.1021/acs.jchemed.1c00370

Lee, E. N., Orgill, M., & Kardash, C. (2020). Supporting English language learners in college science classrooms: Insights from chemistry students. *Multicultural Education, 27*(3), 25–32.

Lee, J., Lee, G.-G., & Hong, H.-G. (2023). Automated assessment of student hand drawings in free-response items on the particulate nature of matter. *Journal of Science Education and Technology, 32*(4), 549–566. https://doi.org/10.1007/s10956-023-10042-3

Li, T., Reigh, E., He, P., & Adah Miller, E. (2023). Can we and should we use artificial intelligence for formative assessment in science? *Journal of Research in Science Teaching, 60*(6), 1385–1389. https://doi.org/10.1002/tea.21867

Lieber, L. S., & Graulich, N. (2020). Thinking in alternatives—A task design for challenging students' problem-solving approaches

in organic chemistry. *Journal of Chemical Education, 97*(10), 3731–3738. https://doi.org/10.1021/acs.jchemed.0c00248

Lieber, L. S., & Graulich, N. (2022). Investigating students' argumentation when judging the plausibility of alternative reaction pathways in organic chemistry. *Chemistry Education Research and Practice, 23*(1), 38–53. https://doi.org/10.1039/D1RP00145K

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022a). Closing the gap of organic chemistry students' performance with an adaptive scaffold for argumentation patterns. *Chemistry Education Research and Practice, 23*(4), 811–828. https://doi.org/10.1039/D2RP00016D

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022b). Students' individual needs matter: A training to adaptively address students' argumentation skills in organic chemistry. *Journal of Chemical Education, 99*(7), 2754–2761. https://doi.org/10.1021/acs.jchemed.2c00213

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching, 53*(2), 215–233. https://doi.org/10.1002/tea.21299

Luykx, A., Lee, O., Mahotiere, M., Lester, B., Hart, J., & Deaktor, R. (2007). Cultural and home language influences on children's responses to science assessments. *Teachers College Record, 109*(4), 897–926. https://doi.org/10.1177/016146810710900403

Lyon, E. G., Bunch, G. C., & Shaw, J. M. (2012). Navigating the language demands of an inquiry-based science performance assessment: Classroom challenges and opportunities for English learners. *Science Education, 96*(4), 631–651. https://doi.org/10.1002/sce.21008

Maerten-Rivera, J., Myers, N., Lee, O., & Penfield, R. (2010). Student and school predictors of high-stakes assessment in science. *Science Education, 94*(6), 937–962. https://doi.org/10.1002/sce.20408

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B., & Krajcik, J. (2021). Using machine learning to score multi-dimensional assessments of chemistry and physics. *Journal of Science Education and Technology, 30*(2), 239–254. https://doi.org/10.1007/s10956-020-09895-9

Martin, P. P., & Graulich, N. (2023). When a machine detects student reasoning: A review of machine learning-based formative assessment of mechanistic reasoning. *Chemistry Education Research and Practice, 24*(2), 407–427. https://doi.org/10.1039/D2RP00287F

Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2023). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21903. Early view article.

Mathew, A., Amudha, P., & Sivakumari, S. (2021). Deep learning techniques: An overview. In A. E. Hassanien, R. Bhatnagar, & A. Darwish (Eds.), *Advanced machine learning technologies and applications: Proceedings of AMLTA 2020* (1141st ed., pp. 599–608). Springer.

MDZ Digital Library team. (2020). *dbmdz German BERT models*. Hugging Face. Retrieved September 21, 2023, from https://huggingface.co/dbmdz/bert-base-german-cased

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*. arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundation of machine learning*. The MIT Press.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology, 21*(1), 183–196. https://doi.org/10.1007/s10956-011-9300-9

Noble, T., Rosebery, A., Suarez, C., Warren, B., & O'Connor, M. C. (2014). Science assessments and English language learners: Validity evidence based on response processes. *Applied Measurement in Education, 27*(4), 248–260. https://doi.org/10.1080/08957347.2014.944309

Novak, G. M., Gavrin, A., Patterson, E., & Christian, W. (1999). *Just-in-time teaching: Blending active learning with web technology*. Prentice Hall.

Noyes, K., McKay, R. L., Neumann, M., Haudek, K. C., & Cooper, M. M. (2020). Developing computer resources to automate analysis of students' explanations of London dispersion forces. *Journal of Chemical Education, 97*(11), 3923–3936. https://doi.org/10.1021/acs.jchemed.0c00445

OpenAI. (2022). *ChatGPT* [Computer program].

OpenAI. (2023). *ChatGPT - Release notes*. OpenAI. Retrieved September 21, 2023, from https://help.openai.com/en/articles/6825453-chatgpt-release-notes

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (32nd ed., pp. 8024–8035). Curran Associates Inc.

Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education, 52*(3), 275–300. https://doi.org/10.1080/15391523.2020.1719943

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 1–24.

Ruder, S. (2019). *Neural transfer learning for natural language processing*. National University of Ireland.

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education, 92*(3), 499–525. https://doi.org/10.1002/sce.20264

Ryu, M. (2019). Mixing languages for science learning and participation: An examination of Korean-English bilingual learners in an after-school science-learning programme. *International Journal of Science Education, 41*(10), 1303–1323. https://doi.org/10.1080/09500693.2019.1605229

Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V., & Boeker, M. (2020). GottBERT: A pure German language model. *arXiv preprint*. arXiv:2012.02110. https://doi.org/10.48550/arXiv.2012.02110

Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint*. arXiv:1511.06709. https://doi.org/10.48550/arXiv.1511.06709

Sevian, H., & Talanquer, V. (2014). Rethinking chemistry: A learning progression on chemical thinking. *Chemistry Education Research and Practice, 15*(1), 10–23. https://doi.org/10.1039/C3RP00111C

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of Big Data, 8*(1), 1–34. https://doi.org/10.1186/s40537-021-00492-0

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553–573. https://doi.org/10.1002/tea.1018

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32*(2), 3–13. https://doi.org/10.3102/0013189x032002003

Soo, K. W. (2019). *The role of granularity in causal learning*. University of Pittsburgh.

Swanson, L. H., Bianchini, J. A., & Lee, J. S. (2014). Engaging in argument and communicating information: A case study of English language learners and their science teacher in an urban high school. *Journal of Research in Science Teaching, 51*(1), 31–64. https://doi.org/10.1002/tea.21124

Taher Pilehvar, M., & Camacho-Collados, J. (2020). *Embeddings in natural language processing: Theory and advances in vector representations of meaning*. Morgan & Claypool Publishers.

Tansomboon, C., Gerard, L. F., Vitale, J. M., & Linn, M. C. (2017). Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education, 27*(4), 729–757. https://doi.org/10.1007/s40593-017-0145-0

Tschisgale, P., Wulff, P., & Kubsch, M. (2023). Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review Physics Education Research, 19*(2), 020123-1–020123-24. https://doi.org/10.1103/PhysRevPhysEducRes.19.020123

Turkan, S., & Liu, O. L. (2012). Differential performance by English language learners on an inquiry-based science assessment. *International Journal of Science Education, 34*(15), 2343–2369. https://doi.org/10.1080/09500693.2012.705046

Urban-Lurain, M., Prevost, L. B., Haudek, K. C., Henry, E. N., Berry, M., & Merrill, J. E. (2013). Using computerized lexical analysis of student writing to support just-in-time teaching in large enrollment STEM courses. *43rd IEEE Frontiers in Education Conference Proceedings* (43rd ed., pp. 1709–1715). IEEE.

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias*. Ablex Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł, & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (30th ed., pp. 5998–6008). Curran Associates Inc.

Vitale, J. M., McBride, E., & Linn, M. C. (2016). Distinguishing complex ideas about climate change: Knowledge integration vs. specific guidance. *International Journal of Science Education, 38*(9), 1548–1569. https://doi.org/10.1080/09500693.2016.1198969

Watts, F. M., Dood, A. J., & Shultz, G. V. (2023). Developing machine learning models for automated analysis of organic chemistry students' written descriptions of organic reaction mechanisms. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (1st ed., pp. 285–303). The Royal Society of Chemistry.

Watts, F. M., Park, G. Y., Petterson, M. N., & Shultz, G. V. (2022). Considering alternative reaction mechanisms: Students' use of multiple representations to reason about mechanisms for a writing-to-learn assignment. *Chemistry Education Research and Practice, 23*(2), 486–507. https://doi.org/10.1039/D1RP00301A

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint.* arXiv:1901.11196. https://doi.org/10.48550/arXiv.1901.11196

Wilson, C. D., Haudek, K. C., Osborne, J. F., Buck Bracey, Z. E., Cheuk, T., Donovan, B. M., Stuhlsatz, M. A. M., Santiago, M. M., & Zhai, X. (2023). Using automated analysis to assess middle school students' competence with scientific argumentation. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21864. Early view article.

Winograd, B. A., Dood, A. J., Finkenstaedt-Quinn, S. A., Gere, A. R., & Shultz, G. V. (2021). Automating characterization of peer review comments in chemistry courses. In C. E. Hmelo-Silver, B. de Wever, & J. Oshima (Eds.), *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning: CSCL 2021* (14th ed., pp. 11–18). International Society of the Learning Sciences.

Winograd, B. A., Dood, A. J., Moon, A., Moeller, R., Shultz, G. V., & Gere, A. R. (2021). Detecting high orders of cognitive complexity in students' reasoning in argumentative writing about ocean acidification. *LAK21: 11th International Learning Analytics and Knowledge Conference* (11th ed., pp. 586–591). Association for Computing Machinery.

Wolf, M. K., Farnsworth, T., & Herman, J. (2008). Validity issues in assessing English language learners' language proficiency. *Educational Assessment, 13*(2–3), 80–107. https://doi.org/10.1080/10627190802394222

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2023). Utilizing a pretrained language model (BERT) to classify preservice physics teachers' written reflections. *International Journal of Artificial Intelligence in Education, 33*(3), 439–466. https://doi.org/10.1007/s40593-022-00290-6

Yik, B. J., Dood, A. J., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2021). Development of a machine learning-based tool to evaluate correct Lewis acid-base model use in written responses to open-ended formative assessment items. *Chemistry Education Research and Practice, 22*(4), 866–885. https://doi.org/10.1039/D1RP00111F

Yik, B. J., Schreurs, D. G., & Raker, J. R. (2023). Implementation of an R Shiny app for instructors: An automated text analysis formative assessment tool for evaluating Lewis acid–base model use. *Journal of Chemical Education, 100*(8), 3107–3113. https://doi.org/10.1021/acs.jchemed.3c00400

Zhai, X., Haudek, K. C., Shi, L., Nehm, R. H., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430–1459. https://doi.org/10.1002/tea.21658

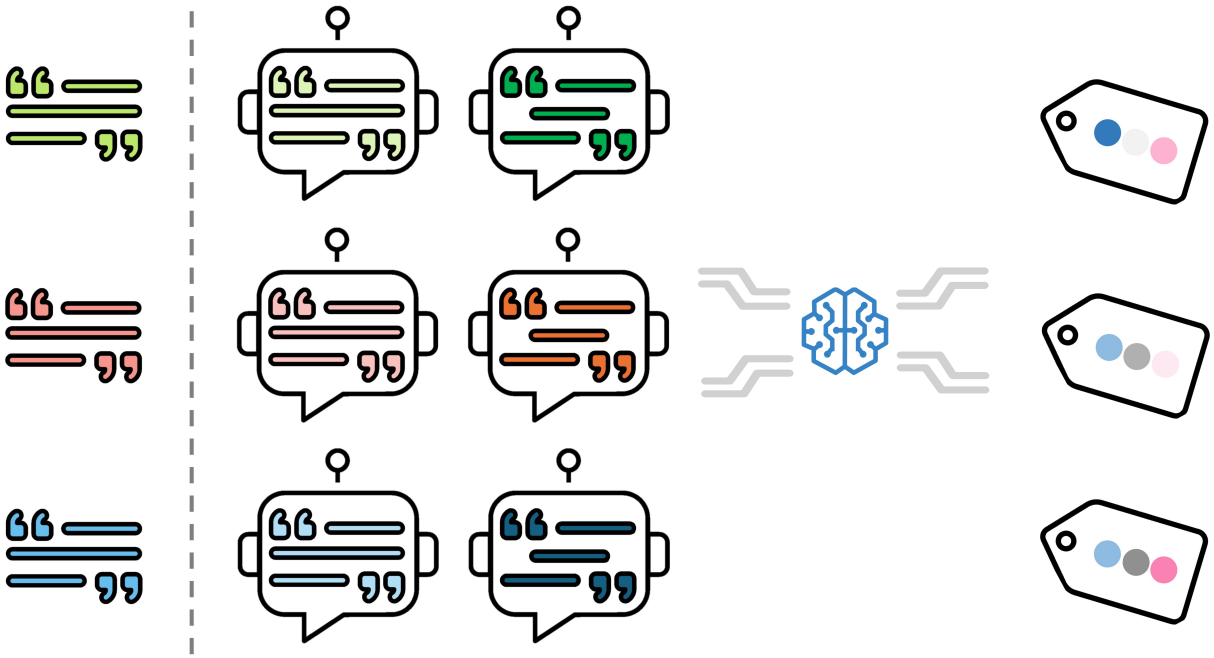
Zhai, X., He, P., & Krajcik, J. (2022). Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching, 59*(10), 1765–1794. https://doi.org/10.1002/tea.21773

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education, 56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5 | Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence

Chapter 5 presents an original research paper. It is reproduced from Martin and Graulich (2024) under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0) (https://creativecommons.org/licenses/by-nc-nd/4.0/). No changes have been made to the original article.



Martin, P. P., & Graulich, N. (2024). Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence. *Computers and Education: Artificial Intelligence*, *7*, Article 100265. https://doi.org/10.1016/j.caeai.2024.100265

Computers and Education: Artificial Intelligence 7 (2024) 100265

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence

Paul P. Martin, Nicole Graulich *

ARTICLE INFO

Keywords:
Assessment
Large language models (LLMs)
Machine learning (ML)
Data augmentation
Science education

ABSTRACT

Machine learning (ML) techniques are commonly seen as an inductive learning procedure, typically involving the identification of patterns in a specific training dataset to make predictions in novel contexts. By doing so, the performance and generalizability of these techniques often rely on the quality and quantity of the available training data. However, gathering a diverse training dataset that captures multiple nuances of students' reasoning poses challenges in educational settings due to resource constraints. We compared three data augmentation strategies to address this issue: collecting additional student data, utilizing chatbots to paraphrase existing responses, and prompting chatbots to generate synthetic responses. We found that leveraging data augmentation significantly improved ML model performance. In detail, combining authentic and/or paraphrased responses with chatbot responses yielded the best machine-human score agreements across various validation conditions. This data augmentation allowed us to expand our applied scoring rubric by introducing a more detailed categorization that better captured the level of causality in undergraduate chemistry students' reasoning about reaction mechanisms. Together, these findings highlight effective possibilities for augmenting the size and heterogeneity of the training data to improve ML model performance and generalizability, introduce a more fine-grained categorization, and reduce human effort in data collection. In the future, these benefits may enhance the scalability of formative assessments that adaptively support students' reasoning in postsecondary chemistry classes.

1. Motivation

Artificial intelligence (AI) techniques, such as machine learning (ML) and natural language processing, can automatically analyze students' open-ended reasoning in formative educational assessments. Harnessing this potential of AI not only speeds up the assessment process but also offers a more continuous evaluation of student reasoning (Swiecki et al., 2022; Zhai, Yin, et al., 2020). Additionally, leveraging AI-driven analysis enables educators to gain deeper insights into students' cognitive processes (Kubsch et al., 2023; Zhai, Haudek, et al., 2020), facilitating targeted guidance to improve learning outcomes (Deeva et al., 2021).

In a current research project, we developed an ML model that automatically evaluates students' reasoning about organic chemistry reaction mechanisms according to a holistic 20-category rubric (Martin et al., 2023). Throughout the rubric design process, we iteratively combined theory-driven human interpretation with data-driven ML techniques to capture students' reasoning reliably and validly. However, some discrepancies appeared during the rubric design process: Our ML model could assign students' reasoning to one of 20 categories, whereas developing a 24-category ML model as proposed in the chemistry education research literature was not possible due to a lack of sufficient sample responses (Martin et al., 2023). From an assessment perspective, including the four not automatically evaluable categories in our analysis would be crucial as they represent unique types of causal reasoning in chemistry (cf., Bodé et al., 2019; Deng & Flynn, 2021; Sevian & Talanquer, 2014). However, because of the tremendous performance decrease from the 20-category model to the 24-category one (Table 1), using the latter was not reasonable (cf., Williamson et al., 2012). Therefore, we looked for ways to increase the performance of the 24-category model.

As the performance of ML-based techniques depends on the quality and quantity of the available data, ML models often show lower accuracy and generalizability with limited or unbalanced training datasets (Fang et al., 2023; Kaldaras & Haudek, 2022). Aiming to boost ML model performance, we applied data augmentation strategies to expand

* Corresponding author. Justus-Liebig-University, Institute of Chemistry Education, Heinrich-Buff-Ring 17, 35392, Giessen, Germany.
E-mail address: Nicole.Graulich@didaktik.chemie.uni-giessen.de (N. Graulich).

https://doi.org/10.1016/j.caeai.2024.100265
Received 27 December 2023; Received in revised form 4 July 2024; Accepted 6 July 2024
Available online 10 July 2024
2666-920X/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Table 1
Machine-human score agreements of a simplified and an extended ML model that evaluates undergraduate chemistry students' reasoning about reaction mechanisms.

Model	Accuracy	Cohen's κ	Weighted F_1 score	Macro F_1 score
20-category	0.87	0.86	0.86	0.80
24-category	0.70	0.68	0.68	0.60

training data size (cf., Bayer et al., 2022; Feng et al., 2021; Shorten et al., 2021). Expanding training data size can, for example, be realized by collecting additional student data (e.g., Noyes et al., 2020; Wang et al., 2021), which may be difficult due to resource constraints. Producing synthetic data, i.e., data that is artificially generated to simulate authentic data, can, thus, be useful to improve ML model performance. Approaches for producing synthetic data include leveraging AI to paraphrase student responses or generate novel samples (e.g., Cochran et al., 2022, 2023; Dai et al., 2023; Fang et al., 2023; Møller et al., 2023; Shi et al., 2020; Ubani et al., 2023; Wu et al., 2019), which may increase ML model performance, generalizability, and granularity.

Taking a broader perspective, examining data augmentation approaches is essential not only for boosting ML model performance but also for strategic data collection planning. For instance, ML model developers could use synthetic responses produced by chatbots to expand their training data (Kieser et al., 2023); hence, reducing the reliance on resource-intensive data collection efforts. Additionally, investigating the impact of these strategies in educational contexts helps establish adaptive learning systems that offer personalized guidance for fine-grained proficiency levels. Given this potential, several data augmentation strategies for educational assessments are validated and compared herein.

2. Background

2.1. Advancements in artificial intelligence

Due to recent technological advancements, ML, a subarea of AI, can adopt activities traditionally associated with human tasks. To perform these tasks, ML focuses on developing algorithms that independently learn from data to make predictions (Bishop, 2006; Mitchell, 1997). Supervised ML algorithms are trained on human-labeled data to discern underlying patterns and predict labels for new data. In recent years, deep learning, an advanced ML technique, has gained remarkable prominence in science and society due to its ability to extract intricate patterns from complex data (Mathew et al., 2021). Deep learning focuses on training deep neural networks with multiple layers of interconnected artificial neurons, inspired by the structure of the human brain (Goodfellow et al., 2016). Over the last few years, several ML techniques, including deep learning, have witnessed a growing application in science education research (e.g., Gerard et al., 2015; Martin & Graulich, 2023b; Zhai, Haudek, et al., 2020; Zhai, Yin, et al., 2020).

A subset of deep learning is generative AI, which focuses on producing human-like data such as text, images, videos, audio, or computer code based on patterns found in the training data. Accordingly, generative AI empowers computers to produce content conventionally seen as a product of human creativity. One of the most notable advancements in generative AI are large language models (LLMs), which are trained on extensive amounts of unlabeled text data to perform various language-related tasks, such as question-answering, text generation, text classification, language translation, or data augmentation (Li et al., 2022; Radford et al., 2019). Compared to previously applied natural language processing techniques, LLMs can grasp complex semantic relationships between words in a sentence (Mikolov et al., 2013; Taher Pilehvar & Camacho-Collados, 2020). Examples of such models are Generative Pre-trained Transformer (GPT) (Brown et al., 2020), Pathways Language Model 2 (PaLM2) (Anil et al., 2023), or Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). These models are trained to capture syntactic nuances, implicit meanings, and the contextual usage of words as well as to produce human-like text. Following that, these LLMs are generally applied to many language-related tasks.

2.2. Approaches for expanding the training data size of an ML model

Training a well-generalizable text classification model that can automatically evaluate students' reasoning across many categories becomes challenging when training data is limited or unbalanced (Fang et al., 2023; Kaldaras & Haudek, 2022). However, limited or unbalanced datasets are common in educational contexts since most students are often grouped into a certain range of scoring categories (e.g., Fang et al., 2023). Because of this, training data for certain minority categories might be scarce, so ML models may face challenges in accurately scoring these categories. This imbalance can result in reduced scoring accuracy, especially affecting those students grouped in the minority categories (cf., Li et al., 2023). Therefore, data augmentation methods might be necessary to expand training datasets and boost ML model performance.

2.2.1. Gathering new data

Several approaches to expanding training data size have been developed so far. Gathering new data is the most common way to increase ML model performance and generalizability. In doing so, researchers aim to gather as heterogeneous data as possible to improve ML model performance, especially for minority categories (Shorten & Khoshgoftaar, 2019). However, this traditional approach involves conducting resource-intensive data collection to provide ML models with a more comprehensive dataset (e.g., Noyes et al., 2020; Wang et al., 2021).

2.2.2. Applying text data augmentation techniques

Moreover, one can also apply semantic modifications to increase the diversity of the data, which is subsumed under the term text data augmentation. Techniques for text data augmentation involve inserting or substituting words, changing word order, adding synonyms, modifying sentence structures, or employing other linguistic transformations. Text data augmentation is experiencing rapid growth in text classification problems as it often helps enhance data heterogeneity, extend the sample size, and improve ML model performance (e.g., Bayer et al., 2022; Feng et al., 2021; Shorten et al., 2021; Wei and Zou, 2019).

For text data augmentation, Sennrich et al. (2015) relied on back-translation techniques, wherein they translated text into another language and subsequently translated it back to the original language. Because of the inherent randomness in the translation process, their augmented text differed from the original text while conceptual consistency was preserved. However, a recent study showed that translation techniques did not significantly increase ML model performance in a science assessment context (Martin & Graulich, 2024a).

LLMs can also be used for text data augmentation. By doing so, Cochran et al. (2022), Shi et al. (2020), and Wu et al. (2019) demonstrated that a fine-tuned BERT model was most effective in augmenting text datasets sensitive to context. Similarly, Dai et al. (2023), Fang et al. (2023), and Ubani et al. (2023) leveraged ChatGPT or the GPT language models to rephrase sentences in the training dataset, producing numerous samples that shared conceptual similarities but showed semantic differences. The rephrased sentences were then used to train an ML model, demonstrating that this approach outperformed other state-of-the-art text data augmentation methods regarding accuracy and transferability. Building upon this research, Møller et al. (2023) explored GPT-4's data augmentation capabilities with minimal prompting, revealing that GPT-4 was specifically useful in classification tasks operating with unbalanced datasets. Analogously, Cochran et al. (2023) reported that utilizing ChatGPT for data augmentation improved the automatic scoring of student essays in scenarios with limited data.

Following this, we built herein upon these proposed text data augmentation methods by using ChatGPT and Google Bard to rephrase student reasoning. During our analysis, the LLM developed by Google was referred to as Bard while it is now known as Gemini.

2.2.3. Prompting chatbots to generate novel responses

Recently, a growing number of research in science education explored the use of generative AI to transform teaching and learning (cf., Alasadi & Baiz, 2023). Here, researchers leveraged pre-trained chatbots like ChatGPT for various reasons, including planning a high-school science lesson (Cooper, 2023; Lee and Zhai, 2024; Martin & Graulich, 2023a), designing assessment tasks (Küchemann et al., 2023), grading assignments (Kortemeyer, 2023b), writing a lab report (Humphry & Fuller, 2023; West et al., 2023a), generating worked solutions in a critiquing activity (Dahlkemper et al., 2023; Exintaris et al., 2023), or solving assessment items (Clark, 2023; Clark et al., 2023; Emenike & Emenike, 2023; Fergus et al., 2023; Kieser et al., 2023; Kortemeyer, 2023a; Leon & Vidhani, 2023; West, 2023a, 2023b; Yeadon et al., 2023). In most cases, ChatGPT was not capable of correctly solving the given assessment items (Emenike & Emenike, 2023; Leon & Vidhani, 2023; Tyson, 2023), especially when they required the application of chemical concepts (Fergus et al., 2023). For this reason, it was inferred that ChatGPT cannot replace human guidance and expertise, yet (Alasadi & Baiz, 2023; Emenike & Emenike, 2023; Leon & Vidhani, 2023).

While this body of research focused on the correctness of the synthetic responses, related studies examined the scientific reasoning abilities of chatbots (Kieser et al., 2023; Talanquer, 2023; Watts et al., 2023b). In particular, Watts et al. (2023b) prompted ChatGPT and Bard to give descriptions and justifications for the occurrence of the acid hydrolysis of a thalidomide molecule and a base-free Wittig reaction, finding that the chatbots rarely discussed a key feature of mechanistic explanations: the movement of electrons. Given this finding, they inferred that chatbots reason more descriptively about organic chemistry reaction mechanisms than undergraduate chemistry students. Moreover, Talanquer (2023) showed that chatbots and chemistry students expressed similar preconceptions and struggled to coherently integrate new information into their reasoning. In detail, chatbots and novice learners often relied on teleological reasoning, i.e., reasoning that attributes chemical processes to the goals, needs, or wants of active agents rather than to the outcome of random interactions constrained by energy and entropy. In a comparable study, Kieser et al. (2023) investigated if ChatGPT can solve a force-related concept inventory in physics from the perspective of different cohorts holding different preconceptions. Their findings indicate that the ChatGPT-generated responses provided no variance when arguing from the perspective of students in different cohorts but varied significantly when adopting well-defined preconceptions. In conjunction with the findings of Talanquer (2023), this highlights that chatbot responses mirror students' actual reasoning abilities, which is why these responses may – at least to some extent – serve as valuable samples for ML model training (Kieser et al., 2023). Consequently, we prompted ChatGPT and Bard to solve our formative assessment in various ways and used the so-created responses to train our ML model across all 24 categories.

3. Research questions

In this study, we aimed to test the impact of three data augmentation strategies on ML model performance: collecting additional student data, applying text data augmentation by paraphrasing existing responses, and prompting ChatGPT and Bard to generate synthetic responses. To evaluate the necessity of leveraging generative AI assistance in data augmentation, we further implemented basic up-sampling augmentation strategies as baseline models without using generative AI. The results not only offer guidance for boosting ML model performance but also discuss the potential of generative AI in augmenting traditional data acquisition strategies. Following that, the research questions examined herein are methodological as they determine the degree to which different approaches for expanding the training data size impact ML model performance (Fig. 1).

1) To what extent does the performance of an ML model trained on student responses increase by adding (i) additional student responses, (ii) chatbot-paraphrased student responses, or (iii) chatbot-generated synthetic responses to the training dataset?
2) To what extent does the performance of an ML model differ by training it on (i) authentic student responses, (ii) chatbot-paraphrased student responses, (iii) chatbot-generated synthetic responses, or (iv) combinations of these data?

4. Context

4.1. Setting of data collection

To support students' reasoning proficiency in organic chemistry, Lieber et al. (2022a, b) and Martin and Graulich (2024b) designed an online setting that prompted students to judge the plausibility of alternative reaction products for Williamson ether synthesis, aldol condensation, Claisen condensation, and enamine hydrolysis (Fig. 2). Engaging in discussions about the plausibility of reaction products requires integrating multiple chemical concepts, which must be weighed to construct evidence-based arguments and counterarguments (Lieber & Graulich, 2020, 2022, Lieber et al., 2022a, b; Watts et al., 2022). Hence, students are encouraged to evaluate the plausibility of reaction products critically, consider various reaction pathways, and link chemical concepts. Lieber et al. (2022a, b) demonstrated an increase in students' chemical concept knowledge and argumentation competence through this task design.

For ML model training, we initially used the 1108 student responses previously gathered within the described setting at a private, research-intensive, liberal arts university in the Northeastern United States in April and May 2021. For demographic information on the participating undergraduate organic chemistry students, see Lieber et al. (2022a).

4.2. Scoring rubric

We designed a two-dimensional holistic rubric consisting of 24 categories to assess students' open-ended reasoning (Fig. 3). This rubric is inspired by two theoretical frameworks pertinent to chemistry education research: the *modes of reasoning* (Sevian & Talanquer, 2014) and the *levels of granularity* (Bodé et al., 2019; Deng et al., 2023; Deng & Flynn, 2021; Soo, 2019). The *modes of reasoning*, including *descriptive*, *relational*, *linear causal*, and *multicomponent causal*, classify the causality of students' open-ended reasoning regarding their ability to establish connections between different chemical concepts and provide well-justified explanations for why phenomena occur (Russ et al., 2008; Sevian & Talanquer, 2014). In particular, the *modes of reasoning* highlight the importance of evaluating students' ability to integrate new information into their existing cognitive network (Sevian & Talanquer, 2014). By contrast, the *levels of granularity* characterize the depth at which phenomena are explained. Generally, different tasks require different *levels of granularity* to adequately reason about the underlying processes (Darden, 2002). In our study, we adopted four *levels of granularity*, namely *structural*, *energetic*, *phenomenological*, and *electronic* (Deng et al., 2023; Deng & Flynn, 2021). For a more detailed account of the rubric development process, see Martin et al. (2023).

For human coding, a research assistant and the first author coded the entire dataset individually in two rounds, assigning one of 24 mutually exclusive codes to the student responses. After that, inter-rater reliability was determined, achieving *almost perfect* agreement (Table 2) (Landis & Koch, 1977). Subsequently, codes were discussed until full agreement was reached and consensus codes were used to train the ML model.

Table 3 illustrates the distribution of the original student responses

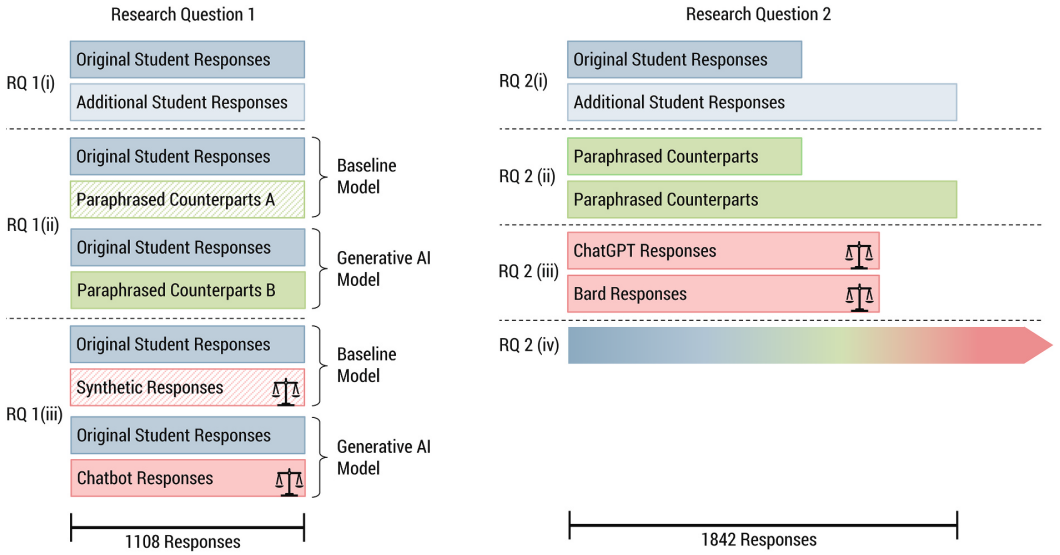


Fig. 1. Comparative overview of the datasets used to train machine learning models for addressing the research questions. *Note:* The scale symbolizes balanced datasets.

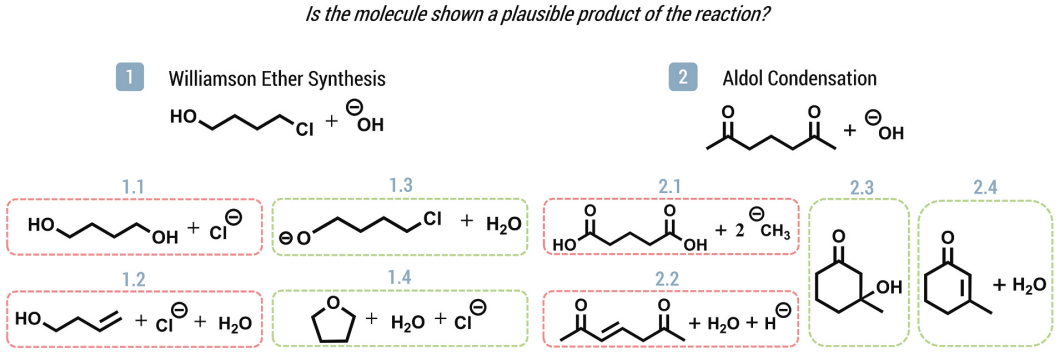


Fig. 2. Students judged the plausibility of four alternative reaction products for different reaction mechanisms, among others, a Williamson ether synthesis and an aldol condensation. Plausible products (1.3, 1.4, 2.3, 2.4) are highlighted in green, while implausible products (1.1, 1.2, 2.1, 2.2) are highlighted in red.

across the 24 categories, showing that students at the same educational level employed diverse *modes of reasoning* to solve identical exercises (Bodé et al., 2019; Deng & Flynn, 2021; Moreira et al., 2019; Weinrich & Talanquer, 2016). In detail, our analysis indicates that students predominantly reasoned in a *descriptive* or *relational mode*, aligning with other research (e.g., Carle et al., 2021; Deng & Flynn, 2021). In contrast, a *linear causal mode of reasoning* was seldom present, posing challenges for ML scoring of responses at a high causality level (Martin et al., 2023).

5. Methods

We performed the ML analysis in Python 3.9.13 (Van Rossum & Drake, 2009) using *PyTorch* as an ML framework (Paszke et al., 2019) and *SpaCy* for natural language processing (Honnibal et al., 2020). Specifically, we leveraged pre-trained, transformer-based deep neural networks for sequence classification to examine students' reasoning. In doing so, we applied *AutoTokenizer* to tokenize the input text and the base variant (*BERT-base-uncased*) of the LLM BERT for natural language processing (Devlin et al., 2018). We added a classification head to predict class labels and established a training loop using the *AdamW* optimizer and a learning rate scheduler to train the models.

We divided our dataset into three subsets: a training, validation, and test set, distributed in a ratio of 65:15:20. The training set was used to train a deep neural network, the validation set helped identify the optimal hyperparameter configuration, and the test set was utilized to assess model performance using four metrics: accuracy, Cohen's κ, weighted F_1 score, and macro F_1 score. Accuracy measures the percentage of agreement between two raters, whereas Cohen's κ quantifies agreement between two raters beyond chance. F_1 scores account for the number and type of prediction errors in unbalanced datasets as the harmonic mean of precision and recall. Weighted F_1 score is the weighted mean of all per-class F_1 scores, while macro F_1 score is their unweighted mean. As hyperparameters, the number of epochs, the learning rate, and the batch size were fine-tuned. Epochs indicate the number of training cycles, the learning rate specifies how quickly the model adapts its parameters to a provided context, and the batch size determines the number of training samples processed simultaneously in a single network pass. We explored values ranging from 1 to 100 for the number of epochs, learning rates of 5e-6 and 1e-5, and batch sizes of 2, 4, 8, 16, and 32. In total, we examined 1000 different hyperparameter

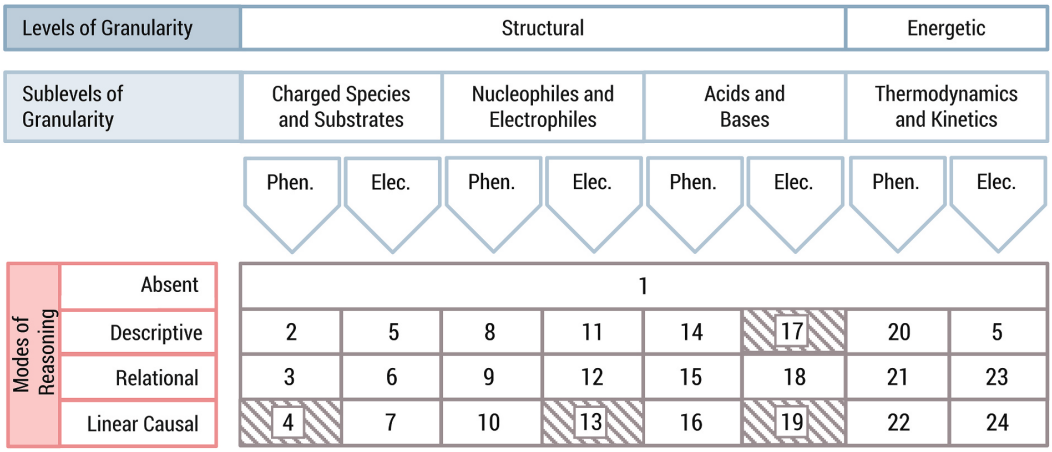


Fig. 3. Holistic 24-category rubric for scoring students' reasoning. The shaded areas show categories that could not be automated in previous analyses. *Note:* Phen. stands for *phenomenological*, Elec. for *electronic*.

Table 2
Human-human inter-rater reliability.

Round	Accuracy	Cohen's κ	Weighted F_1 score	Macro F_1 score
1	0.83	0.82	0.83	0.79
2	0.86	0.85	0.86	0.80

configurations for each model.

We used a combination of internal and external validation to evaluate ML model performance shifts. Internal validation refers to a situation where the characteristics of the training and validation dataset match those of the test dataset. In educational contexts, internal validation means that the characteristics of the data collection procedure influenced by institution, course, content, and tasks are upheld across model training and testing. In other words, the test set is a new dataset, yet it is collected under the same conditions as the training and validation sets. By contrast, external validation, also known as distribution shift, arises when the characteristics of the training and validation set differ from those of the test set (Varshney, 2022). External validation may result in a decline in a model's performance as it may struggle to accurately predict labels for new data that deviates from the training data. Therefore, implementing external validation is key to ensuring the robustness and trustworthiness of ML models in real-world contexts, especially when the encountered data may vary over time or across different conditions (Varshney, 2022).

5.1. Collecting additional student data

To expand the training data size, we collected 1842 additional student responses at two institutions: a large, undergraduate-only university in Western Canada and a large, research-intense university in the Midwestern United States. Data collection in undergraduate organic chemistry courses spanned from the Winter 2020 semester to the Summer 2023 semester. At the former institution, data was collected within the instructional setting described in Section 4.1, which means that these students also had to judge the plausibility of four alternative products for given reaction mechanisms (cf., Fig. 2). At the latter institution, the research instrument used for data collection included two Writing-to-Learn assignments to also support students' chemical concept knowledge (Watts et al., 2023a). Generally, Writing-to-Learn assignments represent an instructional strategy integrating writing tasks into the learning process (e.g., Anderson et al., 2015; Arnold et al., 2017). The primary goal of these writing assignments is to facilitate students' acquisition of domain-specific concepts (Bangert-Drowns et al., 2004; Gere et al., 2019; Gunel et al., 2007), as observed in disciplines like chemistry (Finkenstaedt-Quinn et al., 2017; 2021b). Diverging from traditional writing tasks that predominantly focus on the final draft, Writing-to-Learn highlights the importance of the writing process as a learning tool (Tynjälä et al., 2001). The tasks used for ML model training prompted students to apply their organic chemistry concept knowledge to provide explanations for the acid hydrolysis of thalidomide as well as for a base-free Wittig reaction (Watts et al., 2023a). In doing so, the assignments integrated peer assessment, revision, and collaboration with writing mentors. For demographic information or details on the prompts, design, and implementation of these Writing-to-Learn assignments, refer to Finkenstaedt-Quinn et al. (2021a) and Watts et al. (2023a).

In addition to varying reaction mechanisms, the two instructional settings (Sections 4.1 and 5.1) differed in that the original exercises prompted students to reason about alternative reaction products (cf., Fig. 2); instead, the Writing-to-Learn assignments emphasized the collaborative process of writing chemical explanations. By using these

Table 3
Number of original student responses per rubric category.

	Structural						Energetic	
	Charged Species and Substrates		Nucleophiles and Electrophiles		Acids and Bases		Thermodynamics and Kinetics	
	Phen.	Elec.	Phen.	Elec.	Phen.	Elec.	Phen.	Elec.
Absent	23							
Descriptive	127	70*	47	31	90	18	47	70*
Relational	38	32	64	64	58	51	82	123
Linear Causal	0	36	22	10	20	1	22	32

Note: Phen. stands for *phenomenological*, Elec. for *electronic*. The asterisk shows that categories belong together.

two settings, we envisioned increasing the generalizability of our ML model. Nevertheless, both instructional settings aimed to support students' understanding of organic chemistry concepts by requiring them to evaluate the plausibility of reaction mechanisms. For this reason, the original and additionally collected responses could be scored using the same rubric (Section 4.2). Table 4 provides an overview of the distribution of the additionally collected student responses across all 24 categories. It becomes evident, even more so than with the original dataset, that students predominantly employed *descriptive* reasoning, with other *modes of reasoning* rarely observed. The uneven distribution of these additionally collected student responses across all categories resulted in a highly imbalanced dataset, reflecting a common challenge in training ML models for science assessments (Fang et al., 2023).

5.2. Paraphrasing student responses

To subsequently augment our dataset, we paraphrased all 2950 student responses using ChatGPT with the versions from GPT-3.5 March 23, 2023, to GPT-3.5 September 25, 2023 (Brown et al., 2020; OpenAI, 2022), which were accessed via the free accounts of the first author and two research assistants. In ChatGPT, the written responses were paraphrased in a single chat using the prompt „Please, paraphrase the following responses as much as you can". The paraphrased responses were assigned the same codes as the authentic student responses, so the distribution of the paraphrased data corresponds exactly to the distribution of the student data (Tables 3 and 4). Examples of chatbot-paraphrased student responses are presented in Table 5 (column 2), highlighting the chatbot's ability to precisely grasp the intended meaning of chemical terms within the given context. The paraphrased responses were fed unmodified in the ML model.

Beyond that, we paid attention that we incorporated a student response and its paraphrased counterpart either both into the test set or excluded them from it because otherwise, model performance could be skewed toward higher agreement metrics. In other words, including the original student response and its paraphrased counterpart in both the training and test set would neglect the risk of overfitting the training data, meaning that the ML model would become excessively specialized on a small data excerpt. Potentially, overfitting creates ML models that perform exceptionally well on the training data but fail to make accurate classifications on new data.

5.3. Prompting chatbots to generate synthetic responses

The same ChatGPT versions as used for paraphrasing the student responses and Bard July 13, 2023 (Anil et al., 2023; Google, 2023) were applied to generate synthetic responses based on 24 different prompts, one per rubric category (Appendix), also in a single chat. These prompts were extracted by interpreting the output of a cluster analysis that grouped students' responses into distinct categories (Martin et al., 2023). Each category corresponds to a unique reasoning type. Together, the prompts cover the most prevalent patterns observed in how students typically respond to the assessment items, which means that we could gather an authentic representation of the actual reasoning types at play in this assessment context. To improve the quality of the synthetic responses, we provided the chatbots with information about the alternative reaction products that might form in the given reactions. Examples of synthetic, chatbot-generated responses are listed in Table 5 (column 4), showing the robust language generation capabilities of the chatbots in organic chemistry contexts.

We checked whether the chatbots' output aligned with the rubric category outlined in the prompt to ensure consistency in coding. In some cases, coding adjustments were needed as the output did not match the prompted category, so we recoded some chatbot responses. Table 6 shows the distribution of the chatbot responses across all 24 categories. The dataset of chatbot-generated responses is nearly balanced, whereas the distribution of authentic student responses is highly imbalanced (Tables 3 and 4). We prompted ChatGPT and Bard to produce 1475 synthetic responses each.

5.4. Baseline models without generative AI assistance

To assess the need for generative AI in data augmentation, we compared the quality of the generative AI-augmented data with conventional up-sampling augmentation techniques based on contextual word embeddings. Specifically, we used the LLM *BERT-base-uncased* (Devlin et al., 2018) to generate two synthetic datasets, which correspond to the chatbot-paraphrased student responses and the chatbot-generated synthetic responses. In the first approach, BERT was applied to substitute words in the original student responses with words having semantically similar meanings (Table 5, column 3) to assess the chatbot's capability in paraphrasing student responses. In the second approach, BERT was utilized to substitute *and* insert words at random positions in the student responses (Table 5, column 5) to create a balanced dataset similar to the chatbot-generated synthetic responses. These up-sampling techniques were applied to establish baseline models against which the effectiveness of the generative AI data augmentation could be measured.

As can be seen in Table 5, BERT's language processing and generation capabilities lag behind those of generative AI when augmenting student data within a chemistry context. Particularly, BERT has significant challenges in accurately interpreting chemical terminology. For instance, it paraphrased "no major steric interactions" as "two other sequence elements", a "primary carbon" as "two radicals", a "carbon connected to chlorine" as "mineral dioxide", and "not the lowest energy" as "the lowest risk" (Table 5, column 3). These discrepancies demonstrate BERT's flawed chemical understanding when aiming to generate synthetic data with up-sampling techniques.

6. Results

The initial 24-category ML model developed without data augmentation demonstrated only *substantial* machine-human score agreements (Table 1). Consequently, we compared the effectiveness of three data augmentation strategies to boost ML model performance and better distinguish between different reasoning types in chemistry. To assess performance changes, we applied both internal and external validation.

Table 4
Number of additionally collected student responses per rubric category.

	Structural						Energetic	
	Charged Species and Substrates		Nucleophiles and Electrophiles		Acids and Bases		Thermodynamics and Kinetics	
	Phen.	Elec.	Phen.	Elec.	Phen.	Elec.	Phen.	Elec.
Absent	300							
Descriptive	314	333*	103	60	188	146	60	333*
Relational	36	12	37	35	14	17	3	87
Linear Causal	6	7	11	4	11	9	26	23

Note: Phen. stands for *phenomenological*, Elec. for *electronic*. The asterisk shows that categories belong together.

Table 5
Examples of student responses, chatbot-paraphrased student responses, responses paraphrased without generative AI, chatbot-generated synthetic responses, and synthetic responses created without generative AI. All examples represent *descriptive*, *phenomenological* arguments (Fig. 3) about reaction product 1.1 in Williamson ether synthesis (Fig. 2).

Student Response	Paraphrased with ChatGPT	Paraphrased with BERT	Generated with ChatGPT	Generated with BERT
There are no major steric interactions prohibiting the reaction. The Cl is on a primary carbon and the hydroxide is a small molecule.	The reaction is not hindered by major steric interactions because the Cl atom is attached to a primary carbon and the hydroxide group is a small molecule.	Next are two other sequence elements during the reaction. The Cl is on a primary chain and the other is a small molecule.	The reaction involves 4-chloro-butan-1-ol, which is not sterically hindered. Chlorine is a leaving group replaced by hydroxide. This forms 1,4-butanediol with two hydroxyl groups.	The synthesis is not facilitated by weak steric interactions because either the oxygen atom component represents equivalent to ammonia both primary and the reactive sugar unit is a comparatively small molecule.
S_N2 is possible. The carbon of attack is electrophilic because of the Cl^- group and it is a primary carbon.	S_N2 reaction can occur because the carbon atom being attacked is electrophilic owing to the presence of a Cl^- group and the fact that it is a primary carbon.	S_N2 is possible given carbon of attack is electrophilic because of the Cl^- group. Instead, it always contains two radicals.	In the reaction between 4-chloro-butan-1-ol and hydroxide ions, the hydroxide ion acts as a nucleophile. The carbon atom in the 4-chloro-butan-1-ol molecule is the electrophile.	S_N2 radical reaction can theoretically occur if one polymer short chain as originally added is electrophilic due to the presence of a Cl^- group structure and the fact the carbon backbone is not a primary terminal carbon.
The hydroxide anion can deprotonate the carbon connected to chlorine since the hydroxide anion is a very strong base.	The hydroxide anion is capable of deprotonating the carbon adjacent to chlorine because the hydroxide anion is a potent base with the ability to extract a proton from the carbon close to chlorine.	The resultant anion can deprotonate like mineral dioxide since that same anion is still a very unstable base.	The hydroxide ion acts as a base while the alcohol group of 4-chlorobutan-1-ol acts as an acid, allowing for proton transfer.	The hydroxide ion is capable of attacking exclusively the four carbon molecules next to its chlorine because all the corresponding hydroxide salt groups and a potent base can abstract any fraction from the nucleus.
The products are not very stable. The products are not the lowest energy of this reaction.	The products are not very stable. The products are not at a high level of stability.	The products are both very dangerous. The products have the lowest risk of elimination reaction.	The difference in energy levels between reactants and products indicates a non-favorable energy change.	The products are mostly very stable. However, the products are not without much in critical storage needs but for total stability.

Table 6
Number of chatbot responses per rubric category.

	Structural						Energetic	
	Charged Species and Substrates		Nucleophiles and Electrophiles		Acids and Bases		Thermodynamics and Kinetics	
	Phen.	Elec.	Phen.	Elec.	Phen.	Elec.	Phen.	Elec.
Absent	104							
Descriptive	118	249*	121	120	134	122	124	249*
Relational	126	113	126	145	119	121	125	147
Linear Causal	19	120	120	104	117	106	122	128

Note: Phen. stands for *phenomenological*, Elec. for *electronic*. The asterisk shows that categories belong together.

6.1. Research question 1

To answer the first research question, we implemented text data augmentation by incorporating (i) 1108 additional student responses, (ii) 1108 chatbot-paraphrased student responses, or (iii) 1108 chatbot-generated synthetic responses into the training dataset (Fig. 1). We further established two baseline models to evaluate the necessity of generative AI assistance in data augmentation: one that replaced words based on contextual embedding calculations and another that replaced *and* inserted words at random positions according to these contextual embeddings. Following that, this investigation addresses both the efficacy of the augmentation techniques and the potential of generative AI in data augmentation.

Table 7 presents the performance variations between the ML models augmented with different datasets. A clear trend is evident across all machine-human score agreements: The model trained on a combination of original student data and chatbot-generated data outperformed the rest (Table 7: model 3). This model's superior performance, particularly in macro F_1 score, can be attributed to the high quality and balanced distribution of chatbot-generated data (Tables 5 and 6). Our finding emphasizes the value of augmenting student responses with chatbot-generated data to build ML models that can accurately classify student reasoning in science assessments. The model augmented with student responses ranked second (Table 7: model 1), while the model trained on a mix of student and chatbot-paraphrased responses placed third (Table 7: model 2). Nonetheless, all three augmentation strategies

Table 7
Performance differences of ML models augmented with various text data augmentation approaches.

ID	Original Student Data	Additional Student Data	Rephrased Data	Synthetic Data	Epochs	Learning Rate	Batch Size	Macro F_1 score	Accuracy	Cohen's κ	Weighted F_1 score	Macro F_1 score
								Val. Set	Test Set			
1	X	X			34	1e-5	16	0.8002	0.8671	0.8561	0.8653	0.8267
2B	X		X		51	1e-5	4	0.8259	0.7230	0.7042	0.7255	0.6867
2	X		X		66	1e-5	2	0.9436	0.8468	0.8364	0.8427	0.7737
3B	X			X	84	1e-5	4	0.9090	0.6959	0.6822	0.6860	0.6805
3	X			X	68	1e-5	2	0.8892	0.8941	0.8886	0.8933	0.8884

Note: In the model ID, "B" indicates baseline models without generative AI assistance.

improved the model's performance compared to the original one developed without text data augmentation (Table 1).

Comparisons between models augmented with chatbot-paraphrased and chatbot-generated data against their respective baseline models further proved the usefulness of generative AI in data augmentation. Specifically, the model augmented with chatbot-paraphrased data significantly outperformed the baseline model using BERT-paraphrased responses (Table 7: models 2B and 2), indicating that the improvement achieved through generative AI augmentation stems from the quality of the paraphrased responses (Table 5). Likewise, the model augmented with chatbot-generated data surpassed the model using BERT-generated data (Table 7: models 3B and 3), emphasizing that performance gains cannot be solely attributed to the resolution of data imbalances (Table 6).

To finally examine potential overfitting, we compared the macro F_1 score of the models between validation and testing scenarios. Here, we observed a decrease in the macro F_1 score from validation to test sets in the baseline models as well as in the model that included student data and their paraphrased variants (Table 7: models 2B, 2, and 3B). This suggests that these models learned too intricate details of the training data, decreasing their generalizability. In contrast, the two other models did not overfit the training data as evidenced by the consistent macro F_1 score. Of note, we used the macro F_1 score to assess overfitting as it is the most conservative metric accounting best for imbalances in datasets. In sum, these results highlight the benefits of using generative AI for data augmentation. The remarkable performance of the ML model augmented with chatbot-generated data motivated the analysis discussed in research question 2.

6.2. Research question 2

Building upon these results, we further investigated the effectiveness of the applied text data augmentation strategies. This examination included training an ML model solely on (i) authentic student responses, (ii) chatbot-paraphrased student responses, (iii) chatbot-generated synthetic responses, or (iv) combinations of these data (Fig. 1). Different performance groups emerge when employing internal validation to assess ML model performance: Models exclusively trained on chatbot responses performed best across all agreement metrics (Table 8: models 3, 4, and 5). Moreover, the model trained on a combination of authentic student and chatbot responses demonstrated performance nearly on par with those solely trained on chatbot responses (Table 8: model 7). Besides, the ML model trained on a combination of paraphrased and chatbot responses as well as the one utilizing the entire data range occupied an intermediate position in terms of performance (Table 8: models 8 and 9). In contrast, the models trained exclusively on authentic student data, paraphrased data, or their combination showed the poorest performance across all metrics when applying internal validation (Table 8: models 1, 2, and 6).

We again compared the macro F_1 score on the validation set with the macro F_1 score on the test set to investigate if the models overfitted the training data. In doing so, a decline in the macro F_1 score was observed from the validation to the test set for models including authentic student data and their paraphrased counterparts (Table 8: models 6 and 9). In these instances, the ML model learned too specific nuances of the training data, reducing its generalizability. Conversely, all other models may be resilient against overfitting.

To further detect potential instances of overfitting and assess the efficacy of text data augmentation, we applied external validation. By doing so, introducing novel datasets is essential to enable the validation on non-augmented data. Consequently, we introduced a dataset of 87 expert answers as sample solutions to have a broader spectrum for external validation. The results obtained from external validation significantly diverge from those obtained through internal validation. Precisely, models exclusively trained on chatbot responses demonstrated near-perfect classifications for other chatbot-generated content but struggled with analyzing real-world data (Table 9, Fig. 4: models 3, 4, and 5). This suggests that training an ML model solely on chatbot-generated responses is inadequate for accurately classifying real-world data. Furthermore, the model trained on paraphrased responses outperformed those trained on chatbot responses, albeit ranking as the least accurate among the remaining ones (Table 9, Fig. 4: model 2). Thus, paraphrasing authentic student responses can be an initial step in augmenting datasets; nevertheless, it is crucial to carefully determine the optimal number of paraphrased samples to achieve the best results. Moreover, the model trained on authentic student responses showed a slight improvement as compared to its paraphrased counterpart (Table 9, Fig. 3: model 1). Further performance improvements are observed in the model trained on a combination of authentic student and paraphrased responses (Table 9, Fig. 4: model 6). Finally, the most robust performance comes from models integrating authentic and/or paraphrased responses with chatbot responses (Table 9, Fig. 4: models 7, 8, and 9). These models may generalize well to new contexts, particularly in scenarios where the reasoning types of organic chemistry students differ from the collected training data or evolve over a semester (Martin & Graulich, 2024b). Accordingly, these models are most applicable in large-scale or longitudinal assessments, where the nuances of students' reasoning might not be fully covered by real-world data but can be simulated through chatbot-generated content. The challenge of an ML model being only as accurate as its training data may be mitigated with these augmentation strategies as chatbot-generated responses can increase data heterogeneity.

7. Discussion

In this article, we investigated the impact of different text data augmentation approaches on ML model performance. Our findings reveal that text data augmentation improved the performance of an ML model utilized in postsecondary chemistry classes. Nonetheless, performance variations are evident among the different augmentation approaches (Tables 7–9; Fig. 4): The model trained on original student responses augmented with chatbot-generated data performed best

Table 8
Performance differences of ML models augmented with various text data augmentation approaches assessed by internal validation.

ID	Student Data	Rephrased Data	Bard Data	ChatGPT Data	Epochs	Learning Rate	Batch Size	Macro F_1 score	Accuracy	Cohen's κ	Weighted F_1 score	Macro F_1 score
								Val. Set	Test Set			
1	X				26	1e-5	2	0.8219	0.9102	0.9022	0.9063	0.8142
2		X			33	1e-5	2	0.8073	0.8797	0.8690	0.8762	0.7838
3			X		43	1e-5	16	0.9423	0.9458	0.9430	0.9457	0.9476
4				X	45	5e-6	2	0.9352	0.9631	0.9612	0.9627	0.9589
5			X	X	17	1e-5	16	0.9426	0.9578	0.9558	0.9577	0.9532
6	X	X			17	1e-5	2	0.9291	0.8949	0.8857	0.8940	0.8157
7	X		X	X	82	5e-6	16	0.8901	0.9400	0.9364	0.9398	0.9294
8		X	X	X	77	5e-6	8	0.8821	0.9197	0.9150	0.9190	0.9040
9	X	X	X	X	58	5e-6	2	0.9272	0.9131	0.9074	0.9116	0.8865

Table 9
Performance differences of ML models augmented with various text data augmentation approaches assessed by internal and external validation.

ID	Trained on …				Tested with …							
	Student Data	Rephrased Data	Bard Data	ChatGPT Data	Student Data	Rephrased Data	Chatbot Data	Sample Solution	Accuracy	Cohen's κ	Weighted F_1 score	Macro F_1 score
1	X				X				0.9102	0.9022	0.9063	0.8142
	X					X			0.8661	0.8546	0.8665	0.7735
	X						X		0.7926	0.7823	0.7775	0.7364
	X							X	0.8966	0.8846	0.8997	0.8365
2		X			X				0.8746	0.8633	0.8684	0.7260
		X				X			0.8797	0.8690	0.8762	0.7838
		X					X		0.7858	0.7753	0.7702	0.7235
		X						X	0.8851	0.8722	0.9024	0.7857
3			X		X				0.6095	0.5768	0.6033	0.5587
			X			X			0.5874	0.5539	0.5853	0.5572
			X				X		0.9186	0.9147	0.9188	0.9163
			X					X	0.7125	0.6807	0.7239	0.4993
4				X	X				0.5475	0.5123	0.5547	0.5013
				X		X			0.5712	0.5381	0.5769	0.5407
				X			X		0.9123	0.9081	0.9123	0.9066
				X				X	0.6897	0.6587	0.6718	0.4330
5			X	X	X				0.6288	0.5989	0.6375	0.5692
			X	X		X			0.6271	0.5989	0.6343	0.5697
			X	X			X		0.9578	0.9558	0.9577	0.9532
			X	X				X	0.7011	0.6698	0.6938	0.4741
6	X	X			X				0.9051	0.8968	0.9047	0.8201
	X	X				X			0.8847	0.8747	0.8835	0.8122
	X	X					X		0.7808	0.7701	0.7687	0.7296
	X	X						X	0.9195	0.9102	0.9318	0.8654
7	X		X	X	X				0.9136	0.9059	0.9123	0.8380
	X		X	X		X			0.8593	0.8474	0.8591	0.7923
	X		X	X			X		0.9663	0.9646	0.9661	0.9605
	X		X	X				X	0.9310	0.9228	0.9353	0.8501
8		X	X	X	X				0.8881	0.8781	0.8847	0.8256
		X	X	X		X			0.8898	0.8801	0.8876	0.8252
		X	X	X			X		0.9494	0.9469	0.9484	0.9323
		X	X	X				X	0.9310	0.9229	0.9394	0.8733
9	X	X	X	X	X				0.9085	0.9004	0.9049	0.8373
	X	X	X	X		X			0.8864	0.8766	0.8854	0.7912
	X	X	X	X			X		0.9444	0.9416	0.9433	0.9266
	X	X	X	X				X	0.9080	0.8973	0.9093	0.8456

compared to the other augmentation strategies in the first research question (Table 7: model 3). However, the models exclusively trained on chatbot-generated data demonstrated limited generalizability in real-world scenarios (Table 9: models 3, 4, and 5). Conversely, the models combining authentic and/or paraphrased responses with chatbot responses are the most performant across all validation conditions (Table 9: models 7, 8, and 9). For this reason, it seems helpful to augment ML models trained on student data with chatbot-generated data. This aligns with Fang et al.'s (2023) implication that integrating chatbot responses into training datasets can balance them, improving ML model performance. Here, paraphrasing student data may not even be necessary as the models trained on authentic and chatbot data already showed exceptional performance, so paraphrasing responses may not further increase data heterogeneity.

Regarding the paraphrased responses, models using them generally performed worse than those trained with authentic student responses. Paraphrasing student responses may have altered coding-relevant terms, potentially impacting human coding coherence and complicating machine scoring. Additionally, models trained on authentic and paraphrased data performed slightly poorer due to overfitting.

Surprisingly, the chatbots produced student-like reasoning with a high causality level and incorporated various electronic properties (Table 6; Appendix). Hence, the chatbots helped generate balanced datasets, which is why leveraging chatbots can address the issue of unbalanced data in education. This approach may be a resource-efficient alternative to manual data collection and annotation (Fang et al., 2023). However, the ML models only trained on chatbot responses struggled to classify authentic student responses. This finding suggests that some terms and phrases used by chatbots to explain the underlying processes of chemical phenomena differ from those of undergraduate chemistry students, which resonates with Watts et al.'s (2023b) conclusion that chatbots do not engage in chemistry-related reasoning to the same degree as students.

To further prove this assumption, a research assistant and the first author independently coded a subset of both 1000 student and 1000 chatbot responses in a binary way for correctness (inter-rater reliability: Cohen's $\kappa = 0.91$). Student responses showed 84% chemical correctness, whereas chatbot responses demonstrated only 59% correctness. This outcome aligns with previous studies highlighting that chatbots do not accurately answer chemistry-related assessment items (Emenike & Emenike, 2023; Fergus et al., 2023; Leon & Vidhani, 2023). Thus, the high performance of the ML model trained exclusively on chatbot responses in internal validation should be interpreted with caution.

In sum, our analysis indicates that leveraging chatbots for text data augmentation expanded the number of automatically evaluable rubric categories, created balanced datasets, and improved scoring accuracy. These findings correspond to prior research on text data augmentation, which found that LLMs like GPT can better augment data than more traditional approaches (e.g., Cochran et al., 2022, 2023; Dai et al., 2023; Fang et al., 2023; Møller et al., 2023; Shi et al., 2020; Ubani et al., 2023; Wu et al., 2019). Nevertheless, the effectiveness of text data augmentation is contingent upon various factors, including the inherent characteristics of authentic student responses, the number of synthetic responses generated by chatbots, the nature of the tasks students engage with, the instructional setting, and the available datasets (Fang et al., 2023). In detail, Fang et al. (2023) found that the benefits of text data augmentation were not uniform across two tasks and several agreement metrics. For their first task assigned in a science assessment, a

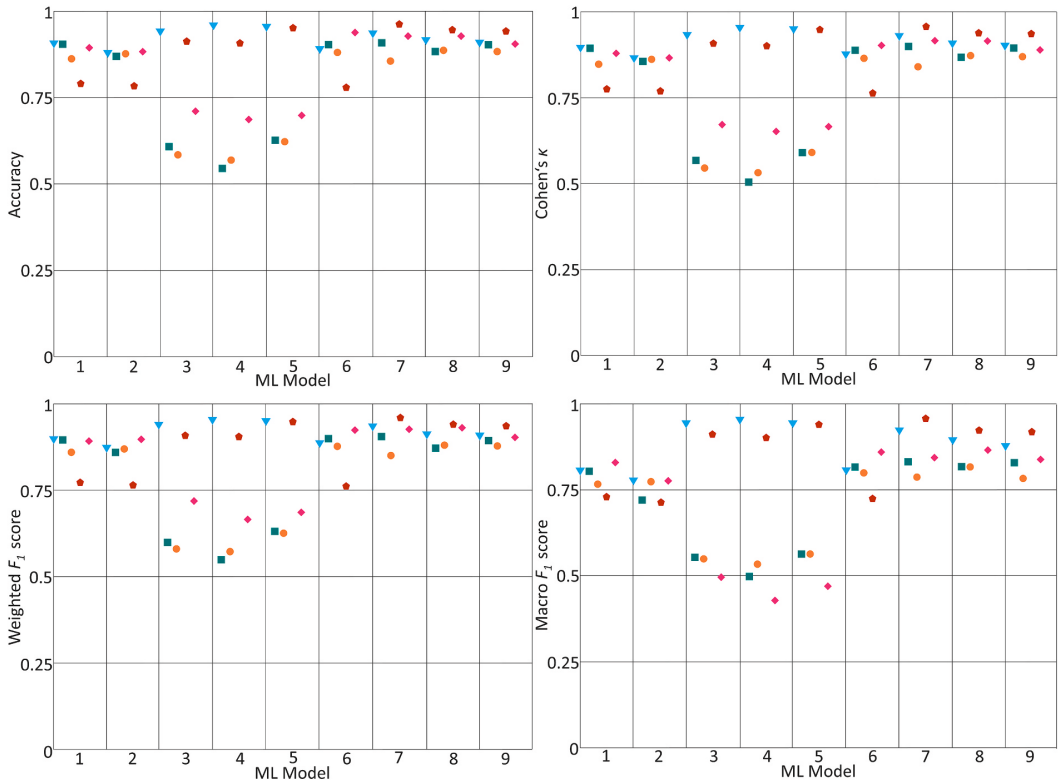


Fig. 4. Performance differences of ML models augmented with various text data augmentation approaches assessed by internal and external validation. *Note:* The ML model numbers (x-axis) correspond to the model IDs listed in Tables 8 and 9. The representation includes blue triangles denoting internal validation (Table 8). In addition, green squares indicate validation using student data, yellow circles using paraphrased data, red pentagons using chatbot data, and pink diamonds using sample solutions (Table 9).

substantial improvement in their ML model's precision, recall, and F_1 score was observed, particularly when augmenting the authentic student responses with only 5% of the GPT-4-generated responses. Significant improvements with data augmentation were also observed for their second task; however, the performance gains reached a saturation point by using 40% of GPT-4-generated data. Notably, authentic student data augmentation appeared more beneficial for certain metrics in the first task than GPT-4 data augmentation, while both augmentation types performed comparably in the second task (Fang et al., 2023). Accordingly, the effectiveness of text data augmentation appears to be context-dependent.

8. Conclusions and implications

As outlined in the Motivation Section, automating the identification of four rubric categories was not possible for us in previous analyses due to the limited number of sample responses in those categories (Fig. 3). Consequently, we had to employ a simplified model in our prior studies that only distinguished between 20 categories, which demonstrated that the accuracy of ML is contingent on the quality of the underlying data. Using data augmentation strategies helped us overcome the challenge posed by the limited sample responses in the four mentioned categories. In particular, text data augmentation using, among others, chatbots significantly improved the performance of our ML model since chatbots can partly mimic the characteristics of student reasoning in organic chemistry.

Precisely, we augmented text data by collecting additional student data, using chatbots to paraphrase students' reasoning, and prompting chatbots to generate synthetic responses. Through internal validation, we determined that using chatbot responses as input data for ML model training yielded the best machine-human score agreements. However, these models could not generalize to real-world datasets as evidenced by the low agreement metrics in external validation (Fig. 4). By contrast, incorporating authentic and/or paraphrased responses with chatbot responses demonstrated the highest performance across all datasets. Hence, a time-efficient strategy for developing ML models with high generalizability is using student data to initially train the algorithm and address imbalances in minority categories subsequently by incorporating chatbot responses.

The findings of our study hold various methodological and pedagogical implications. We prompted chatbots to generate synthetic responses. As prompts, we used the interpretations of a clustering output that grouped student responses based on their reasoning complexity into distinct groups (Martin et al., 2023). The so-identified reasoning types helped ensure validity in generating synthetic responses, so the clustering technique was useful for prompt engineering. Following this, clustering student reasoning can ease the identification of recurring characteristics, which could then be used to prompt chatbots to generate synthetic responses consistently.

Another methodological implication is that relying solely on internal validation, i.e., using a test dataset with characteristics similar to the training and validation sets, may not allow for examining a model's generalizability. In our analysis, only external validation showed that the models exclusively trained on chatbot responses were limited in accurately classifying authentic student responses. Accordingly, it is advisable to employ external validation to evaluate model

generalizability (Varshney, 2022).

Moreover, a still-existing limitation of our ML model is that it assesses structural and energetic reasoning about chemical reactions as distinct categories (Fig. 3). This distinction stems from our data sample, where students often discussed structural and energetic aspects in isolation. While this distinction between structural and energetic accounts is commonly observed (e.g., Bodé et al., 2019; Caspari et al., 2018; Deng & Flynn, 2021; Eckhard et al., 2022; Pölloth et al., 2023), our ML model is currently unable to detect connections between those categories in novel contexts. Looking ahead, leveraging the output of chatbots in future analyses presents the opportunity to introduce responses that bridge structural and energetic categories, allowing for a more integrated analysis of students' reasoning about reaction mechanisms.

Besides, implementing data augmentation strategies may also contribute to developing ML models that evaluate student reasoning with comparable accuracy across different languages. For instance, Martin and Graulich (2024a) compared an English-specific and German-specific deep learning architecture, revealing a 5% decrease in accuracy for the latter. This performance decline is concerning as it implies that the ML model is biased toward the language used, resulting in poorer evaluations of student reasoning in German. Incorporating data augmentation strategies may mitigate such language-related discrepancies, ensuring that ML models maintain their consistency across diverse linguistic contexts.

Taken together, extending the number of automatically evaluable categories allowed the analysis of more nuances in students' chemistry-related reasoning as the four previously challenging categories could be included in the rubric. This augmented analytical capacity could be essential in education as it may lead to more targeted guidance in better scalable assessments, showing the need for establishing highly accurate, fine-grained ML models by navigating the data frontier.

9. Limitations

The ability of chatbots like ChatGPT and Bard to handle scientific queries has been significantly advanced within a relatively short period since their launch (Bubeck et al., 2023). As chatbots progress, the analysis presented herein can only reflect their current capabilities. Due to the rapid pace of technological development in AI, future research could examine how generative AI proceeds to impact data augmentation strategies.

As LLMs like GPT, PaLM2, and BERT are trained on massive amounts of unfiltered language data, they may inadvertently adopt human biases (Caliskan et al., 2017; Kliegr et al., 2021), which can potentially perpetuate discriminatory practices against marginalized groups in science assessments (Cheuk, 2021). Additionally, chatbots tend to hallucinate, meaning they generate factually incorrect responses (Bang et al., 2023). Given these concerns, human experts may continue to be important in evaluating the quality of chatbot responses.

To generate synthetic responses, we employed a zero-shot learning strategy, where we prompted the chatbots to judge the plausibility of alternative reaction products without providing examples of well-crafted arguments. However, integrating a few examples into the generative AI for reference, known as few-shot learning, could improve the quality of the synthetic responses. Notably, LLMs trained with few-shot learning outperformed those trained with zero-shot learning in automatically scoring student-written responses in science assessments (Lee et al., 2024). Hence, using few-shot learning for generating synthetic responses might lead to further improvements in model performance.

Ultimately, although we tried to gather a heterogeneous dataset of undergraduate students' reasoning about organic chemistry reaction mechanisms (cf., Lieber et al. 2022a; Watts et al., 2023a), the extent of transferability of our findings may vary across subjects or educational levels.

Statements on ethics and open data

The Institutional Review Board (IRB) for human subject research approved data collection and secondary data usage. The collection of the original student data at the university in the Northeastern United States was approved by the IRB under Protocol STUDY00001480. A Memorandum of Understanding was established between Nicole Graulich and the Shultz Research Group to access and use student data collected at the large, research-intensive university in the Midwestern United States. Data were collected without identifying or sensitive information, ensuring participant anonymity. Assessments were integrated into normal educational practices. The de-identified data were solely used for ML model training. Informed consent was obtained from all participants and their privacy rights were strictly observed. Students were informed that participation was voluntary and they could withdraw from the study. The data used in this study is available from the corresponding author upon reasonable request.

CRediT authorship contribution statement

Paul P. Martin: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Nicole Graulich:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This publication is part of the first author's doctoral thesis (Dr. rer. nat.) at the Faculty of Biology and Chemistry, Justus-Liebig-University Giessen, Germany. We thank Peter Wulff and David Kranz for their assistance in implementing the ML analysis. Moreover, we thank Leonie Lieber, Ira Caspari-Gnann, and Krenare Ibraj for sharing their research instrument and data. Similarly, we thank Amber Dood, Field Watts, Ginger Shultz, and their research team for providing access to their Writing-to-Learn data. We are also grateful to Nathan Ackroyd and Brett McCollum for permitting us to collect data at their institution. Our further appreciation goes to Lukas Neuser and Felix Blödtner for paraphrasing students' reasoning and generating synthetic responses, as well as to all members of the Graulich group for engaging in productive discussions. Paul P. Martin thanks the Verband der Chemischen Industrie (German Chemical Industry Association) for supporting him with the Kekulé Fellowship. Special acknowledgment is given to all students at the three institutions participating in the studies.

Appendix

Prompts used to generate synthetic responses with chatbots. The prompts started with: "Please, build an argument about the plausibility of the given reaction by …". The shown examples reason about the plausibility of product 1.4 in Williamson ether synthesis (Fig. 2). The listed IDs refer to the category numbers in Fig. 3.

ID	Prompt	Student Response	ChatGPT Response
1	Redescribing the displayed reaction without referring to chemical concepts	The cyclic structure cannot form. Without an alkoxide, the cyclic substitution could not occur.	The chlorine atom in 4-chloro-butanol is replaced, and a series of chemical transformations occur.
2	Identifying leaving groups	The S_N2 reaction mechanism is plausible. The leaving group is stable and was bonded to a primary carbon.	The chlorine atom, being a good leaving group, can depart from the molecule, facilitating the formation of a new bond between carbon and oxygen atoms.
3	Estimating leaving group quality based on base strength	The Cl group leaves as a result of the cyclic formation because it is a good leaving group based on its characteristics as a weak base and halide ion.	Chloride ions are stable and relatively weak bases, which makes them good leaving groups.
4	Explaining how orbital interactions impact substrate stability or leaving group departure	*Not present in the data*	During the reaction, orbital interactions in the transition state between the filled p-orbitals facilitate efficient departure of the chloride ion.
5	Focusing on charges as a surface-level indicator of stability	The Cl^- is a good leaving group because it can stabilize the negative charge and is a stable ion.	Chloride ions are relatively stable because the negative charge is dispersed over the atomic surface.
6	Estimating leaving group quality based on electronegativity	The chlorine is a good leaving group. It is very electronegative, so it can be rather stable when it is in ionic form.	Chlorine is very electronegative, and when the chlorine atom leaves as a chloride ion, it retains a stable configuration with a full valence shell.
7	Estimating leaving group quality based on polarizability	Chloride is a good leaving group. Chloride is a weak base and electronegative, making it well able to stabilize a negative charge.	Chloride ions are relatively stable and can accommodate the negative charge well due to their high polarizability.
8	Identifying nucleophiles and electrophiles	O^- acts as a nucleophile and attacks the alpha carbon due to nucleophilic properties.	The reaction is supported by the presence of a suitable nucleophile, hydroxide ion, and an electrophile, carbon in 4-chloro-butanol.
9	Weighing electrophilic and acidic or nucleophilic and basic properties	Hydroxide is more likely to act as a base than it is to do S_N2 on the primary carbon with the chlorine. Hydroxide is a better base than it is a nucleophile.	The oxygen atom in the hydroxide ion is nucleophilic. Simultaneously, the hydroxide ion can act as a base by abstracting the acidic hydrogen from the substrate.
10	Weighing electrophilic and acidic or nucleophilic and basic properties by considering steric effects	The OH^- can take the proton from the OH on the other molecule, then this negatively charged O can attack the carbon attached to the Cl and eject the leaving group. This would be a proton transfer followed by a nucleophilic attack without steric hindrance.	Hydroxide ions are both nucleophilic and basic. In this case, steric hindrance is not present around the electrophilic carbon, which increases the accessibility of the nucleophile to the reaction site.
11	Identifying electrophiles and nucleophiles based on charges	After deprotonation of alcohol, O^- can act as a nucleophile. Negative charges make good nucleophiles.	The reaction is supported by the electrostatic interaction between the nucleophile and electrophile. The negatively charged hydroxide ion is attracted to the positively charged carbon.
12	Identifying electrophiles and nucleophiles based on electron density	The chlorine atom creates an electrophilic C4-atom due to electron-withdrawing effects. The C4-atom is prone to nucleophilic attack from alkoxide.	The carbon atom becomes electrophilic due to the withdrawal of electron density toward the chlorine.
13	Identifying electrophiles and nucleophiles based on electron density influenced by resonance, inductive effects, or hyperconjugation	The product could be reached by an intramolecular S_N2 reaction between the deprotonated OH and the electrophilic carbon in the C–Cl bond. The C of the C–Cl bond is electrophilic due to the inductive effect of the adjacent electronegative chlorine, so it is vulnerable to nucleophilic attack. The deprotonated anion could then attack this electrophilic carbon.	The hydroxide ion cannot exhibit resonance, increasing its nucleophilic character. The chlorine atom is electron-withdrawing, increasing the electrophilic character of the carbon atom.
14	Identifying acids and bases	OH^- can deprotonate the H from the HO group. The H on the OH group is acidic.	A strong base like hydroxide is more likely to promote the removal of a proton from the substrate. In this case, 4-chloro-butanol undergoes deprotonation by hydroxide ions.
15	Citing pK_a-values for identifying the strongest acid or base	The alcohol on the alkane can be easily deprotonated. Its pK_a is lower than the pK_a of OH^-.	The pK_a value for the hydroxyl group in alcohols, including 4-chloro-butanol, is typically around 16–18. Water molecules have a pK_a value of around 14.
16	Quantifying chemical equilibrium based on pK_a-values	Hydroxide will deprotonate the hydroxyl. The pK_a of hydroxyl is similar to that of hydroxide and they will, thus, exist in a proton transfer equilibrium. When the alkoxide ion is formed, it will react with the nearby electrophile.	The pK_a of the acidic proton in 4-chloro-butanol is higher than the pK_a of water, so the equilibrium will favor the deprotonated form and the subsequent formation of tetrahydrofuran (THF).
17	Invoking the Brønsted or Lewis acid-base theory based on charges	The deprotonation of the alcohol group will occur because of the negatively charged base.	In this reaction, 4-chloro-butanol can be considered a Lewis acid, accepting a pair of electrons.
18	Applying electronic effects to infer acid or base strength	OH^- is most likely to deprotonate the alcohol group. This hydrogen is more acidic than others in the molecule because it is attached to a more electronegative atom.	The presence of the electronegative chlorine atom adjacent to the hydroxyl group induces electron withdrawal, making the acidic proton more positively polarized.
19	Applying electronic effects such as resonance, electron-donating or electron-withdrawing effects, or hyperconjugation to infer acid or base strength	OH^- deprotonates the other OH group due to its high electronegativity. O^- from after the deprotonation will attack the carbon with the leaving group Cl. O^- is very reactive with the extra electron pair and the carbon with electron-withdrawing Cl is a very good target to attack.	The electron-withdrawing effect of chlorine enhances the acidity of the proton in 4-chloro-butanol. The attack by the alcoholate ion is facilitated by its electron-donating character, along with hyperconjugation.
20	Attributing energy levels to molecules	It forms stable products. The Cl^- is more stable than the OH^-, H_2O is more stable than OH^-, and a membered ring with oxygen is relatively stable.	The reaction is supported by the formation of tetrahydrofuran from 4-chloro-butanol and hydroxide ions, involving the creation of more stable bonds in the product.
21	Estimating entropic changes based on the number of reactants and products	This reaction is entropically favorable because 3 products and 2 reactants.	The conversion of 4-chloro-butanol and hydroxide ions to tetrahydrofuran likely involves a reduction in entropy as fewer molecules are present in the product compared to the reactants.
22	Comparing the reaction rate of different reaction types	Intramolecular reactions are faster than intermolecular ones. The rate is dependent on the concentration of substrate. When the reactants are connected, there is	Acid-base reactions, especially those involving strong bases like hydroxide ions, are known to occur very fast.

(*continued on next page*)

(*continued*)

ID	Prompt	Student Response	ChatGPT Response
		essentially limitless substrate and, thus, this reaction can take place quite quickly.	
23	Identifying resonance stabilization or charge localization	The charge is on the most electronegative element. Chlorine is more electronegative than oxygen, so it is the most stable with a negative charge. Leaving it on either oxygen would be less stable because the charge is more localized.	In the reaction between 4-chloro-butanol and hydroxide ions, it is important to note that resonance stabilization of the leaving group is not possible.
24	Locating and explaining resonance stabilization	The OH^- could take the hydrogen from the O on the other side. While both are OH, O is also connected to carbon and can sustain its electronegativity well due to its resonance, whereas the OH^- is looking to be more stable and needs another bind so it can take the hydrogen off that oxygen.	Resonance stabilization typically occurs when a negative charge can be delocalized over multiple atoms through the sharing of electrons. In this case, the chlorine atom is directly bonded to the carbon without any adjacent atoms with available p-orbitals for resonance delocalization.

References

Alasadi, E. A., & Baiz, C. R. (2023). Generative AI in education and research: Opportunities, concerns, and solutions. *Journal of Chemical Education, 100*(8), 2965–2971. https://doi.org/10.1021/acs.jchemed.3c00323

Anderson, P., Anson, C. M., Gonyea, R. M., & Paine, C. (2015). The contributions of writing to learning and development: Results from a large-scale multi-institutional study. *Research in the Teaching of English, 50*(2), 199–235. http://www.jstor.org/stable/24890033

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., … Wu, Y. (2023). PaLM 2: Technical report. *arXiv preprint.* arXiv:2305.10403. https://doi.org/10.48550/arXiv.2305.10403

Arnold, K. M., Umanath, S., Thio, K., Reilly, W. B., McDaniel, M. A., & Marsh, E. J. (2017). Understanding the cognitive processes involved in writing to learn. *Journal of Experimental Psychology: Applied, 23*(2), 115–127. https://doi.org/10.1037/xap0000119

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint.* arXiv:2302.04023. https://doi.org/10.48550/arXiv.2302.04023

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research, 74*(1), 29–58. https://doi.org/10.3102/00346543074001029

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2022). A survey on data augmentation for text classification. *ACM Computing Surveys, 55*(7), Article 146. https://doi.org/10.1145/3544558

Bishop, C. M. (2006). *Pattern recognition and machine learning* (1st ed.). Springer.

Bodé, N. E., Deng, J. M., & Flynn, A. B. (2019). Getting past the rules and to the WHY: Causal mechanistic arguments when judging the plausibility of organic reaction mechanisms. *Journal of Chemical Education, 96*(6), 1068–1082. https://doi.org/10.1021/acs.jchemed.8b00719

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (33rd ed., pp. 1877–1901). Curran Associates, Inc. https://doi.org/10.48550/arXiv.2005.14165

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint.* arXiv:2303.12712. https://doi.org/10.48550/arXiv.2303.12712

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Carle, M. S., El Issa, R. J., Pilote, N., & Flynn, A. B. (2021). *Ten essential delocalization learning outcomes: How well are they achieved? ChemRxiv preprint.*. https://doi.org/10.26434/chemrxiv-2021-c12b9-v2

Caspari, I., Kranz, D., & Graulich, N. (2018). Resolving the complexity of organic chemistry students' reasoning through the lens of a mechanistic framework. *Chemistry Education Research and Practice, 19*(4), 1117–1141. https://doi.org/10.1039/C8RP00131F

Cheuk, T. (2021). Can AI be racist? Color-evasiveness in the application of machine learning to science assessments. *Science Education, 105*(5), 825–836. https://doi.org/10.1002/sce.21671

Clark, T. M. (2023). Investigating the use of an artificial intelligence chatbot with general chemistry exam questions. *Journal of Chemical Education, 100*(5), 1905–1916. https://doi.org/10.1021/acs.jchemed.3c00027

Clark, T. M., Anderson, E., Dickson-Karn, N. M., Soltanirad, C., & Tafini, N. (2023). Comparing the performance of college chemistry students with ChatGPT for calculations involving acids and bases. *Journal of Chemical Education, 100*(10), 3934–3944. https://doi.org/10.1021/acs.jchemed.3c00500

Cochran, K., Cohn, C., Hutchins, N., Biswas, G., & Hastings, P. (2022). Improving automated evaluation of formative assessments with text data augmentation. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial intelligence in education. AIED 2022. Lecture notes in computer science* (23rd ed., pp. 390–401). Cham: Springer. https://doi.org/10.1007/978-3-031-11644-5_32

Cochran, K., Cohn, C., Rouet, J. F., & Hastings, P. (2023). Improving automated evaluation of student text responses using GPT-3.5 for text data augmentation. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial intelligence in education. AIED 2023. Lecture notes in computer science* (24th ed., pp. 217–228). Cham: Springer. https://doi.org/10.1007/978-3-031-36272-9_18

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology, 32*(3), 444–452. https://doi.org/10.1007/s10956-023-10039-y

Dahlkemper, M. N., Lahme, S. Z., & Klein, P. (2023). How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT. *Physical Review Physics Education Research, 19*(1), Article 010142. https://doi.org/10.1103/PhysRevPhysEducRes.19.010142

Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., Liu, N., Li, S., Zhu, D., Cai, H., Sun, L., Li, Q., Shen, D., Liu, T., & Li, X. (2023). AugGPT: Leveraging ChatGPT for text data augmentation. *arXiv preprint.* arXiv:2302.13007. https://doi.org/10.48550/arXiv.2302.13007

Darden, L. (2002). Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward/backward chaining. *Philosophy of Science, 69*(S3), S354–S365. https://doi.org/10.1086/341858

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education, 162*, Article 104094. https://doi.org/10.1016/j.compedu.2020.104094

Deng, J. M., Carle, M. S., & Flynn, A. B. (2023). Students' reasoning in chemistry arguments and designing resources using constructive alignment. In N. Graulich, & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (1st ed., pp. 74–89). The Royal Society of Chemistry. https://doi.org/10.1039/9781839167782-00074

Deng, J. M., & Flynn, A. B. (2021). Reasoning, granularity, and comparisons in students' arguments on two organic chemistry items. *Chemistry Education: Research and Practice, 22*(3), 749–771. https://doi.org/10.1039/D0RP00320D

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-Training of deep bidirectional transformers for language understanding. *arXiv preprint.* arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Eckhard, J., Rodemer, M., Langner, A., Bernholt, S., & Graulich, N. (2022). Let's frame it differently – analysis of instructors' mechanistic explanations. *Chemistry Education Research and Practice, 23*(1), 78–99. https://doi.org/10.1039/D1RP00064K

Emenike, M. E., & Emenike, B. U. (2023). Was this title generated by ChatGPT? Considerations for artificial intelligence text-generation software programs for chemists and chemistry educators. *Journal of Chemical Education, 100*(4), 1413–1418. https://doi.org/10.1021/acs.jchemed.3c00063

Exintaris, B., Karunaratne, N., & Yuriev, E. (2023). Metacognition and critical thinking: Using ChatGPT-generated responses as prompts for critique in a problem-solving workshop (SMARTCHEMPer). *Journal of Chemical Education, 100*(8), 2972–2980. https://doi.org/10.1021/acs.jchemed.3c00481

Fang, L., Lee, G.-G., & Zhai, X. (2023). Using GPT-4 to augment unbalanced data for automatic scoring. *arXiv preprint.* arXiv:2310.18365. https://doi.org/10.48550/arXiv.2310.18365

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A survey of data augmentation approaches for NLP. *arXiv preprint.* arXiv:2105.03075. https://doi.org/10.48550/arXiv.2105.03075

Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating academic answers generated using ChatGPT. *Journal of Chemical Education, 100*(4), 1672–1675. https://doi.org/10.1021/acs.jchemed.3c00087

Finkenstaedt-Quinn, S. A., Halim, A. S., Chambers, T. G., Moon, A., Goldman, R. S., Gere, A. R., & Shultz, G. V. (2017). Investigation of the influence of a writing-to-learn assignment on student understanding of polymer properties. *Journal of Chemical Education, 94*(11), 1610–1617. https://doi.org/10.1021/acs.jchemed.7b00363

Finkenstaedt-Quinn, S. A., Petterson, M., Gere, A., & Shultz, G. V. (2021a). Praxis of writing-to-learn: A model for the design and propagation of writing-to-learn in STEM. *Journal of Chemical Education, 98*(5), 1548–1555. https://doi.org/10.1021/acs.jchemed.0c01482

Finkenstaedt-Quinn, S. A., Polakowski, N., Gunderson, B., Shultz, G. V., & Gere, A. R. (2021b). Utilizing peer review and revision in STEM to support the development of

conceptual knowledge through writing. *Written Communication, 38*(3), 351–379. https://doi.org/10.1177/07410883211006038
Gerard, L. F., Matuk, C., McElhaney, K., & Linn, M. C. (2015). Automated, adaptive guidance for K-12 education. *Educational Research Review, 15*, 41–58. https://doi.org/10.1016/j.edurev.2015.04.001
Gere, A. R., Limlamai, N., Wilson, E., MacDougall Saylor, K., & Pugh, R. (2019). Writing and conceptual learning in science: An analysis of assignments. *Written Communication, 36*(1), 99–135. https://doi.org/10.1177/0741088318804820
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* (1st ed.). MIT Press.
Google. (2023). Google Bard/Gemini [Computer program]. https://www.gemini.com/eu
Gunel, M., Hand, B., & Prain, V. (2007). Writing for learning in science: A secondary analysis of six studies. *International Journal of Science and Mathematics Education, 5*, 615–637. https://doi.org/10.1007/s10763-007-9082-y
Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python [Computer program]. https://spacy.io/
Humphry, T., & Fuller, A. L. (2023). Potential ChatGPT use in undergraduate chemistry laboratories. *Journal of Chemical Education, 100*(4), 1434–1436. https://doi.org/10.1021/acs.jchemed.3c00006
Kaldaras, L., & Haudek, K. C. (2022). Validation of automated scoring for learning progression-aligned Next Generation Science Standards performance assessments. *Frontiers in Education, 7*, Article 968289. https://doi.org/10.3389/feduc.2022.968289
Kieser, F., Wulff, P., Kuhn, J., & Küchemann, S. (2023). Educational data augmentation in physics education research using ChatGPT. *Physical Review Physics Education Research, 19*(2), Article 020150. https://doi.org/10.1103/PhysRevPhysEducRes.19.020150
Kliegr, T., Bahník, Š., & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence, 295*, Article 103458. https://doi.org/10.1016/j.artint.2021.103458
Kortemeyer, G. (2023a). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research, 19*(1), Article 010132. https://doi.org/10.1103/PhysRevPhysEducRes.19.010132
Kortemeyer, G. (2023b). Toward AI grading of student problem solutions in introductory physics: A feasibility study. *Physical Review Physics Education Research, 19*(2), Article 020163. https://doi.org/10.1103/PhysRevPhysEducRes.19.020163
Kubsch, M., Krist, C., & Rosenberg, J. M. (2023). Distributing epistemic functions and tasks—a framework for augmenting human analytic power with machine learning in science education research. *Journal of Research in Science Teaching, 60*(2), 423–447. https://doi.org/10.1002/tea.21803
Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E., & Kuhn, J. (2023). Can ChatGPT support prospective teachers in physics task development? *Physical Review Physics Education Research, 19*(2), Article 020128. https://doi.org/10.1103/PhysRevPhysEducRes.19.020128
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310
Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence, 6*, Article 100213. https://doi.org/10.1016/j.caeai.2024.100213
Lee, G.-G., & Zhai, X. (2024). Using ChatGPT for science learning: A study on pre-service teachers' lesson planning. *arXiv preprint.* arXiv:2402.01674. https://doi.org/10.48550/arXiv.2402.01674
Leon, A. J., & Vidhani, D. (2023). ChatGPT needs a chemistry tutor too. *Journal of Chemical Education, 100*(10), 3859–3865. https://doi.org/10.1021/acs.jchemed.3c00288
Li, B., Hou, Y., & Che, W. (2022). Data augmentation approaches in natural language processing: A survey. *AI Open, 3*, 71–90. https://doi.org/10.1016/j.aiopen.2022.03.001
Li, T., Reigh, E., He, P., & Adah Miller, E. (2023). Can we and should we use artificial intelligence for formative assessment in science? *Journal of Research in Science Teaching, 60*(6), 1385–1389. https://doi.org/10.1002/tea.21867
Lieber, L. S., & Graulich, N. (2020). Thinking in alternatives—A task design for challenging students' problem-solving approaches in organic chemistry. *Journal of Chemical Education, 97*(10), 3731–3738. https://doi.org/10.1021/acs.jchemed.0c00248
Lieber, L. S., & Graulich, N. (2022). Investigating students' argumentation when judging the plausibility of alternative reaction pathways in organic chemistry. *Chemistry Education Research and Practice, 23*(1), 38–53. https://doi.org/10.1039/D1RP00145K
Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022a). Closing the gap of organic chemistry students' performance with an adaptive scaffold for argumentation patterns. *Chemistry Education Research and Practice, 23*(4), 811–828. https://doi.org/10.1039/D2RP00016D
Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022b). Students' individual needs matter: A training to adaptively address students' argumentation skills in organic chemistry. *Journal of Chemical Education, 99*(7), 2754–2761. https://doi.org/10.1021/acs.jchemed.2c00213
Martin, P. P., & Graulich, N. (2023a). Mit Computerintelligenz Lehr-Lern-Prozesse gestalten: Anwendung von ChatGPT zur Unterrichtsplanung [Shaping teaching with artificial intelligence: Using ChatGPT for lesson-planning]. In T. Wilke, & I. Rubner (Eds.), *Digitalisation in Chemistry Education: Vol. Workshop. Friedrich-Schiller-Universität Jena, Institut für Anorganische und Analytische Chemie. DiCE-Tagung.* https://doi.org/10.22032/dbt.59418
Martin, P. P., & Graulich, N. (2023b). When a machine detects student reasoning: a review of machine learning-based formative assessment of mechanistic reasoning. *Chemistry Education Research and Practice, 24*(2), 407–427. https://doi.org/10.1039/D2RP00287
Martin, P. P., & Graulich, N. (2024a). Beyond language barriers: allowing multiple languages in postsecondary chemistry classes through multilingual machine learning. *Journal of Science Education and Technology, 33*(3), 333–348. https://doi.org/10.1007/s10956-023-10087-4
Martin, P. P., & Graulich, N. (2024b). *Lehre in der organischen Chemie individualisieren [Individualized teaching in organic chemistry]. Nachrichten aus der Chemie, 72*(3), 8–11. https://doi.org/10.1002/nadc.20244141003
Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2023). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching, Early View Article*, 1–36. https://doi.org/10.1002/tea.21903
Mathew, A., Amudha, P., & Sivakumari, S. (2021). Deep learning techniques: An overview. In A. E. Hassanien, R. Bhatnagar, & A. Darwish (Eds.), *Advanced machine learning technologies and applications: Proceedings of AMLTA 2020* (1141st ed., pp. 599–608). Springer. https://doi.org/10.1007/978-981-15-3383-9_54
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint.* arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781
Mitchell, T. M. (1997). *Machine learning* (1st ed.). McGraw-Hill
Møller, A. G., Dalsgaard, J. A., Pera, A., & Aiello, L. M. (2023). The parrot dilemma: Human-labeled vs. LLM-augmented data in classification tasks. *arXiv preprint.* arXiv:2304.13861. https://doi.org/10.48550/arXiv.2304.13861
Moreira, P., Marzabal, A., & Talanquer, V. (2019). Using a mechanistic framework to characterise chemistry students' reasoning in written explanations. *Chemistry Education: Research and Practice, 20*(1), 120–131. https://doi.org/10.1039/C8RP00159F
Noyes, K., McKay, R. L., Neumann, M., Haudek, K. C., & Cooper, M. M. (2020). Developing computer resources to automate analysis of students' explanations of London dispersion forces. *Journal of Chemical Education, 97*(11), 3923–3936. https://doi.org/10.1021/acs.jchemed.0c00445
OpenAI. (2022). ChatGPT [Computer program]. https://chat.openai.com/
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintale, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (32nd ed., pp. 8024–8035). Curran Associates, Inc. https://dl.acm.org/doi/abs/10.5555/3454287.3455008
Pölloth, B., Diekemper, D., & Schwarzer, S. (2023). What resources do high school students activate to link energetic and structural changes in chemical reactions? – A qualitative study. *Chemistry Education: Research and Practice, 24*(4), 1153–1173. https://doi.org/10.1039/D3RP00068K
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 1–24. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf
Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education, 92*(3), 499–525. https://doi.org/10.1002/sce.20264
Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint.* arXiv:1511.06709. https://doi.org/10.48550/arXiv.1511.06709
Sevian, H., & Talanquer, V. (2014). Rethinking chemistry: A learning progression on chemical thinking. *Chemistry Education: Research and Practice, 15*(1), 10–23. https://doi.org/10.1039/C3RP00111C
Shi, L., Liu, D., Liu, G., & Meng, K. (2020). AUG-BERT: An efficient data augmentation algorithm for text classification. In Q. Liang, W. Wang, X. Liu, Z. Na, M. Jia, & B. Zhang (Eds.), *Communications, signal processing, and systems, CSPS 2019. Lecture notes in electrical engineering* (8th ed., pp. 2191–2198). Springer. https://doi.org/10.1007/978-981-13-9409-6_266
Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), Article 60. https://doi.org/10.1186/s40537-019-0197-0
Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of Big Data, 8*(1), Article 101. https://doi.org/10.1186/s40537-021-00492-0
Soo, K. W. (2019). *The role of granularity in causal learning.* University of Pittsburgh.
Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., … Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence, 3*, Article 100075. https://doi.org/10.1016/j.caeai.2022.100075
Taher Pilehvar, M., & Camacho-Collados, J. (2020). Embeddings in natural language processing: Theory and advances in vector representations of meaning (1st ed.). *Morgan & Claypool Publishers*.
Talanquer, V. (2023). Interview with the chatbot: How does it reason? *Journal of Chemical Education, 100*(8), 2821–2824. https://doi.org/10.1021/acs.jchemed.3c00472
Tynjälä, P., Mason, L., & Lonka, K. (2001). *Writing as a learning tool: Integrating theory and practice* (1st ed.). Netherlands: Springer.
Tyson, J. (2023). Shortcomings of ChatGPT. *Journal of Chemical Education, 100*(8), 3098–3101. https://doi.org/10.1021/acs.jchemed.3c00361
Ubani, S., Polat, S. O., & Nielsen, R. (2023). ZeroShotDataAug: Generating and augmenting training data with ChatGPT. *arXiv preprint.* arXiv:2304.14334. https://doi.org/10.48550/arXiv.2304.14334
Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual* (2nd ed.). CreateSpace.
Varshney, K. R. (2022). *Trustworthy machine learning* (1st ed.) Independently published.

Wang, C., Liu, X., Wang, L., Sun, Y., & Zhang, H. (2021). Automated scoring of Chinese grades 7-9 students' competence in interpreting and arguing from evidence. *Journal of Science Education and Technology, 30*(2), 269–282. https://doi.org/10.1007/s10956-020-09859-z

Watts, F. M., Dood, A. J., & Shultz, G. V. (2023a). Developing machine learning models for automated analysis of organic chemistry students' written descriptions of organic reaction mechanisms. In N. Graulich, & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (1st ed., pp. 285–303). The Royal Society of Chemistry. https://doi.org/10.1039/9781839167782-00285

Watts, F. M., Dood, A. J., Shultz, G. V., & Rodriguez, J.-M. G. (2023b). Comparing student and generative artificial intelligence chatbot responses to organic chemistry writing-to-learn assignments. *Journal of Chemical Education, 100*(10), 3806–3817. https://doi.org/10.1021/acs.jchemed.3c00664

Watts, F. M., Park, G. Y., Petterson, M. N., & Shultz, G. V. (2022). Considering alternative reaction mechanisms: Students' use of multiple representations to reason about mechanisms for a writing-to-learn assignment. *Chemistry Education: Research and Practice, 23*(2), 486–507. https://doi.org/10.1039/D1RP00301A

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint.* arXiv:1901.11196. https://doi.org/10.48550/arXiv.1901.11196

Weinrich, M. L., & Talanquer, V. (2016). Mapping students' modes of reasoning when thinking about chemical reactions used to make a desired product. *Chemistry Education: Research and Practice, 17*(2), 394–406. https://doi.org/10.1039/C5RP00208G

West, C. G. (2023a). Advances in apparent conceptual physics reasoning in GPT-4. *arXiv preprint.* arXiv:2303.17012v3. https://doi.org/10.48550/arXiv.2303.17012

West, C. G. (2023b). *AI and the FCI: Can ChatGPT project an understanding of introductory physics? arXiv preprint.* arXiv:2303.01067 https://doi.org/10.48550/arXiv.2303.01067

West, J. K., Franz, J. L., Hein, S. M., Leverentz-Culp, H. R., Mauser, J. F., Ruff, E. F., & Zemke, J. M. (2023). An analysis of AI-generated laboratory reports across the chemistry curriculum and student perceptions of ChatGPT. *Journal of Chemical Education, 100*(11), 4351–4359. https://doi.org/10.1021/acs.jchemed.3c00581

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wu, X., Lv, S., Zang, L., Han, J., & Hu, S. (2019). Conditional BERT contextual augmentation. In J. M. F. Rodrigues, P. J. S. Cardoso, J. Monteiro, R. Lam, V. V. Krzhizhanovskaya, M. H. Lees, J. J. Dongarra, & P. M. A. Sloot (Eds.), *Computational science. ICCS 2019. Lecture notes in computer science* (19th ed., pp. 84–95). Cham: Springer. https://doi.org/10.1007/978-3-030-22747-0_7

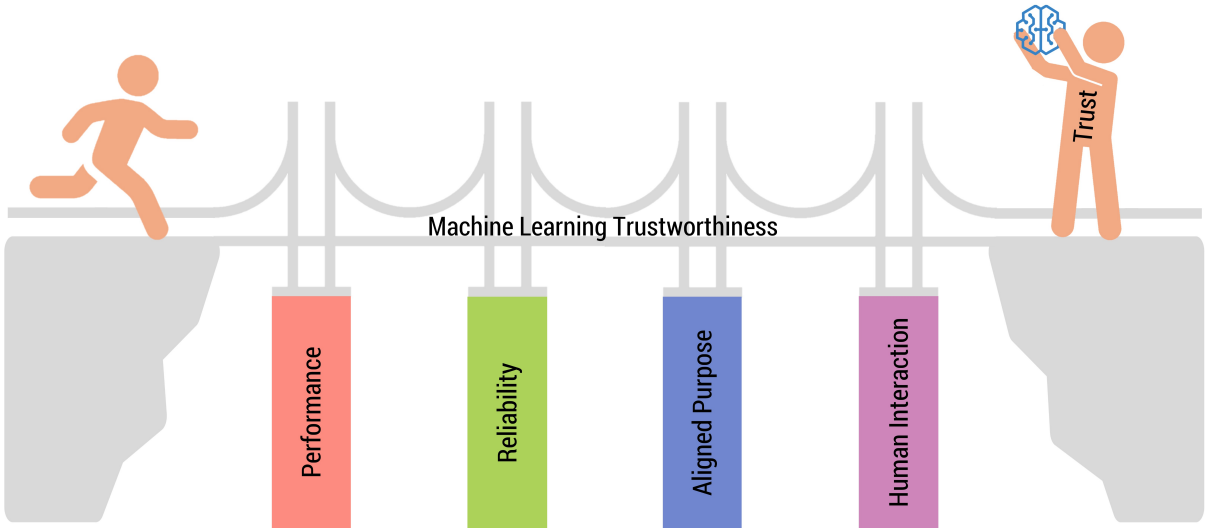
Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education, 58*(3), Article 035027. https://doi.org/10.1088/1361-6552/acc5cf

Zhai, X., Haudek, K. C., Shi, L., Nehm, R. H., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430–1459. https://doi.org/10.1002/tea.21658

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education, 56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757

6 | To what extent can we trust a machine? Establishing trustworthiness guidelines to evaluate machine learning models across chemistry contexts

Chapter 6 presents a research manuscript by Martin, Dood, Watts, Shultz, and Graulich that is currently under review at the *Journal of Research in Science Teaching* but has not yet been accepted for publication.



RESEARCH ARTICLE

To what extent can we trust a machine? Establishing trustworthiness guidelines to evaluate machine learning models across chemistry contexts

Paul P. Martin[1], Amber J. Dood[2], Field M. Watts[3], Ginger V. Shultz[2], Nicole Graulich[1,*]

Paul P. Martin and Amber J. Dood contributed equally to this manuscript.

Abstract

Students' learning in science is heterogeneous. To address this heterogeneity, educators can use machine learning (ML) models to automatically score students' open-ended responses and deliver adaptive support tailored to individual needs. However, there are key concerns that ML models fail to adequately reflect diverse student needs, which may impact the extent to which educators can rely on an ML model's suggestions. Therefore, guidelines to establish trustworthiness in educational ML applications have become more relevant to ensure their valid implementation. Inspired by trustworthy naturalistic inquiry, this article discusses four attributes of trustworthy ML-driven automated scoring: *performance*, *reliability*, *aligned purpose*, and *human interaction*. After reviewing the implementation of these attributes in ML-assisted science education research, we applied them to a case study looking at the automated analysis of undergraduate organic chemistry students' mechanistic reasoning to showcase how the proposed attributes can be used in a real-world scenario. Specifically, we shared and compared data collected from multiple institutions, contexts, tasks, and reaction mechanisms to investigate the extent of trustworthiness of two ML models. Findings indicate that including data from various contexts advances ML model generalizability at the cost of their high performance in the original context, highlighting trade-offs between different attributes of trustworthy ML. Moreover, high overall performance can conceal scoring issues for specific subtasks, potentially compromising reliable scoring in all contexts. Together, the proposed attributes showcase that simply reporting performance metrics is inadequate for establishing trustworthy ML; instead, we must investigate the consistency of ML models across contexts, their alignment with the purpose of the respective assessment, and their level of interpretability.

KEYWORDS artificial intelligence, machine learning, trustworthiness, formative assessment, organic chemistry

Affiliation

[1]Institute of Chemistry Education, Justus-Liebig-University, Heinrich-Buff-Ring 17, 35392 Giessen, Germany
[2]University of Michigan, Department of Chemistry, 930 North University Ave, Ann Arbor, MI 48109, USA
[3]University of Wisconsin—Milwaukee, Department of Chemistry & Biochemistry, 3210 N. Cramer Street, Milwaukee, WI 53211, USA
[*]E-Mail corresponding author: Nicole.Graulich@didaktik.chemie.uni-giessen.de

1 | Introduction

Learning across science domains requires students to integrate crosscutting concepts, disciplinary core ideas, and science practices to understand the natural and engineered world (NRC, 2012). The inherent complexity of these dimensions poses challenges for students in science learning, leading to substantial heterogeneity. For instance, Crandell et al. (2019) identified six areas of heterogeneity in undergraduate chemistry. They observed that the complexity of students' reasoning about acid-base reactions varied across (1) course types, (2) curricula, (3) chemical contexts, (4) within a single cohort, (5) between cohorts, and (6) over time, demonstrating the dynamic and multifaceted nature of learning in chemistry. Further studies also highlight substantial heterogeneity in students' learning outcomes across science subjects (e.g., Braun & Graulich, 2024; Deng & Flynn, 2021; Sevian & Talanquer, 2014; Southard et al., 2017).

Students' heterogeneity challenges educators to design formative assessments that adapt to diverse student needs (Dumont & Ready, 2023). One potential solution is to leverage machine learning (ML) for automated scoring of open-ended responses to enable adaptive learning approaches, such as adaptive and continuous formative assessments (Figure 1) (Swiecki et al., 2022; Xie et al., 2019). Several studies have implemented such ML-assisted adaptive support to address student heterogeneity, for instance, when supporting students' mechanistic reasoning in biology or chemistry (e.g., Ariely et al., 2024; Dood et al., 2020b). By offering tailored support that meets the unique needs of individual students, such ML-assisted adaptivity promises to reduce student heterogeneity (e.g., Kulik & Fletcher, 2016; Viberg et al., 2024).

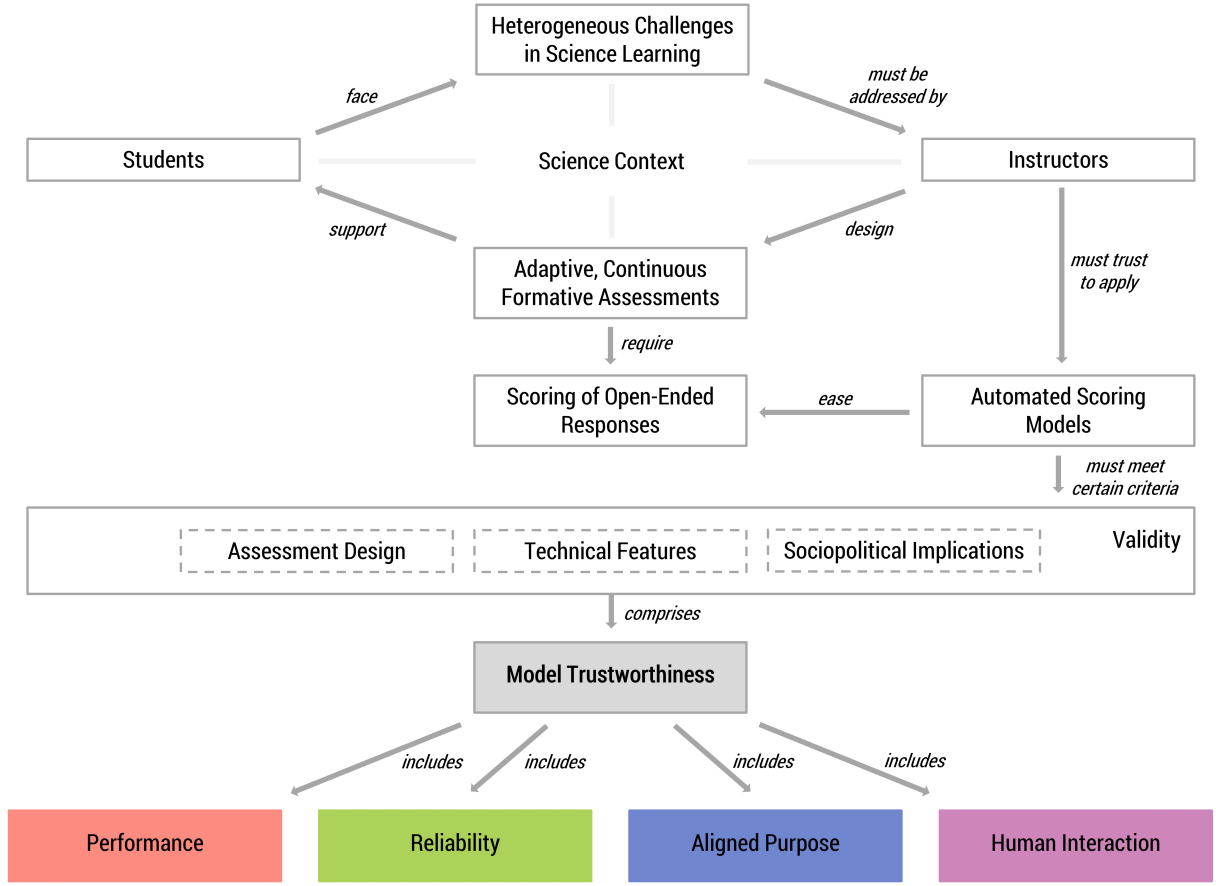


FIGURE 1 Using machine learning-driven automated scoring to support students' science learning in adaptive and continuous formative assessments.

However, implementing ML-driven automated scoring for adaptive learning in real-world scenarios implies a transition from the initial development and training of ML models to their practical deployment. When developing ML models, the focus is on defining the tasks to be performed, gathering data, and training a model to replicate existing patterns or behaviors (Krist & Kubsch, 2023). While the mere development of a model using past data does not affect student learning, deploying ML models as predictive tools to evaluate new data in future applications may impact student learning. For instance, ML-assisted adaptivity may influence students' attitudes, motivation, and learning progression. This transition from simply describing the status quo to proactively shaping the future educational landscape highlights a growing emphasis on designing, evaluating, and deploying ML-driven automated scoring with a focus on validity (Bejar et al., 2016; Clauser et al., 2002; Rupp, 2018; Williamson et al., 2012).

Prior research has conceptualized guidelines that ML-driven automated scoring must meet to ensure the validity of educational assessments (Figure 1). These guidelines can be broadly categorized into three areas: (1) assessment design (Kubsch et al., 2022; Martin & Graulich, 2023; Mislevy, Steinberg, et al., 2003; Pellegrino et al., 2016; Rupp, 2018; Williamson et al., 2012), (2) technical features (Bejar et al., 2016; Rupp, 2018; Williamson et al., 2012), and (3) sociopolitical implications (Cheuk, 2021; Grimm, Steegh, Çolakoğlu, et al., 2023; Krist & Kubsch, 2023). These areas help gather the evidence necessary to construct a validity argument for using ML-driven automated scoring in educational assessments (Kane, 2006).

Examining the validity of ML-driven automated scoring with a focus on assessment design can be approached in line with evidence-centered design (e.g., Mislevy, Almond, et al., 2003; Mislevy, Steinberg, et al., 2003; Pellegrino et al., 2016). Kubsch et al. (2022) and Martin and Graulich (2023) have extensively explored this aspect of validity in ML-driven automated scoring. First, assessment design should begin with a thorough analysis of the cognitive domain targeted by the assessment. The specific competencies students are expected to acquire within this domain must be clearly defined. By considering the characteristics of the target population, assessment designers must then create the tasks to gather the necessary evidence of competence acquisition, identify the elements in students' responses that show this evidence, and operationalize how to analyze the evidence. This process involves developing scoring rubrics, training human raters, and establishing automated scoring models. Notably, the evidence-centered design approach to educational assessments is an iterative, non-linear process.

Besides these considerations for assessment design, there are also technical threats when deploying ML-driven automated scoring in educational assessments. These include, among others, measurement errors or inefficient hyperparameter choices (Krist & Kubsch, 2023). Due to the inherent complexity of ML models like deep neural networks and the opaque nature of their decision-making, many advanced ML models are also black boxes humans cannot interpret, resulting in limited opportunities to understand or explain the ML output (Grimm, Steegh, Kubsch, et al., 2023). Challenges in communicating the ML model's decisions to stakeholders, such as policymakers, administrators, parents, or students, can hinder instructors from adopting ML in their teaching practices. Beyond that, students could game an ML model by exploiting the model's vulnerabilities to achieve a desired outcome in unintended ways (e.g., Yik et al., 2021).

Finally, sociopolitical implications must be addressed when implementing ML-driven automated scoring in educational assessments. Here, ML model development must avoid biases like privileging individuals with majority-driven

characteristics or using classification variables that directly or indirectly perpetuate discrimination against marginalized groups (Cheuk, 2021; Grimm, Steegh, Çolakoğlu, et al., 2023; T. Li et al., 2023). Further sociopolitical threats in ML model development include, among others, prioritizing certain stakeholder perspectives, over- or underrepresenting specific groups in the data, using historically or socially biased data, or lacking institutions accountable for algorithmic harm (Grimm, Steegh, Kubsch, et al., 2023; Krist & Kubsch, 2023). If access to necessary technology is uneven, using ML in science education can also exacerbate socioeconomic inequalities (Holstein & Doroudi, 2022).

As indicated by the preceding paragraphs, prior research has broadly conceptualized the design, evaluation, and deployment of ML-driven automated scoring with a focus on validity. Despite various conceptual frameworks, applying the proposed guidelines remains nontrivial (Rupp, 2018). Previous studies have predominantly implemented the proposed guidelines in the context of automated essay scoring for large-scale summative assessments, often emphasizing a conceptual approach that exceeds the practical level suggested in this work (Bejar et al., 2016; Rupp, 2018; Williamson et al., 2012). Also, automated essay grading is distinct from automated scoring of short-answer responses in science contexts due to the latter's emphasis on content details, rather than writing skills such as grammar, coherence, and style. Further, the way students reason in science disciplines largely depends on the domain, tasks, and contexts; this context-dependency may not be present in an automated analysis of students' writing skills. Consequently, using ML-driven automated scoring in science assessments poses unique challenges to the validity of ML models.

In this work, we start to address this gap by providing actionable guidelines for implementing aspects of validity within ML-driven automated scoring of science assessments. Our emphasis on ML-driven automated scoring stems from its ability to facilitate adaptive and continuous formative assessments that may offer individualized support to help students improve their science learning outcomes (Ariely et al., 2024; Dood et al., 2020b). In our analysis, we focus on the concept of trustworthiness, which integrates the three validity areas outlined above to equip users with guidelines needed to gather evidence that fosters appropriate trust in ML models (Figure 1). Our contribution to the literature is manifold: (1) We adopt a framework for the trustworthiness of ML models and map its attributes to the widely acknowledged criteria of trustworthy educational research as conceptualized in naturalistic inquiry. Specifically, the adopted framework focuses on four attributes of ML model trustworthiness: *performance*, *reliability*, *aligned purpose*, and *human interaction* (Figure 1). (2) We provide actionable guidelines to investigate the degree of trustworthiness of ML models applied for automated scoring of adaptive and continuous formative assessments that may support students' science learning. (3) The adopted framework integrates multiple attributes of ML model trustworthiness to consolidate the fragmented applications of ML trustworthiness in prior science education research. (4) We address novel requirements in ML research to facilitate human interaction with the output of ML models.

2 | Conceptual framework

Adaptive and continuous assessments that support students' science learning often require the scoring of open-ended responses—a process eased by ML-driven automated scoring. However, instructors must trust these ML models to fully leverage their potential in advancing students' learning. Thus, the social concept of trust is essential for establishing guidelines that allow researchers and educators to evaluate the extent of trustworthiness of ML-driven

automated scoring (e.g., Glikson & Woolley, 2020; Jacovi et al., 2021; Mehrotra et al., 2023; Thiebes et al., 2021). Generally, trust involves the relationship between a trustor and a trustee, where the trustor places confidence in the trustee. Although various conceptualizations of trust in artificial intelligence (AI) methodology exist, a definition from organizational management is often applied in AI research (e.g., Jacovi et al., 2021; Thiebes et al., 2021; Toreini et al., 2020; Varshney, 2022; Visser et al., 2023). This definition conceptualizes trust as the willingness of one party to be vulnerable to the actions of another party based on the belief that the latter will act in the best interests of the former (Mayer et al., 1995). Following that, establishing trust aims to increase the predictability of the trustee's behaviors, reduce uncertainty in social interactions, and minimize the risks associated with collaboration (Mayer et al., 1995).

Mayer et al.'s (1995) definition applied to AI suggests that trusting an AI system requires two essential prerequisites: On the one hand, users must consciously choose to be vulnerable to an AI system; on the other hand, users must be able to anticipate the AI's actions based on pattern recognition or AI literacy skills (Jacovi et al., 2021). Consequently, *vulnerability* and *anticipation* are key components of trust in AI methodology; trustworthiness can be conceptualized based on these components (Mayer et al., 1995). Regarding *vulnerability*, an AI system is labeled trustworthy if it is benevolent towards humans and adheres to socially acknowledged principles. Concerning *anticipation*, an AI system earns trustworthiness through consistently high performance across scenarios. This association between trust and trustworthiness emphasizes their interrelated yet distinct nature (Rempel et al., 1985). Trust is a human attitude, intention, or behavior, whereas AI systems are deemed trustworthy if they possess attributes that warrant trust (Ashoori & Weisz, 2019; Toreini et al., 2020). In simpler terms, trust relates to a user's mindset and actions while trustworthiness is a property of an AI system. This distinction between both concepts indicates that trustworthiness does not automatically lead to trust (Mayer et al., 1995). Even the most trustworthy ML model may not gain the trust of every person in every circumstance; conversely, individuals may—and often do—trust untrustworthy models (Chen et al., 2023).

Considering these and similar conceptualizations of trust and trustworthiness, numerous researchers have identified multiple characteristics of trustworthy AI as fundamental requirements for trust in AI methodology (Ashoori & Weisz, 2019; B. Li et al., 2023; Liu et al., 2022; Thiebes et al., 2021; Toreini et al., 2020). Although these frameworks differ in their specifics, they share many common attributes essential for establishing trust in AI systems. Commonly proposed attributes emphasize the agreement between human and machine scores, the generalizability of scores across tasks and assessments, the alignment of AI systems with societal decision-making, and the interpretability or explainability of AI decisions. These shared attributes are reflected in several conceptual frameworks, for example, in those proposed by Varshney (2022) and Williamson et al. (2012).

In this work, we draw upon the conceptual framework developed by Varshney (2022) that synthesizes commonly addressed attributes of trustworthy ML. Varshney (2022) defines four attributes of trustworthy ML: *performance*, *reliability*, *aligned purpose,* and *human interaction* (Figure 2). We adopt this framework because it resonates with the broadly recognized criteria of trustworthy educational research as conceptualized in naturalistic inquiry (Guba, 1981; Guba & Lincoln, 1981, 1982; Lincoln & Guba, 1985). Trustworthy naturalistic inquiry, mostly focusing on qualitative research, must fulfill four highly interrelated criteria to ensure the integrity of the research findings: *credibility*,

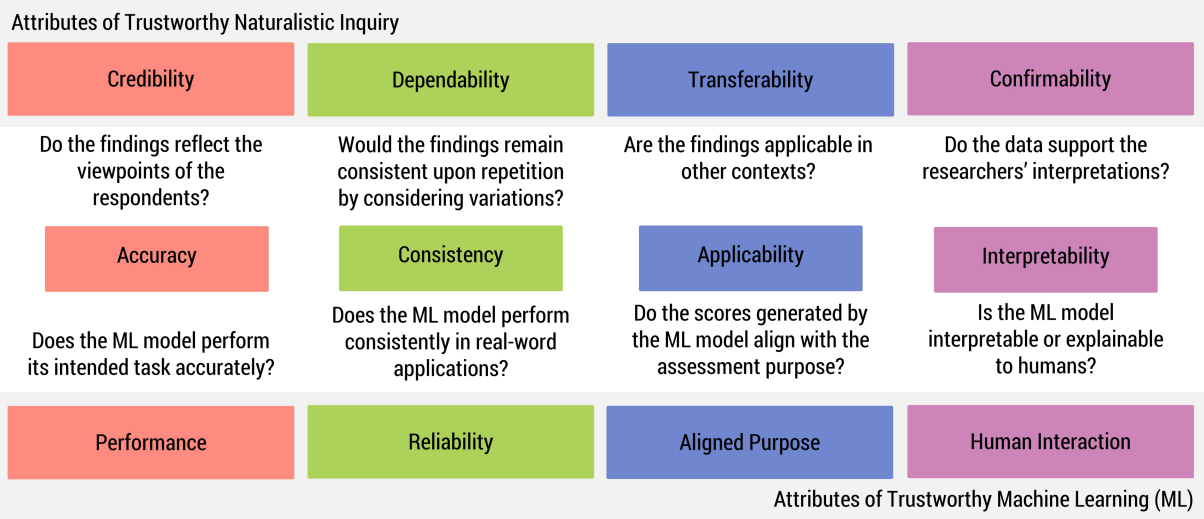


FIGURE 2 Shared attributes of trustworthy naturalistic inquiry and trustworthy machine learning.

dependability, *transferability*, and *confirmability* (Figure 2).

The subsequent paragraphs characterize the criteria of trustworthy naturalistic inquiry, drawing upon Guba's (1981) discussion. We link each criterion to the corresponding attribute of trustworthy ML to provide a familiar foundation for evaluating the extent of ML model trustworthiness. We include brief literature overviews on how trustworthy ML has already been implemented practically.

2.1 Accuracy

2.1.1 Credibility as an attribute of trustworthy naturalistic inquiry

Credibility addresses the extent to which research findings represent the participants' viewpoints. To establish *credibility*, qualitative researchers often seek feedback or validation from the participants of their study, a practice called member-checking. Using triangulation techniques such as methodological, data, investigator, or theoretical triangulation can also help cross-verify sources of evidence and reduce bias. A prolonged engagement—in essence, spending an extended period in the research setting—can further advance a profound familiarity with essential characteristics of the study context. Finally, storing documents, films, videotapes, audio recordings, pictures, and other unprocessed real-life materials during the research process can ensure that sufficient contextual evidence is available to validate the findings.

2.1.2 Performance as an attribute of trustworthy ML

Drawing from *credibility* in naturalistic inquiry, developing scoring models with acceptable *performance* is an essential attribute of trustworthy ML (Clauser et al., 2002; Rupp, 2018; Varshney, 2002; Williamson et al., 2012). For example, Williamson et al. (2012) define the *performance* of automated scoring models in terms of their agreement with human scores and emphasize several criteria: the quality of human scores as a basis for scoring model evaluation, the level of agreement between human and machine scores using metrics like quadratic-weighted κ or Pearson correlations, the degradation in agreement compared to human-human agreement, and standardized mean score differences. Rupp (2018) situates the *performance* of automated scoring models within a broader context and emphasizes key design decisions, including the selection of linguistic features for automated scoring, the ML model's design and structure, and the

approaches for model validation. Varshney (2022) uses *performance* to analyze how accurately an ML model accomplishes its intended tasks in the setting where it was initially developed based on the original input data, compared to human scorers. Since model *performance* depends on the consistency of human scores, researchers should iteratively apply, discuss, and revise a rubric until a satisfactory inter-rater agreement is reached (Watts & Finkenstaedt-Quinn, 2021). Then, the inter-rater agreement can serve as a benchmark against which the model's performance is critically evaluated.

Researchers routinely report ML model *performance*, including metrics like accuracy, precision, recall, Cohen's κ, F_1 scores, or Matthews correlation coefficient (Williamson et al., 2012; Zhai, Shi, et al., 2021; Zhai, Yin, et al., 2020). Deciding on which metrics and corresponding cut-off values to report depends on the nature of the task at hand; nonetheless, it is common to include Cohen's κ with a cut-off value of 0.7 when aiming to apply an ML model in formative science assessments (Williamson et al., 2012). In a meta-analysis, Zhai, Shi, et al. (2021) identified the main categories impacting ML model *performance*. The most impactful determinants of model *performance* were assessment external features, such as the assessment format and the subject domain, and technical features, such as the applied algorithm. These findings could assist in ML model optimization.

Taken together, the *performance* of ML models is analogous to the attribute of *credibility* in naturalistic inquiry due to their common interest in the accuracy of the results. Specifically, *performance* involves the accuracy of ML models in making classifications while *credibility* refers to the accuracy of the study's findings. Besides, ML model developers aim to improve the *performance* of their models by refining algorithms and validating classifications against the ground truth—just as qualitative researchers seek to advance *credibility* through rigorous methodologies, member-checking, and triangulation.

2.2 Consistency

2.2.1 Dependability as an attribute of trustworthy naturalistic inquiry

As a second attribute of trustworthy naturalistic inquiry, *dependability* relates to the consistency of the research findings when a study is repeated. Following this, *dependability* indicates the extent to which another researcher could replicate the study and attain similar results, accounting for the unique characteristics of the original context and the specific decisions made by the original researchers. To strengthen *dependability*, researchers should chronologically document the activities within their research process in a *dependability audit trail*. This audit trail includes all steps in data collection and preparation, justifications of methodological choices, and every personal decision. Furthermore, stepwise replication—a splitting technique where researchers independently investigate data in different groups—can facilitate the development of theories appropriately explaining the data. Ultimately, peer debriefing—that is, sharing data and interpretations among researchers—allows others to evaluate the analysis process and validate the adequacy of the researchers' data interpretations.

2.2.2 Reliability as an attribute of trustworthy ML

Similar to *dependability* in naturalistic inquiry, *reliability* in ML model development and deployment corresponds to the consistency of ML models across real-world contexts. For example, Varshney (2022) and Williamson et al. (2012) define the *reliability* of ML models as the extent to which model performance remains consistent across tasks and data sources. This

includes evaluating the model's alignment with human scores when applied to alternate assessment forms. Hence, a reliable ML model is expected to uphold performance across settings or when faced with new data (Rupp, 2018; Varshney, 2022; Williamson et al., 2012; Zhai, 2024; Zhai, Krajcik, et al., 2021).

Evaluating *reliability* can involve testing the ML model's performance across diverse datasets (Nehm et al., 2012; Yik et al., 2021), applying a model at different institutions or with heterogeneous populations (Dood et al., 2022; Ha et al., 2011; Noyes et al., 2020; Wulff et al., 2023), establishing standardized procedures for reporting dataset information (Gebru et al., 2021), explaining programming code for publication (Thimbleby, 2003), or sharing models and data among researchers (Bond et al., 2024). For example, Nehm et al. (2012) and Yik et al. (2021) employed two testing approaches to examine ML scoring *reliability*. First, they trained their ML model with student responses and tested it using similar ones; second, different sets of student responses were used for model training and testing. Despite both approaches yielding satisfactory Cohen's *κ* values, a decrease in scoring accuracy was observed in the latter approach. In contrast to using various datasets for ML model testing, Dood et al. (2022), Noyes et al. (2020), and Wulff et al. (2023) demonstrated a great generalizability of their ML models across institutions, disciplines, and courses. Conversely, Ha et al. (2011) observed a slight decrease in the performance of their traditional ML model for biological explanations across colleges, which was attributed to distinct language patterns, imbalanced datasets, misspellings, and informal vocabulary. Collectively, the reported studies assessed their ML model performance beyond agreement metrics, offering further information regarding ML trustworthiness.

The *reliability* of ML models is akin to the attribute of *dependability* in naturalistic inquiry in their shared emphasis on consistency. *Reliability* in ML model deployment includes demonstrating consistent performance across various settings. Similarly, *dependability* in naturalistic inquiry addresses the consistency of the research findings upon replication of the study.

2.3 Applicability

2.3.1 Transferability as an attribute of trustworthy naturalistic inquiry

As the next attribute of trustworthy naturalistic inquiry, *transferability* addresses the extent to which a study's findings are applicable in other contexts by accounting for the unique characteristics of the original context. Unlike *dependability*, *transferability* focuses on the applicability, rather than the consistency, of research findings. Given that researchers might not know the locations interested in adopting the findings, it is crucial to offer *thick descriptions* by thoroughly characterizing the study context. For this reason, ensuring *transferability* involves more than just describing the respondents' statements. Naturalistic inquiry must provide a firsthand experience of the social context as well as of the values and emotions surrounding the subject to uncover detailed information about the setting and sample. In addition, selecting heterogeneous samples to maximize the breadth of the collected data also increases a study's *transferability*.

2.3.2 Aligned purpose as an attribute of trustworthy ML

Expanding on *transferability* in naturalistic inquiry, *aligned purpose* defines the degree to which ML models are applied in accordance with acknowledged values or societal wants, ensuring that the models fulfill their designated objectives in an ethically responsible manner (Holmes et al., 2022; Varshney, 2022). In educational contexts, the purpose of an assessment—what it is designed to measure and how its results will be

used—guides the requirements for ML model trustworthiness (Pellegrino et al., 2016; Williamson et al., 2012). For instance, automated scoring models intended for high-stakes decisions must adhere to more rigorous trustworthiness standards than those used for low-stakes purposes since high-stakes decisions carry significant consequences for various stakeholders (Bejar et al., 2016; Rupp, 2018; Williamson et al., 2012). Accordingly, the level of trustworthiness required for an ML model depends on the intended use of its results.

The importance of aligning the implementation of an ML model with its intended purpose is evident in the case of e-rater—an automated grading system developed by the Educational Testing Service to assist in scoring the written portion of the GRE and TOEFL exams. In its initial application, e-rater assigned higher scores to certain student demographics (Chodorow & Burstein, 2004; Lee et al., 2008; Ramineni et al., 2012; Ramineni & Williamson, 2018), indicating a misalignment of the ML model to responsibly grade student writing, especially given that the scores of GRE and TOEFL impact high-stakes decisions, such as university admissions. In response, the Educational Testing Service adopted a dual scoring approach, involving both e-rater and a human scorer for each essay (Ramineni & Williamson, 2018). If the discrepancy between the two scores is too high, the essay is flagged and a second human scorer is consulted. By integrating the ML model into a human-machine collaboration, this example illustrates how to align the purpose of an ML model with its outcomes in an iterative process.

Applying ML models in educational assessments requires that the trustworthiness of these models is aligned with the intended use of the assessment results, the implications for stakeholders, and the broader educational goals they seek to support (Krist & Kubsch, 2023; Williamson et al., 2012). The emphasis on applicability, which considers the social context of intended use, connects the *aligned purpose* of ML models to the criteria of *transferability* in naturalistic inquiry.

2.4 Interpretability

2.4.1 Confirmability as an attribute of trustworthy naturalistic inquiry

The last criterion of trustworthy naturalistic inquiry is *confirmability*, which is attained when the data interpretation is transparent to others. In contrast to the focus of *dependability* on the consistency and replicability of findings, *confirmability* allows peers to interpret the influence of potential biases and applied methodologies on research findings. Here, by practicing reflexivity, researchers can become more aware of a study's underlying epistemological assumptions, rationales, and biases. To mitigate potential biases, researchers must consciously set aside their predispositions during the research process to differentiate between the observations made and the interpretations derived. While doing so, researchers must provide a transparent account of the hypothesis, methods, and decisions in a *confirmability audit trail*, proving that each finding can be retraced and explained through all analysis stages.

2.4.2 Human interaction as an attribute of trustworthy ML

Since ML models make decisions that impact human lives, the interpretability and explainability of these decisions are of utmost importance (Khosravi et al., 2022; Lo Piano, 2020). Especially in educational contexts, instructors may seek some understanding of an ML model before relying on it for executing its intended tasks (Conati et al., 2018; Polak et al., 2022), such as delivering students timely, formative feedback (e.g., Watts et al., 2023). *Human interaction* focuses on the transparency and comprehensibility of an ML model's decision-making (Khosravi et al., 2022; Saltelli &

Funtowicz, 2014; Zhai, Krajcik, et al., 2021) and aims to ensure that users can understand how the model comes to its decisions (Zhai & Nehm, 2023). In this manner, Varshney (2022) defines *human interaction* as the extent to which the decisions of an ML model are accessible, that is, interpretable or explainable to humans. According to Williamson et al. (2012), a high degree of *human interaction* includes communicating the strengths, limitations, and intended uses of automated scoring models.

Although sometimes used interchangeably, there is a distinction between model interpretability and explainability (Gilpin et al., 2018; Lipton, 2018). Humans can *interpret* models without the need for additional tools. An instance of an interpretable algorithm is a regression model with known parameters, where it is straightforward to identify which features contribute to specific classifications. Several researchers applied interpretable ML techniques such as discriminant analysis or logistic regression to score students' open-ended responses, focusing on domains such as acid-base chemistry (Dood et al., 2018; Haudek et al., 2009, 2012), thermodynamics (Haudek et al., 2009; Prevost et al., 2012a, 2012b), or nucleophilic substitutions (Dood et al., 2020a). Applying discriminant analysis or logistic regression provides numerical markers, e.g., discriminant function coefficients, regression coefficients, or odds ratios, that indicate how much a lexical category contributes to the ML model output. These numerical markers make a model's decision-making interpretable as they reveal how specific word groups influence the classification into a certain category.

By contrast, other ML models are too complex for human interpretation and demand additional tools, often termed explainable AI, to provide insights into the model's decision-making (e.g., Arrieta et al., 2020; Khosravi et al., 2022; Miller, 2019). For example, deep learning—an advanced ML type—is typically more accurate than traditional ML algorithms (Goodfellow et al., 2016) but presents challenges for human interpretation. Hence, additional algorithms are required to explain model output, such as Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) or SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017). These post hoc explainability techniques can clarify how a model arrives at its classifications. For instance, Gombert et al. (2023) and Martin et al. (2024) relied on SHAP to explore the internal workings of their models. By doing so, the studies analyzed their models at different levels: Gombert et al. (2023) operated on a local level by calculating the importance scores of different tokens for an example sentence—that is, how specific sequences of letters impact the classification of an individual student response. This fine-grained approach helped them determine how the words of a response contributed to the ML model output. In contrast, Martin et al. (2024) worked globally by aggregating the importance scores for each rubric category across all student responses, extracting the words that impacted the ML model output most. This strategy enabled a holistic understanding of how different words influenced the model's classifications across all responses. Collectively, these studies show that ML decision-making can be partly explained to facilitate human-ML interaction (Khosravi et al., 2022).

While interpretable models are often preferred (Liu & Udell, 2020; Rudin, 2019; Saltelli & Funtowicz, 2014), the simplicity of such models can decrease performance for complex scoring tasks in science assessments (e.g., Gombert et al., 2023; Martin et al., 2024). Tradeoffs between performance, interpretability, and simplicity require careful consideration (Watson & Floridi, 2021). For example, if a model is used to make high-stakes decisions, it is essential to understand the decision-making because of the potential impact on human lives (D'Ignazio & Klein, 2020; Rudin, 2019). By contrast, opting for a less interpretable but more accurate model is

deemed acceptable in formative assessments, considering the low-stakes nature of the model's classifications (Rudin, 2019).

Together, *human interaction* in ML model development and deployment centers on ensuring transparency in ML decision-making to allow humans to interpret or explain ML model outputs. Additionally, *human interaction* questions whether ML and human decision-making are similar. These issues correspond to ensuring *confirmability* in naturalistic inquiry, where peers should be able to confirm researchers' interpretations based on data. Thus, *human interaction* and *confirmability* are committed to the critical validation of decision-making.

3 | Research questions

ML-driven automated scoring of open-ended responses can streamline the implementation of adaptive and continuous formative assessments that provide students with individualized learning experiences and may advance their learning. For such scoring to be trusted, it must meet certain validity criteria, including considerations related to assessment design, technical features, and sociopolitical implications (Figure 1). So far, this article introduced Varshney's (2022) actionable framework to investigate the *performance*, *reliability*, *aligned purpose*, and *human interaction* of ML models (Figure 2). A thorough evaluation of these attributes within a science education context has not yet been reported. However, the science context influences the requirements for automated analyses, which subsequently affects how trustworthiness is operationalized. For instance, when a fine-grained evaluation of students' open-ended reasoning over time is needed, it becomes essential for an ML model to demonstrate high *reliability*, particularly in terms of generalizability across contexts and heterogeneous responses. Therefore, applying the proposed attributes within a science-specific context is valuable for understanding how these attributes interact with students' reasoning in science.

In the following sections, we apply the proposed attributes within a case study in chemistry to offer explicit guidance for exploring the extent of ML model trustworthiness. Particularly, we evaluate two ML models designed to analyze undergraduate students' reasoning about reaction mechanisms in organic chemistry. The research questions (RQs) examined herein align with the four attributes of the discussed conceptual framework and can provide practitioners with insights on investigating and advancing the trustworthiness of their ML-driven automated scoring. We introduce various institutions, contexts, tasks, and mechanisms to provide multiple pieces of evidence that support certain trustworthiness attributes of our ML models (Figure 3).

1) **Performance:** How accurately can ML models evaluate organic chemistry students' reasoning about reaction mechanisms in the settings where the models were initially developed?
2) **Reliability:** How consistently can ML models evaluate organic chemistry students' reasoning across contexts and tasks?
3) **Aligned purpose:** To what extent is a context-independent ML model evaluating organic chemistry students' reasoning applicable across reaction mechanisms?
4) **Human interaction:** To what extent is the ML model's output explainable to humans?

4 | Settings of data collection

Organic chemistry involves identifying, describing, or analyzing mechanisms to explain, predict, or control the outcome of chemical reactions, which requires *mechanistic reasoning* about entities and activities (Goodwin, 2003, 2008). In organic chemistry, entities encompass

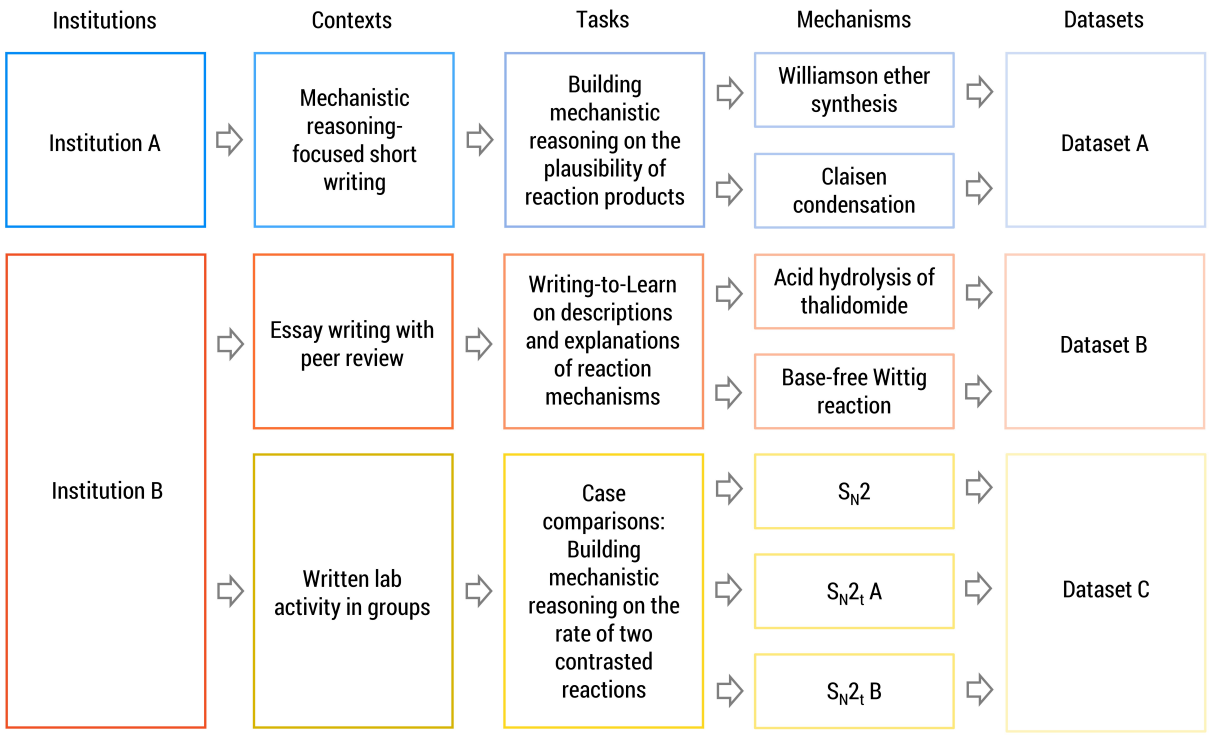


FIGURE 3 Overview of the institutions, contexts, tasks, and organic chemistry reaction mechanisms used for ML model training. *Note:* A summary of the different ML models trained within these settings is provided in Table 1. S_N2 = bimolecular nucleophilic substitution. S_N2_t = bimolecular nucleophilic substitution at a tertiary carbon atom.

electrons, atoms, ions, and molecules while activities involve electron movements leading to bonding changes. Automatically analyzing students' open-ended mechanistic reasoning can allow instructors to offer students individualized support within this organic chemistry practice (Martin & Graulich, 2023). However, to encourage instructors to adopt automated scoring models that support student learning, research must demonstrate that these models possess attributes that warrant trust. We shared data across two research groups to assess two ML models that automatically evaluate undergraduate organic chemistry students' mechanistic reasoning across two institutions, three contexts, three tasks, and seven organic chemistry reaction mechanisms.

Figure 3 provides an overview of the data collection settings described below. Authors PPM and NG initially used the mechanistic reasoning-focused short writing (hereafter referred to as *reasoning-focused short writing*) assigned at Institution A to develop their ML model (Section 4.1). Separately, authors AJD, FMW, and GVS used students' essays collected within the mechanistic reasoning-focused Writing-to-Learn (WTL) activity at Institution B to initially develop their ML model (Section 4.2). Subsequently, these two research groups shared data to investigate and increase their ML models' trustworthiness. Ultimately, both research groups included the dataset on the mechanistic reasoning-focused case comparison activity assigned at Institution B to boost ML model generalizability (Section 4.3).

4.1 Reasoning-focused short writing at Institution A (dataset A)

4.1.1 Setting of data collection at Institution A

To train an ML model that can automatically evaluate students' mechanistic reasoning on a reaction product's plausibility, authors PM and NG used data collected at a private, research-intensive, liberal arts university in the

Northeastern United States during April and May 2021. The Institutional Review Board approved data collection under protocol #STUDY00001480 and all students consented to participate. A total of 64 undergraduate students enrolled in Organic Chemistry II took part in the study, earning extra credit for completing the assignments.

The online-implemented tasks prompted students to reason about alternative reaction products. The two task sets were presented on two separate days within three weeks. For every task set, students sequentially judged the plausibility of four alternative reaction products based on the claim-evidence-reasoning model (McNeill et al., 2006). Students could construct more than one statement per product. The average length of the English-written responses was 27 words (*SD* = 12.7, *max* = 138, *min* = 5) while 1108 statements were collected. For a more detailed description of the study setting, refer to Lieber et al. (2022a).

4.1.2 Implementation of the research instrument on a product's plausibility

To assess and support students' mechanistic reasoning, we used an adaptive scaffold that prompted students to judge the plausibility of alternative reaction products for an intramolecular Williamson ether synthesis and a Claisen condensation (Lieber et al., 2022a, 2022b) (Figure 4). This instrument guided students to claim whether the shown product is plausible, evidence their claim based on chemical concepts, and establish a logical connection between the claim and evidence by considering electronic, steric, or energetic effects. Lieber et al. (2022a, 2022b) demonstrated that this instrument significantly improved students' mechanistic reasoning.

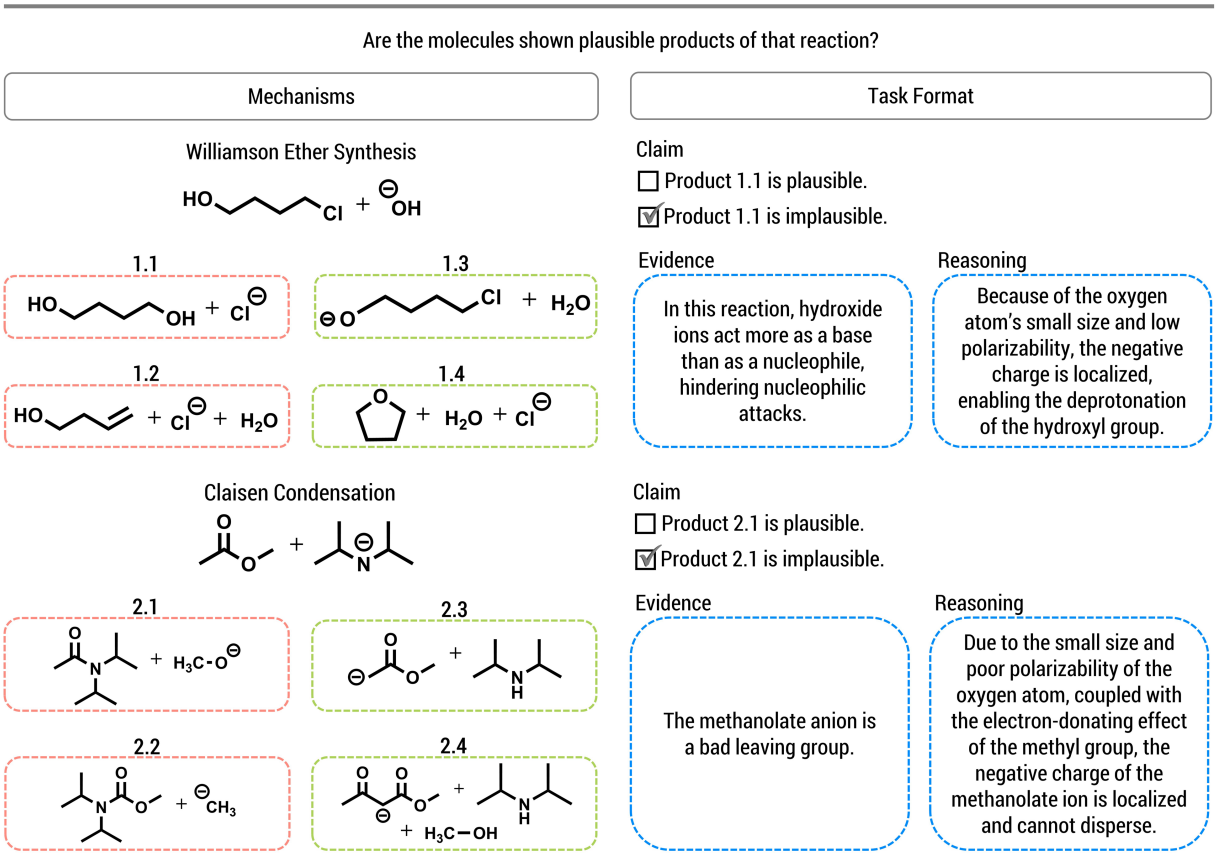


FIGURE 4 Reasoning-focused short writing on the plausibility of reaction products (dataset A). Plausible products are highlighted in green (1.3, 1.4, 2.3, 2.4) and implausible ones in red (1.1, 1.2, 2.1, 2.2). Data used for ML model training is highlighted in blue.

4.1.3 Rubric to score students' mechanistic reasoning on a product's plausibility

To evaluate students' mechanistic reasoning with ML, authors PPM and NG designed a holistic two-dimensional rubric drawing from the *modes of reasoning* (Sevian & Talanquer, 2014) and *levels of granularity* (Bodé et al., 2019; Deng & Flynn, 2021) (Figure 5). As a first dimension, Sevian and Talanquer's (2014) *modes of reasoning* classify the causality of students' mechanistic reasoning (Russ et al., 2008; Sevian & Talanquer, 2014). In our study, we applied three *modes of reasoning—descriptive*, *relational*, and *linear causal*—to assess the increasing causality in how students link the properties of entities or activities. As the second dimension, the *levels of granularity* pertain to the scalar levels at which students explain a phenomenon. In our study, we employed four *levels of granularity—structural*, *energetic*, *phenomenological*, and *electronic*—as used by Deng and Flynn (2021) and Deng et al. (2023) to classify the types of properties students integrate in their mechanistic reasoning. Further details on the rubric can be found in Martin and Graulich (2024a, 2024c) and Martin et al. (2024).

4.2 Essay writing with peer review at Institution B (dataset B)

4.2.1 Setting of data collection at Institution B

Authors AJD, FMW, and GVS collected data for ML model training within a WTL activity in a second-year introductory organic chemistry laboratory course at a large, research-intensive university in the Midwestern United States during the Winter 2020 semester. The Institutional Review Board approved data collection under protocol #HUM00079234 and all students whose data were analyzed in this study consented to participate.

The undergraduate organic chemistry students completed three WTL assignments, which included writing an initial draft, engaging with peer review, and revising the draft. For this study, we used students' final drafts of two assignments, one related to thalidomide hydrolysis and another related to a base-free Wittig reaction. While the assignments ranged from two to three pages and encompassed answers to multiple questions, we only analyzed sections focusing on students' explanations of reaction mechanisms. For machine scoring, students' English-written mechanistic reasoning was evaluated at a sentence level. The mean length of these sentences was 23 words (*SD* = 11.1, *max* = 80, *min* = 4), with 848 explanations

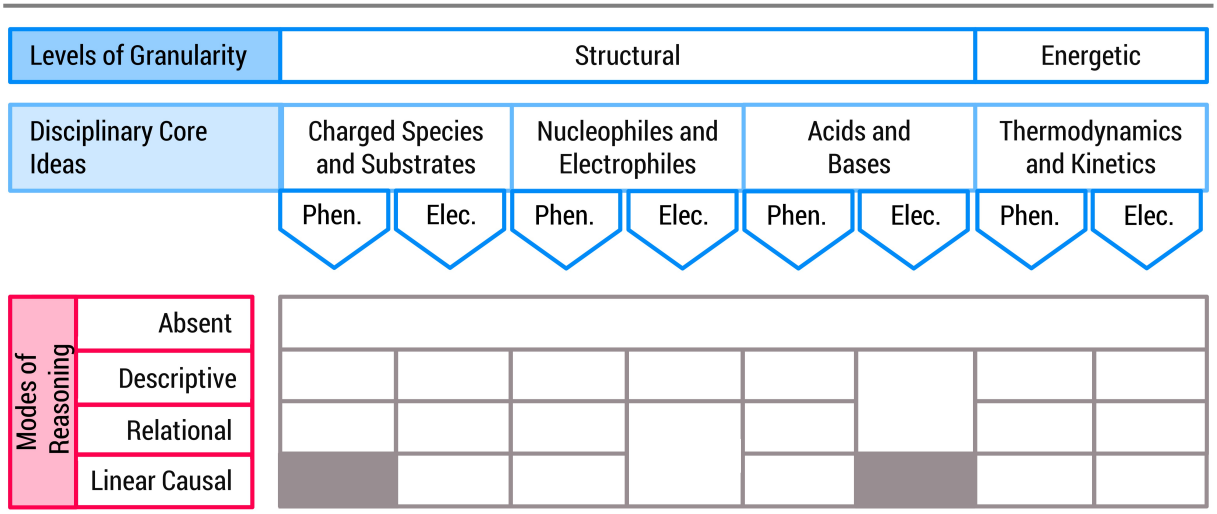


FIGURE 5 Holistic twenty-category rubric to score students' mechanistic reasoning on alternative reaction products (Martin et al., 2024).

analyzed herein. For additional information about the study setting, see Watts, Dood, et al. (2022).

4.2.2 Implementation of the research instrument in the WTL context

The research instrument included two WTL assignments that prompted students to apply their organic chemistry concept knowledge to describe and explain unfamiliar reaction mechanisms (Bangert-Drowns et al., 2004; Gere et al., 2019; Gunel et al., 2007): the acid hydrolysis of thalidomide and a base-free Wittig reaction (Figure 6). In doing so, the assignments incorporated peer review and revision with the support of writing fellows (Finkenstaedt-Quinn et al., 2021). First, students submitted initial drafts, followed by a mandatory peer review process during which students provided content-focused feedback to about three peers. After peer review, students revised their drafts based on the peer feedback. Finally, students had access to writing fellows—undergraduates who completed the course and underwent specific training to support peers in writing. The assignments were intended to assist students in expanding their concept knowledge and organizing their problem-solving approach (Finkenstadt-Quinn et al., 2023; Watts, Park, et al., 2022). In contrast to traditional writing assignments that primarily concentrate on the final draft, these assignments emphasized the process of writing as a learning tool (Tynjälä et al., 2001).

4.2.3 Rubric to score students' essays in the WTL context

To analyze students' final drafts of the two WTL assignments, authors AJD, FMW, and GVS applied Russ et al.'s (2008) mechanistic reasoning framework—rooted in the philosophy of science—to categorize the entities, activities, and properties addressed in students' mechanistic reasoning, along with the connections established between them. For machine scoring, this framework was used as outlined by Watts et al. (2020) to identify the presence or absence of ten mechanistic reasoning features (Figure 7): *reaction medium*, *connectivity*, *charges*, *stereochemistry*, *electron movement*, *non-electronic mechanism*, *bond breaking and making*, *implicit properties*, *stereochemistry formation*, and *electronic causal reasoning.*

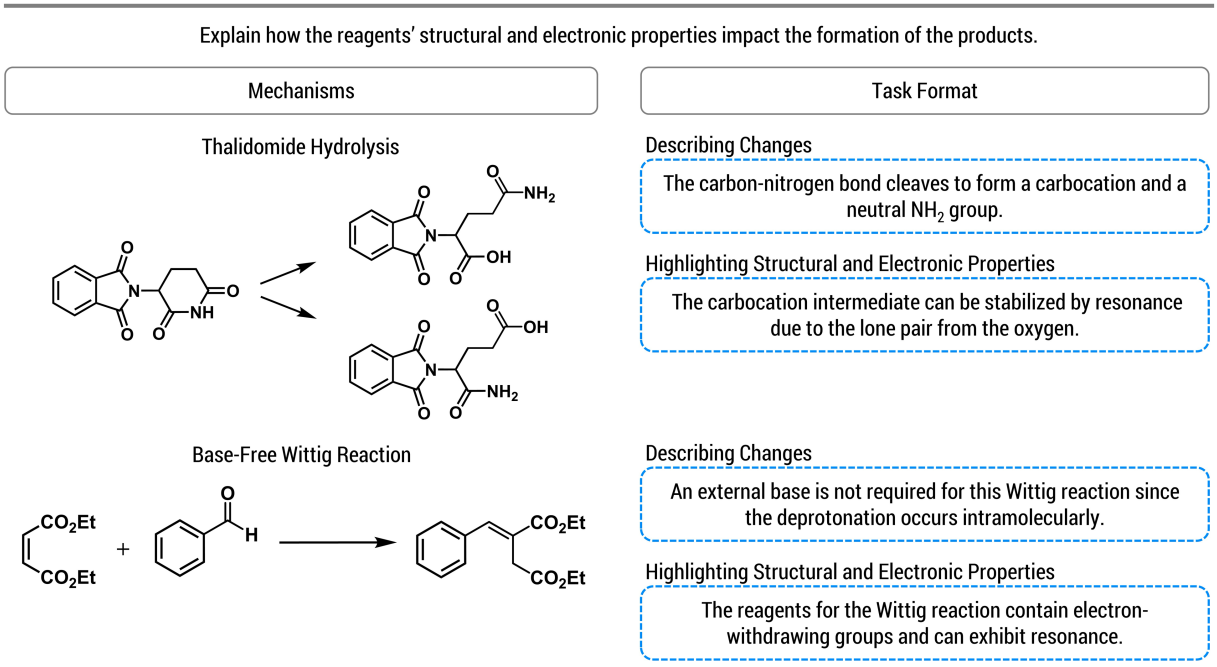


FIGURE 6 Essay writing in a WTL context on explanations of organic chemistry reaction mechanisms (dataset B). Data used for ML model training is highlighted in blue.

Category	Label	present	absent
Identifying setup conditions	Reaction medium		
Identifying explicit properties of entities	Connectivity of molecules		
	Charges		
	Stereochemistry		
Identifying activities	Electron movement		
	Non-electronic mechanism		
	Bond breaking and making		
Identifying implicit properties of entities	Implicit properties		
Organization of entities	Formation of stereochemistry		
Causal reasoning	Electronic causal reasoning		

FIGURE 7 Analytic rubric to score students' essay writing.

4.3 Written laboratory activity at Institution B (dataset C)

4.3.1 Setting of data collection at Institution B

The data collection setting within this activity mirrored that described in the first paragraph of Section 4.2.1. For more information on this setting, the task design, and sample solutions, see Haas et al. (2024), Kranz et al. (2023), and Watts et al. (2021). Details on the scaffold used to elicit student responses are published in Caspari and Graulich (2019) and Graulich and Caspari (2021).

The data used in this study comprised students' written responses to three scaffolded case comparisons (Haas et al., 2024). Each activity included a worksheet with the same scaffolded prompts. Students answered the scaffolded questions in online group work by identifying structural differences of the reactants, contrasting their chemical properties, detailing changes during the chemical reaction, and determining which reaction occurs at a higher rate (Graulich & Caspari, 2021). In total, 1,167 student responses were used for ML model training; the mean length of the English-written responses was 52 words (*SD* = 30.0, *max* = 307, *min* = 4). Student responses were scored with the rubrics outlined in Sections 4.1.3 and 4.2.3.

4.3.2 Implementation of the research instrument on case comparisons

By supporting the identification of explicit or implicit structural differences (Bussey et al., 2013; Lo & Marton, 2012), case comparisons engage students in comparing the properties of molecules (Graulich & Schween, 2018). Many studies in organic chemistry education evidenced that directing students' attention to structural differences facilitates the explanation of how these differences influence the outcome of reactions (e.g., Caspari et al., 2018; Eckhard et al., 2022; Kranz et al., 2023; Rodemer et al., 2020; Watts et al., 2021). The case comparisons in this study prompted students to determine which of the two reactions proceed at a higher rate (Figure 8). The reactions differ in one or two key structural features to help students make informed claims about the reactions' rate.

5 | METHODS

We used different datasets to evaluate ML model performance across settings (Section 4). The authors shared this data to provide multiple

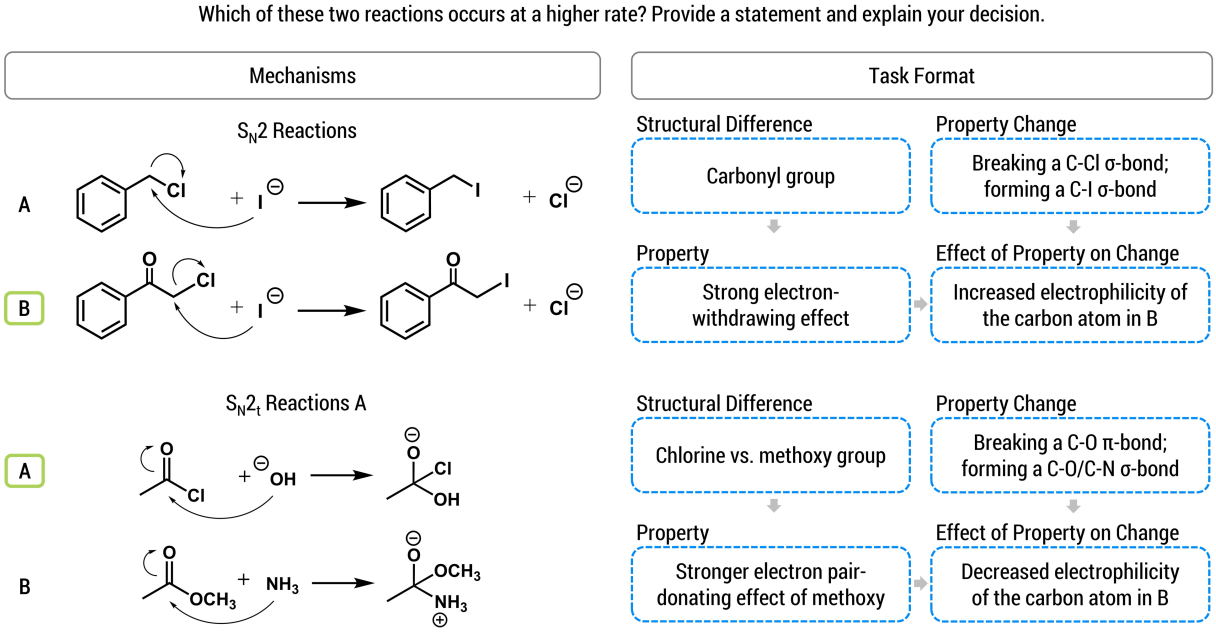


FIGURE 8 Written lab activity on case comparisons about the rate of two contrasted reactions (dataset C). Correct solutions are colored green. Data used for ML model training is highlighted in blue.

pieces of evidence that support certain trustworthiness attributes of the ML models. Table 1 summarizes the validation scenarios for each of the two ML models. For RQ 1, the characteristics of the training set match those of the test set. This means that even though the test set is a new dataset, it is collected under the same conditions as the training set. In contrast, the characteristics of the test set differed from those of the training set in RQ 2 and 3. This discrepancy can lead to reduced model performance as it may struggle with new data that deviates from the training data. Such external validation is crucial for assessing the reliability of ML models across real-world scenarios (Varshney, 2022). In the validation ID column (Table 1), the first string denotes the rubric upon which the ML model is trained (HR: holistic rubric; AR: analytic rubric), the second string indicates the dataset used for model training, and the third string shows the dataset used for model testing. For example, the string HR-A-A refers to the holistic ML model (section 4.1.3) trained and tested on dataset A. Separately, the string AR-BC-A refers to the analytic ML model (section 4.2.3) trained on datasets B and C but tested on dataset A. We used a range of datasets for training and testing to improve ML model generalizability and investigate the impact of rubric, context, and task format on model performance.

5.1 Human scoring

To develop supervised ML models, humans must code data before training the algorithm. For a previous analysis (Martin et al., 2024), author PPM and a research assistant individually coded dataset A on the plausibility of alternative reaction products in a mutually exclusive way using the developed holistic twenty-category rubric (Section 4.1.3). By contrast, authors AJD and FMW coded in a prior analysis dataset B on students' WTL essays at the sentence level non-mutually exclusively for the ten mechanistic reasoning features (Section 4.2.3) (Watts, Dood, et al., 2022). Table 2 (HR-A and AR-B) shows the human-human inter-rater reliability for these previous analyses.

To develop ML models evaluating students' mechanistic reasoning across contexts, author PPM and a research assistant subsequently

TABLE 1 Overview of the validation conditions to test the performance of the ML models across organic chemistry contexts.

Validation ID	Trained on…			Tested with…		
	Dataset A	**Dataset B**	**Dataset C**	**Dataset A**	**Dataset B**	**Dataset C**
Holistic Rubric (HR) (Section 4.1.3)						
RQ 1						
HR-A-A	X			X		
RQ 2						
HR-A-B	X				X	
HR-A-C	X					X
HR-AB-A	X	X		X		
HR-AB-B	X	X			X	
HR-AB-C	X	X				X
HR-AC-A	X		X	X		
HR-AC-C	X		X			X
HR-AC-B	X		X		X	
RQ 3 (balanced datasets, broken down by reaction mechanism)						
HR-ABC-ABC	X	X	X	X	X	X
Analytic Rubric (AR) (Section 4.2.3)						
RQ 1						
AR-B-B		X			X	
RQ 2						
AR-B-A		X		X		
AR-B-C		X				X
AR-AB-B	X	X			X	
AR-AB-A	X	X		X		
AR-AB-C	X	X				X
AR-BC-B		X	X		X	
AR-BC-C		X	X			X
AR-BC-A		X	X	X		
RQ 3 (balanced datasets, broken down by reaction mechanism)						
AR-ABC-ABC	X	X	X	X	X	X

coded datasets B and C (Figure 3) employing the holistic twenty-category rubric. In contrast, authors AJD and FMW coded datasets A and C (Figure 3) using their non-mutually exclusive approach for coding the ten mechanistic reasoning features. Coding was done in three rounds, with the datasets divided into subsets in a 25:25:50 ratio. During each coding round, the authors independently coded the data, assessed inter-rater reliability (Table 2), discussed any disagreements, and refined their coding guidelines. Ultimately, consensus codes were applied for ML model training. Tables 3 and 4 show the distribution of consensus codes split by dataset.

5.2 Training the ML models

We applied pre-trained, transformer-based deep neural networks for sequence classification to analyze students' mechanistic reasoning,

TABLE 2 Human-human inter-rater reliability.

Human Coding ID	Accuracy	Cohen's κ	Human Coding ID	Accuracy	Fleiss' κ
HR-A	0.8413	0.8276	AR-B	0.7469	0.6985
HR-B-1	0.7255	0.6868	AR-A-1	0.9460	0.8131
HR-B-2	0.7512	0.7025	AR-A-2	0.9495	0.8321
HR-B-3	0.8396	0.8070	AR-A-3	0.9592	0.8526
HR-C-1	0.8350	0.7436	AR-C-1	0.9430	0.8362
HR-C-2	0.8667	0.7965	AR-C-2	0.9443	0.8427
HR-C-3	0.9233	0.8655	AR-C-3	0.9490	0.8539

Note: In the human coding ID, the first string denotes the rubric used to code student responses, the second string indicates the coded dataset, and the number shows the coding round.

TABLE 3 Distribution of consensus codes split by dataset using the holistic rubric.

	Structural						Energetic	
	Charged species and substrates		Nucleophiles and electrophiles		Acids and bases		Thermodynamics and kinetics	
	Phen.	Elec.	Phen.	Elec.	Phen.	Elec.	Phen.	Elec.
Dataset A								
Absent				23				
Descriptive	127	70*	47	31	90	70	47	70*
Relational	38	32	64	74	58		82	123
Linear causal	0	36	22		20	0	22	32
Dataset B								
Absent				130				
Descriptive	140	247*	11	21	89	121	11	247*
Relational	6	11	1	9	0		0	37
Linear causal	0	1	0		0	0	0	13
Dataset C								
Absent				11				
Descriptive	2	15*	108	69	0	2	5	15*
Relational	1	16	6	650	0		0	250
Linear causal	0	3	0		0	0	0	29

Note: The asterisk (*) symbolizes that these categories were combined for automated scoring since they reflect the same mechanistic reasoning pattern.

applying the *PyTorch* deep learning framework implemented in Python (Paszke et al., 2019). We used the *AutoTokenizer* to tokenize the input text. We then employed the base-variant (*BERT-base-uncased)* of the large language model Bidirectional Encoder Representations from Transformers (BERT) to generate word embeddings (Devlin et al., 2018) with an additional classification head to predict class labels. Finally, we set up a training loop with the *AdamW* optimizer and a learning rate scheduler to train the models.

Hyperparameter adjustments were made for the number of epochs, learning rate, and batch size. Epochs refer to the training cycles an ML model undergoes, with more epochs typically improving model performance on the training data up to a saturation point. However, training a model on too many epochs can lead to overfitting, where a

TABLE 4 Distribution of consensus codes split by dataset using the analytic rubric.

		Dataset A		Dataset B		Dataset C	
Category	**Label**	**present**	**absent**	**present**	**absent**	**present**	**absent**
Identifying setup conditions	Reaction medium	17	1091	257	591	0	1167
Identifying explicit properties of entities	Connectivity of molecules	72	1036	779	69	210	957
	Charges	250	858	797	51	162	1005
	Stereochemistry	100	1008	212	636	14	1153
Identifying activities	Electron movement	51	1057	750	98	266	901
	Non-electronic mechanism	792	316	848	0	1149	18
	Bond breaking and making	76	1032	827	21	270	897
Identifying implicit properties of entities	Implicit properties	370	738	677	171	177	990
Organization of entities	Formation of stereochemistry	5	1103	152	696	5	1162
Causal reasoning	Electronic causal reasoning	280	828	387	461	467	700

model learns patterns specific to the training data so well that it performs poorly on new, unseen data. We varied this hyperparameter within a range from 1 to 100. The learning rate dictates the speed at which an optimizer updates the model parameters during training, which influences the ML model's adaptability to different contexts. In this study, we tested two learning rates: 5e-6 and 1e-5. Lastly, the batch size determines the number of concurrent training examples processed in a single network pass. We explored batch sizes of 2, 4, 8, 16, and 32. In total, each model underwent testing with a set of 1000 different hyperparameter configurations.

We partitioned our datasets into a training, validation, and test set with a distribution of 65:15:20. The training set was used to train the models; the validation set helped determine the optimal hyperparameter configurations. Furthermore, we used the test set to assess the models' performance using four metrics: accuracy, precision, recall, and Cohen's κ or Fleiss' κ. Accuracy measures the percentage of agreement between two raters. Precision is the ratio of correct positives to the total number of positive classifications; recall is the ratio of correct positives to the total number of actual positive instances. In education, low precision demonstrates an overestimation of students' achievement in certain responses while low recall reveals an underestimation of their achievement. We computed the unweighted mean—the macro average—of precision and recall across rubric categories. Besides, Cohen's κ assesses agreement beyond chance, whereas Fleiss' κ with Jaccard distance is particularly useful for multi-class analytic scoring tasks

(Section 4.2.3). Of note, Fleiss' κ and Krippendorf's α show comparable applicability in multi-class problems with binary classifications of nominal data (Zapf et al., 2016); for clarity, we report only Cohen's κ for the holistic rubric and Fleiss' κ for the analytic one.

6 | Results

6.1 RQ 1: Performance of the ML models in their original context

Initially, authors PPM and NG (Martin et al., 2024), as well as authors AJD, FMW, and GVS (Watts, Dood et al., 2022), individually developed ML models to automatically evaluate organic chemistry students' mechanistic reasoning within a single context. Authors PPM and NG applied a holistic rubric to classify students' mechanistic reasoning about the plausibility of alternative reaction products (Section 4.1). Authors AJD, FMW, and GVS adapted an analytic rubric to classify students' essays within the WTL activity (Section 4.2). In doing so, A's and E's holistic rubric classified students' mechanistic reasoning by highlighting their most prominent strengths and challenges in two dimensions: the *modes of reasoning* (Sevian & Talanquer, 2014) and *levels of granularity* (Bodé et al., 2019; Deng & Flynn, 2021). By contrast, the analytic approach of B, C, and D provided insights into students' mechanistic reasoning with each category specifying the presence of a specific mechanistic reasoning feature (Russ et al., 2008).

Table 5 provides the performance metrics of the two ML models in their original setting. The analytic ML model performed better in accuracy and recall, whereas its holistic counterpart exhibited higher precision and Cohen's κ. The performance differences can be partly attributed to the inherent nature of these models. The analytic model evaluates the presence or absence of ten mechanistic reasoning features. For this reason, the model must select between two options for each feature in every response—assuming a balanced dataset, an untrained model would have a 50% chance of making a correct decision. Conversely, the holistic model categorizes every student response into one of twenty distinct categories, introducing a larger pool of choices during decision-making. Consequently, the chance of correctly guessing decreases to only 5% in a balanced dataset.

Whether to use an analytic or holistic rubric is not strictly an empirical question as predicting the performance of models trained on these different types of rubrics might be possible in advance based on human-human inter-rater reliability, data distribution, and category numbers. Also, the choice between analytic and holistic rubrics depends on the specific context and the desired outcomes. For instance, the detailed breakdown of complex categories in an analytic rubric can simplify their evaluation, which may assist in precisely identifying students' strengths and challenges (Jescovitch et al., 2019, 2021). By contrast, a holistic approach may facilitate a more continuous analysis of students' skills, which becomes important when students are expected to progress over time (Martin & Graulich, 2024b). Following that, the

TABLE 5 Performance of the ML models in their original contexts.

Validation ID	Epochs	Learning Rate	Batch Size	Accuracy	Macro Precision	Macro Recall	Cohen's κ, Fleiss' κ
HR-A-A	38	1e-5	8	0.8711	0.8529	0.8076	0.8610
AR-B-B	90	5e-6	2	0.9402	0.8135	0.8797	0.8334

Note: In the validation ID, the first string denotes the rubric type, the second indicates the training set, and the third specifies the test set.

pedagogical purpose of applying ML in science education should guide rubric type selection. Both analytic and holistic rubrics can demonstrate high performance in scoring students' reasoning—a combination of these approaches may also be beneficial (Brookhart, 2018; Harsch & Martin, 2013; Martin et al., 2025; Tomas et al., 2019).

6.2 RQ 2: Reliability of the ML models across contexts

To determine the extent of *reliability* of our ML models, we applied them across institutions, contexts, tasks, and reaction mechanisms. For clarity, we remain at the context and task level within this RQ (Figure 3). In each subsection, we compare the performance of the holistic and analytic models trained on specific datasets to derive methodological and pedagogical implications for increasing ML model *reliability* in our context. In detail, subquestion 2a examines the generalizability of the original models (Section 6.1) in unfamiliar contexts, subquestion 2b explores how data sharing between the two original contexts impacts model generalizability, and subquestion 2c assesses how integrating an entirely new dataset impacts generalizability.

6.2.1 RQ 2a: Performance of the original ML models in novel contexts

For RQ 1, we assessed the performance of the ML models in their original context (Table 5: HR-A-A, AR-B-B). Nevertheless, a reliable model should generalize well to new contexts, so we also evaluated the ability of these models to score data reliably in other contexts (Table 6).

First, we applied the model originally developed to classify students' mechanistic reasoning on a product's plausibility (dataset A) in other contexts. Validation HR-A-B (Table 6) shows the performance of this model when scoring students' WTL essays (dataset B), indicating only *fair* machine-human score agreement (Landis & Koch, 1977). The differing learning objectives of the two task formats explain the ML model's low generalizability. The reasoning-focused short writing on a product's plausibility was designed to guide students in weighing multiple chemical concepts to justify their claims causally. The applied rubric values causal reasoning that includes electronic properties of the involved molecules (Section 4.1.3). On the contrary, the WTL activities prompted students to communicate the structure and reactivity of the involved molecules to a non-expert audience, encouraging them not to rely heavily on organic chemistry jargon. Consequently, students primarily used a *descriptive mode of reasoning* (Table 4) to explain a synthesis pathway comprehensibly. Therefore, the differing learning objectives of these activities led to variations in the structure, causality, and granularity of students' responses, resulting in reduced ML model generalizability.

TABLE 6 Performance of the original ML models in novel contexts.

Validation ID	Epochs	Learning Rate	Batch Size	Accuracy	Macro Precision	Macro Recall	Cohen's κ, Fleiss' κ
HR-A-B	38	1e-5	8	0.3809	0.3979	0.5216	0.3242
HR-A-C	38	1e-5	8	0.7781	0.5526	0.4893	0.6661
AR-B-A	90	5e-6	2	0.9053	0.5575	0.6796	0.7128
AR-B-C	90	5e-6	2	0.8826	0.7715	0.6568	0.7040

Note: In the validation ID, the first string denotes the rubric type, the second indicates the training set, and the third specifies the test set.

Given the poor generalizability of the holistic model to the WTL activity, we also validated the performance of this model when evaluating students' mechanistic reasoning regarding the rate of contrasted reactions (dataset C). The holistic model performed notably better in scoring responses from that activity (Table 6: HR-A-C). The higher generalizability of the holistic ML model in the case comparison activity can be attributed to the shared characteristics of the two task formats: The case comparisons and the reasoning-focused short writing explicitly contrast alternative reaction products to guide students in integrating electronic properties into their decision-making. Following this, students are encouraged to reason more causally, leading to comparable mechanistic reasoning patterns (Table 4). Due to these shared characteristics, the model developed to analyze students' mechanistic reasoning on a product's plausibility generalized better to the case comparison activity than to the WTL activity.

Furthermore, we assessed the generalizability of the analytic ML model trained to classify students' mechanistic reasoning in the WTL activity (dataset B). Compared to the holistic ML model, this model generalized better to novel contexts (Table 6: AR-B-A, AR-B-C). In particular, the accuracy of this model in unfamiliar contexts was around 90%, indicating that it correctly scores students' mechanistic reasoning in new contexts in 9 out of 10 cases. However, precision and recall separated by rubric labels reveal that this high level of accuracy does not equally apply to all mechanistic reasoning features (Section 4.2.3). For the data on a reaction's plausibility (dataset A), the model had significant challenges in predicting instances of *stereochemistry* (precision = 0.1792; recall = 0.7308), the *formation of stereochemistry* (precision = 0.0000; recall = 0.0000), and *electronic-causal* reasoning (precision = 0.2745; recall = 0.4855), probably due to the imbalanced nature of these rubric categories (Table 4). This finding resonates with the task design of the reasoning-focused short writing that did not intentionally prompt students to consider stereochemistry, leading to an underrepresentation of these instances in the data. By contrast, high performance could be achieved for labels with more balanced training data, such as identifying *charges* (precision = 0.8993; recall = 0.9959) and *non-electronic* explanations (precision = 0.8399; recall = 0.9901). These two labels are equally reflected in both rubrics: a focus on *charges* in the analytic rubric relates to *electronic*, *descriptive* accounts in the holistic one; *non-electronic* explanations in the analytic rubric are termed *phenomenological* in its holistic counterpart. Due to this overlap, the analytic ML model could accurately score responses within these labels.

The analytic ML model also encountered challenges for certain labels when generalizing to the case comparison activity (dataset C) (Table 6: AR-B-C). Again, the labels *stereochemistry* (precision = 0.5000; recall = 0.4000) and *formation of stereochemistry* (precision = undefined; recall = 0.0000) showed the lowest machine-human score agreements since the case comparison activity was not designed to capture mechanistic reasoning about stereochemistry (Table 4). Conversely, the model performed best in identifying *electron movement* (precision = 0.8182; recall = 0.8417) and *non-electronic* explanations (precision = 0.9889; recall = 0.9897). The identification of resonance, hyperconjugation, and inductive effects—properties that impact *electron movement*—was essential to solving the case comparisons correctly. Lower- and higher-achieving students differed in how they integrated electronic properties: lower-achieving students relied on *non-electronic* explanations while higher-achieving students reasoned about *electron movement* (Haas et al., 2024; Kranz et al., 2023; Watts et al., 2021). The high agreement metrics for these two rubric categories reflect their usefulness in distinguishing between these achievement groups.

In summary, ML models developed within a single context often struggled to generalize to unfamiliar contexts. For instance, the holistic ML model partly failed to analyze responses elicited in the WTL activity. Generalizability improved when tasks or rubrics shared certain features, such as the instructional strategy of comparing molecules or beneficial overlaps in labels among rubrics. This result highlights the need for careful rubric and task design when developing ML models (Kubsch et al., 2022; Martin & Graulich, 2023). Moreover, high overall performance does not necessarily apply uniformly to all labels as demonstrated by the low precision and recall for some labels despite the high overall accuracy of the analytic model. These findings motivated us to increase ML model generalizability by expanding the training data size (Sections 6.2.2 and 6.2.3).

6.2.2 RQ 2b: Performance of the ML models trained on shared data

Within this RQ, we expanded the training data size of our ML models by adding WTL data (dataset B) to the holistic model and data on a product's plausibility (dataset A) to its analytic counterpart. We examined the impact of the increased data heterogeneity on model performance (Table 7).

To determine whether an expanded training dataset advances ML model performance in its original context, we compared the ML models' initial performance—as discussed in Section 6.1—with their updated performance based on a larger training dataset (Table 7: HR-AB-A; AR-AB-B). Particularly, we evaluated whether adding WTL data (dataset B) to the holistic model improved its performance in classifying students' mechanistic reasoning on a product's plausibility (dataset A). Similarly, we tested whether incorporating data on a product's plausibility (dataset A) boosted the analytic model's performance in categorizing WTL data (dataset B). To our surprise, adding WTL essays to the training data degraded the performance of the holistic ML model in its original context of alternative reaction products (Table 5: HR-A-A; Table 7: HR-AB-A). Similarly, including students' mechanistic reasoning on alternative reaction products did not significantly improve the analytic model's performance in the WTL activity (Table 5: AR-B-B; Table 7: AR-AB-B). The non-improving, partly lowered performance of the ML models may be explained by different response lengths, increased data heterogeneity, data imbalances in categories, and differences in data distributions (Tables 3 and 4), which can cause classification errors (Groh, 2022; Wu et al., 2019). Additionally, the underlying relationships between semantic features and machine scores may have changed with the new data, making the model's training in its original context partly ineffective (Allogmany & Josyula, 2022)—a

TABLE 7 Performance of the ML models trained on shared data.

Validation ID	Epochs	Learning Rate	Batch Size	Accuracy	Macro Precision	Macro Recall	Cohen's κ, Fleiss' κ
HR-AB-A	89	5e-6	8	0.8286	0.8249	0.7967	0.8156
HR-AB-B	89	5e-6	8	0.9171	0.8450	0.9343	0.8980
HR-AB-C	89	5e-6	8	0.8886	0.7715	0.6862	0.8273
AR-AB-B	59	5e-6	2	0.9430	0.8226	0.8644	0.8390
AR-AB-A	59	5e-6	2	0.9378	0.6309	0.7995	0.7919
AR-AB-C	59	5e-6	2	0.9013	0.7950	0.7526	0.7423

Note: In the validation ID, the first string denotes the rubric type, the second indicates the training set, and the third specifies the test set.

phenomenon known as concept drift (Hashmani et al., 2020; Lu et al., 2019). For this reason, further model fine-tuning is necessary (Martin & Graulich, 2024c).

Beyond that, we investigated whether the expanded training data improved model performance in contexts for which the ML models were not originally developed. Specifically, we applied the expanded holistic model to analyze the WTL data (dataset B) and the case comparison data (dataset C), and the expanded analytic model to analyze data on a product's plausibility (dataset A) and the case comparison data (dataset C). Validation HR-AB-B shows that incorporating data from the WTL activity significantly improved model performance in classifying responses within that activity, reaching *almost perfect* machine-human score agreements (Table 6: HR-A-B; Table 7: HR-AB-B). Similarly, integrating data from the WTL activity also advanced the model's ability to categorize responses from the case comparison activity—although not part of the training data (Table 6: HR-A-C; Table 7: HR-AB-C). A comparable trend can be found for the analytic ML model. Adding data on a product's plausibility increased the model's ability to categorize students' mechanistic reasoning on these alternative reaction products (Table 6: AR-B-A; Table 7: AR-AB-A). However, macro precision reveals that certain labels remain challenging to evaluate automatically, primarily due to the imbalanced data structure (Table 4). Additionally, the model performance increased in the case comparison activity not included in the training set (Table 6: AR-B-C; Table 7: AR-AB-C). Hence, adding a second context to the training data increased model generalizability to novel contexts. This result was anticipated as expanding training data heterogeneity can advance ML model generalizability (e.g., Dood et al., 2022; Noyes et al., 2020; Wang et al., 2021; Wulff et al., 2023).

6.2.3 RQ 2c: ML model performance when integrating an entirely new dataset

To assess the impact of different task formats on improving ML model generalizability, we expanded the training data of the original models by including data from the case comparison activity (dataset C). Within this RQ, the holistic model is trained on data on a product's plausibility (dataset A) and case comparison data (dataset C) while its analytic counterpart is trained on WTL data (dataset B) and case comparison data (dataset C). Again, we investigated how increasing the training data size affected ML model performance (Table 8).

The findings that could be drawn from RQ 2b also apply to this RQ: First, expanding the training data size by incorporating data from novel contexts did not advance model performance in its original context. In detail, the holistic model expanded with case comparison data performed

TABLE 8 ML model performance when integrating an entirely new dataset.

Validation ID	Epochs	Learning Rate	Batch Size	Accuracy	Macro Precision	Macro Recall	Cohen's κ, Fleiss' κ
HR-AC-A	64	1e-5	2	0.8364	0.8336	0.8035	0.8232
HR-AC-C	64	1e-5	2	0.9627	0.8182	0.7502	0.9358
HR-AC-B	64	1e-5	2	0.4929	0.4407	0.5820	0.4263
AR-BC-B	67	5e-6	4	0.9378	0.8101	0.8689	0.8273
AR-BC-C	67	5e-6	4	0.9706	0.9182	0.9735	0.9237
AR-BC-A	67	5e-6	4	0.9258	0.6143	0.6395	0.7551

Note: In the validation ID, the first string denotes the rubric type, the second indicates the training set, and the third specifies the test set.

worse in its original context than the original holistic model (Table 5: HR-A-A; Table 8: HR-AC-A). Likewise, including data from the case comparison activity did not boost the performance of the analytic ML model in its original context of WTL (Table 5: AR-B-B; Table 8: AR-BC-B). Second, incorporating data from the case comparison activity significantly improved the performance of the holistic and analytic models within that activity, as indicated by their *almost perfect* machine-human score agreements (Table 6: HR-A-C, AR-B-C; Table 8: HR-AC-C, AR-BC-C). Integrating data from the case comparison activity also boosted the generalizability of the holistic model to the WTL activity (Table 6: HR-A-B; Table 8: HR-AC-B) as well as the generalizability of the analytic model to the activity on a product's plausibility (Table 6: AR-B-A; Table 8: AR-BC-A). However, the holistic model continued to perform only *moderately* in the WTL context.

Interestingly, the developed ML models showed higher generalizability when classifying students' mechanistic reasoning in activities involving contrasted reactions (Table 7: HR-AB-C, AR-AB-C; Table 8: AR-BC-A). By contrast, generalizability to the WTL context on explanations of reaction mechanisms was rather poor (Table 8: HR-AC-B). The differences between the highly structured prompts of the case comparison activities and the more open-ended nature of essay writing may explain this finding, which reinforces the effectiveness of case comparisons in organic chemistry classes to scaffold students' mechanistic reasoning (Graulich & Lieber, 2024). In particular, by engaging in comparisons, case comparisons prompt students to recognize similarities and differences in chemical and physical properties (Bussey et al., 2013; Lo & Marton, 2012), which helps students structure their responses. The ML models can better analyze these more structured responses.

6.3 RQ 3: Performance differences among reaction mechanisms

We developed ML models to automatically analyze students' mechanistic reasoning in chemistry classes over time, potentially allowing for adaptive support in different settings. However, since the provided reaction mechanisms influence the complexity of students' mechanistic reasoning (Deng & Flynn, 2021) and, thus, machine-human score agreements, we investigated the performance of the ML models across reaction mechanisms. By doing so, we used a balanced dataset of 349 student responses per mechanism. The Supplementary Information accompanying the online article offers a more detailed analysis of the ML models' performance across reaction mechanisms with performance broken down by each label of the scoring rubrics.

Table 9 shows the performance of the ML models across the seven reaction mechanisms. As indicated by accuracy, recall, and Cohen's κ, the performance of the holistic ML model improves in the following order: Claisen condensation, thalidomide hydrolysis, intramolecular Williamson ether synthesis, base-free Wittig reaction, S_N2, S_N2_t A, S_N2_t B. Precision gives the following order of improved performance: thalidomide hydrolysis, Claisen condensation, intramolecular Williamson ether synthesis, S_N2, base-free Wittig reaction, S_N2_t B, S_N2_t A. For the analytic model, a tentative trend is evident across the performance metrics: The model performed the worst for thalidomide hydrolysis and base-free Wittig reaction, moderately for S_N2_t A and S_N2_t B, and best for Claisen condensation, Williamson ether synthesis, and S_N2.

The relatively poor performance of the holistic ML model in Claisen condensation and Williamson ether synthesis, and the analytic ML model in thalidomide hydrolysis and base-free Wittig reaction is unexpected, considering these models were precisely trained for classifying

TABLE 9 Performance of the ML models across reaction mechanisms.

Validation ID	Epochs	Learning Rate	Batch Size	Accuracy	Macro Precision	Macro Recall	Cohen's κ, Fleiss' κ
HR-ABC-WR	23	1e-5	2	0.8971	0.8101	0.7733	0.8602
HR-ABC-TH	23	1e-5	2	0.8261	0.6418	0.6982	0.7805
HR-ABC-CC	23	1e-5	2	0.7059	0.6893	0.6304	0.6802
HR-ABC-ES	23	1e-5	2	0.8451	0.7707	0.7672	0.8299
HR-ABC-SN	23	1e-5	2	0.9231	0.7817	0.8325	0.8807
HR-ABC-SA	23	1e-5	2	0.9714	0.8452	0.8541	0.9455
HR-ABC-SB	23	1e-5	2	0.9846	0.8125	0.8750	0.9752
AR-ABC-WR	92	5e-6	2	0.9443	0.7449	0.7672	0.7845
AR-ABC-TH	92	5e-6	2	0.9251	0.7569	0.6930	0.7960
AR-ABC-CC	92	5e-6	2	0.9614	0.9263	0.9408	0.8895
AR-ABC-ES	92	5e-6	2	0.9634	0.8778	0.8895	0.8817
AR-ABC-SN	92	5e-6	2	0.9704	0.9200	0.9173	0.8959
AR-ABC-SA	92	5e-6	2	0.9437	0.8774	0.7372	0.8128
AR-ABC-SB	92	5e-6	2	0.9494	0.8107	0.7264	0.8375

Note: WR = base-free Wittig reaction, TH = thalidomide hydrolysis, CC = Claisen condensation, ES = Williamson ether synthesis, SN = S_N2, SA = S_N2_t A, SB = S_N2_t B.

responses within these reaction mechanisms. Once more, the lowered performance of the ML models may be due to the increased data heterogeneity, category imbalances, and differences in data distributions (Tables 3 and 4).

Besides, both ML models performed relatively poorly in thalidomide hydrolysis. Several factors explain this underperformance: The data for thalidomide hydrolysis contains noisy responses that do not address the topics defined in the rubric, which is why the ML models struggled to classify them accurately. On top of that, the geometry of thalidomide is more branched (Figure 6), which poses challenges for students in identifying, describing, and interpreting relevant features of this representation (Braun et al., 2022). Consequently, decoding the molecule and identifying interacting structural features might have induced greater challenges (Gupte et al., 2021), resulting in higher data variability and lower machine-human score agreements. In contrast, the ML models performed better in the WTL assignment on the base-free Wittig reaction. This activity might have encouraged students to integrate implicit properties into their mechanistic reasoning (Zaimi et al., 2024), inducing more structured responses that the model classified more accurately. Also, thalidomide hydrolysis was the first assignment in the semester of data collection while the base-free Wittig reaction was the second. Consequently, students might have been more adept at writing about reaction mechanisms in the base-free Wittig reaction. For evaluating a product's plausibility, students rated the Williamson ether synthesis easier than the Claisen condensation (Lieber et al., 2022a)—a trend similarly reflected by the increased machine-human score agreements for the ether synthesis.

Ultimately, the three reaction mechanisms from the case comparison activity consistently performed well across both ML models. It seems that case comparisons prompt students to better structure their responses, which makes them easier to classify for the ML models. Here, the ML models' high performance cannot be solely attributed to the S_N2 portions of the mechanisms

since the Williamson ether synthesis also involves an S_N2-type reaction. Instead, it may result from the combination of S_N2—rather familiar to students—and the case comparison task format, which guides students' comparisons and promotes coherent responses. The high performance of the ML models on these reaction mechanisms also shows the potential of using automatically evaluable case comparisons (Figure 8) to support and monitor students' mechanistic reasoning over time.

6.4 RQ 4: Assessing model explainability to increase human interaction

While deep neural networks have shown promising results in various classification tasks, their inherent lack of interpretability poses challenges for *human interaction*. To ease *human interaction* with the ML output, we validated ML decision-making using the SHAP model explainer—a method for transparently examining ML model output (Lundberg & Lee, 2017). In doing so, we calculated the importance scores of the ten words that impacted the ML output most significantly for every response and averaged them within each rubric category. For clarity, results are only reported for the holistic ML model trained on the balanced dataset of every reaction mechanism (Section 6.3). Figure 9 illustrates the top ten words that most significantly influence the ML output for three exemplary rubric categories. Here, blue scores indicate an increase in the likelihood of a positive classification within the relevant category when the respective word is included in a student response. On the contrary, red scores reduce the likelihood of a positive classification when the respective word is included.

The category of *descriptive*, *phenomenological* mechanistic reasoning on *nucleophiles and electrophiles* exemplifies a case where the ML model aligned exactly with human scoring guidelines (Figure 9a). Notably, the ML model scored responses in this category when incorporating terms such as "electrophilic," "electrophile," "nucleophilic," and "nucleophile," corresponding to the topic of this category. Conversely, terms like "electron-loving," "charge," and "polar" decreased the likelihood of a positive classification as these

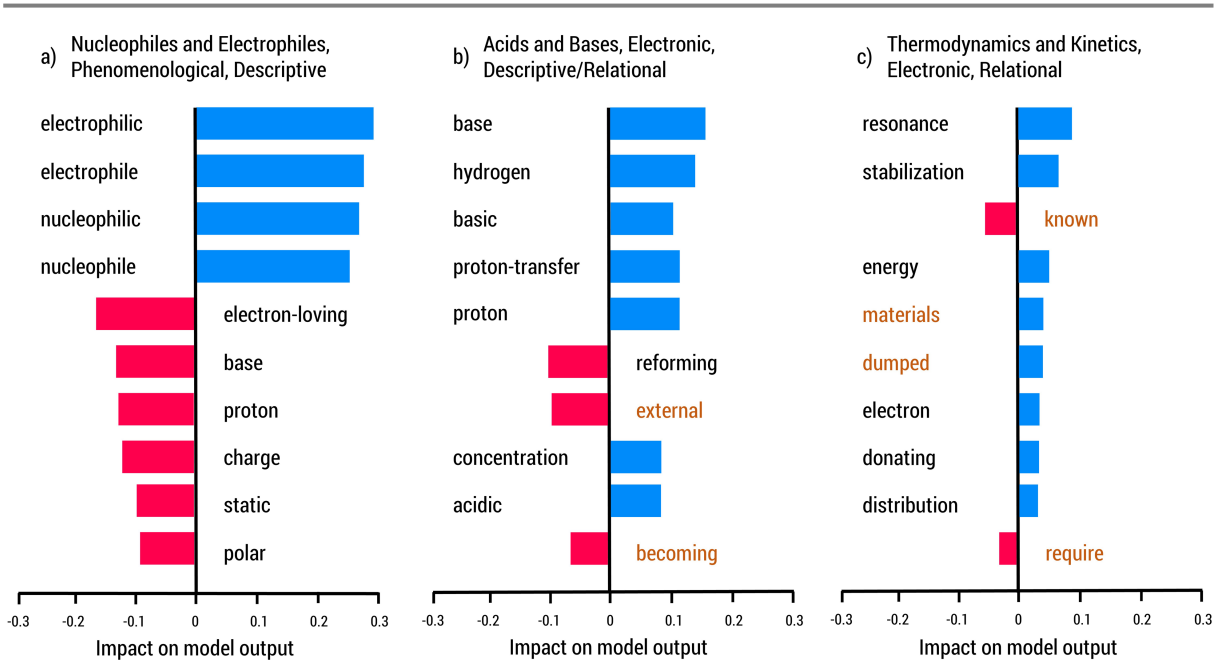


FIGURE 9 Top ten words that impact the output of the holistic ML model most significantly for three exemplary rubric categories. Misalignments between human and machine scoring guidelines are highlighted in orange.

terms represent *electronic*, rather than *phenomenological* mechanistic reasoning. Moreover, terms like "base" and "proton" underscore a strong emphasis on acids and bases while the term "static" introduces steric considerations; thus, diverging from reasoning about *nucleophiles and electrophiles*.

The category comprising *electronic* mechanistic reasoning about *acids and bases* is an example where eight alignments between human and machine scoring were observed (Figure 9b). Terms like "base," "hydrogen," "basic," "proton-transfer," "proton," "concentration," and "acidic" clearly show a focus on *acids and bases*. As a negative predictor, "reforming" suggests associations with organic chemistry mechanisms involving functional group rearrangements or ring formation; therefore, not necessarily addressing the disciplinary core idea of *acids and bases*. However, "external" and "becoming" lack relevance in this category, indicating misalignments.

Finally, the category delineating *linear causal*, *electronic* mechanistic reasoning on *thermodynamics and kinetics* illustrates a case where only six alignments between human and machine scoring emerged (Figure 9d). Terms such as "resonance," "stabilization," "energy," "electron," "donating," and "distribution" resonate with mechanistic reasoning about thermodynamics on an electronic level. However, the significant impact of words like "known," "materials," "dumped," and "require" lack meaning within this category. These misalignments highlight that although the ML models performed exceptionally in certain categories, the ML models incorporated words to score responses that do not adhere to human scoring guidelines (Martin et al., 2024).

To generalize this claim, we counted the number of words aligned between human and machine scoring for every rubric category (Table 10). For example, only six out of ten words aligned between human and machine scoring in three rubric categories, showcasing that the scoring approaches do not necessarily rely on the same decision-making.

7 | Discussion

ML-driven automated scoring can provide individualized support to improve students' learning outcomes. However, instructors must trust these scoring models to adopt them (Figure 1). In this article, we applied a conceptual framework designed to investigate and advance the trustworthiness of ML models by focusing on four attributes: *performance*, *reliability*, *aligned purpose*, and *human interaction* (Figure 2). Our case study shows that these attributes represent distinct categories that form a structured framework to analyze ML model trustworthiness.

7.1 Applying the trustworthiness framework within a case study in chemistry

Within our case study, we first evaluated the *performance* of our ML models in their original settings. We found that the ML models performed well in the context in which they were originally developed (Section 6.1). This finding aligns with Zhai, Shi, et al.'s (2020) review on the use of ML in science education, which

TABLE 10 Number of holistic rubric categories with a specific count of word alignments between human and machine scoring based on the top ten words.

Number of words aligned between coding approaches	10	9	8	7	6
Number of rubric categories	5	6	4	2	3

demonstrates that ML models can achieve high *performance* in their original science context. However, we only assessed how well the ML models classified student responses across demographics, which is not without critique. Neglecting to analyze ML model *performance* for specific subgroups can introduce bias since marginalized groups may receive lower scores (Cheuk, 2021; Cheuk et al., 2019; Grimm, Steegh, Çolakoğlu, et al., 2023; T. Li et al., 2023). Following that, employing universal cut-off values to measure *performance* without considering demographics may be inappropriate in education, where decisions substantially impact student learning. Varshney (2022) indirectly addresses this concern by suggesting that "good performance" in real-world tasks requires measurement tailored to the specifics of the problem, rather than solely focusing on high accuracy. Specifics in education could be that a model maintains high accuracy across subgroups, rather than showing high average accuracy. These considerations emphasize that the attributes of trustworthy ML are closely related to sociopolitical implications.

To investigate the extent of *reliability* of our ML models, we applied them across contexts and tasks. Results indicate that the ML models trained on a single context demonstrated poor generalizability across contexts (Section 6.2.1). This finding is consistent with the studies by Nehm et al. (2012) and Yik et al. (2021), which show that ML models trained on a single dataset often exhibit reduced performance when applied to a different dataset. Nonetheless, including a second context in the training set greatly advanced generalizability—even in unfamiliar contexts (Sections 6.2.2 and 6.2.3), suggesting that including two contexts in the training dataset is crucial for developing more generalizable models. This resonates with the studies by Dood et al. (2022), Noyes et al. (2020), and Wulff et al. (2023), which demonstrate that ML models trained across multiple institutions show higher generalizability. However, trade-offs exist among different dimensions of trustworthy ML. For example, improving *reliability*—that is, generalizability across contexts—slightly decreased the high *performance* of our ML models in their original contexts (Sections 6.2.2 and 6.2.3). Further, the *reliability* of an ML model also relies on the assessment design, including features such as prompt structure, rubric development, and sample heterogeneity (Martin & Graulich, 2023; Raker et al., 2022). We posit that investigating an ML model's *reliability* as generalizability across contexts represents an initial step in decoding the complex concept of *reliability*. However, this simplification of *reliability* does not free researchers from adhering to established standards before ML model development. Assessing the impact of assessment design on *reliability* may remain challenging (Watts & Finkenstaedt-Quinn, 2021).

After exploring the *reliability* of our ML models, we investigated whether they demonstrated their *aligned purpose* of accurately analyzing students' reasoning in organic chemistry across reaction mechanisms. We found that the high overall performance of the ML models concealed scoring problems for certain reaction mechanisms (Section 6.3), undermining the ML models' purpose. Given these performance differences, our ML models may not be considered fair as they evaluate students' reasoning for certain mechanisms more accurately than others. However, this prompts the question of what constitutes fairness in ML model development and which variables should be equitably assessed (Grimm, Steegh, Kubsch, et al., 2023). For example, the extent to which other variables, such as gender, should factor into determining the fairness of an ML model remains uncertain. The answers to these ethical questions cannot be established scientifically; they must be addressed through political and societal negotiations (Grimm, Steegh, Kubsch, et al., 2023).

As a last step in investigating the trustworthiness of our ML models, we applied post hoc explainability techniques to support *human interaction* by extracting the words that impacted the ML output most significantly (Section 6.4). Results indicate that machine scoring occasionally relied on other decision-making processes than humans, which aligns with other studies using post hoc explainability techniques to analyze ML decision-making (e.g., Martin et al., 2024). On the one hand, a different way of ML decision-making may be advantageous due to the high consistency, scalability, and efficiency of scoring; on the other hand, a lack of contextual understanding or the limited adaptability to unexpected inputs may compromise model trustworthiness. Post hoc explainability techniques help humans understand which words influence the ML output; nevertheless, human intervention is still necessary to correct classification errors.

Together, our analysis indicates that evaluating the trustworthiness of ML models extends beyond reporting predictive accuracy. It also involves examining the consistency of these models across contexts, their alignment with the purpose of the assessment, and their level of interpretability. The presented analysis could be an initial step toward developing more trustworthy ML applications. However, our results also emphasize potential trade-offs between different attributes of trustworthy ML. For instance, improving generalizability across contexts may compromise model performance in its original context. Similarly, a high model performance may come at the cost of reduced model explainability. Therefore, researchers must consciously prioritize the attributes of trustworthiness that are most important in a given context.

7.2 Relationship between trustworthiness and trust

While trustworthiness is a crucial aspect of ML model development, it must be questioned whether trustworthiness is sufficient for increasing *trust* in ML methodology. Our conceptual framework outlined that trustworthiness and trust are interconnected yet distinct concepts (Section 2). Although our analysis demonstrated that our ML models possess certain attributes of trustworthy ML, the extent to which users trust these models remains influenced by individual attitudes, intentions, or behaviors. Therefore, the concept of trust empowers the *user* to ultimately determine whether an ML model is deemed trustworthy.

The extent of user-based trust does not necessarily align with an ML's factual trustworthiness, which is why Schlicker and Langer (2021) distinguish between an ML's perceived and actual trustworthiness. Data on an ML's actual trustworthiness can potentially bridge the gap between both facets; nonetheless, users must possess the competence to interpret this data (Schlicker & Langer, 2021). The next steps in increasing trust in ML methodology could involve educating users about ML technologies, including their capabilities, limitations, and biases. Furthermore, transparent communication about the workings, objectives, and alternatives of ML models may be essential for building trust among users.

8 | Implications

Our case study highlights how attributes of trustworthiness can be implemented practically in science education, which is an important but challenging endeavor given the inherent context-specificity and considerable heterogeneity of students' reasoning across science domains. These science-specific challenges underscore

that developing trustworthy ML-driven automated scoring models is not straightforward in science education. For instance, it remains uncertain whether to develop task-specific ML models that analyze students' reasoning in depth within a single context, or generalizable models that track student achievement across a sequence of similar tasks—benefiting from an expanded training dataset that spans multiple tasks. To what degree the heterogeneity in students' content-specific responses should be captured is also unclear. Furthermore, whether an ML model designed to predict students' science achievement should prioritize a high model *performance* or transparent communication with stakeholders through *human interaction* is questionable. Similarly, if the purpose of an ML model is to identify at-risk students for timely support, it might prioritize higher recall over precision to ensure that no at-risk student is overlooked. However, this approach may result in some students being erroneously labeled as at-risk, which could negatively impact students' future achievement and self-perception. By clarifying the purpose of ML use, future studies could provide explicit descriptions of their approach to strike a balance between different attributes of trustworthiness in ML model development—an increasingly challenging task. As the complexity of ML techniques increases, ensuring trustworthiness may become even more complex, which necessitates continuous adaptation of the discussed conceptual framework.

Furthermore, we thoroughly examined the trustworthiness of the developed ML models in our case study; however, the reported metrics still represent a limited subset of metrics useful for ML model evaluations. We assessed neither model explainability using metrics such as faithfulness, nor model fairness by calculating disparate impact ratio or average odds difference, nor adversarial robustness through measuring empirical robustness or CLEVER score (Varshney, 2022). Recognizing these gaps, future investigations could incorporate a more extensive array of metrics to provide a more nuanced understanding of the trustworthiness of ML models.

To close the gap between the perceived and actual trustworthiness of ML models, researchers and practitioners should offer users information about an ML model's actual trustworthiness. Key details can be provided in line with the four attributes of trustworthy ML: For *performance*, researchers should disclose details about their datasets, the ML model development process, and the performance metrics. Regarding *reliability*, they could hypothesize or assess the model's performance on unseen data. To ensure an *aligned purpose*, researchers could describe the intended use of an ML model and outline the model's impact, benefits, and limitations. For *human interaction*, they could advance ML model interpretability and specify the extent of intended human-AI interactions. A more comprehensive overview of guidelines to increase trust in ML methodology is outlined by Roszel et al. (2021). Some of this information can also be integrated into the design of user interfaces. When interacting with an ML model, it could be beneficial to display the features an ML model considers for decision-making, classification examples, predictive accuracy, the uncertainty associated with specific classifications, or the rationale behind its recommendations (Schlicker & Langer, 2021).

9 | Limitations

To investigate the trustworthiness of ML models developed to assess students' heterogeneous mechanistic reasoning in chemistry, we relied on ML-assisted automated scoring in our case study. Because of their distinct technological characteristics, more tailored frameworks may be needed to examine the trustworthiness of other AI applications, such as unsupervised learning or generative AI. For instance, we

defined *reliability* in terms of consistency across different cohorts. However, for generative AI, it might be more useful to think of *reliability* as consistency across multiple runs—essentially, how consistent the AI's outputs are when given the same prompt. Therefore, future research could extend the generalizability of the conceptual framework beyond supervised learning tasks.

Similarly, our analysis of students' written responses was exclusively conducted using BERT while we did not employ zero- or few-shot learning approaches using generative AI. However, Latif and Zhai (2024) found that a fine-tuned GPT-3.5 model outperformed BERT in automatically scoring some science assessment items. Accordingly, using generative AI for automated scoring may further increase model *performance*. Besides, generative AI may advance *human interaction* by offering comprehensible descriptions of complex model structures, producing counterfactual explanations that clarify how modifications to input text affect model classifications, or detecting biases in decision-making (Wu et al., 2024). Future research must investigate to what extent generative AI can meet the trustworthiness attributes outlined in this study (Sadler et al., 2024).

Moreover, techniques like SHAP offer advanced opportunities to increase ML transparency; nevertheless, additional investigation is required to gain a more detailed understanding of the inner workings of advanced ML models. Adversarial examples—purposefully designed inputs that cause a model to misclassify them—could, for instance, be used to probe the *reliability* and *human interaction* of ML models further (Balda et al., 2020). Such examples can be created by making small, strategically calculated perturbations to the input data, such as text or images, which are difficult for humans to recognize but sufficient to mislead an ML model. In educational settings, students could use adversarial examples to manipulate automated grading systems intentionally (Yik et al., 2021). Understanding and mitigating the impact of adversarial examples is crucial for ensuring the trustworthiness of ML applications in education and beyond.

Finally, we acknowledge that large language models like BERT and GPT are trained on raw language data sourced, among others, from Wikipedia and various online platforms. Accordingly, these models may internalize stereotypical biases (Caliskan et al., 2017; Kliegr et al., 2021), which can perpetuate discriminatory practices against racially marginalized groups in science assessment (Cheuk, 2021). Mitigating such discriminatory practices entails strategies like refining the ML training process or implementing subject-specific interventions to counteract discrimination (Grimm, Steegh, Kubsch, et al., 2023), which is not within the scope of this article.

10 | Conclusions

Given the possibility of leveraging ML-driven automated scoring of open-ended responses to address student heterogeneity and implement adaptive learning approaches, establishing trustworthiness in ML methodology has emerged as a critical research area in education. In this article, we discussed a framework to investigate and advance the trustworthiness of ML models. The outlined attributes showcase that solely reporting performance metrics is insufficient for establishing trustworthiness in ML methodology; instead, the consistency of ML models across contexts, their alignment with the purpose of the respective assessment, and their level of interpretability should be investigated.

However, an analysis of the trustworthiness of ML models designed to evaluate undergraduate students' reasoning about reaction mechanisms in organic chemistry revealed that meeting multiple trustworthiness attributes simultaneously is challenging. This difficulty

arises from potential trade-offs between the proposed attributes. For example, prioritizing *performance* in ML model development can negatively impact other trustworthiness attributes, such as *reliability* or *human interaction*. Developing an ML model to support student learning is, thus, a complex task that requires thoughtful planning, rigorous evaluation, and continuous refinement.

Acknowledgments

We thank Marcus Kubsch for his thoughtful discussion of the manuscript and Dane DeSutter for his inspiring talk at the Biennial Conference on Chemical Education 2024. Authors PPM and NG thank Peter Wulff and David Kranz for their help in implementing the ML analysis. Furthermore, PPM and NG thank Leonie Lieber, Ira Caspari-Gnann, and Krenare Ibraj for providing access to their data and Felix Blödtner for scoring students' responses. PPM thanks the German Chemical Industry Association (Verband der Chemischen Industrie) for supporting him with the Kekulé Fellowship. Many thanks go to all the students who participated in the three studies. This publication is part of PPM's doctoral thesis (Dr. rer. nat.) at the Faculty of Biology and Chemistry, Justus-Liebig-University Giessen, Germany.

Conflict of interest statement

There are no conflicts of interest to declare.

References

Allogmany, B., & Josyula, D. (2022). An approach to dealing with incremental concept drift in personalized learning systems. *Proceedings of the 2022 IEEE 4th International Conference on Cognitive Machine Intelligence* (pp. 139-147). IEEE.

Ariely, M., Nazaretsky, T., & Alexandron, G. (2024). Causal-mechanical explanations in biology: Applying automated assessment for personalized learning in the science classroom. *Journal of Research in Science Teaching*, *61*(8), 1858-1889.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115.

Ashoori, M., & Weisz, J. D. (2019). In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. *arXiv preprint*, Article arXiv:1912.02675, 1-10.

Balda, E. R., Behboodi, A., & Mathar, R. (2020). Adversarial examples in deep neural networks: An overview. In W. Pedrycz & S.-M. Chen (Eds.), *Deep learning: Algorithms and applications* (pp. 31-65). Springer International Publishing.

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research*, *74*(1), 29-58.

Bejar, I. I., Mislevy, R. J., & Zhang, M. (2016). Automated scoring with validity in mind. In A. A. Rupp & J. P. Leighton (Eds.), *Handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 226-246). Wiley.

Bodé, N. E., Deng, J. M., & Flynn, A. B. (2019). Getting past the rules and to the WHY: Causal mechanistic arguments when judging the plausibility of organic reaction mechanisms. *Journal of Chemical Education*, *96*(6), 1068-1082.

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, *21*(1), Article 4, 1-41.

Braun, I., & Graulich, N. (2024). Exploring diversity: student's (un-) productive use of resonance in organic chemistry tasks through the lens of the coordination class theory. *Chemistry Education Research and Practice*, *25*(3), 643-671.

Braun, I., Langner, A., & Graulich, N. (2022). Let's draw molecules: Students' sequential drawing processes of resonance structures in organic

chemistry. *Frontiers in Education*, *7*, Article 1055280, 1-24.

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, *3*, Article 22, 1-12.

Bussey, T. J., Orgill, M., & Crippen, K. J. (2013). Variation theory: A theory of learning and a useful theoretical framework for chemical education research. *Chemistry Education Research and Practice*, *14*(1), 9-22.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Caspari, I., & Graulich, N. (2019). Scaffolding the structure of organic chemistry students' multivariate comparative mechanistic reasoning. *International Journal of Physics and Chemistry Education*, *11*(2), 31-43.

Caspari, I., Kranz, D., & Graulich, N. (2018). Resolving the complexity of organic chemistry students' reasoning through the lens of a mechanistic framework. *Chemistry Education Research and Practice*, *19*(4), 1117-1141.

Chen, V., Liao, Q. V., Wortman Vaughan, J., & Bansal, G. (2023). Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction* (pp. 1-32), *7*(CSCW2). Association for Computing Machinery.

Cheuk, T. (2021). Can AI be racist? Color-evasiveness in the application of machine learning to science assessments. *Science Education*, *105*(5), 825-836.

Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating e-rater®'s performance on TOEFL® essays. *ETS Research Report Series*, *2004*(1), i-38.

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, *15*(4), 413-432.

Conati, C., Porayska-Pomsta, K., & Mavrikis, M. (2018). AI in education needs interpretable machine learning: Lessons from open learner modelling. *arXiv preprint*, Article arXiv:1807.00154, 1-7.

Crandell, O. M., Kouyoumdjian, H., Underwood, S. M., & Cooper, M. M. (2019). Reasoning about reactions in organic chemistry: Starting it in general chemistry. *Journal of Chemical Education*, *96*(2), 213-226.

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT press.

Deng, J. M., Carle, M. S., & Flynn, A. B. (2023). Students' reasoning in chemistry arguments and designing resources using constructive alignment. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (pp. 74-89). The Royal Society of Chemistry.

Deng, J. M., & Flynn, A. B. (2021). Reasoning, granularity, and comparisons in students' arguments on two organic chemistry items. *Chemistry Education Research and Practice*, *22*(3), 749-771.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint,* Article arXiv: 1810.04805, 1-16.

Dood, A. J., Dood, J. C., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2020a). Analyzing explanations of substitution reactions using lexical analysis and logistic regression techniques. *Chemistry Education Research and Practice*, *21*(1), 267-286.

Dood, A. J., Dood, J. C., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2020b). Using the research literature to develop an adaptive intervention to improve student explanations of an S_N1 reaction mechanism. *Journal of Chemical Education*, *97*(10), 3551-3562.

Dood, A. J., Fields, K. B., & Raker, J. R. (2018). Using lexical analysis to predict Lewis acid-base model use in response to an acid-base proton-transfer reaction. *Journal of Chemical Education*, *95*(8), 1267-1275.

Dood, A. J., & Watts, F. M. (2022). Mechanistic reasoning in organic chemistry: A scoping review of how students describe and explain mechanisms in the chemistry education research literature. *Journal of Chemical Education*, *99*(8), 2864-2876.

Dood, A. J., & Watts, F. M. (2023). Students' strategies, struggles, and successes with mechanism problem solving in organic chemistry: A scoping review of the research literature. *Journal of Chemical Education*, *100*(1), 53-68.

Dood, A. J., Watts, F. M., Connor, M. C., & Shultz, G. V. (2024). Automated text analysis of organic chemistry students' written hypotheses. *Journal of Chemical Education*, *101*(3), 807-818.

Dood, A. J., Winograd, B. A., Finkenstaedt-Quinn, S. A., Gere, A. R., & Shultz, G. V. (2022). PeerBERT: Automated characterization of peer review comments across courses. *Proceedings of the 12th International Conference on Learning Analytics and Knowledge* (pp. 492-499). Association for Computing Machinery.

Dumont, H., & Ready, D. D. (2023). On the promise of personalized learning for educational equity. *npj Science of Learning*, *8*(1), Article 26, 1-6.

Eckhard, J., Rodemer, M., Langner, A., Bernholt, S., & Graulich, N. (2022). Let's frame it differently—Analysis of instructors' mechanistic explanations. *Chemistry Education Research and Practice*, *23*(1), 78-99.

Finkenstaedt-Quinn, S. A., Petterson, M., Gere, A., & Shultz, G. (2021). Praxis of Writing-to-Learn: A model for the design and propagation of Writing-to-Learn in STEM. *Journal of Chemical Education*, *98*(5), 1548-1555.

Finkenstaedt-Quinn, S. A., Watts, F. M., Shultz, G. V., & Gere, A. R. (2023). A portrait of MWrite as a research program: A review of research on Writing-to-Learn in STEM through the MWrite program. *International Journal for the Scholarship of Teaching and Learning*, *17*(1), Article 18.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

Gere, A. R., Limlamai, N., Wilson, E., MacDougall Saylor, K., & Pugh, R. (2019). Writing and conceptual learning in science: An analysis of assignments. *Written Communication*, *36*(1), 99-135.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics* (pp. 80-89). IEEE.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660.

Gombert, S., Di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., Grimm, A., Bohm, I., Neumann, K., & Drachsler, H. (2023). Coding energy knowledge in constructed responses with explainable NLP models. *Journal of Computer Assisted Learning*, *39*(3), 767-786.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Goodwin, W. M. (2003). Explanation in organic chemistry. *Annals of the New York Academy of Sciences*, *988*(1), 141-153.

Goodwin, W. M. (2008). Structural formulas and explanation in organic chemistry. *Foundations of Chemistry*, *10*(2), 117-127.

Graulich, N. (2015). The tip of the iceberg in organic chemistry classes: how do students deal with the invisible? *Chemistry Education Research Practice*, *16*(1), 9-21.

Graulich, N., & Caspari, I. (2021). Designing a scaffold for mechanistic reasoning in organic chemistry. *Chemistry Teacher International*, *3*(1), 19-30.

Graulich, N., & Lieber, L. S. (2024). Why comparing matters – on case comparisons in organic chemistry. *Frontiers in Education*, *9*, Article 1374793, 1-10.

Graulich, N., & Schween, M. (2018). Concept-oriented task design: Making purposeful case comparisons in organic chemistry. *Journal of Chemical Education*, *95*(3), 376-383.

Grimm, A., Steegh, A., Çolakoğlu, J., Kubsch, M., & Neumann, K. (2023). Positioning responsible learning analytics in the context of STEM identities of under-served students. *Frontiers in Education*, *7*, Article 1082748, 1-12.

Grimm, A., Steegh, A., Kubsch, M., & Neumann, K. (2023). Learning analytics in physics education: Equity-focused decision-making lacks guidance! *Journal of Learning Analytics*, *10*(1), 71-84.

Groh, M. (2022). Identifying the context shift between test benchmarks and production data. *arXiv preprint*, Article arXiv:2207.01059, 1-10.

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology Journal*, *29*(2), 75-91.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation*. Jossey-Bass.

Guba, E. G., & Lincoln, Y. S. (1982). Epistemological and methodological bases of naturalistic inquiry.

Educational Communication and Technology Journal, *30*(4), 233-252.

Gunel, M., Hand, B., & Prain, V. (2007). Writing for learning in science: A secondary analysis of six studies. *International Journal of Science and Mathematics Education*, *5*, 615-637.

Gupte, T., Watts, F. M., Schmidt-McCormack, J. A., Zaimi, I., Gere, A. R., & Shultz, G. V. (2021). Students' meaningful learning experiences from participating in organic chemistry Writing-to-Learn activities. *Chemistry Education Research and Practice*, *22*(2), 396-414.

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: Prospects and limitations. *CBE - Life Sciences Education*, *10*(4), 379-393.

Haas, D. B., Watts, F. M., Dood, A. J., & Shultz, G. V. (2024). Analysis of organic chemistry students' developing reasoning elicited by a scaffolded case comparison activity. *Chemistry Education Research and Practice*, *25*(3), 742-759

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice*, *20*(3), 281-307.

Hashmani, M. A., Jameel, S. M., Rehman, M., & Inoue, A. (2020). Concept drift evolution in machine learning approaches: A systematic literature review. *International Journal on Smart Sensing and Intelligent Systems*, *13*(1), 1-16.

Haudek, K. C., Moscarella, R. A., Urban-Lurain, M., Merrill, J. E., Sweeder, R. D., & Richmond, G. (2009). *Using lexical analysis software to understand student knowledge transfer between chemistry and biology*. Paper presented at the Annual Conference of the National Association of Research in Science Teaching, Garden Grove, CA.

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J. E., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. *CBE - Life Sciences Education*, *11*(3), 283-293.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, *32*(3), 504-526.

Holstein, K., & Doroudi, S. (2022). Equity and artificial intelligence in education. In W. Holmes & K. Porayska-Pomsta (Eds.), The ethics of artificial intelligence in education: Practices, challenges, and debates (pp. 151-173). Routledge.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 624-635). Association for Computing Machinery.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Doherty, J. H., Wenderoth, M. P., Merrill, J. E., Urban-Lurain, M., & Haudek, K. C. (2019). Deconstruction of holistic rubrics into analytic bins for large-scale assessments of students' reasoning of complex science concepts. *Practical Assessment, Research & Evaluation*, *24*(7), 1-13.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J. E., Urban-Lurain, M., Doherty, J. H., & Haudek, K. C. (2021). Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression. *Journal of Science Education and Technology*, *30*(2), 150-167.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17-64). American Council on Education/Praeger.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, Article 100074, 1-22.

Kliegr, T., Bahník, Š., & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, *295*, Article 103458, 1-33.

Kranz, D., Schween, M., & Graulich, N. (2023). Patterns of reasoning—Exploring the interplay of students' work with a scaffold and their conceptual knowledge in organic chemistry. *Chemistry Education Research and Practice*, *24*(2), 453-477.

Krist, C., & Kubsch, M. (2023). Bias, bias everywhere: A response to Li et al. and Zhai and Nehm. *Journal*

of Research in Science Teaching, *60*(10), 2395-2399.

Kubsch, M., Czinczel, B., Lossjew, J., Wyrwich, T., Bednorz, D., Bernholt, S., Fiedler, D., Strauß, S., Cress, U., Drachsler, H., Neumann, K., & Rummel, N. (2022). Toward learning progression analytics—Developing learning environments for the automated analysis of learning using evidence centered design. *Frontiers in Education*, *7*, Article 981910, 1-15.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, *86*(1), 42-78.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, Article 100210, 1-10.

Lee, Y.-W., Gentile, C., & Kantor, R. (2008). Analytic scoring of TOEFL® CBT essays: Scores from humans and e-rater®. *ETS Research Report Series*, *2008*(1), i-71.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy ML: From principles to practices. *ACM Computing Surveys*, *55*(9), Article 177, 171-146.

Li, T., Reigh, E., He, P., & Adah Miller, E. (2023). Can we and should we use artificial intelligence for formative assessment in science? *Journal of Research in Science Teaching*, *60*(6), 1385-1389.

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022a). Closing the gap of organic chemistry students' performance with an adaptive scaffold for argumentation patterns. *Chemistry Education Research and Practice*, *23*(4), 811-828.

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022b). Students' individual needs matter: A training to adaptively address students' argumentation skills in organic chemistry. *Journal of Chemical Education*, *99*(7), 2754-2761.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM Queue*, *16*(3), 31-57.

Liu, B., & Udell, M. (2020). Impact of accuracy on model interpretations. *arXiv preprint*, Article arXiv:2011.09903, 1-7.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A., & Tang, J. (2022). Trustworthy ML: A computational perspective. *ACM Transactions on Intelligent Systems and Technology*, *14*(1), Article 4, 1-59.

Lo, M. L., & Marton, F. (2012). Towards a science of the art of teaching. *International Journal for Lesson and Learning Studies*, *1*(1), 7-22.

Lo Piano, S. (2020). Ethical principles in machine learning and artificial intelligence: Cases from the field and possible ways forward. *Humanities and Social Sciences Communications*, *7*(1), Article 9, 1-7.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, *31*(12), 2346-2363.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765-4774). Curran Associates, Inc.

Martin, P. P., & Graulich, N. (2023). When a machine detects student reasoning: a review of machine learning-based formative assessment of mechanistic reasoning. *Chemistry Education Research and Practice*, *24*(2), 407-427.

Martin, P. P., & Graulich, N. (2024a). Beyond language barriers: Allowing multiple languages in postsecondary chemistry classes through multilingual machine learning. *Journal of Science Education and Technology*, *33*(3), 333-348.

Martin, P. P., & Graulich, N. (2024b). Lehre in der organischen Chemie individualisieren [Individualized teaching in organic chemistry]. *Nachrichten aus der Chemie*, *72*(3), 8-11.

Martin, P. P., & Graulich, N. (2024c). Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence. *Computers & Education: Artificial Intelligence*, *7*, Article 100265, 1-15.

Martin, P. P., Kranz, D., & Graulich, N. (2025). Revealing Rubric Relations: Investigating the Interdependence of a Research-Informed and a Machine Learning-Based Rubric in Assessing

Student Reasoning in Chemistry. *International Journal of Artificial Intelligence in Education*, 1-39.

Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2024). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*, *61*(8), 1757-1792.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709-734.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences*, *15*(2), 153-191.

Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C. M., & Tielman, M. L. (2023). A systematic review on fostering appropriate trust in human-AI interaction. *arXiv preprint,* Article arXiv:2311.06305, 1-39.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3-62.

National Research Council [NRC]. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* National Academies Press.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, *21*(1), 183-196.

Noyes, K., McKay, R. L., Neumann, M., Haudek, K. C., & Cooper, M. M. (2020). Developing computer resources to automate analysis of students' explanations of London dispersion forces. *Journal of Chemical Education*, *97*(11), 3923-3936.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates Inc.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, *51*(1), 59-81.

Polak, S., Schiavo, G., & Zancanaro, M. (2022). Teachers' perspective on artificial intelligence education: An initial investigation. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-7, Article 431). Association for Computing Machinery.

Prevost, L. B., Haudek, K. C., Merrill, J. E., & Urban-Lurain, M. (2012a). *Deciphering student ideas on thermodynamics using computerized lexical analysis of student writing.* Paper presented at the ASEE Annual Conference & Exposition, San Antonio, TX.

Prevost, L. B., Haudek, K. C., Merrill, J. E., & Urban-Lurain, M. (2012b). Examining student constructed explanations of thermodynamics using lexical analysis. *Proceedings of the 2012 IEEE 42nd Frontiers in Education Conference* (pp. 1-6). IEEE.

Raker, J. R., Yik, B. J., & Dood, A. J. (2022). Development of a generalizable framework for machine learning-based evaluation of written explanations of reaction mechanisms from the post-secondary organic chemistry curriculum. In N. Graulich & G. V. Shultz (Eds.), *Student Reasoning in Organic Chemistry: Research Advances and Evidence-based Instructional Practices* (pp. 304-319). The Royal Society of Chemistry.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). Evaluation of the e-rater® scoring engine for the TOEFL® independent and integrated prompts. *ETS Research Report Series*, *2012*(1), i-51.

Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the e-rater® automated scoring engine and humans for demographically based groups in the GRE®

general test. *ETS Research Report Series*, *2018*(1), 1-31.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95-112.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint*, Article arXiv:1606.05386, 1-5.

Rodemer, M., Eckhard, J., Graulich, N., & Bernholt, S. (2020). Decoding case comparisons in organic chemistry: Eye-tracking students' visual behavior. *Journal of Chemical Education*, *97*(10), 3530-3539.

Roszel, M., Norvill, R., Hilger, J., & State, R. (2021). Know your model (KYM): Increasing trust in AI and machine learning. *arXiv preprint*, Article arXiv:2106.11036, 1-11.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Rupp, A. A. (2018). Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied Measurement in Education*, *31*(3), 191-214.

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education*, *92*(3), 499-525.

Sadler, T. D., Mensah, F. M., & Tam, J. (2024). Artificial intelligence and the Journal of Research in Science Teaching. *Journal of Research in Science Teaching*, *61*(4), 739-743.

Saltelli, A., & Funtowicz, S. (2014). When all models are wrong. *Issues in Science and Technology*, *30*(2), 79-85.

Schlicker, N., & Langer, M. (2021). Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. *Proceedings of Mensch und Computer* (pp. 325-329). Association for Computing Machinery.

Sevian, H., & Talanquer, V. (2014). Rethinking chemistry: A learning progression on chemical thinking. *Chemistry Education Research and Practice*, *15*(1), 10-23.

Southard, K. M., Espindola, M. R., Zaepfel, S. D., & Bolger, M. S. (2017). Generative mechanistic explanation building in undergraduate molecular and cellular biology. *International Journal of Science Education*, *39*(13), 1795-1829.

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, Article 100075, 1-10.

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, *31*(2), 447-464.

Thimbleby, H. (2003). Explaining code for publication. *Software: Practice and Experience*, *33*(10), 975-1001.

Tomas, C., Whitt, E., Lavelle-Hill, R., & Severn, K. (2019). Modeling holistic marks with analytic rubrics. *Frontiers in Education*, *4*, Article 89, 1-19.

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & Van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 272-283). Association for Computing Machinery.

Tynjälä, P., Mason, L., & Lonka, K. (2001). *Writing as a learning tool: Integrating theory and practice*. Springer Netherlands.

Varshney, K. R. (2022). *Trustworthy machine learning*. Independently published.

Viberg, O., Kizilcec, R. F., Wise, A. F., Jivet, I., & Nixon, N. (2024). Advancing equity and inclusion in educational practices with AI-powered educational decision support systems (AI-EDSS). *British Journal of Educational Technology*, *55*(5), 1974-1981.

Visser, R., Peters, T. M., Scharlau, I., & Hammer, B. (2023). Trust, distrust, and appropriate reliance in (X)AI: A survey of empirical evaluation of user trust. *arXiv preprint*, Article arXiv:2312.02034, 1-50.

Wang, C., Liu, X., Wang, L., Sun, Y., & Zhang, H. (2021). Automated scoring of Chinese grades 7-9 students' competence in interpreting and arguing from evidence. *Journal of Science Education and Technology*, *30*(2), 269-282.

Watson, D. S., & Floridi, L. (2021). The explanation game: A formal framework for interpretable machine learning. In L. Floridi (Ed.), *Ethics, governance, and policies in artificial intelligence* (pp. 185-219). Springer.

Watts, F. M., Dood, A. J., & Shultz, G. V. (2023). Automated, content-focused feedback for a Writing-to-Learn assignment in an undergraduate organic chemistry course. *Proceedings of the 13th International Conference on Learning Analytics and Knowledge* (pp. 531-537). Association for Computing Machinery.

Watts, F. M., Dood, A. J., & Shultz, G. V. (2022). Developing machine learning models for automated analysis of organic chemistry students' written descriptions of organic reaction mechanisms. In N. Graulich & G. V. Shultz (Eds.), *Student Reasoning in Organic Chemistry: Research Advances and Evidence-based Instructional Practices* (pp. 285-303). The Royal Society of Chemistry.

Watts, F. M., & Finkenstaedt-Quinn, S. A. (2021). The current state of methods for establishing reliability in qualitative chemistry education research articles. *Chemistry Education Research and Practice*, *22*(3), 565-578.

Watts, F. M., Park, G. Y., Petterson, M. N., & Shultz, G. V. (2022). Considering alternative reaction mechanisms: Students' use of multiple representations to reason about mechanisms for a writing-to-learn assignment. *Chemistry Education Research and Practice*, *23*(2), 486-507.

Watts, F. M., Schmidt-McCormack, J. A., Wilhelm, C. A., Karlin, A., Sattar, A., Thompson, B. C., Gere, A. R., & Shultz, G. V. (2020). What students write about when students write about mechanisms: Analysis of features present in students' written descriptions of an organic reaction mechanism. *Chemistry Education Research and Practice*, *21*(4), 1148-1172.

Watts, F. M., Zaimi, I., Kranz, D., Graulich, N., & Shultz, G. V. (2021). Investigating students' reasoning over time for case comparisons of acyl transfer reaction mechanisms. *Chemistry Education Research and Practice*, *22*(2), 364-381.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2-13.

Wu, X., Zhao, H., Zhu, Y., Shi, Y., Yang, F., Liu, T., Zhai, X., Yao, W., Li, J., & Du, M. (2024). Usable XAI: 10 strategies towards exploiting explainability in the LLM era. *arXiv preprint*, Article arXiv:2403.08946, 1-56.

Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., & Fu, Y. (2019). Large scale incremental learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 374-382). IEEE.

Wulff, P., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2023). Enhancing writing analytics in science education research with machine learning and natural language processing—Formative assessment of science and non-science preservice teachers' written reflections. *Frontiers in Education*, *7*, Article 1061461, 1-18.

Xie, H., Chu, H. C., Hwang, G. J., & Wang, C. C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Computers & Education*, *140*, Article 103599, 1-16.

Yik, B. J., Dood, A. J., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2021). Development of a machine learning-based tool to evaluate correct Lewis acid-base model use in written responses to open-ended formative assessment items. *Chemistry Education Research and Practice*, *22*(4), 866-885.

Zaimi, I., Dood, A. J., & Shultz, G. V. (2024). The evolution of an assignment: how a Writing-to-Learn assignment's design shapes organic chemistry students' elaborations on reaction mechanisms. *Chemistry Education Research and Practice*, *25*(1), 327-342.

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data—Which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, *16*(1), 93-103.

Zhai, X. (2024). AI and machine learning for Next Generation Science Assessments. In H. Jiao & R. W. Lissitz (Eds.), *Machine learning, natural language processing, and psychometrics* (pp. 201-218). Information Age Publisher.

Zhai, X., Haudek, K. C., Shi, L., Nehm, R. H., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching*, *57*(9), 1430-1459.

Zhai, X., Krajcik, J., & Pellegrino, J. W. (2021). On the validity of machine learning-based Next Generation Science Assessments: A validity

inferential network. *Journal of Science Education and Technology*, *30*(2), 298-312.

Zhai, X., & Nehm, R. H. (2023). AI and formative assessment: The train has left the station. *Journal of Research in Science Teaching*, *60*(6), 1390-1398.

Zhai, X., Shi, L., & Nehm, R. H. (2021). A meta-analysis of machine learning-based science assessments: Factors impacting machine-human score agreements. *Journal of Science Education and Technology*, *30*(3), 361-379.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education*, *56*(1), 111-151.

SUPPORTING INFORMATION: PERFORMANCE OF THE ML MODELS BY MECHANISMS AND LABELS

We developed ML models that automatically evaluate students' mechanistic reasoning in organic chemistry to provide adaptive support across reaction mechanisms. To ensure accurate automated formative assessment, these ML models must maintain consistent performance across reaction mechanisms and scoring rubric labels. Therefore, we assessed the performance of both ML models across these two variables.

Tables S1 and S2 present the precision and recall of the holistic ML model—trained and tested on the balanced dataset (Section 6.3)—broken down by mechanism type and scoring rubric label. For precision, among the 20 labels, 10 exhibit *almost perfect* machine-human score agreement, 5 *substantial* agreement, 3 *moderate* agreement, 1 *fair* agreement, and 1 *slight* agreement. For recall, 12 labels achieve *almost perfect* agreement, 3 show *substantial* agreement, 1 *moderate* agreement, 3 *fair* agreement, and 1 *slight* agreement. Labels that perform well in precision and recall and are represented in at least three different reaction mechanisms are *nucleophiles and electrophiles—phenomenological—descriptive*, *nucleophiles and electrophiles—electronic—descriptive*, *nucleophiles and electrophiles—electronic—relational or linear causal*, and *acids and bases—phenomenological—descriptive*. Labels with lower performance are *charged species and substrates—electronic—relational*, *charged species and substrates—electronic—linear causal*, and *thermodynamics and kinetics—phenomenological—descriptive*.

For the holistic ML model, the following trends emerge from precision and recall (Tables S1 and S2): The ML model demonstrates the highest performance for the disciplinary core ideas of *nucleophiles and electrophiles* as well as *thermodynamics and kinetics*, slightly lower performance for *acids and bases*, and significantly lower performance for *charged species and substrates*. Moreover, performance declines as students' mechanistic reasoning transitions from *descriptive* to *relational* and further to *linear causal* reasoning. Additionally, the model performs slightly better for *phenomenological* reasoning than for *electronic* reasoning. These trends are influenced by the number and heterogeneity of responses in each rubric category. Particularly, the fewest student responses addressed the disciplinary core idea of *charged species and substrates*, which complicates the automated analysis of this idea. Similarly, significantly more student responses show *descriptive* or *relational* reasoning than *linear causal* reasoning, making the latter most difficult to automate. Besides, the vocabulary used in *phenomenological* responses tends to be less diverse than that in *electronic* responses, which facilitates the automated analysis of *phenomenological* responses.

Tables S3 and S4 provide the precision and recall of the analytic ML model—trained and tested on the balanced dataset (Section 6.3)—broken down by mechanism type and scoring rubric label. Among the 10 labels, precision reveals that 5 achieve *almost perfect* machine-human score agreement, 3 demonstrate *substantial* agreement, and 2 show *moderate* agreement. For recall, 7 labels achieve *almost perfect* agreement, 1 shows *substantial* agreement, 1

TABLE S1 Precision of the holistic machine learning model trained and tested on the balanced dataset separated by mechanism type and scoring rubric label.

Label	AVE	WR	TH	CC	ES	SN	SA	SB
Absent	0.720	0.800	0.700	0.000	*n.a.*	1.000	1.000	0.500
CS-PHEN-DES	0.744	0.833	0.800	0.556	0.769	*n.a.*	*n.a.*	*n.a.*
CS-PHEN-REL	0.833	*n.a.*	0.000	1.000	1.000	*n.a.*	0.000	*n.a.*
CS/TK-ELEC-DES	0.887	0.968	0.824	0.625	1.000	1.000	0.000	1.000
CS-ELEC-REL	0.286	*n.a.*	0.000	0.250	0.000	1.000	*n.a.*	*n.a.*
CS-ELEC-CAU	0.000	*n.a.*	0.000	0.000	0.000	0.000	*n.a.*	*n.a.*
NE-PHEN-DES	0.968	1.000	*n.a.*	1.000	1.000	*n.a.*	0.917	1.000
NE-PHEN-REL	0.600	*n.a.*	*n.a.*	0.500	0.500	*n.a.*	1.000	*n.a.*
NE-PHEN-CAU	0.667	*n.a.*	*n.a.*	1.000	0.000	*n.a.*	*n.a.*	*n.a.*
NE-ELEC-DES	0.905	1.000	1.000	1.000	1.000	0.600	1.000	1.000
NE-ELEC-REL/CAU	0.992	1.000	0.500	1.000	1.000	1.000	1.000	1.000
AB-PHEN-DES	0.926	0.500	1.000	0.750	1.000	*n.a.*	*n.a.*	*n.a.*
AB-PHEN-REL	0.667	*n.a.*	*n.a.*	0.000	1.000	*n.a.*	*n.a.*	*n.a.*
AB-PHEN-CAU	0.500	*n.a.*	*n.a.*	0.000	1.000	*n.a.*	*n.a.*	*n.a.*
AB-ELEC-DES/REL	0.969	1.000	0.952	1.000	1.000	0.000	*n.a.*	*n.a.*
TK-PHEN-DES	0.500	0.000	*n.a.*	1.000	1.000	*n.a.*	*n.a.*	0.000
TK-PHEN-REL	0.909	*n.a.*	*n.a.*	1.000	0.833	*n.a.*	*n.a.*	*n.a.*
TK-PHEN-CAU	1.000	*n.a.*	*n.a.*	*n.a.*	1.000	*n.a.*	*n.a.*	*n.a.*
TK-ELEC-REL	0.847	1.000	0.000	0.727	0.000	0.903	1.000	1.000
TK-ELEC-CAU	0.800	0.000	*n.a.*	1.000	*n.a.*	0.750	*n.a.*	1.000

Note: AVE = average across all mechanisms, WR = base-free Wittig reaction, TH = thalidomide hydrolysis, CC = Claisen condensation, ES = Williamson ether synthesis, SN = S_N2, SA = S_N2_t A, SB = S_N2_t B. CS = charged species and substrates, NE = nucleophiles and electrophiles, AB = acids and bases, TK = thermodynamics and kinetics, PHEN = phenomenological, ELEC = electronic, DES = descriptive, REL = relational, CAU = linear causal. The abbreviation *n.a.* (not assigned) indicates that a label was not used within a specific mechanism.

moderate agreement, and 1 *slight* agreement. Labels that perform well in precision and recall include *charges*, *electron movement*, and *non-electronic mechanism*. These labels align across the rubrics: *Non-electronic* explanations in the analytic rubric are termed *phenomenological* in the holistic one; *electron movement* in the analytic rubric refers to *electronic* reasoning in the holistic one. Further, the focus on *charges* in the analytic rubric corresponds to *electronic*, *descriptive* accounts in the holistic rubric. Conversely, labels such as *stereochemistry* and *formation of stereochemistry* perform poorly. The ML model faced challenges in classifying instances related to stereochemistry due to the imbalanced nature of these categories (Table 4). This imbalance reflects the task design of the reasoning-focused short writing and the case comparison activity, which did not explicitly prompt students to address stereochemistry, resulting in its underrepresentation in the data.

TABLE S2 Recall of the holistic machine learning model trained and tested on the balanced dataset separated by mechanism type and scoring rubric label.

Label	AVE	WR	TH	CC	ES	SN	SA	SB
Absent	0.857	0.800	1.000	0.000	*n.a.*	1.000	1.000	1.000
CS-PHEN-DES	0.879	0.909	0.667	0.833	1.000	*n.a.*	*n.a.*	*n.a.*
CS-PHEN-REL	0.625	*n.a.*	0.000	0.333	1.000	*n.a.*	0.000	*n.a.*
CS/TK-ELEC-DES	0.873	1.000	0.875	0.625	0.500	1.000	0.000	1.000
CS-ELEC-REL	0.400	*n.a.*	0.000	1.000	0.000	1.000	*n.a.*	*n.a.*
CS-ELEC-CAU	0.000	*n.a.*	0.000	0.000	0.000	0.000	*n.a.*	*n.a.*
NE-PHEN-DES	0.968	1.000	*n.a.*	1.000	0.875	*n.a.*	1.000	1.000
NE-PHEN-REL	1.000	*n.a.*	*n.a.*	1.000	1.000	*n.a.*	1.000	*n.a.*
NE-PHEN-CAU	0.250	*n.a.*	*n.a.*	0.500	0.000	*n.a.*	*n.a.*	*n.a.*
NE-ELEC-DES	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
NE-ELEC-REL/CAU	0.977	1.000	1.000	1.000	1.000	0.944	0.979	1.000
AB-PHEN-DES	0.926	0.500	0.909	1.000	1.000	*n.a.*	*n.a.*	*n.a.*
AB-PHEN-REL	1.000	*n.a.*	*n.a.*	1.000	1.000	*n.a.*	*n.a.*	*n.a.*
AB-PHEN-CAU	0.333	*n.a.*	*n.a.*	0.000	1.000	*n.a.*	*n.a.*	*n.a.*
AB-ELEC-DES/REL	0.775	0.857	0.833	0.250	1.000	0.000	*n.a.*	*n.a.*
TK-PHEN-DES	0.455	0.000	*n.a.*	0.333	0.667	*n.a.*	*n.a.*	0.000
TK-PHEN-REL	0.909	*n.a.*	*n.a.*	0.833	1.000	*n.a.*	*n.a.*	*n.a.*
TK-PHEN-CAU	1.000	*n.a.*	*n.a.*	*n.a.*	1.000	*n.a.*	*n.a.*	*n.a.*
TK-ELEC-REL	0.943	0.667	0.000	0.889	0.000	0.966	1.000	1.000
TK-ELEC-CAU	0.800	0.000	*n.a.*	0.750	*n.a.*	0.750	*n.a.*	1.000

Note: AVE = average across all mechanisms, WR = base-free Wittig reaction, TH = thalidomide hydrolysis, CC = Claisen condensation, ES = Williamson ether synthesis, SN = S_N2, SA = S_N2_t A, SB = S_N2_t B. CS = charged species and substrates, NE = nucleophiles and electrophiles, AB = acids and bases, TK = thermodynamics and kinetics, PHEN = phenomenological, ELEC = electronic, DES = descriptive, REL = relational, CAU = linear causal. The abbreviation *n.a.* (not assigned) indicates that a label was not used within a specific mechanism.

TABLE S3 Precision of the analytic machine learning model trained and tested on the balanced dataset separated by mechanism type and scoring rubric label.

Label	AVE	WR	TH	CC	ES	SN	SA	SB
Reaction medium	0.750	*n.d.*	*n.d.*	*n.d.*	*n.d.*	*n.d.*	*n.d.*	0.750
Connectivity of molecules	0.780	0.500	0.429	0.920	0.600	0.786	0.807	0.783
Charges	0.938	1.000	0.955	0.778	0.895	0.833	1.000	0.917
Stereochemistry	0.522	0.500	0.500	*n.d.*	*n.d.*	*n.d.*	0.667	0.600
Electron movement	0.915	0.750	0.500	1.000	0.900	0.889	0.909	1.000
Non-electronic mechanism	0.932	0.870	0.922	0.971	0.958	0.986	0.870	0.750
Bond breaking and making	0.879	0.429	0.800	1.000	0.818	0.963	0.867	0.905
Implicit properties	0.788	0.767	0.800	0.750	0.857	0.857	0.655	0.903
Formation of stereochemistry	0.500	*n.d.*	*n.d.*	*n.d.*	*n.d.*	*n.d.*	*n.d.*	0.500
Electronic causal reasoning	0.826	0.833	0.765	0.844	0.750	0.886	1.000	0.000

Note: AVE = average across all mechanisms, WR = base-free Wittig reaction, TH = thalidomide hydrolysis, CC = Claisen condensation, ES = Williamson ether synthesis, SN = S_N2, SA = S_N2_t A, SB = S_N2_t B. The abbreviation *n.d.* (not defined) is used when neither true positives nor false positives exist.

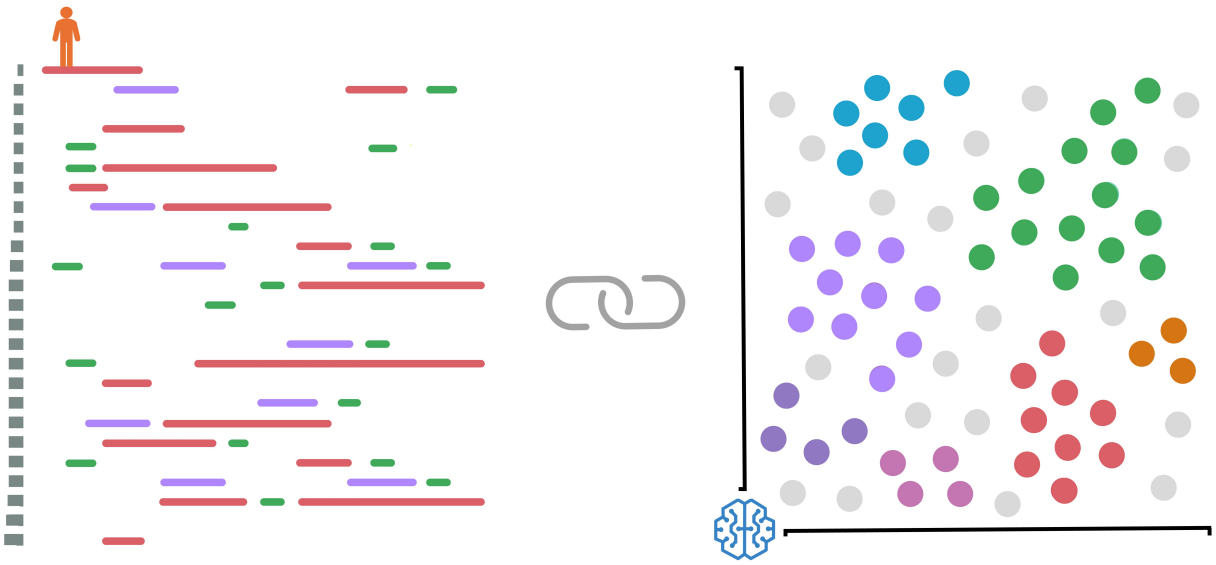
TABLE S4 Recall of the analytic machine learning model trained and tested on the balanced dataset separated by mechanism type and scoring rubric label.

Label	AVE	WR	TH	CC	ES	SN	SA	SB
Reaction medium	1.000	*n.d.*	*n.d.*	*n.d.*	*n.d.*	*n.d.*	*n.d.*	1.000
Connectivity of molecules	0.859	0.500	0.429	0.958	0.600	0.846	0.962	0.900
Charges	0.982	1.000	0.429	1.000	0.895	1.000	1.000	1.000
Stereochemistry	0.600	0.833	0.333	*n.d.*	*n.d.*	*n.d.*	1.000	0.500
Electron movement	0.949	0.750	0.667	1.000	1.000	0.889	0.952	1.000
Non-electronic mechanism	0.988	1.000	0.979	1.000	1.000	1.000	0.952	0.857
Bond breaking and making	0.936	1.000	0.800	0.929	0.900	1.000	0.813	1.000
Implicit properties	0.885	1.000	0.333	0.750	0.750	0.600	0.826	0.933
Formation of stereochemistry	0.200	*n.d.*	0.000	*n.d.*	*n.d.*	*n.d.*	0.000	0.333
Electronic causal reasoning	0.731	0.588	0.722	0.771	0.750	0.838	0.167	0.000

Note: AVE = average across all mechanisms, WR = base-free Wittig reaction, TH = thalidomide hydrolysis, CC = Claisen condensation, ES = Williamson ether synthesis, SN = S_N2, SA = S_N2_t A, SB = S_N2_t B. The abbreviation *n.d.* (not defined) is used when neither true positives nor false negatives exist.

7 | Revealing rubric relations: Investigating the interdependence of a research-informed and a machine learning-based rubric in assessing student reasoning in chemistry

Chapter 7 presents an original research paper. It is reproduced from Martin, Kranz, and Graulich (2024) under the terms of a Creative Commons Attribution 4.0 International License (CC BY 4.0) (https://creativecommons.org/licenses/by/4.0/). No changes have been made to the original article.



Martin, P. P., Kranz, D., & Graulich, N. (2024). Revealing rubric relations: Investigating the interdependence of a research-informed and a machine learning-based rubric in assessing student reasoning in chemistry. *International Journal of Artificial Intelligence in Education*, 1-39. https://doi.org/10.1016/j.caeai.2024.100265

International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-024-00440-y

ARTICLE



Revealing Rubric Relations: Investigating the Interdependence of a Research-Informed and a Machine Learning-Based Rubric in Assessing Student Reasoning in Chemistry

Paul P. Martin[1] · **David Kranz**[1] · **Nicole Graulich**[1]

Accepted: 6 November 2024
© The Author(s) 2024

Abstract

Valid rubrics facilitate assessing the level of complexity in students' open-ended responses. To design a valid rubric, it is essential to thoroughly define the types of responses that represent evidence of varying complexity levels. Formulating such evidence statements can be approached deductively by adopting predefined criteria from the research literature or inductively by detecting topics, for example, based on data-driven machine learning (ML) techniques. Investigating the interdependence of such research-informed and ML-based rubrics is key to validating ML-based approaches and enhancing their applicability in formative assessments. This study quantitatively compares a research-informed and an ML-based rubric designed to capture the complexity of students' reasoning on the relative rate of contrasted reactions in undergraduate organic chemistry. Specifically, we leveraged an ML-based clustering technique to inductively develop a holistic fifteen-category rubric to evaluate students' open-ended reasoning. Subsequently, we performed a quantitative analysis to examine whether the ML-based rubric and its research-informed counterpart are significantly associated. Our findings indicate that research-informed and ML-based rubrics assess students' reasoning comparably. Thus, both rubric types are valid for categorizing students' reasoning, underscoring the applicability of integrating ML techniques into rubric development. Nevertheless, aligning ML-based rubrics with the respective assessment objectives remains crucial. A well-aligned, evidence-based rubric may ultimately ease the analysis of student reasoning.

Keywords Machine learning · Validity · Formative assessment · Rubric design · Organic chemistry

Paul P. Martin and David Kranz contributed equally to this manuscript.

Extended author information available on the last page of the article

Published online: 10 December 2024

Introduction

Developing valid rubrics is essential in educational research and instruction because they can provide detailed insights into student performance, clarify expectations, standardize scoring, and enable educators to offer constructive feedback (Brookhart & Chen, 2015; Jonsson & Svingby, 2007). Valid rubrics are especially important when evaluating the complexity of students' ambiguous open-ended reasoning since they can offer clear criteria for assessing varied responses. Researchers and instructors can approach the development of valid rubrics in line with evidence-centered design (e.g., Mislevy, 2016; Mislevy, Almond et al., 2003; Pellegrino et al., 2016). Before rubric development, this strategy for educational assessment design demands a clear definition of the skills to be acquired (Mislevy, 2016; Pellegrino et al., 2016). Then, researchers or instructors must determine which elements in students' responses constitute evidence of skill acquisition, how these elements are interpreted and coded, and what techniques are used to validate these interpretations (Mislevy, Almond et al., 2003; Rupp et al., 2012). Finally, rubrics should be adjusted depending on the tasks at hand and the characteristics of the population that is going to respond to these tasks (Kubsch et al., 2022; Mislevy, Almond et al., 2003; Mislevy & Haertel, 2007; Pellegrino et al., 2016). Collectively, these considerations highlight that rubric design is an integral part of developing assessments that help draw coherent conclusions from students' responses to their skills (Mislevy, 2016; Mislevy, Almond et al., 2003; Mislevy & Haertel, 2007; Mislevy, Steinberg et al., 2003; Pellegrino et al., 2016).

Nonetheless, designing valid rubrics is a complex endeavor (Panadero & Jonsson, 2020). Rubrics must align with learning objectives, assessment methods, and instructional strategies, which can be challenging. Furthermore, many learning objectives involve complex skills that are difficult to define, complicating their quantification within a rubric. Also, rubrics must strike a balance between being specific enough to provide clear criteria for evaluating students' reasoning and flexible enough to accommodate diverse responses. Ultimately, designing, testing, and refining rubrics requires time and effort, which can pose a barrier for educators.

Alongside these overarching challenges, domain-specific obstacles also exist when developing valid rubrics, for example, in undergraduate organic chemistry. Organic chemistry involves predicting the selectivity of transformations, designing new synthesis routes, and critically evaluating the plausibility of reactions—tasks necessitating *mechanistic reasoning* (Bhattacharyya & Bodner, 2005). Mechanistic reasoning is a science practice that goes beyond simply describing the outcome of chemical reactions (Dood & Watts, 2022; Graulich, 2015). It entails deducing the underlying processes of observed phenomena to explain *how* and *why* changes occur at the molecular level (Cooper et al., 2016; Russ et al., 2008). When chemists use mechanistic reasoning, they aim to rationalize the mechanistic pathway from reactants to products (Machamer et al., 2000). This includes analyzing the movement of electrons, the formation and breaking of chemical bonds, and the energetic level of the reagents (Goodwin, 2003, 2008).

Following that, mechanistic reasoning is a multifaceted science practice, and capturing its complexity in a rubric is not straightforward (Caspari et al., 2018).

In addition to the intricate nature of mechanistic reasoning, the challenges students face when reasoning about mechanisms make it also difficult to develop valid rubrics that capture various skill levels (Raker et al., 2022). Students' challenges include establishing cause-effect relations (Crowder et al., 2024; Frost et al., 2023; Weinrich & Talanquer, 2016; Yik, Dood et al., 2023), integrating implicit properties in their reasoning (Graulich et al., 2019), or connecting structural and energetic accounts to chemical reactions (Caspari et al., 2018; Pölloth et al., 2023). Students can only overcome these challenges and demonstrate more complex mechanistic reasoning when these skills are consistently assessed and supported (DeGlopper et al., 2022; Stowe & Cooper, 2017; Stowe et al., 2021). Hence, valid assessments are necessary to capture the complexity of students' mechanistic reasoning, requiring rubrics that accommodate diverse student challenges.

To address these design obstacles, rubrics can be developed deductively by incorporating predefined criteria from the research literature or inductively by identifying topics in students' responses. Advances in machine learning (ML) have facilitated the inductive development of rubrics, alleviating educators from the time-consuming task of designing rubrics from scratch. However, little is known about the validity of using ML to define rubrics, whether research-informed and ML-based rubrics map skill levels similarly, and the degree to which these rubric types are associated.

This study compares a research-informed and an ML-based rubric for analyzing undergraduate organic chemistry students' mechanistic reasoning to determine whether ML-based rubrics are valid for capturing students' mechanistic reasoning and whether it is appropriate to ground formative assessments exclusively on these ML-based rubrics. Examining the interdependence of both rubric types helps integrate their benefits: On the one hand, ML techniques could be used to confirm frameworks proposed in education research because these techniques offer additional data-driven insights into students' reasoning (Sherin, 2013). On the other hand, research-informed frameworks could inform the interpretation of an ML analysis since existing frameworks add depth to data-driven classifications (Nelson, 2020). This study explores whether integrating ML techniques into rubric development advances the validity of formative assessments.

Background

Methodological Basics: Applying Machine Learning in Educational Assessment

Artificial intelligence (AI) encompasses a range of technologies that simulate human abilities, allowing machines to perform human tasks (Bellmann, 1978; Haugeland, 1989). In education, AI holds great potential to transform teaching and learning (Kubsch et al., 2023; Martin & Graulich, 2023; Zhai, Haudek et al., 2020; Zhai, Yin et al., 2020l). For instance, intelligent tutoring systems

can assess students' learning needs, offer real-time feedback, and deliver personalized exercises (Deeva et al., 2021), reducing the workload for instructors and enabling faster assessment processes (Urban-Lurain et al., 2013).

ML—a subarea of AI—is centered around creating algorithms that allow computers to learn from data (Bishop, 2006; Mitchell, 1997; Mohri et al., 2012; Samuel, 1959). Over the past fifteen years, ML has brought significant advancements in education (Deeva et al., 2021; Gerard et al., 2015; Martin & Graulich, 2023; Zhai, Haudek et al., 2020; Zhai, Yin et al., 2020), for example, in adaptive learning (e.g., Lim et al., 2023; Sailer et al., 2023). One type of ML is unsupervised learning, which involves detecting patterns or relationships within unlabeled data. Unlike supervised learning, which relies on labeled data, unsupervised ML algorithms are used in educational studies to identify similarities, group data points, reveal underlying patterns, and inductively extract data-driven topics (e.g., Anderson et al., 2020; Haudek et al., 2015; Prevost et al., 2012; Rosenberg & Krist, 2021; Sherin, 2013; Tschisgale et al., 2023; Wulff et al., 2022; Wulff, Westphal et al., 2023; Zehner et al., 2016). This approach is especially useful when dealing with large, unstructured datasets because it eases data exploration and pattern identification. Despite potential benefits, unsupervised ML is still rarely applied in educational studies (Zhai, Haudek et al., 2020; Zhai, Yin et al., 2020), particularly in studies on mechanistic reasoning in chemistry (Martin & Graulich, 2023).

To evaluate human language with ML, text data must undergo preprocessing using appropriate natural language processing (NLP) techniques. NLP covers various methods that enable computers to analyze and generate human language. In recent years, large language models have emerged as a significant milestone in NLP (Taher Pilehvar & Camacho-Collados, 2020). These large language models are trained on massive amounts of unlabeled text data—often billions of sentences—to undertake various language-related tasks, such as question-answering, text classification, or text generation (Radford et al., 2019). Large language models are also increasingly applied in education research since they can process words related to each other, even when they are far apart in a sentence (e.g., Dood et al., 2022, 2024; Gombert et al., 2023; Martin et al., 2024; Winograd, Dood, Moon et al., 2021; Winograd, Dood, Finkenstaedt-Quinn et al., 2021; Wulff, Mientus et al., 2023). Additionally, the pre-training of these large language models on extensive text corpora improves model generalizability and accuracy, while less training data is required for model training (Vaswani et al., 2017).

Due to these recent advancements, incorporating ML and NLP techniques in teaching and learning holds promise but a successful application relies on the design of valid rubrics that clarify what responses are representative of different complexity levels (Kubsch et al., 2022; Martin & Graulich, 2023). Educators can develop such rubrics deductively using predefined research-informed criteria or inductively, for example, by identifying categories through data-driven ML techniques. However, data-driven categories require human interpretation to capture the specifics of a context. For this reason, a methodological framework that outlines how to integrate human qualitative interpretation into ML techniques guides the analysis presented herein.

Methodological Framework: *Computational Grounded Theory*

Computational grounded theory specifies how unsupervised clustering can inform the development of a rubric (Carlsen & Ralund, 2022; Nelson, 2020). Generally, *computational grounded theory* builds on the traditional grounded theory approach, which involves creating theories about social phenomena by closely examining data (Glaser & Strauss, 1999). This approach allows topics to emerge inductively from data, rather than being imposed on it, to provide data-driven insights into social phenomena that are not directly observable (Charmaz, 2014). However, because grounded theory relies on subjective decisions regarding data interpretation, it can lead to biased outcomes (Saldana, 2015), produce hypotheses that are difficult to replicate (Biernacki, 2012), and be less effective for analyzing large, unstructured datasets (Bail, 2014). To address these limitations, *computational grounded theory* integrates qualitative research with computational techniques to analyze large datasets systematically for theory development (Kubsch et al., 2023; Nelson, 2020; Rosenberg & Krist, 2021). This approach helps identify patterns in unstructured data, inductively generate theories about the phenomena under study, and validate these theories through quantitative testing (Nelson, 2020).

In practice, *computational grounded theory* outlines a three-step process to guide researchers in rubric design and application (Nelson, 2020): The initial step, the *pattern detection step*, involves computational methods to convert complex, content-rich text into interpretable categories. Specifically, unsupervised ML algorithms can be applied in this step to uncover reproducible data patterns. Inductively detecting patterns can be an initial step in developing ML-based rubrics since it allows researchers to derive data-driven categories from student reasoning, pre-structure conceptually different ideas into distinct categories, and extract student examples representing each category. However, pattern detection only assigns numerical labels to student responses with each label corresponding to a specific category. Humans must then identify the common characteristics among the responses sharing the same label. Consequently, human experts must interpret the data-driven categories using research-informed thematic content analysis in the following *pattern refinement step*. This step allows the finalization of the pre-structured rubric since human experts can refine the data-driven categories. Besides, merging conceptually related categories or introducing new ones is also possible. In the final *pattern confirmation step*, computational methods like supervised ML can test the reliability and generalizability of the inductively derived patterns. This step is useful for applying the rubric in practice because ML models can be developed that automatically evaluate students' responses based on the established rubric.

This article examines the validity of the first two steps of *computational grounded theory* as little is known about how these steps interact. In doing so, we applied an unsupervised clustering technique to extract categories from undergraduate organic chemistry students' reasoning about reaction mechanisms. Next, we interpreted these categories based on the concepts students integrated into their mechanistic reasoning. We then quantitatively investigated the association between this ML-based rubric and a traditional research-informed approach. In other words, we examined how well the categories extracted by an unsupervised ML algorithm correspond to a

research-informed rubric based on specific types of reasoning in organic chemistry. Developing an ML classifier that automatically evaluates student reasoning within this approach is beyond the scope of this article but is covered in detail elsewhere (e.g., Martin et al., 2024; Rosenberg & Krist, 2021; Tschisgale et al., 2023).

Literature Review: Comparing Rubric Types in Educational Assessment

Our comparison of ML-based and research-informed rubrics contributes to the existing literature on the similarities and differences between various rubric types in educational assessment, primarily focusing on analytic and holistic rubrics (e.g., Franovic et al., 2023; Haudek et al., 2015; Jescovitch, Doherty et al., 2019; Jescovitch, Scott et al., 2019; Jescovitch et al., 2021; Kaldaras et al., 2022; Liu et al., 2014; Noyes et al., 2022; Wang et al., 2021; Wilson et al., 2023; Wulff, Westphal et al., 2023). These two rubric types are frequently applied for manual or ML-based scoring of open-ended assessment tasks. Analytic rubrics analyze responses in multiple binary categories, where each analytic component represents a unique idea. By contrast, holistic rubrics classify responses based on their overall quality, making them polytomous with each scoring level being mutually exclusive from the others. As indicated by previous research, choosing either an analytic or holistic rubric can impact the reliability and validity of the assessment results, human coding effort, and the performance of ML models (Jescovitch, Doherty et al., 2019; Jescovitch et al., 2021; Kaldaras et al., 2022; Wang et al., 2021). Therefore, understanding the similarities and differences between these two approaches can contribute to developing improved coding methods.

So far, some researchers merged analytic and holistic rubrics into one another. For example, Franovic et al. (2023) and Noyes et al. (2022) found that beginning with an analytic rubric allowed for a fine-grained analysis in the initial stages of evaluating students' reasoning on protein-ligand binding. Their analytic rubric—developed through multiple iterations—comprised three essential ideas derived from an ideal explanation. These three essential ideas allowed them to combine the analytic categories into three holistic performance levels with increasing complexity. In contrast, Jescovitch, Doherty et al. (2019), Jescovitch, Scott et al. (2019), Jescovitch et al. (2021), Kaldaras et al. (2022), Liu et al. (2014), and Wilson et al. (2023) broke down a holistic rubric into a set of analytic categories to clarify rubric criteria for machine coding, aiming to develop more accurate ML models than with holistic scoring. After developing reliable scoring models for each analytic category, they recombined these categories to form an overall holistic score using logic operators. These composite scores matched the initial holistic scores but produced higher machine-human score agreements than using a holistic rubric alone (Jescovitch, Doherty et al., 2019; Jescovitch et al., 2021; Kaldaras et al., 2022; Wilson et al., 2023). Consequently, they inferred that analytic rubrics identify additional complexity in students' responses, which was rationalized by the reduced ambiguity in human coding (Jescovitch, Doherty et al., 2019; Jescovitch et al., 2021; Wang et al., 2021). Combining both analytic and holistic elements in rubrics may, thus, increase

their validity and provide a more nuanced analysis of student responses (Jescovitch, Doherty et al., 2019; Noyes et al., 2022).

However, integrating analytic and holistic rubrics can pose challenges. When transforming a holistic rubric into analytic categories, it is challenging to incorporate all relevant categories and maintain a category number that still reflects the essence of the holistic rubric (Liu et al., 2014). Introducing distinct analytic categories that do not oversimplify the evaluation of complex constructs is especially challenging (Jescovitch, Doherty et al., 2019; Jescovitch, Scott et al., 2019; Kaldaras & Haudek, 2022; Kaldaras et al., 2022). Conversely, holistic rubrics have been shown to complicate the reliable scoring of student responses that contain scientifically normative and non-normative ideas since human raters must decide how to penalize errors (Liu et al., 2014). Effectively navigating these challenges is critical for creating valid rubrics that accurately categorize students' reasoning.

Overall, these findings suggest that no rubric type consistently outperforms the other (Brookhart, 2018). Previous ML-related studies in chemistry education research successfully constructed scoring models using either dichotomous analytic rubrics (e.g., Dood et al., 2018; Watts, Dood et al., 2022; Winograd, Dood, Moon et al., 2021; Yik et al., 2021; Yik, Schreurs et al., 2023) or multi-level holistic ones (e.g., Donnelly et al., 2015; Dood et al., 2020; Haudek et al., 2009, 2012, 2019; Liu et al., 2014; Maestrales et al., 2021; Noyes et al., 2020; Tansomboon et al., 2017; Vitale et al., 2016). Even a combination of both approaches can be useful (Harsch & Martin, 2013; Tomas et al., 2019). However, deciding between an analytic or holistic rubric is not strictly an empirical question since it may be possible to calculate the performance of ML models trained on these different rubric types in advance by considering factors such as inter-rater reliability, data distribution, and the number of rubric categories.

Similar validity and reliability issues arise from the increasing combination of research-informed and ML-based rubrics. For this reason, we explored whether an ML-based, exclusively data-driven rubric reflects research-informed considerations. One can assume that ML-based rubrics do not mirror existing frameworks per se since the corresponding unsupervised ML techniques detect clusters in data based on statistical similarity measured by the distance between data points in a high-dimensional space. Because of this emphasis on statistical values, applying ML in formative assessments could lead to neglecting evidence-based teaching practices, potentially harming student learning (Li et al., 2023). Comparing research-informed and ML-based rubrics may help clarify whether these concerns are reasonable and whether ML techniques are valid for rubric development.

Research Questions

This study investigates the association between research-informed and ML-based rubrics to analyze students' mechanistic reasoning in undergraduate organic chemistry. We prompted students in open-ended items to explain which of two contrasted reactions occurs at a higher rate and developed rubrics to capture the complexity of

students' mechanistic reasoning (Kranz et al., 2023, under review)[1]. We then examined whether an ML-based rubric reflects the numerous theory-rich considerations that led to the development of the initial research-informed rubric. Our analysis was guided by two research questions (RQs).

1. What clusters can be revealed by applying unsupervised ML to develop a rubric that captures students' mechanistic reasoning?
2. To what degree are the research-informed and ML-based rubrics associated?
    a Are there statistically significant associations between the computationally revealed clusters and research-informed categories? If so, what are the specific associations among the rubrics?
    b Do these associations remain when merging the research-informed categories? If so, what are the specific associations among the rubrics?

The research-informed rubric is detailed in Sect. 4.3 ("Research-Informed Rubric") and its ML-based counterpart is described in Sect. 5.1 ("Detecting ML-Based Clusters"). Initially, the research-informed rubric featured three *analytic* categories, whereas the ML-based rubric is *holistic*. To compensate for these design variations, we merged the three analytic categories into a holistic research-informed rubric (Sect. 4.5: "Quantitative Comparisons between the Research-Informed and ML-Based Rubrics"). We then investigated the associations between the computationally revealed clusters and the merged research-informed categories (RQ 2b). Transforming the research-informed analytic categories into a holistic rubric may improve the validity and robustness of our findings as both the merged research-informed and ML-based rubric use a mutually exclusive approach to scoring student reasoning.

Two potential outcomes are possible: If the two rubrics are not associated, the ML-based rubric would overlook research-informed considerations, neglecting valuable evidence-based contributions from education research and potentially reducing the validity of the applied ML technique. In contrast, if the two rubrics are associated, the ML-based rubric would naturally incorporate research-informed considerations, bolstering confidence in both approaches (Sherin, 2013). Following this, exploring these RQs advances our understanding of ML's potential to complement traditional formative assessments.

Methods

Setting of Data Collection

This study was conducted in the Winter 2021 and Summer 2022 semesters in four courses across three German universities and is part of a larger intervention study

[1] As the manuscript Kranz et al. (under review) had not yet been published, we provide a detailed description of the research-informed rubric in Sect. 4.3 ("Research-Informed Rubric"). The analysis presented herein can be fully understood without access to Kranz et al. (under review).

(Kranz et al., under review). The intervention aimed to support students' mechanistic reasoning through rubric-based formative assessments in undergraduate organic chemistry without initially applying ML techniques. The tasks were presented in a digital format. A total of 122 undergraduate students participated in this study, representing various backgrounds: 62 were chemistry bachelor students, 33 were preservice chemistry teachers, and 27 were enrolled in chemistry-related study programs such as food chemistry or materials science. The age of the participants ranged from 18 to 40 years ($M = 21.09$, $SD = 2.91$), with 62 identifying as male, 59 as female, and one as non-binary. All but four students were native German speakers. A total of 454 German-written responses were collected ($M_{wordcount} = 47.01$, $SD = 32.32$). Student responses and the entire data analysis were translated from German to English only for publication.

The data collection and analysis adhered to the General Data Protection Regulation of the European Union (2016). While German universities do not require Institutional Review Board approval, the data collection followed ethical guidelines (Deutsche Forschungsgemeinschaft, 2022). Data were collected without identifying or sensitive information to ensure participant anonymity. Informed consent was obtained from all participants and their privacy rights were strictly observed. Students were informed that participation was voluntary and they could withdraw from the study. Students were informed that their anonymized writing and data would be analyzed and discussed by the research group and used for publication. Only the authors had access to the data, which was stored securely on a local hard drive.

Task Format: Case Comparisons

Case comparisons have been used as a task format in this study to elicit students' mechanistic reasoning (Chin et al., 2016; Graulich & Schween, 2018). This task format involves presenting students with two or more cases carefully designed to highlight distinct structural features. The primary goal of case comparisons is to make differing task features more salient and easier to identify for students (Bussey et al., 2013; Lo & Marton, 2012). By focusing students' attention on these differing features, explaining their impact on the outcome of organic chemistry reactions can be supported (Bodé et al., 2019; Caspari & Graulich, 2019; Caspari et al., 2018; Kranz et al., 2023; Watts et al., 2021). In a meta-analysis, Alfieri et al. (2013) found that students working with case comparisons could identify significantly more features than students working with single cases ($d = 0.60$, 95% CI [0.47; 0.72]), which emphasizes that case comparisons highlight essential structural features that students might otherwise overlook (Caspari & Graulich, 2019).

The two case comparisons used in this study are presented in Figs. 1 and 2. The given reactions present a chemical process, where a negatively charged atom leaves a molecule and a positively charged structure forms—the beginning stage of a specific kind of reaction called a unimolecular nucleophilic substitution (S_N1). In technical language, the case comparisons represent the first step of an S_N1 reaction, showing the departure of a leaving group and the formation of a carbocation. In contrast to the

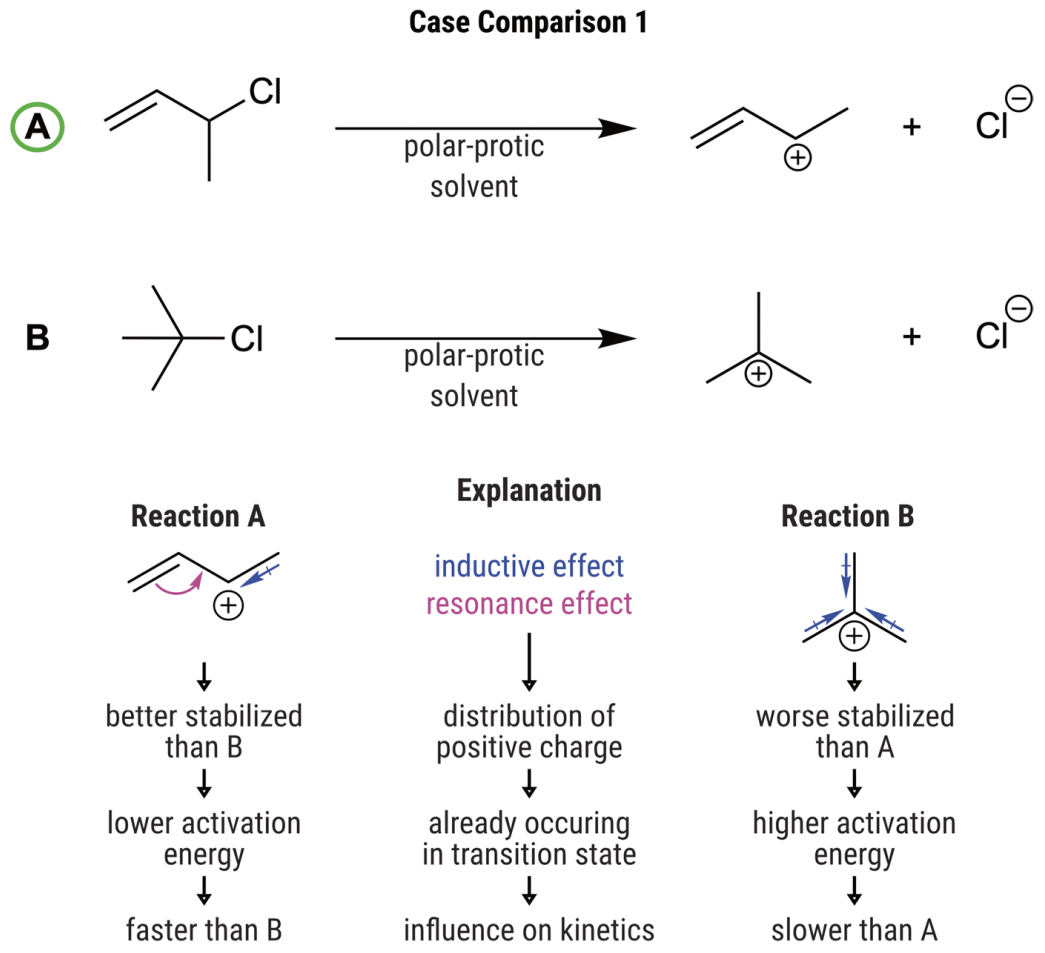


Fig. 1 Case comparison 1 with a sample solution included: The column on the left argues for reaction A, while the column on the right argues for reaction B. The column in the middle illustrates the general reasoning path. The faster reaction is highlighted in green

first task, the second task introduces two different leaving groups, causing students to analyze how ion size affects electron distribution. Students were prompted to explain which of the two reactions occurs faster, and we developed a research-informed rubric to evaluate students' responses (Kranz et al., under review).

The case comparisons are designed to reinforce a specific reasoning path: identifying structural differences, comparing their chemical properties, outlining changes in the reaction process, analyzing how differing properties affect those changes, connecting structural and energetic features, and determining which reaction occurs at a higher rate (Caspari & Graulich, 2019; Graulich & Caspari, 2021). Despite the similarity in sample solutions, students' mechanistic reasoning varied significantly across the case comparisons (Table 1). Besides, students sometimes included unproductive resources in their mechanistic reasoning, such as assuming that higher electronegativity always improves a leaving group's quality (Table 1: ID: 2). Hence, although the two tasks share some characteristics, they elicited a wide range of mechanistic reasoning.

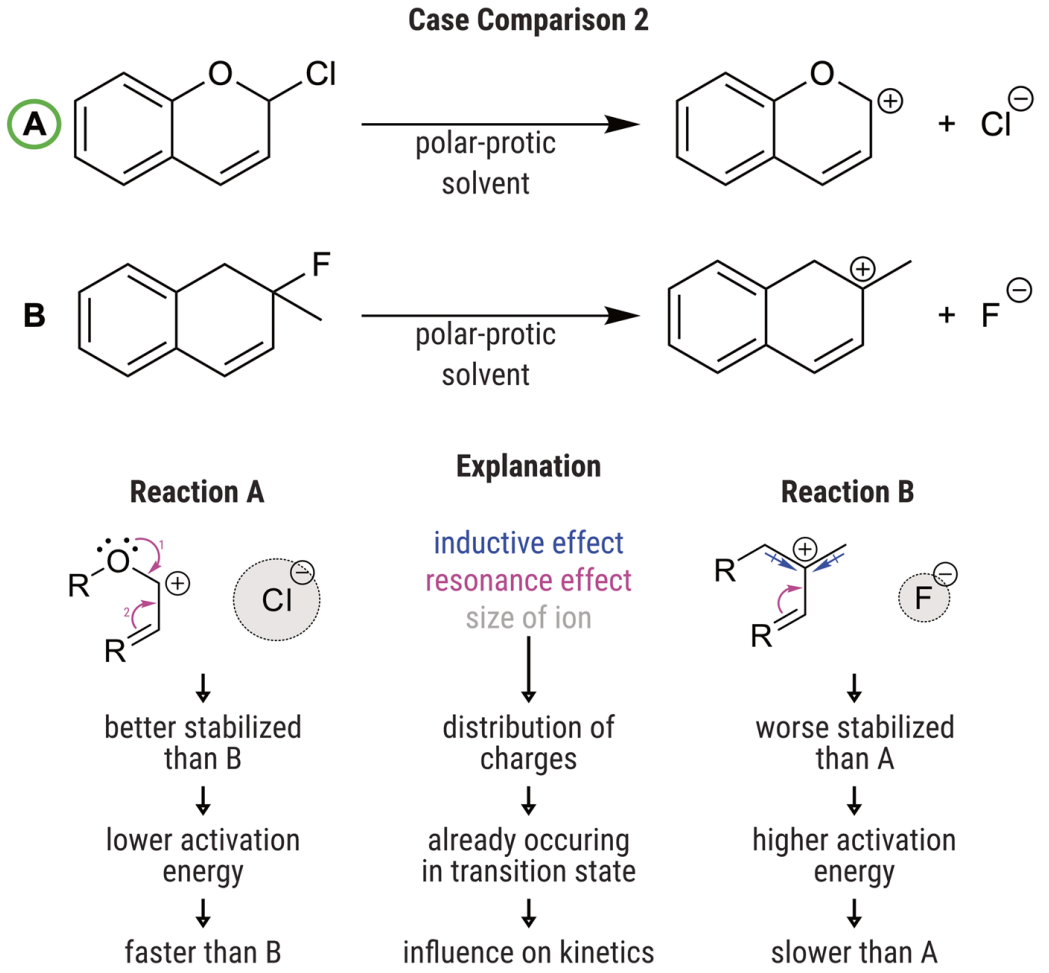


Fig. 2 Case comparison 2 with a sample solution included: The column on the left argues for reaction A, while the column on the right argues for reaction B. The column in the middle illustrates the general reasoning path. The faster reaction is highlighted in green

Research-Informed Rubric

The research-informed rubric (Fig. 3) is based on previous work by Kranz et al. (2023, under review). This rubric evaluates whether a student reasons in a *multivariate*, *comparative*, and *electronic* way. The three categories were applied to analyze students' mechanistic reasoning on case comparisons, demonstrating that using this task format leads to a significantly higher learning gain for organic chemistry students with low prior knowledge as compared to single cases (Kranz et al., under review). We refined the original rubric by removing the *productive* code, which only evaluated the correctness rather than the complexity of students' mechanistic reasoning. Beyond that, we renamed the *weighing process* code to *comparative* reasoning to better align with existing literature (Bodé et al., 2019; Deng & Flynn, 2021).

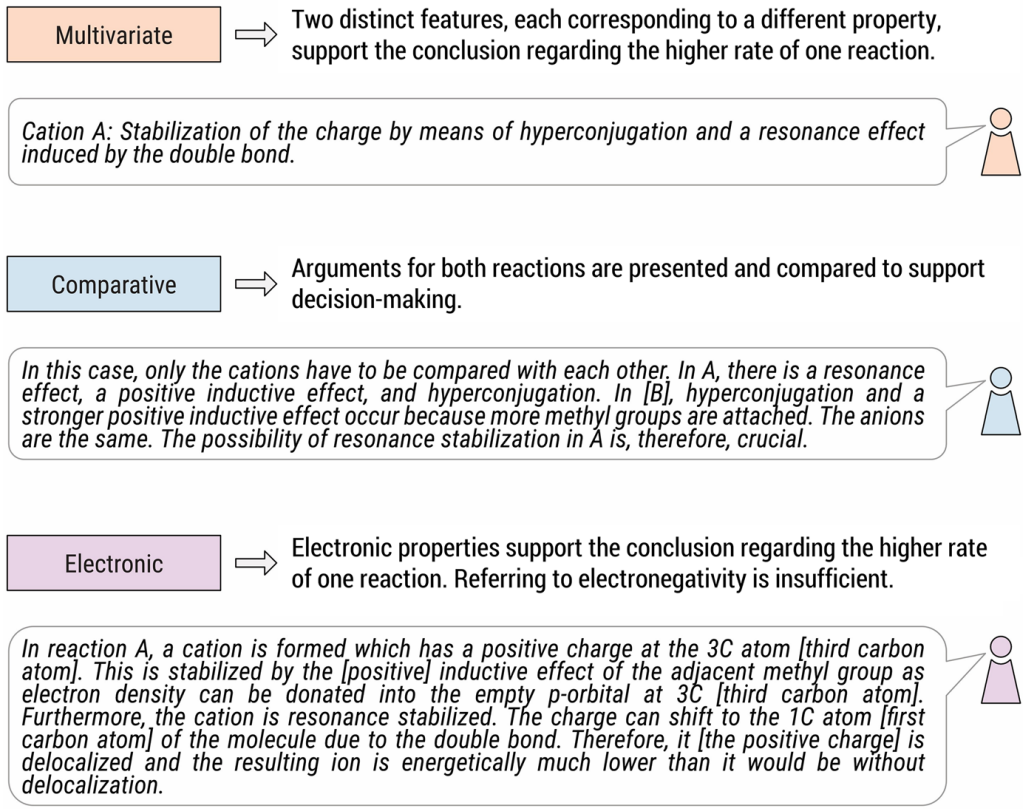


Fig. 3 Research-informed reasoning types with definitions and student-written examples included

As a first research-informed category, *multivariate* reasoning identifies multiple chemical properties, analyzes their interactions, and explains how these interactions impact the plausibility of chemical reactions (Kraft et al., 2010; Kranz et al., 2023; Sevian & Talanquer, 2014). Following that, *multivariate* reasoning unpacks the influence of multiple properties driving a phenomenon (Kranz et al., 2023; Sevian & Talanquer, 2014). As indicated in many studies, students often struggle to build *multivariate* mechanistic reasoning (Deng & Flynn, 2021; Deng et al., 2022; Frost et al., 2023; Kranz et al., 2023; Yik, Schreurs et al., 2023). In alignment with our prior research (Kranz et al., 2023, under review), students' mechanistic reasoning was considered *multivariate* if a minimum of two chemical properties were outlined for at least one of the given reactions. Instead, the response was considered *univariate* if only one chemical property was included in a student response.

Another frequently studied reasoning type is *comparative* reasoning, which comprises the second category of our research-informed rubric. *Comparative* reasoning involves contrasting the molecules or their resulting reactivity to focus on similarities and differences (Bodé et al., 2019; Caspari et al., 2018; Deng & Flynn, 2021). Such comparisons are necessary to determine differing properties of the given molecules across two or more reactions. Furthermore, comparing is essential for scientific argumentation as it helps support claims and refute counterarguments (Kuhn & Udell, 2003; Toulmin, 2003). While analyzing students' *comparative* reasoning, Bodé et al.

(2019) and Deng and Flynn (2021) found that most students established partial or complete comparisons in organic chemistry. In our analysis, students' mechanistic reasoning was considered *comparative* if they argued about a similar property in both displayed reactions to decide which one occurs faster.

The third category of our research-informed rubric addresses *electronic* reasoning, which considers the movement and distribution of electrons in entities (Becker et al., 2016; Bodé et al., 2019; Caspari et al., 2018; Cooper et al., 2016; Crandell et al., 2019; Deng & Flynn, 2021). In other words, *electronic* reasoning describes how electron density influences the properties of molecules, which means that this reasoning type requires incorporating electronic effects, such as resonance, hyperconjugation, or inductive effects. Generally, *electronic* reasoning allows students to explain the underlying causes of a mechanism, so it is highly valued across science subjects (Krist et al., 2019). Whether students include electronic properties in their reasoning depends on the context and prompt (Deng & Flynn, 2021). We coded a students' mechanistic reasoning as *electronic* if it addressed the movement or distribution of electrons within at least one molecule.

Author DK coded all student responses in a binary way based on the three research-informed categories. After that, author PPM independently coded a randomly selected subset of 20% of the data to assess inter-rater reliability. A Cohen's κ of 0.73 was achieved in the first coding round. After codes were discussed and refined, Cohen's κ on the initial 20% of coded data increased to 0.84 (95% CI [0.78, 0.91]), indicating an *almost perfect* agreement (Landis & Koch, 1977). Finally, author DK revisited the entire dataset and re-evaluated the codes considering the discussion.

ML-Based Rubric

To investigate whether an ML-based rubric mirrors the research-informed considerations outlined above, we used unsupervised ML to design a novel rubric capturing students' mechanistic reasoning. By doing so, we initially removed stopwords, i.e., filler words, to boost the performance of the applied unsupervised ML technique. To accomplish this, we downloaded a generic, pre-defined list of stopwords accessed via the programming environment R and made minor adjustments, creating a final list of 586 words excluded from the analysis. After that, we applied a large language model to convert students' written accounts into contextualized embeddings (Fig. 4a), which are high-dimensional numerical representations of

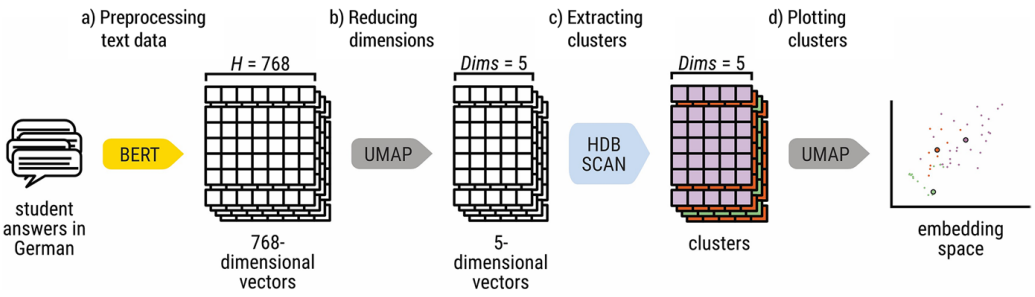


Fig. 4 Analysis steps for the development of the ML-based rubric

words. This transformation typically improves the accuracy of ML-based methods (Zehner et al., 2016). We utilized the large language model *BERT-base-German-cased* to preprocess our German-written responses. This large language model has become a valuable resource for education researchers analyzing German-written reasoning because of its strong language understanding capabilities (Wulff et al., 2022; Wulff, Mientus et al., 2023; Wulff, Westphal et al., 2023). From a technical perspective, *BERT-base-German-cased* is a specific version of Bidirectional Encoder Representations from Transformers (BERT)—a powerful large language model developed to analyze the context of words in a sentence (Devlin et al., 2018).

After preprocessing students' responses, we applied Uniform Manifold Approximation and Projection (UMAP) to reduce the dimensionality of the contextualized embeddings (McInnes, Healy, Saul et al., 2018). UMAP reduces high-dimensional data while preserving local structure. By reducing the 768-dimensional vectors to only five dimensions (Fig. 4b), we could computationally ease clustering. Generally, such a dimensionality reduction is appropriate because complex data typically stores most information in only a few dimensions (Brunton & Kutz, 2019; Zehner et al., 2016). The number of neighbors was set to 15 in our analysis because this configuration produced easily interpretable results in previous studies in science education (Martin et al., 2024; Tschisgale et al., 2023; Wulff et al., 2022; Wulff, Westphal et al., 2023) and other domains (Grootendorst, 2020).

Finally, we used the unsupervised ML algorithm Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to group data points into compact, mutually exclusive clusters (Fig. 4c) (McInnes et al., 2017). We applied HDBSCAN because it can identify clusters of arbitrary shape, automatically determine the number of clusters, and handle noisy language data—as common when analyzing students' reasoning in science subjects. HDBSCAN handles noisy language by extracting noise points, which represent responses that encompass non-normative or mixed ideas as well as topics that do not meet the threshold for the minimum cluster size. Excluding noise enabled a more concise examination of students' mechanistic reasoning.

To ensure the robustness of our findings, we qualitatively evaluated 37 HDBSCAN solutions. These solutions were generated by varying three hyperparameters: minimum cluster size, minimum sample size, and dimension number. The minimum cluster size specifies the smallest group size considered a cluster, the minimum sample size determines the portion of data points classified as noise, and the dimension number refers to the dimensions of the word embedding space after dimensionality reduction. Different hyperparameter settings can significantly influence the cluster solutions, making it crucial to compare various outcomes carefully. Rather than selecting the optimal cluster solution based on statistical measures, we assessed the solutions based on two qualitative criteria: extracted clusters should have avoided merging conceptually distinct mechanistic reasoning types and additional clusters should have introduced new mechanistic reasoning types. Based on these criteria, we selected a solution with fifteen clusters (Table 1). The clusters encompass various organic chemistry reasoning types, such as *structural*, *energetic*, *phenomenological*, and *electronic* reasoning (Bodé et al., 2019; Deng & Flynn, 2021; Martin et al., 2024). While cluster solutions with more groups did not yield additional

insights into students' mechanistic reasoning, solutions with fewer clusters covered important nuances. The hyperparameters of the selected cluster solution were as follows: The minimum cluster size was set to nine, meaning that HDBSCAN extracted a cluster only if it contained at least nine student responses. The minimum sample size was set to six to achieve a balance between minimizing the exclusion of student responses as noise and ensuring the extraction of concise clusters. Additionally, the dimension number was set to five as mentioned above. We performed further dimensionality reduction to visualize the clusters in a two-dimensional space (Fig. 4d). Throughout the analysis, we used the Euclidean distance metric. A step-by-step guide for implementing our analysis is available in Grootendorst (2020).

Of note, we also applied two other clustering algorithms—*k*-means clustering (MacQueen, 1967) and *k*-medoids clustering (Kaufman & Rousseeuw, 1990). These algorithms merged conceptually distinct mechanistic reasoning types into the same clusters, so these algorithms were less effective in revealing nuanced differences in students' reasoning about reaction mechanisms. Accordingly, we used HDBSCAN as discussed above to identify finer distinctions in students' mechanistic reasoning. This choice aligns with previous research showing that HDBSCAN generates interpretable, specific, and robust clusters in science education contexts (Wulff et al., 2022), which is why several studies have adopted this technique to develop their rubrics (e.g., Martin et al., 2024; Tschisgale et al., 2023; Wulff et al., 2022; Wulff, Westphal et al., 2023). Although this analysis was conducted in German, Martin et al. (2024) performed a similar study in English, showing that HDBSCAN can identify nuanced differences in students' English-written responses. We hypothesize that the clustering technique might perform even better in English than in German, given the generally superior performance of the English-specific large language models in science assessments (Martin & Graulich, 2024a).

The analysis was performed in Python 3.9.13 (Van Rossum & Drake, 2009) using *scikit-learn* as an ML framework (Pedregosa et al., 2011), *spacy* for NLP (Honnibal et al., 2020), and *umap* for dimensionality reduction (McInnes, Healy, Saul et al., 2018; McInnes, Healy, Melville et al., 2018). We also applied *numpy* for generating multidimensional arrays (Harris et al., 2020), *matplotlib* and *seaborn* for data visualization (Hunter, 2007; Waskom et al., 2020), and *pandas* for further data processing (McKinney, 2010). For all other operations, we used the standard Python libraries (Van Rossum & Drake, 2009).

Quantitative Comparisons between the Research-Informed and ML-Based Rubrics

We quantitatively compared each of the three research-informed categories (Sect. 4.3: "Research-Informed Rubric") with the ML-based rubric. For human coding, each research-informed category was applied in a binary way, indicating the presence or absence of a specific reasoning type. Based on the binary codes and the 15 detected clusters (Sect. 5.1: "Detecting ML-Based Clusters"), a contingency table containing all code frequencies in every cluster was created for each of the three categories (Fig. 5a).

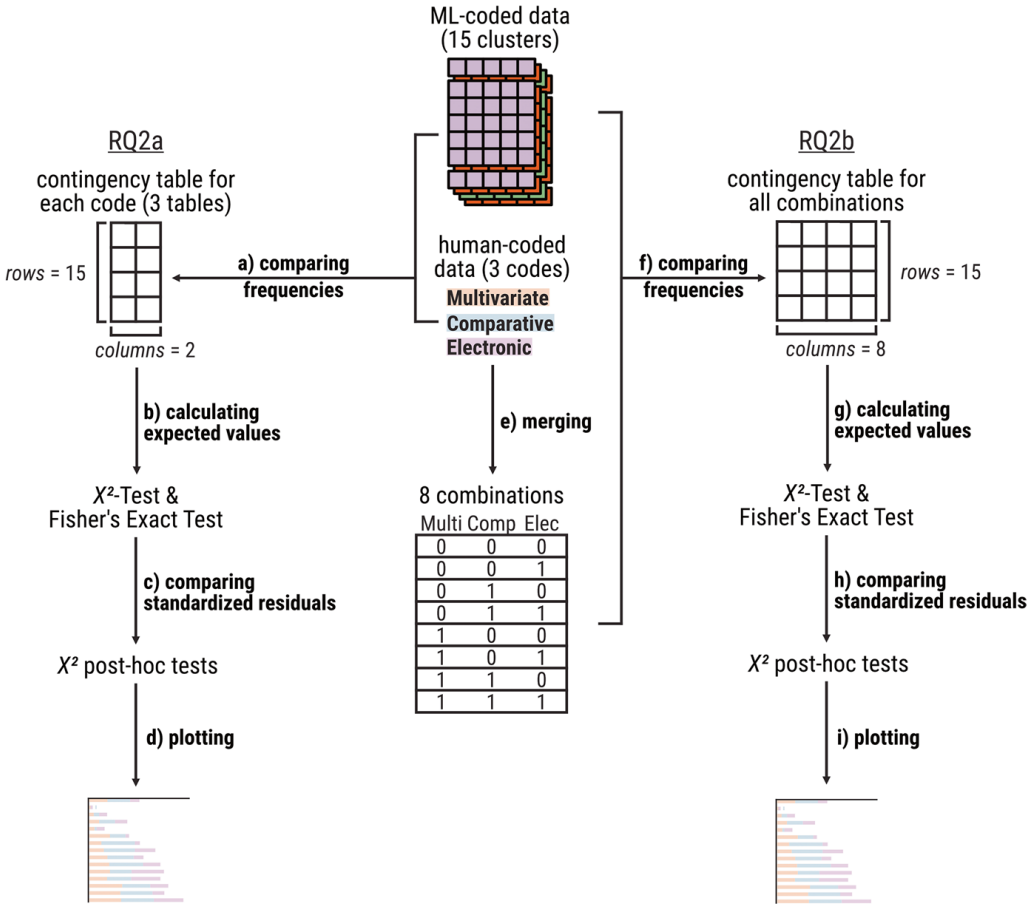


Fig. 5 Steps of the quantitative analysis. Letters a) – i) represent the sequence of the applied analysis steps

To subsequently test whether the research-informed categories and the ML-based rubric are associated, we performed a χ^2-Test for association (Fig. 5b) (Agresti, 2013, 2018; McHugh, 2013). Two assumptions underlie a χ^2-Test: the independence of data and expected frequencies in the contingency table exceeding five (Howell, 2006; Field et al., 2012). In our study—although students were prompted to reason about two tasks in a pre-post design—the assumption of data independence is met because our analysis is conducted at the response level, rather than the student level. In other words, we did not use the χ^2-Test to identify performance changes from pre- to post-assessment; instead, we investigated the association between the two applied rubrics based on the responses' codes. Since each rubric uniquely codes every response, the responses contribute solely to one cell in the contingency table; thus, the independence assumption is satisfied. However, some expected frequencies in the contingency table fell below five. To address this concern, we also conducted a Fisher's Exact Test (Armitage et al., 2008) using

simulated data with 10^5 replicates to confirm our findings with a more robust statistical test (Patefield, 1981). Ultimately, we performed a post-hoc analysis on the standardized residuals of the χ^2-Test to identify the specific differences among the clusters based on the codes that occurred more or less frequently than statistically expected (Fig. 5c) (Agresti, 2018; Beasley & Schumacker, 1995) and plotted the results (Fig. 5d). The three χ^2-Tests performed within RQ 2a used a Bonferroni-corrected significance level of $p < .017$.

However, due to the solely binary coding of the three categories, a single research-informed category was more likely to correspond to the ML-based rubric than their combination. Therefore, we merged the research-informed codes, resulting in eight combinations (Fig. 5e). For instance, an answer with the presence of the *multivariate* and *comparative* codes but with the absence of the *electronic* code received the code *Multi1Comp1Elec0*. This approach allowed us to convert the three research-informed analytic categories into a holistic rubric. After merging the codes, we also generated a contingency table comprising the frequencies of the eight possible combinations in each cluster (Fig. 5f) and repeated the analysis described above (Fig. 5g-i).

Of note, we intentionally decided to investigate the association between the two rubric types by coding student responses from two tasks collected at two different time points. Including two tasks in the analysis helped derive more context-independent conclusions, ensuring that observed correlations are not only artifacts of a single task's requirements. Similarly, incorporating data from two time points contributed to establishing conclusions independent of students' in-the-moment understanding, which increases the robustness of our findings.

We performed the statistical analysis in R (version 4.3.0) (R Core Team, 2023) with *RStudio* as an Integrated Development Environment (IDE) (RStudio Team, 2023). We used the libraries *gmodels* to perform the χ^2-analysis (Warnes et al., 2023), *chisq.posthoc.test* for the post-hoc analysis (Ebbert, 2019), and *tidyverse*, *reshape2*, and *dplyr* to transform the data (Wickham, 2007; Wickham et al., 2019, 2022). For graphical processing, we applied the libraries *ggplot2*, *gridextra*, *webshot*, and *kableExtra* (Auguie, 2017; Chang, 2017; Wickham, 2016; Zhu et al., 2022). For all other operations, we used the basic R library (R Core Team, 2023).

Results and Discussion

RQ 1: Detecting ML-Based Clusters

To design an ML-based rubric, we leveraged HDBSCAN to identify 14 clusters, an additional noise cluster, their ten most representative words, and sample statements (Table 1). Through a qualitative analysis of the most representative words and sample responses, we described each cluster and visualized them in a two-dimensional space (Fig. 6). Surprisingly, responses to both tasks are similarly reflected in all clusters, so the clusters did not segregate the two tasks.

Table 1 Overview of the extracted clusters

ID	Cluster Size	Description	Top Words	Student Examples
-1	136	Noise-Cluster	hyperconjugation, cation, tertiary, resonance, charge, forms, leaving, group, positive, occurs	Reaction B: fluorine is more reactive than chlorine.
0	17	Referring to stability without including chemical properties	product, occurs, chain, stable, arrangement, similar, substituent, connected, steric, identical	Reaction B is faster because the carbocation is better stabilized.
1	27	Focusing on the number of alkyl substituents as the only impact on carbocation stability	positive, molecule, alkyl, tertiary, charge, inductive, occurs, atom, determining, number	Reaction B occurs faster because the tertiary carbocation can be better stabilized by the alkyl substituents.
2	50	Emphasizing the influence of electronegativity and electron-withdrawing effects on leaving group quality	pulls, electrons, withdrawing, stronger, electronegativity, atom, fluorine, chloride, higher, binding	To determine which reaction occurs faster, one must identify the structural differences first; here, fluoride and chloride as leaving groups. Reaction B proceeds faster because of the more electronegative fluorine that attracts electrons based on its electron-withdrawing effect.
3	18	Referring to resonance on a surface level as the only impact on carbocation stability	double-bond, aromatic, oxygen-atom, total, resonance, structure, occurs, breaking, delocalized, electron-poor	Reaction A occurs faster because the formed product is stabilized by resonance.
4	30	Referring to substrate stability and leaving group quality by focusing on charges	leaving, group, allyl, chloride, charge, double-bond, worse, occurs, fluoride, aromatic	Reaction A occurs faster. Although fluoride is a better leaving group than chloride, the positive charge in reaction A is stabilized across both rings.
5	14	Noticing inductive and resonance effects on a surface level	inductive, resonance, effect, stronger, substituent, alkyl, weighing, positive, hyperconjugation, predominant	In reaction A, the positive charge is stabilized by a resonance effect; in reaction B, only inductive effects are present.
6	14	Weighing inductive and resonance effects mostly on an electronic level	effect, cation, weighing, charge, additional, positive, double-bond, stronger, inductive, resonance	I assume that reaction A occurs faster due to the delocalization of the charge across both rings induced by the oxygen atom. Although reaction B provides more inductive effects, the cation in reaction A is more stable due to resonance; therefore, reaction A occurs faster.

Table 1 (continued)

ID	Cluster Size	Description	Top Words	Student Examples
7	15	Comparing explicit properties of the substrate and the electronegativity of the leaving group	alkyl, fluorine, forming, solvent, occurs, additional, chloride, secondary, higher, tertiary	Reaction B occurs faster because the carbocation is stabilized by the additional methyl group. Moreover, fluorine is more electronegative than chlorine.
8	64	Weighing single or multiple chemical properties coherently by including scientifically normative and non-normative ideas	ion, carbocation, tertiary, secondary, forming, product, adjacent, leaving, group, alkyl	Since the solvent does not differ, it is necessary to consider which carbocation is better stabilized because this will be formed faster. It is also important to note which halogen is cleaved off faster. First of all, both carbocations are stabilized. Carbocation B is stabilized by hyperconjugation of the adjacent methyl group, and carbocation A by resonance, i.e., by the pi-system. Moreover, the oxygen bond in the ring is also an electron-pair donor. The positive charge at the top is, thus, stabilized by the delocalization of the charge over the whole molecule. This is much more efficient than local stabilization by hyperconjugation. For this reason, I would assume that reaction A is faster, even if fluoride is more easily split off because it polarizes the bond to the carbon stronger than chlorine.

Table 1 (continued)

ID	Cluster Size	Description	Top Words	Student Examples
9	10	Weighing stability differences based on resonance and hyperconjugation on an electronic level	hyperconjugation, positive, three, delocalized, charge, distributes, pi, molecule, conjugation, cation	I assume a higher rate of reaction A. The anions are identical and, therefore, irrelevant. I have considered the stabilization of the cations. The positive charge at A is at an atom with two adjacent carbon atoms. The double bond provides the possibility for a further resonance structure. Generally, carbon atoms have an electron-donating and hyperconjugation effect. I chose A because I consider the resonance effect stronger than the inductive and hyperconjugation effect, even though it is present three times in B.
10	21	Describing how resonance delocalizes electrons	atom, double-bond, positive, charge, oxygen-atom, adjacent, delocalized, electron-pair, shifts, electron-withdrawing	Reaction A occurs faster because the positive charge is resonance-stabilized by the allyl double bond. The double bond can shift and the positive charge would be located at the outermost carbon atom. Thus, a pi-electron system distributes the charge across several atoms.
11	18	Relating leaving group quality to ion size or electron distribution and weighing it with other properties	leaving, group, cation, chlorine, negative, compares, fluorine, product, charge, electrons	Reaction A occurs faster. There are significantly more resonance structures in A, which delocalize the positive charge when chloride leaves the molecule. In B, the additional methyl group also compensates for the departure of fluoride and shifts the charge to a tertiary carbon atom. The fluorine atom in B has a higher electronegativity but is smaller than chlorine. In sum, resonance stabilization is stronger in A and chloride leaves more easily due to its size and proximity to oxygen.

Table 1 (continued)

ID	Cluster Size	Description	Top Words	Student Examples
12	9	Emphasizing that resonance effects outweigh inductive and hyperconjugation effects	hyperconjugation, forming, weighing, cation, product, leaving, group, occurs, resonance, predominant	Reaction A occurs faster because this product can form resonance structures due to the allyl substituent. This makes it more stable than the product in B, where only the electron-donating and hyperconjugation effects are present because it is a tertiary carbon atom. In both reactions, the chloride ion is the leaving group so there is no difference.
13	11	Weighing all electronic and non-electronic effects and explaining their impact on electron distribution	molecule, differs, double-bond, ring, chloride, atom, six-membered, binding, reactant, oxygen-atom	The polar protic solvent stabilizes the cations in both reactions but makes no difference because it is identical. Since both reactions involve the same leaving group, there is no difference between the two. The secondary carbocation in A and the tertiary carbocation in B may lead to the preference for B. However, in reaction A, the double bond's proximity allows for the delocalization of electrons in the product. Delocalization lowers the product's energy much more than the hyperconjugation of alkyl groups, favoring reaction A.

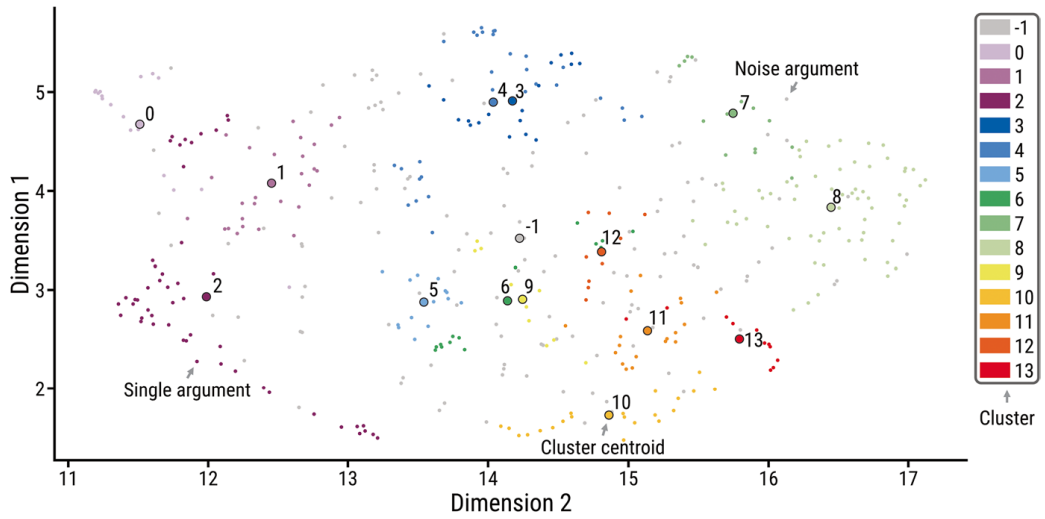


Fig. 6 Embedding space of the extracted clusters. *Note*: Each data point in the embedding space represents a student's written response. Different colors visualize different clusters. The gray points scattered throughout the embedding space show noise arguments that cannot be distinctly assigned to any cluster. The larger points, labeled as cluster centroids, represent the centers of their respective clusters

In the two-dimensional space, clusters sharing similar content are near to one another: Responses in clusters 0, 1, and 3 reasoned descriptively about substrate or carbocation stability while focusing solely on explicit, structural properties. In particular, responses in these clusters primarily relied on one-reason decision-making (Talanquer, 2014) and *non-comparative*, *non-electronic* reasoning. Responses in clusters 2 and 4 focused on different grain sizes on leaving group quality. In clusters 5 and 6, students coherently weighed inductive and resonance effects, which showcases students' strengths in establishing complete comparisons of the most relevant properties. Responses in clusters 7 and 8 included explicit and implicit properties of the substrate and the leaving group, revealing students' competence in weighing multiple features. Furthermore, responses in clusters 9 and 10 combined resonance and other electronic effects, such as hyperconjugation, implying a solid understanding of the electronic properties that impact a reaction's plausibility. In cluster 12, responses weighed various electronic properties like hyperconjugation, resonance, and inductive effects, showing a high complexity level. Lastly, responses in clusters 11 and 13 explained the impact of multiple properties on electron distribution, which means these responses include *multivariate* comparisons. Responses in these clusters are the highest in quality. Generally, we noticed that clusters having lower values in dimension 1 and higher values in dimension 2 (Fig. 6) demonstrated higher complexity.

The detected patterns encompass a range of chemical topics discussed at varying grain sizes, highlighting that the clustering method captured the content *and* complexity of students' mechanistic reasoning. However, interpreting the data-driven clusters required significant subject matter expertise. We hypothesize that this expertise will remain necessary to interpret data-driven clusters in the future. Nevertheless, exactly this type of human-machine collaboration can augment human analytic power (Kubsch et al., 2023): Computational techniques offer breadth, reproducibility, and

scalability in data analysis, while human expertise ensures that categorizations accurately capture the underlying nuances specific to the context.

To investigate whether the extracted clusters also mirror theory-rich considerations, we quantitatively compared the data-driven clusters with our research-informed categorization. Associations between both rubric types would increase their validity because the data-driven clusters would confirm the research-informed categories—and vice versa.

RQ 2a: Comparing the Computationally Revealed Clusters and Research-Informed Categories

To determine if the computationally revealed clusters capture varying degrees of *multivariate*, *comparative*, and *electronic* reasoning (Fig. 7), we performed χ^2-Tests for association for each of the three research-informed categories and verified the results with Fisher's Exact Tests. Subsequently, we used the standardized residuals of the χ^2-Test to pinpoint specific associations. An even distribution of the three research-informed categories across the 15 clusters would imply that the two rubrics are not associated because the classification of a student response into a certain cluster would be irrelevant for the presence of a certain reasoning type. By contrast, an uneven distribution of the research-informed categories across the clusters would indicate associations since some clusters would predict a specific reasoning type. In other words, differences among the clusters regarding the degree of *multivariate*, *comparative*, and *electronic* reasoning imply rubric relations as certain clusters predict specific reasoning types.

For *multivariate* reasoning, the χ^2-Test revealed a significant association between this reasoning type and the distribution of students' responses among the 15 ML-based clusters ($\chi^2(14) = 74.55$, $p < .001$), with a medium effect size (Cramer's $V = 0.41$, 95% CI[0.37; 0.50]) (Cohen, 1988). Fisher's Exact Test confirmed the significant association ($p < .001$). Looking at the standardized residuals, clusters 0 ($n_{expected} = 8.28$, $n_{observed} = 1$, $p = .010$) and 1 ($n_{expected} = 13.14$, $n_{observed} = 4$, $p = .009$) contained significantly fewer instances of *multivariate* reasoning than statistically expected, while cluster 11 ($n_{expected} = 8.76$, $n_{observed} = 17$, $p = .002$) included significantly more instances of *multivariate* reasoning. Responses in cluster 0 did not include any chemical properties. For instance, the student example of cluster 0 only referred to stability without mentioning the corresponding chemical properties (Table 1), which explains the absence of *multivariate* reasoning. In cluster 1, students only counted the number of alkyl groups to estimate carbocation stability (Table 1), leading to *univariate* reasoning. Conversely, responses in cluster 11 established causal links between leaving group quality and substrate stability based on electron distribution (Table 1), which is why these responses are *multivariate*.

For *comparative* reasoning, the χ^2-Test indicated that this reasoning type is also significantly reflected in the computationally revealed clusters ($\chi^2(14) = 132.38$, $p < .001$), even with a large effect size (Cramer's $V = 0.54$, 95% CI[0.50; 0.61]). Again, Fisher's Exact Test confirmed this association ($p < .001$). By comparing the standardized residuals, clusters 0 ($n_{expected} = 10.26$, $n_{observed} = 0$, $p < .001$), 1 ($n_{expected} = 16.30$, $n_{observed} = 4$, $p < .001$), and 3 ($n_{expected} = 10.86$, $n_{observed} = 1$, $p < .001$)

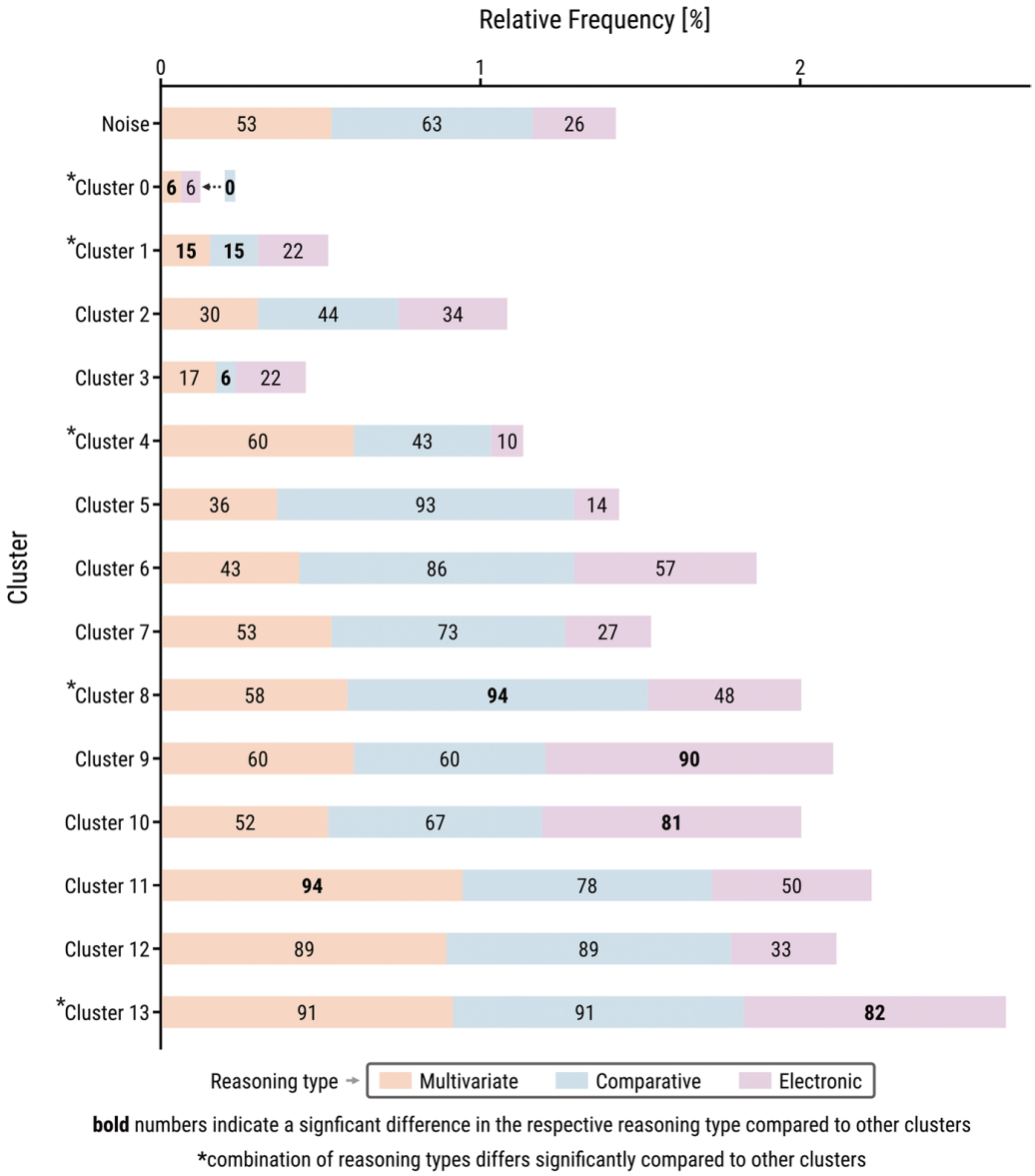


Fig. 7 Relative frequencies of the reasoning types in each cluster

contained examples of *comparative* reasoning less frequently. Responses in these clusters strictly focused on a single property in only one of the two given reactions (Table 1), leading to a lack of *comparative* reasoning. By contrast, responses in cluster 8 ($n_{expected} = 38.63$, $n_{observed} = 60$, $p < .001$) demonstrated *comparative* reasoning more frequently. Here, students weighed in detail how the differing properties impact the rate of the contrasted reactions (Table 1).

Finally, the results of the χ^2-Test were also significant for *electronic* reasoning ($\chi^2(14) = 79.53$, $p < .001$), with a medium effect size (Cramer's $V = 0.42$, 95% CI[0.37; 0.52]). Hence, the ability to demonstrate *electronic* reasoning is reflected in the clusters. This association could be confirmed with a Fisher's Exact Test

($p < .001$). In particular, clusters 9 ($n_{expected} = 3.48$, $n_{observed} = 9$, $p = .006$), 10 ($n_{expected} = 7.31$, $n_{observed} = 17$, $p < .001$), and 13 ($n_{expected} = 3.83$, $n_{observed} = 9$, $p = .016$) showed significantly more instances of *electronic* reasoning than statistically expected. Responses in cluster 9 described the impact of resonance, hyperconjugation, and inductive effects on electron distribution, while responses in cluster 10 referred to electron delocalization based on shifting pi-electrons (Table 1). In cluster 13, students weighed multiple electronic properties like hyperconjugation and resonance, implying a high degree of *electronic* reasoning (Table 1).

In addition to these associations, the extracted clusters imply a hierarchy because higher-number clusters roughly indicate higher complexity (Fig. 7). This hierarchy may arise from the increasing response length across clusters, where longer responses often exhibit more complex mechanistic reasoning (Table 1). The opportunity to use response length as a surface-level indicator of students' mechanistic reasoning complexity may have facilitated the organization of the responses within this hierarchy. The reasons for hierarchically arranging students' mechanistic reasoning complexity are also connected to how HDBSCAN operates—the "H" in HDBSCAN stands for "hierarchical" (McInnes et al., 2017). HDBSCAN identifies clusters by finding groups of points with high density and separating them from less dense regions. It then merges these clusters progressively as the density threshold is lowered, creating a hierarchy of clusters at different density levels. Finally, the algorithm extracts clusters based on their stability across these levels, meaning it looks at how consistently a cluster appears as the density threshold changes. In our analysis, lower-number clusters may be more stable because they consist of shorter responses and have higher within-cluster similarity. As a result, these lower-number clusters remain stable across density thresholds and form the foundation of the ML-based reasoning hierarchy (Table 1).

While this first analysis provided useful insights into the manifold relationships between the two rubric types, we investigated interdependencies for each of the three categories separately. However, since each category was only coded in a binary way, the probability of associations by chance increases. Hence, we also merged the three research-informed categories (Sect. 4.5: "Quantitative Comparisons between the Research-Informed and ML-Based Rubrics") to validate our findings.

RQ 2b: Comparing the Computationally Detected Clusters and the Combination of the Research-Informed Categories

To reveal similarities between the computationally detected clusters and the merged research-informed rubric, we performed another χ^2-Test for association, demonstrating that both rubrics are significantly associated ($\chi^2(98) = 333.39$, $p < .001$). Since the effect size necessarily decreases as the size of the contingency table increases, the calculated effect size can be considered large (Cramer's $V = 0.32$, 95% CI[0.32; 0.40]). This result was confirmed by Fisher's Exact Test ($p < .001$). For this reason, the ML-based rubric also reflects the combination of the research-informed categories.

Afterward, we identified the clusters that contributed to the significant association by analyzing which code combinations appeared more or less frequently than statistically expected (Fig. 7). We found that the code combination *Multi0Comp0Elec0* appeared significantly more frequently in clusters 0 ($n_{expected} = 3.48$, $n_{observed} = 15$, $p < .001$), 1 ($n_{expected} = 5.53$, $n_{observed} = 15$, $p < .001$), and 3 ($n_{expected} = 3.69$, $n_{observed} = 12$, $p < .001$), and significantly less frequently in cluster 8 ($n_{expected} = 13.11$, $n_{observed} = 2$, $p = .025$). Responses assigned to clusters 0, 1, or 3 make a claim about which reaction occurs at a higher rate and justify their claim based on stability, the number of alkyl substituents, or resonance (Table 1). Students incorporated neither multiple chemical properties nor the impact of electronic effects, demonstrating a lack of *multivariate*, *comparative*, and *electronic* reasoning. In contrast, responses in cluster 8 weighed multiple electronic properties across reactions, implying the presence of at least one of the three reasoning types. Moreover, the code *Multi1Comp0Elec0* appeared more frequently in cluster 4 ($n_{expected} = 2.84$, $n_{observed} = 9$, $p = .009$), where students included at least two properties in their reasoning. For instance, the student example of cluster 4 mentions leaving group quality and charge stabilization (Table 1), thus, being *multivariate* but neither *comparative* nor *electronic*. Last, the combination *Multi1Comp1Elec1* was more often detected in cluster 13 ($n_{expected} = 2.08$, $n_{observed} = 9$, $p < .001$). Here, students explained the impact of multiple properties like leaving group quality, hyperconjugation, and resonance on reaction rate and coherently weighed these properties across reactions on an *electronic* level (Table 1), showcasing the presence of all three reasoning types.

Together, our findings indicate an association between the research-informed and ML-based rubrics. Accordingly, ML-based rubrics have the potential to not only summarize students' reasoning content-wise but also to reflect reasoning types discussed in the literature. In other words, the ML-based rubric naturally reflects research-informed considerations so that a classification of a student response into a cluster already allows one to claim the presence of a certain research-informed reasoning type. Further post-hoc tests showed that certain clusters are predictors for specific research-informed categories, giving insights into why the two rubric types share common features. From a broader perspective, our analysis validates the applied clustering technique for developing rubrics in our organic chemistry context, suggesting that it is appropriate to ground formative assessments exclusively on ML-based rubrics in our case.

Conclusions and Implications

Generally, developing a rubric for categorizing students' open-ended reasoning requires subject matter expertise to recognize emerging ideas. Leveraging the output of clustering techniques, such as HDBSCAN, to develop rubrics can reduce human design effort because it replaces much of the preliminary exploratory work typically associated with rubric development (Haudek et al., 2015). Following this, ML techniques can be applied to ground rubrics on data. Herein, we analyzed the associations between three research-informed categories and various ML-based clusters to investigate whether ML-based rubrics are valid for scoring formative assessments.

First, we applied a clustering technique to identify various mechanistic reasoning patterns at different complexity levels arranged in some sort of hierarchy (Figs. 6 and 7). Subsequently, we compared this ML-based rubric with its research-informed counterpart that coded the presence or absence of *multivariate*, *comparative*, and *electronic* reasoning in students' responses. We found that both rubrics shared significant associations, revealing that both rubric types are appropriate for scoring formative assessments. This finding implies that the ML-based rubric reflects theory-rich considerations, which advances the validity of the applied clustering technique and the research-informed categories. Accordingly, researchers and instructors can generally use either a research-informed or ML-based rubric to evaluate students' reasoning validly in organic chemistry. Nonetheless, our findings do not provide conclusive evidence regarding the broader applicability of ML-based rubrics across contexts. Further research is needed to determine if ML-based rubrics are valid foundations for formative assessments in various settings.

Transferring our rubric to a new context, for example, running the unsupervised ML analysis based on different student responses, can be technically challenging for researchers and instructors. Grootendorst (2020) provides open-source code and a tutorial to facilitate this process. Recent advances in generative AI have made implementing such ML analyses also progressively easier. For instance, researchers and instructors could use chatbots to conduct simpler clustering analyses of student data. However, clustering only categorizes student responses, while human experts must interpret these categories. We believe instructors can interpret the clusters using their teaching experience, although familiarity with the research base would likely advance their ability to identify recurring patterns.

Beyond the potential for its replication, the analysis reported herein offers additional merits. By demonstrating that research-informed *and* ML-based rubrics can effectively be used in formative organic chemistry assessments, we provide a solid rationale for applying AI technologies in this context. In other words, our research primarily validates the integration of AI in organic chemistry teaching because it highlights the applicability of ML-based rubrics. Our findings may also generalize to easy-to-implement AI tools like chatbots. Future research should investigate how this claim applies to various contexts and AI techniques, including supervised ML and generative AI. On top of that, while we hypothesize that our findings apply to the English language (Martin et al., 2024; Martin & Graulich, 2024a), further studies could explore their transferability beyond the German language.

Despite their associations, our research-informed and ML-based rubrics relied on different evaluation units. The former was intentionally designed in an analytic way, whereas we interpreted the latter holistically. As a result, the research-informed and ML-based rubrics fulfill different functions. The research-informed rubric incorporates theory-driven learning objectives to evaluate students' *multivariate*, *comparative*, and *electronic* reasoning. Accordingly, this rubric can provide multifaceted information on students' reasoning since every category outlines a unique competence. This detailed information eases identifying students' specific strengths and challenges, allowing for targeted guidance in either *multivariate*, *comparative*, or *electronic* reasoning. Besides, our research-informed rubric enabled us to break down complex open-ended responses into

binary categories. Thus, we believe that the three research-informed categories are highly instructor-friendly since human raters only need to determine whether each of the three reasoning types is present or absent. Conversely, the ML-based rubric is grounded in data, which can be a starting point for proposing future guidance and monitoring students' learning progression over an extended period (Donnelly et al., 2015; Jescovitch et al., 2021; Martin et al., 2024; Tansomboon et al., 2017; Vitale et al., 2016). Hence, the ML-based rubric could continuously analyze when and how students' mechanistic reasoning patterns change, allowing researchers to track students' skills across multiple levels and uncover their most prominent strengths and challenges over time.

Given the benefits of both rubric types, incorporating human qualitative interpretation *and* ML techniques into rubric design can be highly beneficial for developing assessments. Martin et al. (2024), Rosenberg and Krist (2021), and Tschisgale et al. (2023) showed that a rubric that integrated theory- *and* data-driven considerations could accurately evaluate students' reasoning. By combining the strengths of human expertise and ML capabilities, such rubrics offer a valid means of categorizing student performance.

Collectively, the findings indicate that research-informed and ML-based rubrics share significant associations. Both rubric types have their merits and can be used depending on the assessment goal and context. The pedagogical purpose of the assessment, the respective learning objectives, and the nature of the tasks should inform rubric type selection (Kubsch et al., 2022; Martin & Graulich, 2023).

Related Research and Next Steps

We expanded upon the findings of this analysis in a related research project—also using case comparisons to support students' mechanistic reasoning—by developing a more generalizable rubric for mechanistic reasoning in undergraduate organic chemistry. This rubric integrates ML-driven analysis with qualitative human interpretation to capture similar types of organic chemistry reasoning, including the level of granularity and the degree of causality in students' mechanistic reasoning (Martin et al., 2024). Based on this rubric, we developed a supervised ML algorithm that automatically evaluates undergraduate students' reasoning about reaction mechanisms over time (Martin & Graulich, 2024b, c). The ML algorithm allowed us to adaptively support students' mechanistic reasoning and monitor their learning progression. Additionally, we leveraged this rubric to create an ML model that can automatically analyze students' reasoning across different languages (Martin & Graulich, 2024a). Instructors can use these ML applications to support students' mechanistic reasoning and ease effective communication across languages. The analysis presented herein guided these practical ML applications because it demonstrates the validity of integrating ML into assessment development.

Future work could investigate the advantages of generative AI when automatically evaluating students' reasoning using ML-based rubrics. Generative AI can

potentially ease the distribution of ML-assisted formative assessments through self-constructed chatbots, such as custom GPTs in ChatGPT. In this way, the next step in practically applying ML-based rubrics could be developing readily accessible tools that automatically analyze students' reasoning to implement continuous and adaptive formative assessments.

Limitations

This study investigated the interdependence of a research-informed and an ML-based rubric in assessing student reasoning in chemistry. By doing so, we compared three research-informed *analytic* categories with an ML-based *holistic* rubric, so the two applied rubrics differ in their features. While consciously deciding on analytic rubrics for human coding, we did not set the rubric type for ML-based coding. Rather, we understand the topics uncovered by the clustering technique as a holistic rubric since these topics represent different complexity levels. In contrast, Haudek et al. (2015) and Wulff, Westphal et al. (2023) argued that the categories they uncovered with a clustering technique can be used for holistic *and* analytic coding. For this reason, it might as well be possible to apply an ML-based rubric for analytic coding. However, the results of the applied clustering algorithm did not allow for an analytic ML-based coding of students' responses in our analysis because the identified clusters characterize unique mechanistic reasoning patterns, rather than non-mutually exclusive conceptual components. To address these design differences, we combined the three research-informed analytic codes into a holistic rubric, which assigned each student response to a mutually exclusive category (Sect. 4.5: "Quantitative Comparisons between the Research-Informed and ML-Based Rubrics" and 5.3: "Comparing the Computationally Detected Clusters and the Combination of the Research-Informed Categories"). Consequently, comparing the merged research-informed and ML-based rubrics may be more valid because both are holistic.

However, the number of research-informed categories and ML-based clusters differ. The merged research-informed rubric includes eight categories, whereas the ML-based rubric features fifteen. Again, while consciously including three research-informed reasoning types, leading to eight combinations, we did not set the cluster number for ML-based coding in advance since HDBSCAN automatically determines this number.

Moreover, we concentrated on the validity of an ML-based rubric for formative assessments but did not measure the performance of supervised ML classifiers trained on such a rubric. Consequently, the potential for scaling our approach across formative assessments remains unexplored. Nonetheless, other analyses suggest that combining unsupervised rubric development with supervised classification is promising for scoring student reasoning across domains (e.g., Martin et al., 2024; Rosenberg & Krist, 2021; Tschisgale et al., 2023). Future research should investigate the accuracy, generalizability, and interpretability of these classifiers when applied to new data.

From a chemistry perspective, our analysis is limited by the context of this study. We concentrated solely on undergraduate organic chemistry and used a specific

task format—case comparisons—to elicit students' mechanistic reasoning, focusing exclusively on S_N1 mechanisms. Consequently, the administered tasks shared the same prompt, structure, and content, leading to a conceptual overlap in students' responses. Other task types might reveal different aspects of students' reasoning in organic chemistry, potentially leading to modifications in the categories of the ML-based rubric.

Ultimately, the ML-based rubric applies only to chemical reactions including S_N1 mechanisms—like those it is trained on. In contrast, the research-informed rubric demonstrates greater generalizability as it classifies explanations across reaction types. Therefore, the research-informed rubric can be applied to a broader range of chemistry assessments. Future research could build on our current analysis by creating a more generalizable ML-based rubric applicable to a wider range of organic chemistry reactions. Examining how such a rubric relates to well-established research-informed categories would be valuable.

Acknowledgements This publication is part of the doctoral thesis (Dr. rer. nat.) of both co-lead authors at the Faculty of Biology and Chemistry, Justus-Liebig-University Giessen, Germany. We thank Peter Wulff for his help in implementing the machine learning analysis and all members of the Graulich group for fruitful discussions. Paul P. Martin thanks the German Chemical Industry Association (Verband der Chemischen Industrie) for supporting him with the Kekulé Fellowship. David Kranz thanks the German Research Foundation DFG (Deutsche Forschungsgemeinschaft) for funding this research (grant number: 446349713).

Author Contributions Paul P. Martin: Conceptualization, Investigation, Writing - original draft, Methodology, Validation, Writing - review & editing, Resources, Software, Formal analysis

David Kranz: Investigation, Writing - original draft, Methodology, Visualization, Resources, Data curation, Software, Formal analysis

Nicole Graulich: Funding acquisition, Investigation, Writing - review & editing, Supervision, Project administration

Funding Open Access funding enabled and organized by Projekt DEAL. Author Paul P. Martin has received research support from the German Chemical Industry Association (Verband der Chemischen Industrie) and author David Kranz has received research support from the German Research Foundation DFG (Deutsche Forschungsgemeinschaft) (grant number: 446349713). The authors have no relevant financial or non-financial interests to disclose.

Data Availability The data used in this study is available from the corresponding author upon reasonable request.

Declarations

Ethical Approval All data collection procedures are based on the ethical guidelines of the German Research Foundation.

Competing Interests The authors have no conflicts of interest to declare.

Consent to Participate Informed consent was obtained from all participants.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line

to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Agresti, A. (2013). *Categorical data analysis*. Wiley.
Agresti, A. (2018). *An introduction to categorical data analysis*. Wiley.
Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A Meta-Analytic Review. *Educational Psychologist*, *48*(2), 87–113. https://doi.org/10.1080/00461520.2013.775712
Anderson, D., Rowley, B., Stegenga, S., Irvin, P. S., & Rosenberg, J. M. (2020). Evaluating content-related validity evidence using a text-based machine learning procedure. *Educational Measurement: Issues and Practice, 39*(4), 53–64. https://doi.org/10.1111/emip.12314
Armitage, P., Berry, G., & Matthews, J. N. S. (2008). *Statistical methods in medical research*. Wiley.
Auguie, B. (2017). *gridExtra: Miscellaneous functions for grid graphics*. [Computer Program].
Bail, C. A. (2014). The cultural environment: Measuring culture with big data. *Theory and Society*, *43*(3/4), 465–482. https://doi.org/10.1007/s11186-014-9216-5
Beasley, T. M., & Schumacker, R. E. (1995). Multiple regression approach to analyzing contingency tables: Post hoc and planned comparison procedures. *The Journal of Experimental Education*, *64*(1), 79–93. https://doi.org/10.1080/00220973.1995.9943797
Becker, N., Noyes, K., & Cooper, M. (2016). Characterizing students' mechanistic reasoning about London dispersion forces. *Journal of Chemical Education, 93*(10), 1713–1724. https://doi.org/10.1021/acs.jchemed.6b00298
Bellmann, R. (1978). *An introduction to artificial intelligence. Can computers think?* Boyd and Fraser.
Bhattacharyya, G., & Bodner, G. M. (2005). "It gets me to the product": How students propose organic mechanisms. *Journal of Chemical Education, 82*(9), 1402–1407. https://doi.org/10.1021/ed082p1402
Biernacki, R. (2012). *Reinventing evidence in social inquiry: Decoding facts and variables*. Palgrave Macmillan.
Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
Bodé, N. E., Deng, J. M., & Flynn, A. B. (2019). Getting past the rules and to the WHY: Causal mechanistic arguments when judging the plausibility of organic reaction mechanisms. *Journal of Chemical Education, 96*(6), 1068–1082. https://doi.org/10.1021/acs.jchemed.8b00719
Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, *3*(22), 1–12. https://doi.org/10.3389/feduc.2018.00022
Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, *67*(3), 343–368. https://doi.org/10.1080/00131911.2014.929565
Brunton, S. L., & Kutz, J. N. (2019). *Data-Driven Science and Engineering: Machine learning, Dynamical systems, and control*. Cambridge University Press.
Bussey, T. J., Orgill, M., & Crippen, K. J. (2013). Variation theory: A theory of learning and a useful theoretical framework for chemical education research. *Chemistry Education Research and Practice*, *14*(1), 9–22. https://doi.org/10.1039/C2RP20145C
Carlsen, H. B., & Ralund, S. (2022). Computational grounded theory revisited: From computer-led to computer-assisted text analysis. *Big Data & Society, 9*(1), 20539517221080146. https://doi.org/10.1177/20539517221080146
Caspari, I., & Graulich, N. (2019). Scaffolding the structure of organic chemistry students' multivariate comparative mechanistic reasoning. *International Journal of Physics and Chemistry Education*, *11*(2), 31–43. https://doi.org/10.12973/ijpce/211359
Caspari, I., Kranz, D., & Graulich, N. (2018). Resolving the complexity of organic chemistry students' reasoning through the lens of a mechanistic framework. *Chemistry Education Research and Practice*, *19*(4), 1117–1141. https://doi.org/10.1039/C8RP00131F
Chang, W. (2017). *Webshot: Take screenshots of web pages.* [Computer program].
Charmaz, K. (2014). *Constructing grounded theory*. Sage.

Chin, D. B., Chi, M., & Schwartz, D. L. (2016). A comparison of two methods of active learning in physics: Inventing a general solution versus compare and contrast. *Instructional Science*, *44*(2), 177–195. https://doi.org/10.1007/s11251-016-9374-0

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Psychology Press Taylor & Francis.

Cooper, M. M., Kouyoumdjian, H., & Underwood, S. M. (2016). Investigating students' reasoning about acid–base reactions. *Journal of Chemical Education, 93*(10), 1703–1712. https://doi.org/10.1021/acs.jchemed.6b00417

Crandell, O. M., Kouyoumdjian, H., Underwood, S. M., & Cooper, M. M. (2019). Reasoning about reactions in organic chemistry: Starting it in general chemistry. *Journal of Chemical Education, 96*(2), 213–226. https://doi.org/10.1021/acs.jchemed.8b00784

Crowder, C. J., Yik, B. J., Frost, S. J., Cruz-Ramírez de Arellano, D., & Raker, J. R. (2024). Impact of prompt cueing on level of explanation sophistication for organic reaction mechanisms. *Journal of Chemical Education, 101*(2), 398–410. https://doi.org/10.1021/acs.jchemed.3c00710

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, *162*(104094), 1–43. https://doi.org/10.1016/j.compedu.2020.104094

DeGlopper, K. S., Schwarz, C. E., Ellias, N. J., & Stowe, R. L. (2022). Impact of assessment emphasis on organic chemistry students' explanations for an alkene addition reaction. *Journal of Chemical Education, 99*(3), 1368–1382. https://doi.org/10.1021/acs.jchemed.1c01080

Deng, J. M., & Flynn, A. B. (2021). Reasoning, granularity, and comparisons in students' arguments on two organic chemistry items. *Chemistry Education Research and Practice*, *22*(3), 749–771. https://doi.org/10.1039/D0RP00320D

Deutsche Forschungsgemeinschaft. (2022). *Guidelines for safeguarding good research practice, code of conduct*. DFG.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, arXiv:1810.04805, 1–16. https://doi.org/10.48550/arXiv.1810.04805

Donnelly, D. F., Vitale, J. M., & Linn, M. C. (2015). Automated guidance for thermodynamics essays: Critiquing versus revisiting. *Journal of Science Education and Technology, 24*(6), 861–874. https://doi.org/10.1007/s10956-015-9569-1

Dood, A. J., & Watts, F. M. (2022). Mechanistic reasoning in organic chemistry: A scoping review of how students describe and explain mechanisms in the chemistry education research literature. *Journal of Chemical Education, 99*(8), 2864–2876. https://doi.org/10.1021/acs.jchemed.2c00313

Dood, A. J., Fields, K. B., & Raker, J. R. (2018). Using lexical analysis to predict Lewis acid–base model use in response to an acid–base proton-transfer reaction. *Journal of Chemical Education, 95*(8), 1267–1275. https://doi.org/10.1021/acs.jchemed.8b00177

Dood, A. J., Dood, J. C., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2020). Analyzing explanations of substitution reactions using lexical analysis and logistic regression techniques. *Chemistry Education Research and Practice*, *21*(1), 267–286. https://doi.org/10.1039/C9RP00148D

Dood, A. J., Winograd, B. A., Finkenstaedt-Quinn, S. A., Gere, A. R., & Shultz, G. V. (2022). PeerBERT: Automated characterization of peer review comments across courses. In *LAK22: 12th international learning analytics and knowledge conference* (pp. 492–499). Association for Computing Machinery. https://doi.org/10.1145/3506860.3506892

Dood, A. J., Watts, F. M., Connor, M. C., & Shultz, G. V. (2024). Automated text analysis of organic chemistry students' written hypotheses. *Journal of Chemical Education, 101*(3), 807–818. https://doi.org/10.1021/acs.jchemed.3c00757

Ebbert, D. (2019). Chisq.posthoc.test: A post hoc analysis for Pearson's chi-squared test for count data. [Computer program].

European Union (2016). Regulation 2016/679 (General Data Protection Regulation). *Official Journal of the European Union*, *OJ L 119*, 04052016. cor. OJ L 127, 23.5.2018.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage.

Franovic, C. G. C., Noyes, K., Stoltzfus, J. R., Schwarz, C. V., Long, T. M., & Cooper, M. M. (2023). Undergraduate chemistry and biology students' use of causal mechanistic reasoning to explain and predict preferential protein-ligand binding activity. *Journal of Chemical Education, 100*(5), 1716–1727. https://doi.org/10.1021/acs.jchemed.2c00737

Frost, S. J. H., Yik, B. J., Dood, A. J., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2023). Evaluating electrophile and nucleophile understanding: A large-scale study of learners' explanations of reaction mechanisms. *Chemistry Education Research and Practice*, *24*(2), 706–722. https://doi.org/10.1039/D2RP00327A

Gerard, L. F., Matuk, C., McElhaney, K., & Linn, M. C. (2015). Automated, adaptive guidance for K-12 education. *Educational Research Review*, *15*, 41–58. https://doi.org/10.1016/j.edurev.2015.04.001

Glaser, B., & Strauss, A. (1999). *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Gombert, S., Di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., Grimm, A., Bohm, I., Neumann, K., & Drachsler, H. (2023). Coding energy knowledge in constructed responses with explainable NLP models. *Journal of Computer Assisted Learning*, *39*(3), 767–786. https://doi.org/10.1111/jcal.12767

Goodwin, W. M. (2003). Explanation in organic chemistry. *Annals of the New York Academy of Sciences, 988*(1), 141–153. https://doi.org/10.1111/j.1749-6632.2003.tb06093.x

Goodwin, W. M. (2008). Structural formulas and explanation in organic chemistry. *Foundations of Chemistry*, *10*(2), 117–127. https://doi.org/10.1007/s10698-007-9033-2

Graulich, N. (2015). The tip of the iceberg in organic chemistry classes: How do students deal with the invisible? *Chemistry Education Research and Practice*, *16*(1), 9–21. https://doi.org/10.1039/C4RP00165F

Graulich, N., & Caspari, I. (2021). Designing a scaffold for mechanistic reasoning in organic chemistry. *Chemistry Teacher International*, *3*(1), 19–30. https://doi.org/10.1515/cti-2020-0001

Graulich, N., & Schween, M. (2018). Concept-oriented task design: Making purposeful case comparisons in organic chemistry. *Journal of Chemical Education, 95*(3), 376–383. https://doi.org/10.1021/acs.jchemed.7b00672

Graulich, N., Hedtrich, S., & Harzenetter, R. (2019). Explicit versus implicit similarity—exploring relational conceptual understanding in organic chemistry. *Chemistry Education Research and Practice*, *20*(4), 924–936. https://doi.org/10.1039/C9RP00054B

Grootendorst, M. (2020). Topic modeling with BERT. Retrieved 20 April 2023 from https://towardsdatascience.com/topicmodeling-with-bert-779f7db187e6

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles Policy & Practice*, *20*(3), 281–307. https://doi.org/10.1080/0969594X.2012.742422

Haudek, K. C., Moscarella, R. A., Urban-Lurain, M., Merrill, J. E., Sweeder, R. D., & Richmond, G. (2009). Using lexical analysis software to understand student knowledge transfer between chemistry and biology. Paper presented at the National Association of Research in Science Teaching, Annual Conference, Garden Grove, CA.

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J. E., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid–base chemistry in introductory biology. *CBE - Life Sciences Education, 11*(3), 283–293. https://doi.org/10.1187/cbe.11-08-0084

Haudek, K. C., Moscarella, R. A., Weston, M., Merrill, J. E., & Urban-Lurain, M. (2015). Construction of rubrics to evaluate content in students' scientific explanation using computerized text analysis. Paper presented at the National Association of Research in Science Teaching, Annual Conference, Chicago, IL.

Haudek, K. C., Wilson, C. D., Stuhlsatz, M. A. M., Donovan, B., Bracey, Z. B., Gardner, A., Osborne, J. F., & Cheuk, T. (2019). Using automated analysis to assess middle school students' competence with scientific argumentation. Paper presented at the National Conference on Measurement in Education (NCME), Annual Conference, Toronto, ON.

Haugeland, J. (1989). *Artificial intelligence: The very idea*. MIT Press.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength natural language processing in Python.* [Computer program].

Howell, D. C. (2006). *Statistical methods for psychology*. PWS-Kent Publishing Co.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://ieeexplore.ieee.org/document/4160265

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Doherty, J. H., Wenderoth, M. P., Merrill, J. E., Urban-Lurain, M., & Haudek, K. C. (2019). Deconstruction of holistic rubrics into analytic bins for large-scale assessments of students' reasoning of complex science concepts. *Practical Assessment Research & Evaluation, 24*(7), 1–13. https://doi.org/10.7275/9h7f-mp76

Jescovitch, L. N., Doherty, J. H., Scott, E. E., Cerchiara, J. A., Wenderoth, M. P., Urban-Lurain, M., Merrill, J. E., & Haudek, K. C. (2019). Challenges in developing computerized scoring models for principle-based reasoning in a physiology context. Paper presented at the National Association of Research in Science Teaching, Annual Conference, Baltimore, MD.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J. E., Urban-Lurain, M., Doherty, J. H., & Haudek, K. C. (2021). Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression. *Journal of Science Education and Technology, 30*(2), 150–167. https://doi.org/10.1007/s10956-020-09858-0

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Kaldaras, L., & Haudek, K. C. (2022). Validation of automated scoring for learning progression-aligned Next Generation Science Standards performance assessments. *Frontiers in Education, 7*(968289), 1–22. https://doi.org/10.3389/feduc.2022.968289

Kaldaras, L., Yoshida, N. R., & Haudek, K. C. (2022). Rubric development for AI-enabled scoring of three-dimensional constructed-response assessment aligned to NGSS learning progression. *Frontiers in Education*, *7*(983055), 1–15. https://doi.org/10.3389/feduc.2022.983055

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley.

Kraft, A., Strickland, A. M., & Bhattacharyya, G. (2010). Reasonable reasoning: Multi-variate problem-solving in organic chemistry. *Chemistry Education Research and Practice*, *11*(4), 281–292. https://doi.org/10.1039/C0RP90003F

Kranz, D., Schween, M., & Graulich, N. (2023). Patterns of reasoning—exploring the interplay of students' work with a scaffold and their conceptual knowledge in organic chemistry. *Chemistry Education Research and Practice*, *24*(2), 453–477. https://doi.org/10.1039/D2RP00132B

Kranz, D., Martin, P. P., Schween, M., & Graulich, N. (under review). Should we scaffold it? Analysing students' learning gains to evaluate the effect of task format and scaffolding. *Chemistry Education Research and Practice.*

Krist, C., Schwarz, C. V., & Reiser, B. J. (2019). Identifying essential epistemic heuristics for guiding mechanistic reasoning in science learning. *Journal of the Learning Sciences, 28*(2), 160–205. https://doi.org/10.1080/10508406.2018.1510404

Kubsch, M., Czinczel, B., Lossjew, J., Wyrwich, T., Bednorz, D., Bernholt, S., Fiedler, D., Strauß, S., Cress, U., Drachsler, H., Neumann, K., & Rummel, N. (2022). Toward learning progression analytics—developing learning environments for the automated analysis of learning using evidence centered design. *Frontiers in Education*, *7*(981910), 1–15. https://doi.org/10.3389/feduc.2022.981910

Kubsch, M., Krist, C., & Rosenberg, J. M. (2023). Distributing epistemic functions and tasks—A framework for augmenting human analytic power with machine learning in science education research. *Journal of Research in Science Teaching*, *60*(2), 423–447. https://doi.org/10.1002/tea.21803

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development*, *74*(5), 1245–1260. https://doi.org/10.1111/1467-8624.00605

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Li, T., Reigh, E., He, P., & Adah Miller, E. (2023). Can we and should we use artificial intelligence for formative assessment in science? *Journal of Research in Science Teaching*, *60*(6), 1385–1389. https://doi.org/10.1002/tea.21867

Lim, L., Bannert, M., van der Graaf, J., Singh, S., Fan, Y., Surendrannair, S., Rakovic, M., Molenaar, I., Moore, J., & Gašević, D. (2023). Effects of real-time analytics-based personalized scaffolds on students' self-regulated learning. *Computers in Human Behavior*, *139*(107547), 1–18. https://doi.org/10.1016/j.chb.2022.107547

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, *33*(2), 19–28. https://doi.org/10.1111/emip.12028

Lo, M. L., & Marton, F. (2012). Towards a science of the art of teaching. *International Journal for Lesson and Learning Studies*, *1*(1), 7–22. https://doi.org/10.1108/20468251211179678

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25. https://doi.org/10.1086/392759

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. Lecam, & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (pp. 281–297). University of California.

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B., & Krajcik, J. (2021). Using machine learning to score multidimensional assessments of chemistry and physics. *Journal of Science Education and Technology, 30*(2), 239–254. https://doi.org/10.1007/s10956-020-09895-9

Martin, P. P., & Graulich, N. (2023). When a machine detects student reasoning: A review of machine learning-based formative assessment of mechanistic reasoning. *Chemistry Education Research and Practice*, *24*(2), 407–427. https://doi.org/10.1039/D2RP00287F

Martin, P. P., & Graulich, N. (2024a). Beyond language barriers: Allowing multiple languages in post-secondary chemistry classes through multilingual machine learning. *Journal of Science Education and Technology, 33*(2), 333–348. https://doi.org/10.1007/s10956-023-10087-4

Martin, P. P., & Graulich, N. (2024b). Lehre in der Organischen Chemie individualisieren [Individualized teaching in organic chemistry]. *Nachrichten aus der Chemie, 72*(3), 8–11. https://doi.org/10.1002/nadc.20244141003

Martin, P. P., & Graulich, N. (2024c). Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence. *Computers and Education: Artificial Intelligence*, *7*(100265). https://doi.org/10.1016/j.caeai.2024.100265

Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2024). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*, *61*(8), 1757–1792. https://doi.org/10.1002/tea.21903

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia Medica*, *23*(2), 143–149. https://doi.org/10.11613/bm.2013.018

McInnes, L., Healy, J., & Astels, S. (2017). HDBSCAN: Hierarchical density-based clustering. *Journal of Open Source Software*, *2*(11), 205–206. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software, 3*(29), 861–862. https://doi.org/10.21105/joss.00861

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint*, arXiv:1802.03426, 1–63. https://doi.org/10.48550/arXiv.1802.03426

McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 51–56). SciPy 2010. https://doi.org/10.25080/Majora-92bf1922-00a

Mislevy, R. J. (2016). How developments in psychology and technology challenge validity argumentation. *Journal of Educational Measurement, 53*(3), 265–292. https://doi.org/10.1111/jedm.12117

Mislevy, R. J., & Haertel, G. D. (2007). Implications of evidence-centered design for educational testing. *Educational Measurement:Issues and Practice, 25*(4), 6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, (1), i–29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundation of machine learning*. The MIT Press.

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research, 49*(1), 3–42. https://doi.org/10.1177/0049124117729703

Noyes, K., McKay, R. L., Neumann, M., Haudek, K. C., & Cooper, M. M. (2020). Developing computer resources to automate analysis of students' explanations of London dispersion forces. *Journal of Chemical Education, 97*(11), 3923–3936. https://doi.org/10.1021/acs.jchemed.0c00445

Noyes, K., Carlson, C. G., Stoltzfus, J. R., Schwarz, C. V., Long, T. M., & Cooper, M. M. (2022). A deep look into designing a task and coding scheme through the lens of causal mechanistic reasoning. *Journal of Chemical Education, 99*(2), 874–885. https://doi.org/10.1021/acs.jchemed.1c00959

Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review, 30*(100329), 1–19. https://doi.org/10.1016/j.edurev.2020.100329

Patefield, W. (1981). Algorithm AS 159: An efficient method of generating random R×C tables with given row and column totals. *Journal of the Royal Statistical Society*, *30*(1), 91–97. https://doi.org/10.2307/2346669

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J. T., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*(1), 2825–2830. https://doi.org/10.5555/1953048.2078195

Pellegrino, J., DiBello, L., & Goldman, S. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 1–23. https://doi.org/10.1080/00461520.2016.1145550

Pölloth, B., Diekemper, D., & Schwarzer, S. (2023). What resources do high school students activate to link energetic and structural changes in chemical reactions?—A qualitative study. *Chemistry Education Research and Practice*, *24*(4), 1153–1173. https://doi.org/10.1039/D3RP00068K

Prevost, L. B., Haudek, K. C., Merrill, J. E., & Urban-Lurain, M. (2012). Examining student constructed explanations of thermodynamics using lexical analysis. In *42nd Frontiers in Education Conference Proceedings* (pp. 1–6). IEEE. https://doi.org/10.1109/FIE.2012.6462451

R Core Team. (2023). *R: A language and environment for statistical computing*. [Computer Program].

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 1–24. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Raker, J. R., Yik, B. J., & Dood, A. J. (2022). Development of a generalizable framework for machine learning-based evaluation of written explanations of reaction mechanisms from the post-secondary organic chemistry curriculum. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (pp. 304–319). The Royal Society of Chemistry. https://doi.org/10.1039/9781839167782-00304

Rosenberg, J. M., & Krist, C. (2021). Combining machine learning and qualitative methods to elaborate students' ideas about the generality of their model-based explanations. *Journal of Science Education and Technology, 30*(2), 255–267. https://doi.org/10.1007/s10956-020-09862-4

RStudio Team. (2023). *RStudio: Integrated development environment for R*. [Computer Program].

Rupp, A. A., Levy, R., Dicerbo, K. E., Sweet, S. J., Crawford, A. V., Caliço, T., Benson, M., Fay, D., Kunze, K. L., Mislevy, R. J., & Behrens, J. T. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining, 4*(1), 49–110. https://doi.org/10.5281/zenodo.3554643

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education, 92*(3), 499–525. https://doi.org/10.1002/sce.20264

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction*, *83*(101620), 1–10. https://doi.org/10.1016/j.learninstruc.2022.101620

Saldana, J. (2015). *The coding manual for qualitative researchers*. Sage.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 211–229. https://doi.org/10.1147/rd.33.0210

Sevian, H., & Talanquer, V. (2014). Rethinking chemistry: A learning progression on chemical thinking. *Chemistry Education Research and Practice*, *15*(1), 10–23. https://doi.org/10.1039/C3RP00111C

Sherin, B. (2013). A computational study of commonsense science: An exploration in the automated analysis of clinical interview data. *Journal of the Learning Sciences, 22*(4), 600–638. https://doi.org/10.1080/10508406.2013.836654

Stowe, R. L., & Cooper, M. M. (2017). Practicing what we preach: Assessing critical thinking in organic chemistry. *Journal of Chemical Education, 94*(12), 1852–1859. https://doi.org/10.1021/acs.jchemed.7b00335

Stowe, R. L., Scharlott, L. J., Ralph, V. R., Becker, N. M., & Cooper, M. M. (2021). You are what you assess: The case for emphasizing chemistry on chemistry assessments. *Journal of Chemical Education, 98*(8), 2490–2495. https://doi.org/10.1021/acs.jchemed.1c00532

Taher Pilehvar, M., & Camacho-Collados, J. (2020). *Embeddings in natural language processing: Theory and advances in vector representations of meaning*. Morgan & Claypool.

Talanquer, V. (2014). Chemistry education: Ten heuristics to tame. *Journal of Chemical Education, 91*(8), 1091–1097. https://doi.org/10.1021/ed4008765

Tansomboon, C., Gerard, L. F., Vitale, J. M., & Linn, M. C. (2017). Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education, 27*(4), 729–757. https://doi.org/10.1007/s40593-017-0145-0

Tomas, C., Whitt, E., Lavelle-Hill, R., & Severn, K. (2019). Modeling holistic marks with analytic rubrics. *Frontiers in Education, 4*(89), 1–19. https://doi.org/10.3389/feduc.2019.00089

Toulmin, S. E. (2003). *The uses of argument* (Rev. ed.). Cambridge University Press.

Tschisgale, P., Wulff, P., & Kubsch, M. (2023). Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review Physics Education Research*, *19*(2), 020123. https://doi.org/10.1103/PhysRevPhysEducRes.19.020123

Urban-Lurain, M., Prevost, L. B., Haudek, K. C., Henry, E. N., Berry, M., & Merrill, J. E. (2013). Using computerized lexical analysis of student writing to support Just-in-Time Teaching in large enrollment STEM courses. In *43rd Frontiers in Education Conference Proceedings* (pp. 1709–1715). IEEE. https://doi.org/10.1109/FIE.2013.6685130

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł, & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc. https://doi.org/10.48550/arXiv.1706.03762

Vitale, J. M., McBride, E., & Linn, M. C. (2016). Distinguishing complex ideas about climate change: Knowledge integration vs. specific guidance. *International Journal of Science Education*, *38*(9), 1548–1569. https://doi.org/10.1080/09500693.2016.1198969

Wang, C., Liu, X., Wang, L., Sun, Y., & Zhang, H. (2021). Automated scoring of Chinese grades 7–9 students' competence in interpreting and arguing from evidence. *Journal of Science Education and Technology*, *30*(2), 269–282. https://doi.org/10.1007/s10956-020-09859-z

Warnes, G. R., Bolker, B., Lumley, T., & Johnson, R. C. (2023). *gmodels: Various R programming tools for model fitting*. [Computer program].

Waskom, M., Gelbart, M., Botvinnik, O., Ostblom, J., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., & Warmenhoven, J. (2020). *seaborn.* [Computer program].

Watts, F. M., Zaimi, I., Kranz, D., Graulich, N., & Shultz, G. V. (2021). Investigating students' reasoning over time for case comparisons of acyl transfer reaction mechanisms. *Chemistry Education Research and Practice*, *22*(2), 364–381. https://doi.org/10.1039/D0RP00298D

Watts, F. M., Dood, A. J., & Shultz, G. V. (2022). Developing machine learning models for automated analysis of organic chemistry students' written descriptions of organic reaction mechanisms. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-base*.

Weinrich, M. L., & Talanquer, V. (2016). Mapping students' modes of reasoning when thinking about chemical reactions used to make a desired product. *Chemistry Education Research and Practice*, *17*(2), 394–406. https://doi.org/10.1039/C5RP00208G

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. https://doi.org/10.18637/jss.v021.i12

Wickham, H. (2016). Data analysis. In H. Wickham (Ed.), *ggplot2: Elegant graphics for data analysis* (pp. 189–201). Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4_9

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1–6. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *dplyr: A grammar of data manipulation.* [Computer program].

Wilson, C. D., Haudek, K. C., Osborne, J. F., Bracey, B., Cheuk, Z. E., Donovan, T., Stuhlsatz, B. M., Santiago, M. A. M., M. M., & Zhai, X. (2023). Using automated analysis to assess middle school

students' competence with scientific argumentation. *Journal of Research in Science Teaching*, *61*(1), 38–69. https://doi.org/10.1002/tea.21864

Winograd, B. A., Dood, A. J., Finkenstaedt-Quinn, S. A., Gere, A. R., & Shultz, G. V. (2021). Automating characterization of peer review comments in chemistry courses. In C. E. Hmelo-Silver, B. De Wever, & J. Oshima (Eds.), *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning: CSCL 2021* (pp. 11–18). International Society of the Learning Sciences. https://doi.org/10.22318/cscl2021.11

Winograd, B. A., Dood, A. J., Moon, A., Moeller, R., Shultz, G. V., & Gere, A. R. (2021). Detecting high orders of cognitive complexity in students' reasoning in argumentative writing about ocean acidification. In *11th International Learning Analytics and Knowledge Conference* (pp. 586–591). Association for Computing Machinery. https://doi.org/10.1145/3448139.3448202

Wulff, P., Buschhüter, D., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2022). Bridging the gap between qualitative and quantitative assessment in science education research with machine learning—A case for pretrained language models-based clustering. *Journal of Science Education and Technology, 31*(4), 490–513. https://doi.org/10.1007/s10956-022-09969-w

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2023). Utilizing a pretrained language model (BERT) to classify preservice physics teachers' written reflections. *International Journal of Artificial Intelligence in Education, 33*(3), 439–466. https://doi.org/10.1007/s40593-022-00290-6

Wulff, P., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2023). Enhancing writing analytics in science education research with machine learning and natural language processing—formative assessment of science and non-science preservice teachers' written reflections. *Frontiers in Education, 7*(1061461), 1–18. https://doi.org/10.3389/feduc.2022.1061461

Yik, B. J., Dood, A. J., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2021). Development of a machine learning-based tool to evaluate correct Lewis acid–base model use in written responses to open-ended formative assessment items. *Chemistry Education Research and Practice, 22*(4), 866–885. https://doi.org/10.1039/D1RP00111F

Yik, B. J., Schreurs, D. G., & Raker, J. R. (2023). Implementation of an R shiny app for instructors: An automated text analysis formative assessment tool for evaluating Lewis acid-base model use. *Journal of Chemical Education, 100*(8), 3107–3113. https://doi.org/10.1021/acs.jchemed.3c00400

Yik, B. J., Dood, A. J., Frost, S. J. H., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2023). Generalized rubric for level of explanation sophistication for nucleophiles in organic chemistry reaction mechanisms. *Chemistry Education Research and Practice, 24*(1), 263–282. https://doi.org/10.1039/D2RP00184E

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic Coding of Short text responses via clustering in Educational Assessment. *Educational and Psychological Measurement*, *76*(2), 280–303. https://doi.org/10.1177/0013164415590022

Zhai, X., Haudek, K. C., Shi, L., Nehm, R. H., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430–1459. https://doi.org/10.1002/tea.21658

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education, 56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757

Zhu, H., Travison, T., Tsai, T., Beasley, W., Xie, Y., Yu, G., Laurent, S., Shepherd, R., & Sidi, Y. (2022). *kableExtra: Construct complex table with kable and pipe syntax*. [Computer Program].

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

International Journal of Artificial Intelligence in Education

Authors and Affiliations

Paul P. Martin[1] · David Kranz[1] · Nicole Graulich[1]

✉ Nicole Graulich
Nicole.Graulich@didaktik.chemie.uni-giessen.de

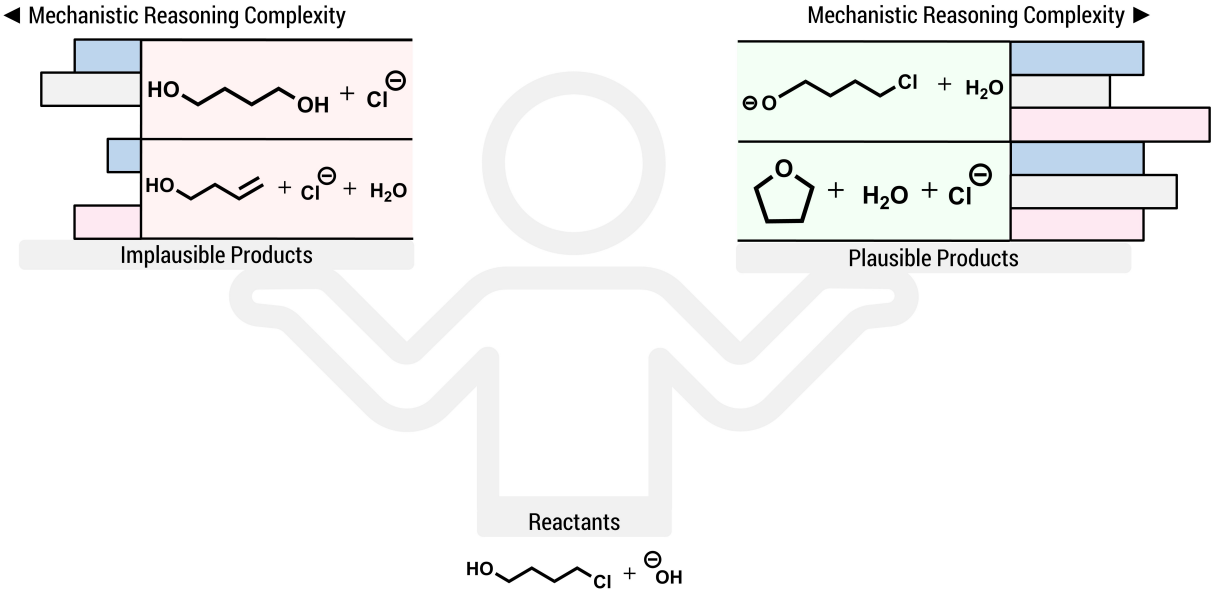
Paul P. Martin
Paul.Martin@didaktik.chemie.uni-giessen.de

David Kranz
David.Kranz@didaktik.chemie.uni-giessen.de

[1] Institute of Chemistry Education, Justus-Liebig-University, Heinrich-Buff-Ring 17, 35392 Giessen, Germany

8 | Mechanistic reasoning about alternatives: How students reason about plausible and implausible mechanisms in chemistry

Chapter 8 presents a book chapter by Martin, Yik, Burlingham, and Graulich that is accepted for publication in the book *Mechanistic Reasoning in Science Education* (eds. R. W. Bachtiar, M. Haskel-Ittah, & N. Graulich; published by Springer) but has not yet been in press.



Chapter 1.3

Mechanistic Reasoning About Alternatives: How Students Reason About Plausible and Implausible Mechanisms in Chemistry

Paul P. Martin[1], Brandon J. Yik[2], Benjamin T. Burlingham[3], Nicole Graulich[1,*]

Abstract Mechanistic reasoning is a critical thinking skill in science disciplines like chemistry because it helps students understand the underlying processes driving mechanisms. Given the complex interplay of the involved entities and the occurring activities, explicitly guiding students to explain why mechanisms occur can advance their mechanistic reasoning. Nonetheless, encouraging students not only to explain why a mechanism is plausible but also why related ones may be rather implausible might further enhance their mechanistic reasoning. This study compares the complexity of undergraduate organic chemistry students' mechanistic reasoning when judging the plausibility of alternative mechanisms using quantitative analyses. The findings reveal certain characteristics that contribute to a higher mechanistic reasoning complexity, with the initial claim about whether a mechanism is plausible or implausible being a central predictor of students' mechanistic reasoning complexity. Considering these findings, supporting students' claim-making can be a first step in advancing their mechanistic reasoning. From a broader perspective, such support can highlight that specific interpretations of mechanistic outcomes allow for a certain degree of ambiguity, which provides transferable implications to other science disciplines.

---

[1]Institute of Chemistry Education
Justus-Liebig-University. Giessen, 35392
[*]E-mail: Nicole.Graulich@didaktik.chemie.uni-giessen.de

[2]Department of Chemistry
University of Georgia. Athens, 30602

[3]Department of Chemistry
Indiana University. Bloomington, 47405-7102

1.3.1 Introduction

Mechanisms of phenomena can be conceptualized as a dynamic interplay of entities and activities organized in a way to rationalize consistent changes from start to finish conditions (Machamer et al., 2000). Given this conceptualization, every mechanism contains four basic features: a phenomenon requiring explanation, the involved entities, the properties of those entities, and the resulting activities (Craver, 2007; Illari & Williamson, 2012). In organic chemistry, explaining phenomena encompasses how and why chemical reactions occur. Here, entities are the atoms and molecules involved in a mechanism, which are defined by specific properties, including their atom types, connectivity, spatial orientation, and electron distribution. Activities refer to the dynamic processes that entities exhibit during a mechanism, for instance, a nucleophilic substitution. A primary driving force for initiating these activities is to attain a lower energy state through bond rearrangements. Consequently, mechanisms in organic chemistry provide a model of continuous electron flow to illustrate, describe, and explain the underlying processes that drive chemical reactions (Goodwin, 2003, 2012).

Reasoning about mechanisms is crucial as it allows chemists to anticipate the selectivity of transformations, consider alternative pathways, design novel synthesis routes, and critically evaluate proposed reactions (Bhattacharyya & Bodner, 2005). Given these affordances, instruction must explicitly support chemistry students' mechanistic reasoning so they can consider multiple molecular properties to predict the plausibility of various products (Grove & Bretz, 2010). This explicit support might help students move beyond a simplistic view of mechanisms as a linear sequence of events; instead, students can learn to integrate the complex interactions of entities and activities in their mechanistic reasoning (e.g., Weinrich & Sevian, 2017).

When dealing with mechanisms of varying plausibility, students need additional support to handle the arising ambiguity in the reasoning process (Cruz-Ramírez de Arellano & Towns, 2014; Grove & Bretz, 2010; Lieber et al., 2022a, 2022b). Therefore, instruction should guide students to shift their perspective from believing that a mechanism yields only one defined product to understanding that reagents can produce various products and byproducts with differing probabilities (Grove & Bretz, 2010). This shift could be essential as it can contribute to a more sophisticated perception of a mechanism as an outcome of interrelated interactions constrained by energy and entropy, rather than a pre-defined outcome (Talanquer, 2013). Hence, this study investigates differences in students' mechanistic reasoning complexity when judging the plausibility of alternative mechanisms. By evaluating students' reasoning about alternative mechanisms, we aim to support their mechanistic reasoning as well as their understanding that multiple outcomes of a mechanism can coexist in future formative assessments.

1.3.2 Conceptual Background

Mechanistic reasoning can be conceptualized as a form of causal reasoning that explains the sequential stages, from input to output, of the underlying causal events leading to a phenomenon (Krist et al., 2019; Machamer et al., 2000; Russ et al., 2008). Consequently, mechanistic reasoning extends beyond simply describing a phenomenon. It offers a theoretical perspective on the regularities underlying empirical observations, aiming to propose and reason through possible mechanisms that provide testable hypotheses on how and why phenomena occur (Darden & Craver, 2002; Russ et al., 2008). By systematically identifying, unpacking, and linking the underlying properties that impact entities and activities, mechanistic reasoning is a valuable strategy for predicting and explaining the outcome of scientific phenomena (Krist et al., 2019).

This study builds upon the *essential epistemic heuristics for mechanistic reasoning* by Krist et al. (2019) to categorize undergraduate organic chemistry students' reasoning about alternative mechanisms (Fig. 1.3.1). Krist et al. (2019) specify three essential epistemic heuristics for evaluating the extent of students' engagement in mechanistic reasoning. The first essential epistemic heuristic involves reasoning across different scalar levels, especially considering the scalar level below the target phenomenon (Krist et al., 2019). In other words, reasoning about what is happening at a scale below the target phenomenon forms the foundation for complex mechanistic reasoning that characterizes the causes of underlying processes. In our organic chemistry context, we used two *levels of granularity* to classify whether a student considered the scalar level below the target phenomenon (Fig. 1.3.1): *phenomenological* and *electronic* (Bodé et al., 2019; Deng & Flynn, 2021). *Phenomenological* descriptions refer to mechanistic reasoning about measurable properties of entities or activities like pK_a values, the direction of chemical equilibrium, or reaction rate, without examining electron distribution. In this way, *phenomenological* descriptions of organic chemistry reactions tend to consider only the target phenomenon while disregarding the lower scalar level. A lower scalar level encompasses *electronic* properties. *Electronic* mechanistic reasoning discusses the distribution and movement of electrons in chemical compounds through electronic effects, such as resonance, hyperconjugation, or inductive effects. Therefore, the *electronic* level represents a lower scalar level than the *phenomenological* level, making *electronic* mechanistic reasoning more complex than *phenomenological* mechanistic reasoning.

Once students examine the lower scalar level, they must identify and unpack the relevant entities, activities, and properties that drive the phenomenon (Krist et al., 2019). Identifying and unpacking relevant entities, activities, and properties requires knowledge of disciplinary core ideas (Grove et al., 2012). Our analysis used disciplinary core ideas related to *structural* and *energetic* features to determine the entities and activities students reasoned about (Bodé et al., 2019; Deng & Flynn, 2021). Disciplinary core ideas related to *structural* features concentrate on the

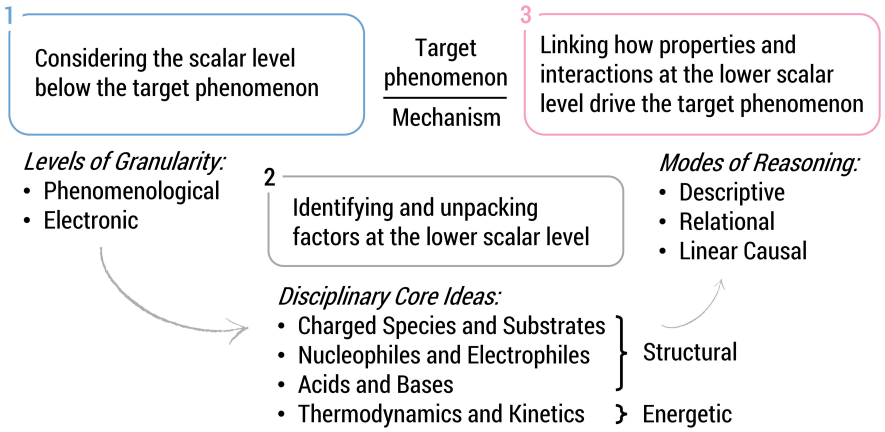


Fig. 1.3.1 Essential epistemic heuristics (Krist et al., 2019) to classify students' mechanistic reasoning adapted to an organic chemistry context

properties of entities. In chemistry contexts, *structural* features like atom connectivity, atom sizes, or charges govern the entities' properties. To categorize the entities and their corresponding properties in students' mechanistic reasoning, we disaggregated this *structural* level into three disciplinary core ideas according to the properties students assigned to the entities: *charged species and substrates*, *nucleophiles and electrophiles*, and *acids and bases* (Martin et al., 2024). Instead, disciplinary core ideas related to *energetic* features focus on the properties of activities. In chemistry, these properties of activities include thermodynamic and kinetic changes like the reaction rate, any influence of entropic effects on the reaction process, and whether a reaction favors reactants or products. We introduced the disciplinary core idea of *thermodynamics and kinetics* to capture students' *energetic* mechanistic reasoning (Fig. 1.3.1). Together, the disciplinary core ideas identify and unpack the properties of entities and activities in students' mechanistic reasoning about chemical reactions.

After identifying and unpacking the relevant entities, activities, and properties, these factors must be organized to form a sequence of causes over time or space. Accordingly, the third essential epistemic heuristic focuses on how well the parts of the underlying mechanism explain the observed phenomenon (Krist et al., 2019). This reasoning requires linking the identified properties to the phenomenon at its original scalar level and examining whether the unpacked entities and activities fully explain the phenomenon. To evaluate the degree of causality in this third heuristic, we applied three *modes of reasoning* (Fig. 1.3.1): *descriptive*, *relational*, and *linear causal* (Sevian & Talanquer, 2014). *Descriptive* reasoning only identifies explicit properties of single entities without providing any connections between those entities. The next level, *relational* reasoning, starts to unpack the implicit properties of entities, establishes relationships between a constrained number of those entities, and builds correlative links to the target phenomenon. Ultimately,

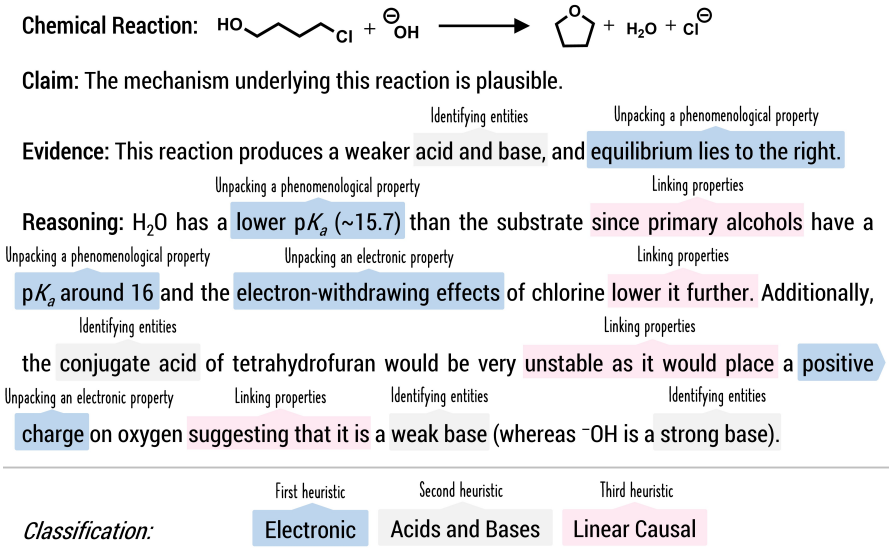


Fig. 1.3.2 Applying the essential epistemic heuristics to an exemplary student response

linear causal reasoning unpacks the explicit and implicit properties of most entities and creates causal links to the target phenomenon. In sum, these *modes of reasoning* evaluate the causality of students' mechanistic reasoning. We regard *linear causal* mechanistic reasoning as the most complex form.

Fig. 1.3.2 illustrates how the three essential epistemic heuristics may appear in students' mechanistic reasoning when they judge the plausibility of a mechanism within our organic chemistry context (Section 1.3.4). As shown in the student example, reasoning about products and byproducts with varying plausibility may require students to unpack a range of *phenomenological* and *electronic* properties. Combining the *levels of granularity* to classify the scalar level of students' mechanistic reasoning in the first heuristic with the *modes of reasoning* to operationalize students' causality in the third heuristic served as the foundation for evaluating the complexity of students' mechanistic reasoning. Accordingly, this analysis focuses on students' application of the first and third essential epistemic heuristics. Variations in how students apply the second heuristic are not examined in this article since the use of disciplinary core ideas is context-dependent (Deng & Flynn, 2021). For instance, an acid-base reaction may naturally prompt students to reason about acids and bases, so we hypothesize that students' use of disciplinary core ideas varies substantially depending on the specific mechanism. Consequently, longitudinal comparisons of disciplinary core ideas may only be partially valid and are, thus, not the focus of this article.

1.3.3 Research Question

Given this conceptualization, it becomes evident that mechanistic reasoning in organic chemistry is a complex practice. Besides, the possibility that the same reagents can produce multiple products of varying plausibility is not straightforward for students during decision-making (Grove & Bretz, 2010). Following this, the complexity of students' mechanistic reasoning may vary depending on their evaluation of a mechanism's plausibility. To identify the characteristics that facilitate complex mechanistic reasoning when confronted with alternatives, we compared the mechanistic reasoning complexity of undergraduate organic chemistry students for multiple sets of alternative mechanisms. Our analysis is guided by the following research question: To what degree does the complexity of students' mechanistic reasoning differ among response characteristics when reasoning about the plausibility of mechanisms? We examined the impact of three response characteristics on students' mechanistic reasoning complexity: the given reaction (Context), the actual plausibility or implausibility of a mechanism (Plausibility), and a student's claim about a mechanism's plausibility (Claim).

1.3.4 Research Setting

This study was conducted at a public university with very high research activity in the Midwestern United States during the Winter 2023 semester. The Institutional Review Board for the Social and Behavioral Sciences of a university in the mid-Atlantic region of the United States approved the study under Protocol #5937. Informed consent was obtained from all participants whose data were analyzed. At the beginning of the semester, 314 students from the Organic Chemistry II course consented to participate in this longitudinal study. Students received course credit for completing the assignments, with 80% awarded for task completion and 20% for the correctness of some closed-response questions. The closed-response questions are not further examined in this study. In total, 266 students completed all four assignments, resulting in 9,221 written responses. Data from students who did not complete all assignments were excluded from further analysis.

Of the 266 students, 19 opted not to provide demographic information. Of the remaining 247 participants, 149 identified as a woman, 93 as a man, and five as other gender identities. The students majored in various science-related fields; they were, on average, in their second year of undergraduate studies; and their ages ranged from 18 to 24 years.

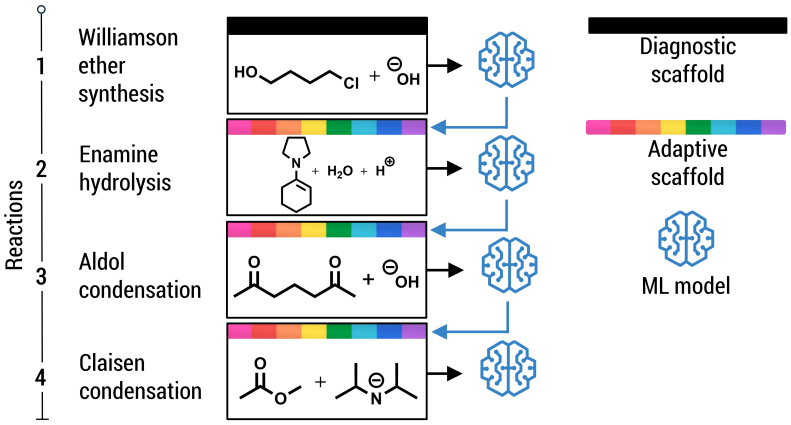


Fig. 1.3.3 Applying machine learning to adaptively support Organic Chemistry II students' mechanistic reasoning over one semester

1.3.4.1 Research Instrument

We designed and implemented four online scaffolds on the Qualtrics survey platform to support students' mechanistic reasoning. The scaffolds prompted students to judge the plausibility of alternative mechanisms for a Williamson ether synthesis, enamine hydrolysis, aldol condensation, and Claisen condensation (Fig. 1.3.3) (Lieber et al., 2022a, 2022b). Each of the four chemical reactions included two plausible and two implausible mechanisms. Students were sequentially presented with the corresponding alternative reaction products associated with these mechanisms but had to identify the underlying mechanisms that lead to these products independently.

Each scaffold consisted of three parts: closed-response tasks to guide students in structuring arguments within domain-general and chemistry-specific scenarios, closed-response tasks to activate students' chemical concept knowledge, and open-ended tasks to prompt students' reasoning about a mechanism's plausibility. The first scaffold served as a baseline for evaluating students' argumentation skills and chemical concept knowledge. Students' open-ended responses on a mechanism's plausibility were automatically scored using a machine learning (ML) algorithm—precisely a transformer-based deep neural network for sequence classification leveraging a pre-trained large language model for natural language processing (Martin et al., 2024). Before conducting this study, a text corpus of 3,430 student-written mechanistic reasoning statements was human-coded with an inductively developed rubric (Section 1.3.4.2) and served as training data for the ML algorithm (Martin & Graulich, 2024). The ML algorithm then identified the underlying patterns between students' open-ended reasoning and the human-assigned codes in the training corpus and automatically predicted the codes for the data gathered

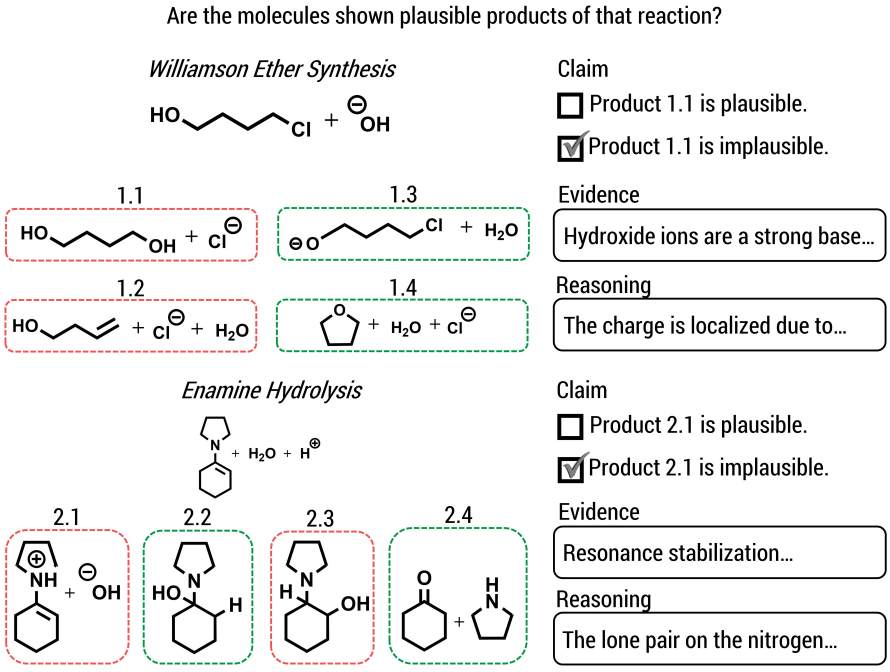


Fig. 1.3.4 Students judged the plausibility of four alternative mechanisms for four reactions, among others, a Williamson ether synthesis and an enamine hydrolysis. In doing so, students made a claim about the plausibility of the mechanism, justified the claim with evidence, and linked the claim and evidence based on reasoning. Implausible products (1.1, 1.2, 2.1, 2.3) are highlighted in red, and plausible ones in green (1.3, 1.4, 2.2, 2.4)

within this study. Thus, the ML algorithm replaced human coding by taking students' open-ended reasoning as input and generating a score representing the complexity of their mechanistic reasoning as output (Section 1.3.4.2). We then averaged students' complexity scores generated by the ML algorithm and manually assigned individualized feedback and a tailored set of exercises in Qualtrics to each student aligned with the complexity level demonstrated in their previous responses. This approach allowed us to individualize the second scaffold. After completing the exercises related to enamine hydrolysis, the ML algorithm automatically assessed students' open-ended answers again so that subsequent scaffolds could be continually individualized (Fig. 1.3.3). Our ML analysis has the potential for fine-grained diagnostic analysis of students' mechanistic reasoning (Martin & Graulich, 2023). Besides, our ML-based scoring relates to further projects presented in this book, which include ML applications to quantitatively analyze students' mechanistic reasoning (Watts et al., Chapter 1.1) along with alternative strategies for leveraging scaffolding and ML to deliver feedback and characterize students' mechanistic reasoning (Cooper et al., Chapter 2.2).

To scaffold students' reasoning about alternative mechanisms, we used a simplified version of Toulmin's argumentation pattern: the claim-evidence-reasoning model (McNeill & Krajcik, 2011; McNeill et al., 2006). A claim serves as a statement and requires causal support. Causal support comes from explanations or scientific data, known as evidence. This evidence draws on information from various sources to substantiate the claim. The reasoning establishes the cause-effect relation between the claim and evidence. In our setting, students claimed whether the displayed product is (im-)plausible, provided evidence explaining why the product might be (im-)plausible, and connected the claim and evidence through reasoning statements by incorporating electronic, steric, or energetic effects (Fig. 1.3.4). Sample solutions and student examples can be found in Table 1.3.3 and elsewhere (Lieber et al., 2022a, 2022b; Martin et al., 2024). Students were given a week to complete each scaffold. Completing a scaffold took 75–90 minutes. The scaffolds were scheduled in weeks 6, 8, 12, and 13 of the semester.

1.3.4.2 Rubric for Machine Scoring

To quantify students' mechanistic reasoning complexity—hereafter simply referred to as reasoning complexity—we developed a two-dimensional holistic rubric drawing from the *levels of granularity* (Bodé et al., 2019; Deng & Flynn, 2021) and *modes of reasoning* (Sevian & Talanquer, 2014) (Section 1.3.2). This rubric categorized every response holistically into one reasoning pattern. Based on the

		Structural						Energetic	
Disciplinary Core Ideas		Charged Species and Substrates		Nucleophiles and Electrophiles		Acids and Bases		Thermo-dynamics and Kinetics	
Levels of Granularity		Phen.	Elec.	Phen.	Elec.	Phen.	Elec.	Phen.	Elec.
Modes of Reasoning	Absent	0							
	Descriptive	2	3	2	3	2	3	2	3
	Relational	3	4	3	4	3	4	3	4
	Linear Causal	4	5	4	5	4	5	4	5

Fig. 1.3.5 Rubric to classify students' reasoning complexity. Numbers represent scores students received for their mechanistic reasoning. *Note:* Phen. stands for *phenomenological*; Elec. for *electronic*. Martin et al. (2024) provided a detailed account of the rubric development process

applied *level of granularity* (*absent* = 0, *phenomenological* = 1, *electronic* = 2) and *mode of reasoning* (*absent* = 0, *descriptive* = 1, *relational* = 2, *linear causal* = 3), every student-written response was machine-scored for increasing complexity between zero and five (Fig. 1.3.5). The disciplinary core ideas did not affect this scoring process; instead, students' mechanistic reasoning score was solely determined by whether their responses reflected *phenomenological* or *electronic granularity* along with the applied *mode of reasoning*.

1.3.5 Methods

We constructed hierarchical linear models (HLMs) with random effects to address the research question. HLMs are statistical models that account for data nested within multiple levels and allow for a flexible analysis of hierarchically structured repeated-measures data (Hox et al., 2017; Raudenbush & Bryk, 2002). Put differently, HLMs extend linear regression by incorporating multiple levels of data—so-called nested data structures—to capture individual and group-level influences on an outcome. In our case, students evaluated the plausibility of alternative mechanisms at four distinct time points during a semester. The analysis is performed at the response level, where the 9,221 student-written responses (level 1) are nested within the 266 students (level 2). We specified a two-level model with responses nested within students and time as a varying predictor due to the longitudinal nature of the data.

Each student response is classified according to three categorical predictor variables: the reaction to which it belonged (dummy-coded using the orthogonal contrasts defined in Table 1.3.2), the actual plausibility or implausibility of the given mechanism (implausible mechanism = 0, plausible mechanism = 1), and a student's claim regarding a mechanism's plausibility (student claim: implausible = 0, student claim: plausible = 1). For example, the students reasoned about mechanism 1.1 (Fig. 1.3.4), an implausible mechanism that belongs to Williamson ether synthesis; each student had to judge the mechanism's plausibility independently. We investigated the impact of these three response characteristics on students' reasoning complexity classified by the rubric discussed in Sections 1.3.2 and 1.3.4.2.

Normalization or standardization was not required because of the categorical nature of the predictor variables. The maximum-likelihood estimator was used due to the large sample size. We performed the statistical analysis in R (version 4.3.2) (R Core Team, 2023) with RStudio as an Integrated Development Environment (RStudio Team, 2023) and used the package *nlme* for multilevel modeling (Pinheiro et al., 2011).

1.3.5.1 Step-by-Step Analysis with Hierarchical Linear Models

Analyzing HLMs involves several steps: assessing the need for an HLM, adding fixed effects, introducing random slopes, and adding interaction terms (Field et al., 2012; Hox et al., 2017; Raudenbush & Bryk, 2002). To evaluate the necessity of an HLM, we first fitted a baseline model that included only an intercept and subsequently allowed intercepts to vary across students, i.e., a random intercept model (Table 1.3.1: ID 1 and 2). We then compared the two model fits using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)—statistical metrics used to assess the quality of statistical models. Lower values for these criteria indicate better-fitting models, meaning these models capture the underlying data patterns more accurately while minimizing statistical error. We also performed a likelihood ratio test that compares the deviance—a measure of goodness of fit—of a simpler model to that of a more complex one based on a χ^2-distribution. In our analysis, the more complex model fitted the data better (Table 1.3.1: ID 2). The adjusted intraclass correlation (ICC) was 0.17 indicating that 17% of the variation in reasoning complexity is due to differences between students, which suggests that an HLM is appropriate in this context.

We hypothesized that students' reasoning complexity differed for mechanisms of varying plausibility, so we added the three response characteristics described above as fixed effects to the random-intercept model. Adding these characteristics as fixed effects significantly improved model fit (Table 1.3.1: IDs 3, 4, and 5).

Afterward, we tested whether adding a random slope further improved model fit. As typically done for longitudinal data, we allowed the effect of the reaction, i.e., the time point of data collection, to vary across students (Field et al., 2012). A comparison of AIC and BIC values demonstrated a better fit for the random-slopes model (Table 1.3.1: ID 6). For this reason, the complexity of students' mechanistic reasoning followed different progressions across the semester.

Subsequently, we added interaction terms between the response characteristics to the random-slopes model. Most importantly, the interaction between all three response characteristics was statistically significant (Table 1.3.1: ID 10), implying that students' reasoning complexity depended on the interplay of all response characteristics. We interpret this best-fitting final model (pseudo-R^2 = 0.18 according to Raudenbush and Bryk (2002)) in Section 1.3.6 using orthogonal contrasts (Table 1.3.2). For example, contrast C1 compares the Williamson ether synthesis with the combined effects of the other three mechanisms. Additionally, we report Cohen's f^2 as an effect size, with values greater than 0.02 indicating small effects, values greater than 0.15 indicating medium effects, and values greater than 0.35 indicating large effects (Cohen, 1988).

Random effects of the final model suggest that the relationship between response characteristics and reasoning complexity differed significantly in intercepts across students, $SD = 0.45$, 95% CI [0.38, 0.49], $\chi^2(1) = 1115.37$, $p < .001$. Similarly, slopes varied across students, $SD = 0.13$, 95% CI [0.10, 0.16], $\chi^2(9) = 154.68$, $p < .001$. Slopes and intercepts were negatively correlated, $r = –0.40$, 95% CI [–0.58, –0.18]. The responses' reasoning complexity showed high variability, $SD = 0.96$.

Table 1.3.1 Model fit statistics

ID	Model	*df*	AIC	BIC	Deviance	Test	*L.Ratio*	*p*-value
Assessing the Need for an HLM								
1	Intercept-Only	2	27973	27987	27969			
2	Random Intercept	3	26860	26881	26854	1 vs 2	1115	< .001
Adding Fixed Effects								
3	Context	6	26732	26774	26720	2 vs 3	134	< .001
4	Claim	7	26687	26737	26673	3 vs 4	47	< .001
5	Plausibility	8	26674	26731	26658	4 vs 5	15	.001
Introducing Random Slopes								
6	Random Slope	17	26538	26659	26504	5 vs 6	155	< .001
Adding Two-Way Interactions								
7	Context × Claim	20	26537	26680	26497	6 vs 7	6	.106
8	Context × Plausibility	23	26506	26670	26460	7 vs 8	37	< .001
9	Claim × Plausibility	24	26502	26673	26454	8 vs 9	6	.016
Adding Three-Way Interactions								
10	Context × Claim × Plausibility	27	26478	26671	26424	9 vs 10	30	< .001

Note: Models build upon each other; for instance, Model 4 includes all effects of Model 3 with *Claim* as an additional fixed effect. *df* = degrees of freedom, AIC = Akaike information criterion, BIC = Bayesian information criterion, *L.Ratio* = likelihood ratio

Table 1.3.2 Orthogonal contrasts used for dummy-coding the categorical variable *Context* to estimate and report the models

Variable	Contrast	Reference Group	Target Group
Context	C1	Ether Synthesis	Enamine Hydrolysis + Aldol Condensation + Claisen Condensation
Context	C2	Enamine Hydrolysis	Aldol Condensation + Claisen Condensation
Context	C3	Aldol Condensation	Claisen Condensation

1.3.5.1 Assumptions Check

We checked whether our model met the necessary assumptions (Hox et al., 2017). Therefore, we initially examined the assumptions related to the least-square residuals resulting from a linear regression separated by student. Visual inspection of a histogram confirmed the normal distribution of these residuals. However, a scatterplot revealed that homogeneity of variance was not achieved. Moreover, there were some outliers, which is expected given the 9,221 student responses. As these outliers were distributed evenly across students, they were not further investigated.

The residuals of the random intercepts and slopes followed a normal distribution as indicated by Shapiro-Wilk tests (p = .248 for random intercepts, p = .058 for random slopes). A few outliers were observed for random intercepts and slopes; however, these outliers only represented students with particularly high or low reasoning complexity.

We also conducted a Rainbow test ($F(4611, 4604) = 0.80$, p = .999), which confirms that the data can be modeled linearly due to the insignificant test statistic. Variance inflation factors ranged from 1.016 to 1.363, indicating the data is not multicollinear. Finally, we examined studentized residuals, Cook's distance, hat values, and the covariance ratio, and found no irregularities.

As not all assumptions were met, we also performed a robust HLM using the *robustlmm* package (Koller, 2016). The results of the robust HLM were consistent with those of the traditional HLM; hence, further details are not discussed.

1.3.6 Results

Fig. 1.3.6 presents the distribution of students' responses across combinations of the *levels of granularity* and *modes of reasoning*. This distribution reveals that students within the same cohort exhibited diverse mechanistic reasoning patterns with varying complexity, highlighting substantial heterogeneity in how students reason about mechanisms. Moreover, the applied *levels of granularity* and *modes of reasoning* differed across contexts, implying that students may face context-dependent challenges in demonstrating complex mechanistic reasoning. Table 1.3.3 presents examples of students' mechanistic reasoning about the plausibility of product 1.4 in the ether synthesis (Fig. 1.3.4) categorized by the six combinations of the *levels of granularity* and *modes of reasoning*.

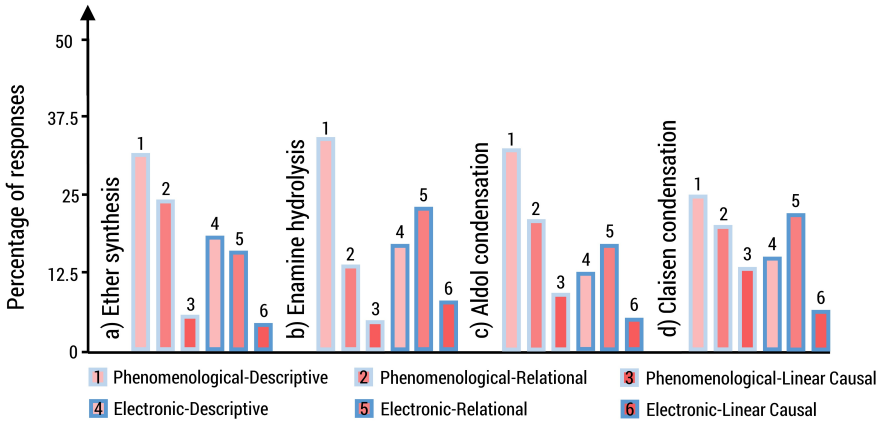


Fig. 1.3.6 Distribution of students' mechanistic reasoning across combinations of the *levels of granularity* and *modes of reasoning* separated by mechanism. *Note:* Responses coded as *absent* are excluded from this figure

To investigate how students' reasoning complexity differed among mechanisms of varying plausibility, we interpreted the best-fitting HLM (Table 1.3.1: ID 10). This model includes a significant three-way interaction, implying that a multifaceted interplay of response characteristics impacts students' reasoning complexity (Table 1.3.4). Due to the significant three-way interaction, main effects and lower-level interactions should be interpreted cautiously.

Table 1.3.3 Examples of students' mechanistic reasoning on the plausibility of product 1.4 in the ether synthesis by *levels of granularity* and *modes of reasoning*

Level of Granularity	Mode of Reasoning	Student Example
Phenomenological	Descriptive	The hydroxide ion deprotonates the alcohol. The acid-base reaction will proceed due to the strong base.
Phenomenological	Relational	The hydroxide ion can remove the acidic proton of the alcohol. The pK_a values of H_2O and a primary alcohol are about the same, so the hydroxyl may be deprotonated.
Phenomenological	Linear Causal	The hydroxide ion deprotonates the alcohol. The conjugate acid of the $^-$OH has a pK_a of 15.7 and the alcohol has a pK_a of 16. The $^-$OH will act as a base. Protons can be exchanged in equilibrium since the pK_a-values are so close.

Table 1.3.3 (continued)

Level of Granularity	Mode of Reasoning	Student Example
Electronic	Descriptive	The ⁻OH is likely to deprotonate the alcohol to become more stable. An ⁻OH group is not stable due to the negative charge.
Electronic	Relational	The oxygen of the alcohol becomes negatively charged after the ⁻OH attacks (…). The negatively charged oxygen is electron-rich due to electron-donating effects, so it attacks the carbon. This causes the chlorine to leave, resulting in a more stable structure based on molecular principles.
Electronic	Linear Causal	This reaction starts with the deprotonation of the hydroxyl group of the primary alcohol. The hydroxyl group is the most acidic site. Its deprotonation leads to a localized negative charge on the oxygen due to the high electronegativity of the oxygen and the electron-donating effect of the alkyl group. Finally, intramolecular substitution results in the ejection of the chloride ion and the formation of a low-strain ring structure.

First, the context, i.e., the reaction that a response addressed, significantly predicted students' reasoning complexity (Table 1.3.4: ID 10.2). Students' mechanistic reasoning was less complex in the ether synthesis compared to the combination of all other reactions (Table 1.3.4: ID 10.2a), comparable in the enamine hydrolysis and condensation reactions (Table 1.3.4: ID 10.2b), and more complex in Claisen condensation than in the aldol condensation (Table 1.3.4: ID 10.2c). These differences show that students' reasoning complexity differed across the four reactions (Fig. 1.3.7a). Furthermore, students' individual claim regarding the plausibility of a mechanism also predicted their reasoning complexity (Table 1.3.4: ID 10.3). Here, the reasoning complexity was higher when students argued for the plausibility of a mechanism (Fig. 1.3.7b). Conversely, the actual plausibility of a mechanism did not significantly influence students' reasoning complexity (Table 1.3.4: ID 10.4; Fig. 1.3.7c).

Table 1.3.4 Parameters of the best-fitting final model (Table 1.3.1: ID 10)

				95% CI				
ID	Fixed Effects	Contrast	*b*	*LL*	*UL*	*t*-value	*p*-value	Cohen's f^2
10.1	Intercept	-	2.883	2.817	2.949	85.417	< .001	0.816
10.2a	Context	C1	0.092	0.066	0.118	6.944	< .001	0.005
10.2b	Context	C2	–0.025	–0.059	0.010	–1.570	.117	< 0.001
10.2c	Context	C3	0.154	0.105	0.203	6.155	< .001	0.004
10.3	Claim	-	0.121	0.043	0.199	3.045	.002	0.001
10.4	Plausibility	-	0.016	–0.101	0.133	0.268	.789	< 0.001
10.5a	Context × Claim	C1	–0.086	–0.124	–0.049	–4.507	< .001	0.002
10.5b	Context × Claim	C2	–0.023	–0.086	0.039	–0.733	.464	< 0.001
10.5c	Context × Claim	C3	0.043	–0.081	0.167	0.683	.494	< 0.001
10.6a	Context × Plausibility	C1	–0.155	–0.198	–0.111	–6.987	< .001	0.006
10.6b	Context × Plausibility	C2	–0.038	–0.072	–0.003	–2.122	.034	0.001
10.6c	Context × Plausibility	C3	0.107	–0.039	0.253	1.431	.152	< 0.001
10.7	Claim × Plausibility	-	0.248	0.127	0.368	4.032	< .001	0.002
10.8a	Context × Claim × Plausibility	C1	0.155	0.098	0.211	5.348	< .001	0.003
10.8b	Context × Claim × Plausibility	C2	0.031	–0.064	0.126	0.642	0.521	< 0.001
10.8c	Context × Claim × Plausibility	C3	–0.138	–0.331	0.056	–1.395	0.163	< 0.001

Note: The definitions of contrasts C1-C3 are provided in Table 1.3.2. *Degrees of freedom* = 8940, *b* = regression coefficient, *LL* = lower limit, *UL* = upper limit

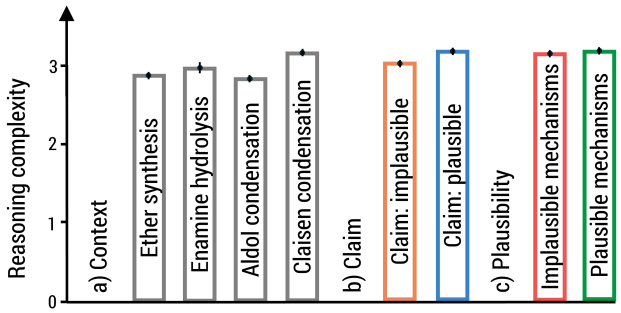


Fig. 1.3.7 Error bar charts visualizing how the context, students' claims, and a mechanism's actual plausibility impact students' reasoning complexity

Moreover, we investigated the two-way interactions, which assess the combined effects of two of the three response characteristics on students' reasoning complexity. Here, the context and students' individual claims are related (Table 1.3.4: ID 10.5), implying that the increased reasoning complexity when arguing *for* the plausibility of a mechanism depended significantly on the reaction. Precisely, students' reasoning complexity was higher—in part insignificantly—when arguing *for*, rather than *against* a mechanism's plausibility for enamine hydrolysis, aldol condensation, and Claisen condensation; this difference was even more pronounced for Williamson ether synthesis (Fig. 1.3.8a). Orthogonal contrasts confirm that variations in students' reasoning complexity considering their claims were more pronounced for Williamson ether synthesis compared to the other reactions (Table 1.3.4: ID 10.5a), whereas variations were similar among enamine hydrolysis, aldol condensation, and Claisen condensation (Table 1.3.4: IDs 10.5b and 10.5c). Similarly, the context and a mechanism's actual plausibility are related (Table 1.3.4: ID 10.6), indicating that students' reasoning complexity for plausible and implausible mechanisms differed significantly across reactions. Students' reasoning complexity was comparable for plausible and implausible mechanisms in the aldol condensation and Claisen condensation, higher for plausible mechanisms in Williamson ether synthesis, but significantly lower for plausible mechanisms in enamine hydrolysis (Fig. 1.3.8b). Orthogonal contrasts confirm that variations in students' reasoning complexity regarding a mechanism's actual plausibility were more pronounced for the ether synthesis compared to the combined effect of the other mechanisms (Table 1.3.4: ID 10.6a). Furthermore, students demonstrated significantly more complex mechanistic reasoning about implausible mechanisms in enamine hydrolysis compared to the condensation reactions (Table 1.3.4: IDs 10.6b); variations in reasoning complexity between plausible and implausible mechanisms were comparable for the aldol and Claisen condensations (Table 1.3.4: ID 10.6c). Beyond that, a student's claims and a mechanism's actual plausibility are also related (Table 1.3.4: ID 10.7). This interaction shows that the reasoning complexity increased when students' individual claims about a mechanism's plausibility aligned with its actual plausibility, which indicates that students' mechanistic reasoning was more complex when their claim was correct (Fig. 1.3.8c).

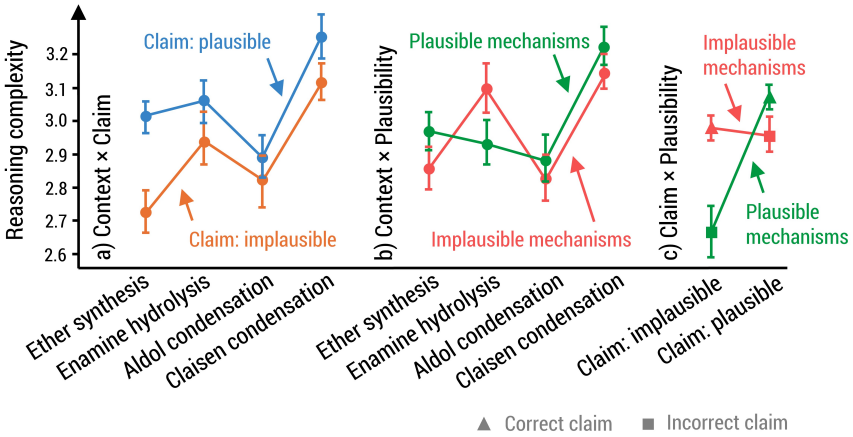


Fig. 1.3.8 Line graphs visualizing how two of three response characteristics impact students' reasoning complexity. *Note:* Blue indicates that students claim a mechanism's plausibility and orange that they claim a mechanism's implausibility; green symbolizes plausible mechanisms and red implausible ones; triangles visualize correct claims and squares incorrect claims

Most importantly, the interaction between all three response characteristics is also statistically significant (Table 1.3.4: ID 10.8). This three-way interaction reveals that a correct claim did not equally increase students' reasoning complexity across all reactions. For example, students' mechanistic reasoning was more complex with correct claims in enamine hydrolysis, aldol condensation, and Claisen condensation (Fig. 1.3.9b, 1.3.9c, and 1.3.9d). By contrast, students' reasoning was more complex when they argued *for* the plausibility of a mechanism in the ether synthesis, irrespective of a mechanism's actual plausibility (Fig. 1.3.9a). Orthogonal contrasts confirm that the trend of students having more complex reasoning when their claims were correct does not hold for Williamson ether synthesis (Table 1.3.4: IDs 10.8a) but applies to the other three mechanisms (Table 1.3.4: IDs 10.8b and 10.8c). Overall, this result suggests that choosing the correct claim is typically the key response characteristic causing complex mechanistic reasoning in most mechanisms—except for the ether synthesis.

1.3.7 Discussion

Our study revealed how the complexity of students' mechanistic reasoning classified by the first and third essential epistemic heuristics varies across different response characteristics when evaluating a mechanism's plausibility. The analysis showed that students' reasoning complexity ranged roughly from 2.25 to 3.25, indicating that the responses examined are mostly constructed at a *descriptive* (44%)

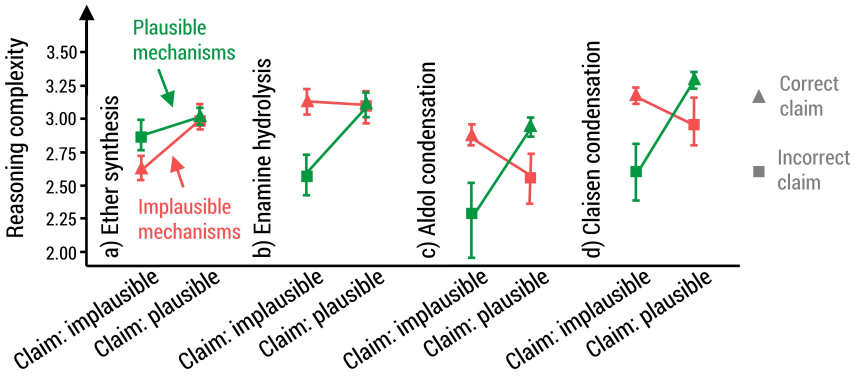


Fig. 1.3.9 Line graphs visualizing how the three response characteristics impact students' reasoning complexity. *Note:* Green symbolizes plausible mechanisms and red implausible ones; triangles visualize correct claims and squares incorrect claims

or *relational* (35%) *mode of reasoning* and a *phenomenological* (56%) *level of granularity* (Fig. 1.3.6). This finding aligns with our prior analyses classifying students' reasoning on alternative mechanisms in different settings (Martin et al., 2024). The lack of causal connections in students' mechanistic reasoning is a common observation across disciplines (e.g., Flowers et al., 2023). Nevertheless, students' reasoning complexity differed in our setting significantly across the three examined response characteristics, which implies that certain characteristics promote complex reasoning about a mechanism's plausibility.

First, our findings demonstrate that students' mechanistic reasoning is more complex when arguing *for*, rather than *against* a mechanism's plausibility (Fig. 1.3.7b). This finding can be rationalized with common instructional practices in organic chemistry that emphasize only the most plausible mechanism as determined by the lowest energy path (Caspari et al., 2018; Goodwin, 2012). As a result, students are more often prompted to argue *for*, rather than *against* a given outcome, so they tend to concentrate only on the most plausible products, neglecting potential byproducts (Lieber & Graulich, 2022; Lieber et al., 2022b; Popova & Bretz, 2018). This common practice can lead to a lower reasoning complexity when arguing *against* a mechanism's plausibility. The observed difference in reasoning complexity was even more pronounced in the Williamson ether synthesis (Fig. 1.3.8a). This reaction was the first context in our study where students encountered implausible mechanisms, so one can assume that students were less familiar with justifying a mechanism's implausibility, which decreased students' reasoning complexity. Continually practicing to argue not only *for* but also *against* potential mechanisms may help students recognize that chemical reactions can yield multiple products of different plausibility, unpack the key factors influencing a mechanism's likelihood, and link the emerging properties to support their claims. Integrating implausible alternatives and contrary viewpoints more explicitly into instruction may, thus, support students' mechanistic reasoning (Larrain et al., 2019).

While students' claims impacted their reasoning complexity, the actual plausibility of a mechanism did not significantly influence their reasoning complexity (Fig. 1.3.7c). This is not surprising as a mechanism's actual plausibility was not accessible to the students in our setting since they did not receive feedback on the correctness of their claims. Surprisingly, students' mechanistic reasoning was more complex for implausible mechanisms in enamine hydrolysis while this trend was reversed for the other three mechanisms (Fig. 1.3.8b). The implausible mechanisms in enamine hydrolysis arose from an acid-base reaction and an electrophilic addition in the anti-Markovnikov position, i.e., a typically used rule for asymmetric alkene addition, which were probably clearly implausible to the students. Future analyses need to investigate this difference for enamine hydrolysis in detail.

As further shown, students' reasoning complexity was most importantly influenced by the correctness of their claims, with correct claims leading to more complex mechanistic reasoning (Fig. 1.3.8c). This finding corresponds to other qualitative and quantitative analyses: For example, Cruz-Ramírez de Arellano and Towns (2014) found that students with correct claims about alkyl halide reactions could more thoroughly weigh basic and nucleophilic properties of substances, accurately argue about intermediates and transition states, and build chemically sound reasoning statements to connect entities, activities, and their properties. Also, students' *modes of reasoning* were significantly higher for students with correct claims than for those with incorrect claims when judging which of three bases can best force the acid-base equilibrium to the product side (Deng & Flynn, 2021). By contrast, a similar analysis showed that the proportion of the *modes of reasoning* did not vary significantly between responses with correct and incorrect claims when contrasting the plausibility of two unimolecular nucleophilic substitutions (Bodé et al., 2019). Our analysis builds on this work by demonstrating that a correct claim may lead to a higher reasoning complexity (Fig. 1.3.8c); however, this trend does not apply to the Williamson ether synthesis (Fig. 1.3.9). The reactions evaluated by Bodé et al. (2019) and the Williamson ether synthesis analyzed herein both involve nucleophilic substitution (S_N) mechanisms, so unsurprisingly, these reactions deviate from the observed trend. Students' familiarity with the S_N mechanisms—which are introduced early in undergraduate organic chemistry and follow a clear reaction pattern—may reduce the role of claim-making in students' decision-making. Instead of engaging in mechanistic reasoning, students likely relied on their existing knowledge and recalled the S_N portions of the reactions (Grove et al., 2012). This may have allowed the students to arrive at conclusions more confidently without a time-intensive, complex decision-making process. Nonetheless, the significant impact of making a correct claim on students' reasoning complexity could encourage instruction to guide the process of constructing and justifying claims more consciously (McNeill & Pimentel, 2010).

Ultimately, students' reasoning complexity differed significantly across mechanisms and tended to increase across the semester (Fig. 1.3.7a). This increase was expected due to continuous adaptive guidance in our setting that helped students use chemical concepts to reason causally. A previous comparable adaptive training also provided students effective support in building scientific arguments and applying chemical concept knowledge (Lieber et al., 2022a, 2022b), so it is

unsurprising that students' reasoning complexity tends to increase over time. However, aldol condensation seems particularly challenging for students due to the complex interplay of acid-base equilibria, carbonyl chemistry, and elimination reactions (Watts et al., 2022). This context-dependency of students' mechanistic reasoning transcends disciplinary boundaries as exemplified by an analysis of students' mechanistic reasoning in molecular and cellular biology (Southard et al., 2017).

1.3.8 Conclusions and Implications

To investigate students' reasoning complexity when judging the plausibility of alternative mechanisms, we operationalized the *essential epistemic heuristics for mechanistic reasoning* (Krist et al., 2019). Specifically, we compared the complexity of students' mechanistic reasoning related to granularity and causality—reflecting the first and third essential epistemic heuristics—to identify the response characteristics that support complex mechanistic reasoning about alternative mechanisms. Our findings hold various implications for guiding undergraduate students' mechanistic reasoning in science classes.

First, our analysis revealed that students are better at arguing *for* than *against* the plausibility of a mechanism. To support students' mechanistic reasoning in science classes, instruction can progressively increase the difficulty of exercises by initially presenting students with plausible mechanisms and gradually incorporating implausible ones throughout the semester. In this context, discussions about mechanisms of different plausibility should occur explicitly. For instance, when discussing competitive mechanisms such as substitution and elimination reactions, instruction should highlight that there can be multiple, sometimes conflicting interpretations of a mechanism's outcome, rather than viewing this outcome as strictly right or wrong with little room for ambiguity (Grove & Bretz, 2010). Advancing students' ability to critically evaluate conflicting information, for example, mechanistic hypotheses, may finally enrich the complexity of students' reasoning in everyday life.

Furthermore, students demonstrated greater complexity in reasoning across scalar levels and linking properties at lower scalar levels if their claims were correct. This finding suggests that correct claim-making is crucial for the ability to transition between scalar levels and establish causal links, which means correct claim-making and mechanistic reasoning benefit each other (e.g., Choi et al., 2013; Grove et al., 2012; Lieber et al., 2022a; Walker et al., 2019). Consequently, instruction should explicitly support students' claim-making by introducing plausible *and* implausible mechanisms.

In sum, our findings may inform the development of a step-by-step instructional sequence that guides students in reasoning about multiple mechanistic alternatives. From a broader perspective, encouraging students to consider alternative hypotheses or explanations can help them evaluate evidence, weigh different

possibilities, and make informed judgments about scientific phenomena (e.g., Driver et al., 2000). Such guidance can support students in looking beyond a simplistic view of mechanisms as a linear sequence of events; instead, it can encourage them to reason about the dynamic interplay of entities and activities by weighing various alternatives.

1.3.9 Limitation

To identify response characteristics that support complex mechanistic reasoning, we examined how the complexity of students' mechanistic reasoning differed across mechanisms of varying plausibility. Our analysis focused on the response level. Further insights into students' mechanistic reasoning could be provided by centering our analysis at the student level. Examining each student, rather than every response, may provide a more holistic view of students' reasoning complexity. This allows, on the one hand, to model students' mechanistic reasoning paths across the four reactions. On the other hand, we may identify the cognitive and affective variables that favor complex mechanistic reasoning. Accordingly, future analysis could adopt a student-centered perspective to uncover productive mechanistic reasoning paths and explore their causes.

1.3.10 References

Bhattacharyya, G., & Bodner, G. M. (2005). "It Gets Me to the Product": How Students Propose Organic Mechanisms. *Journal of Chemical Education*, *82*(9), 1402-1407.

Bodé, N. E., Deng, J. M., & Flynn, A. B. (2019). Getting Past the Rules and to the WHY: Causal Mechanistic Arguments When Judging the Plausibility of Organic Reaction Mechanisms. *Journal of Chemical Education*, *96*(6), 1068-1082.

Caspari, I., Weinrich, M., Sevian, H., & Graulich, N. (2018). This mechanistic step is "*productive*": organic chemistry students' backward-oriented reasoning. *Chemistry Education Research and Practice*, *19*(1), 42-59.

Choi, A., Hand, B., & Greenbowe, T. (2013). Students' Written Arguments in General Chemistry Laboratory Investigations. *Research in Science Education*, *43*(5), 1763-1783.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Vol. 2). Hillsdale, NJ: Erlbaum.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience* (Vol. 1). Oxford: Clarendon Press.

Cruz-Ramírez de Arellano, D., & Towns, M. H. (2014). Students' understanding of alkyl halide reactions in undergraduate organic chemistry. *Chemistry Education Research and Practice*, *15*(4), 501-515.

Darden, L., & Craver, C. (2002). Strategies in the interfield discovery of the mechanism of protein synthesis. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *33*(1), 1-28.

Deng, J. M., & Flynn, A. B. (2021). Reasoning, granularity, and comparisons in students' arguments on two organic chemistry items. *Chemistry Education Research and Practice*, *22*(3), 749-771.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, *84*(3), 287-312.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R* (Vol. 1). Los Angeles: Sage.

Flowers, S., Holder, K. H., & Gardner, S. M. (2023). Here Is the Biology, Now What is the Mechanism? Investigating Biology Undergraduates' Mechanistic Reasoning within the Context of Biofilm Development. *Journal of Microbiology & Biology Education*, *24*(2), e00221-00222.

Goodwin W. M., (2003), Explanation in Organic Chemistry, *Annals of the New York Academy of Sciences*, *988*(1), 141–153.

Goodwin, W. M. (2012). Mechanisms and chemical reaction. In A. I. Woody, R. F. Hendry, & P. Needham (Eds.), *Philosophy of chemistry* (Vol. 6, pp. 309-327). Amsterdam: North-Holland.

Grove, N. P., & Bretz, S. L. (2010). Perry's Scheme of Intellectual and Epistemological Development as a framework for describing student difficulties in learning organic chemistry. *Chemistry Education Research and Practice*, *11*(3), 207-211.

Grove, N. P., Cooper, M. M., & Cox, E. L. (2012). Does Mechanistic Thinking Improve Student Success in Organic Chemistry? *Journal of Chemical Education*, *89*(7), 850-853.

Hox, J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel Analysis: Techniques and Applications* (Vol. 3). New York: Routledge.

Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*(1), 119-135.

Koller, M. (2016). *robustlmm: Robust Linear Mixed Effects Models* [Computer Program].

Krist, C., Schwarz, C. V., & Reiser, B. J. (2019). Identifying Essential Epistemic Heuristics for Guiding Mechanistic Reasoning in Science Learning. *Journal of the Learning Sciences*, *28*(2), 160-205.

Larrain, A., Freire, P., López, P., & Grau, V. (2019). Counter-Arguing During Curriculum-Supported Peer Interaction Facilitates Middle-School Students' Science Content Knowledge. *Cognition and Instruction*, *37*(4), 453-482.

Lieber, L. S., & Graulich, N. (2022). Investigating students' argumentation when judging the plausibility of alternative reaction pathways in organic chemistry. *Chemistry Education Research and Practice*, *23*(1), 38-53.

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022a). Closing the gap of organic chemistry students' performance with an adaptive scaffold for argumentation patterns. *Chemistry Education Research and Practice*, *23*(4), 811-828.

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022b). Students' Individual Needs Matter: A Training to Adaptively Address Students' Argumentation Skills in Organic Chemistry. *Journal of Chemical Education*, *99*(7), 2754-2761.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking About Mechanisms. *Philosophy of Science*, *67*(1), 1-25.

Martin, P. P., & Graulich, N. (2023). When a machine detects student reasoning: a review of machine learning-based formative assessment of mechanistic reasoning. *Chemistry Education Research and Practice*, *24*(2), 407-427.

Martin, P. P., & Graulich, N. (2024). Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with

generative artificial intelligence. *Computers and Education: Artificial Intelligence*, *7*, Article 100265.

Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2024). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*, *61*(8), 1757-1792.

McNeill, K. L., & Krajcik, J. (2011). *Supporting Grade 5-8 Students in Constructing Explanations in Science: The Claim, Evidence, and Reasoning Framework for Talk and Writing* (Vol. 1). London: Pearson Allyn & Bacon.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting Students' Construction of Scientific Explanations by Fading Scaffolds in Instructional Materials. *Journal of the Learning Sciences*, *15*(2), 153-191.

McNeill, K. L., & Pimentel, D. S. (2010). Scientific discourse in three urban classrooms: The role of the teacher in engaging high school students in argumentation. *Science Education*, *94*(2), 203-229.

Pinheiro, J., Bates, D., & R Core Team. (2011). *nlme: Linear and Nonlinear Mixed Effects Models* [Computer Program].

Popova, M., & Bretz, S. L. (2018). "It's Only the Major Product That We Care About in Organic Chemistry": An Analysis of Students' Annotations of Reaction Coordinate Diagrams. *Journal of Chemical Education*, *95*(7), 1086-1093.

R Core Team (2023). *R: A Language and Environment for Statistical Computing* [Computer Program].

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (Vol. 1). London: Sage.

RStudio Team (2023). *RStudio: Integrated Development Environment for R* [Computer Program].

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing Mechanistic Reasoning in Student Scientific Inquiry: A Framework for Discourse Analysis Developed From Philosophy of Science. *Science Education*, *92*(3), 499-525.

Sevian, H., & Talanquer, V. (2014). Rethinking chemistry: a learning progression on chemical thinking. *Chemistry Education Research and Practice*, *15*(1), 10-23.

Southard, K. M., Espindola, M. R., Zaepfel, S. D., & Bolger, M. S. (2017). Generative mechanistic explanation building in undergraduate molecular and cellular biology. *International Journal of Science Education*, *39*(13), 1795-1829.

Talanquer, V. (2013). When Atoms Want. *Journal of Chemical Education*, *90*(11), 1419-1424.

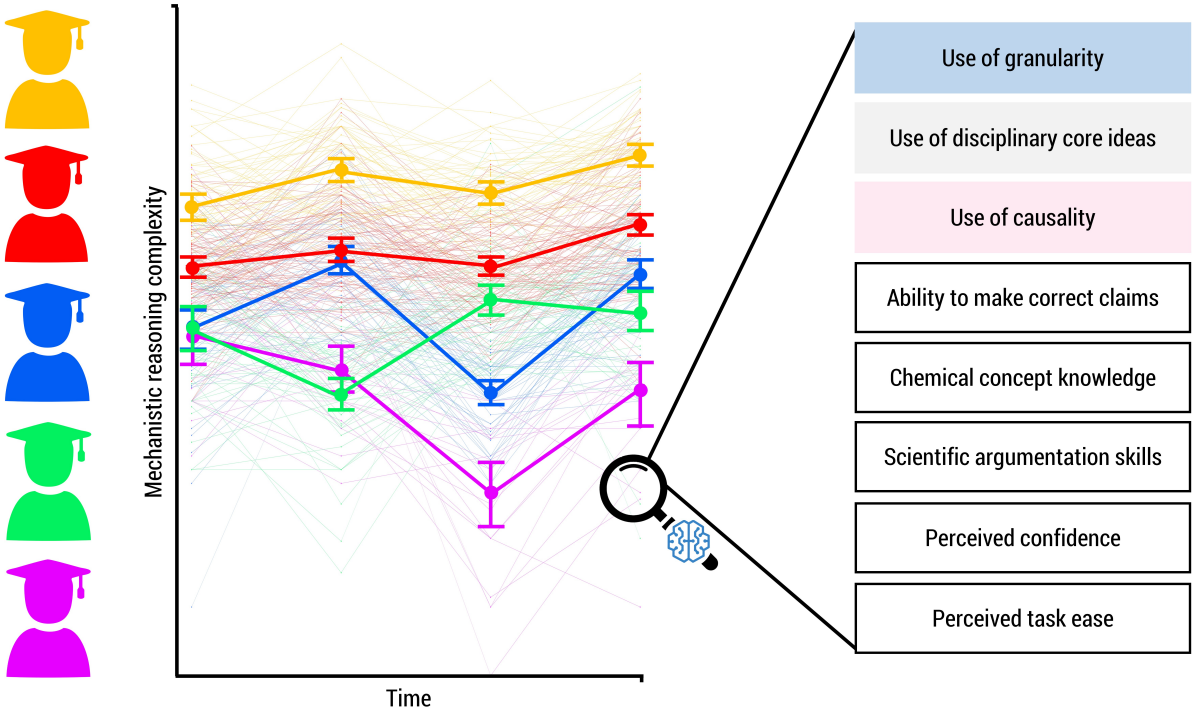
Walker, J. P., Van Duzor, A. G., & Lower, M. A. (2019). Facilitating Argumentation in the Laboratory: The Challenges of Claim Change and Justification by Theory. *Journal of Chemical Education*, *96*(3), 435-444.

Watts, F. M., Park, G. Y., Petterson, M. N., & Shultz, G. V. (2022). Considering alternative reaction mechanisms: Students' use of multiple representations to reason about mechanisms for a writing-to-learn assignment. *Chemistry Education Research and Practice*, *23*(2), 486-507.

Weinrich, M. L., & Sevian, H. (2017). Capturing students' abstraction while solving organic reaction mechanism problems across a semester. *Chemistry Education Research and Practice*, *18*(1), 169-190.

9 | Tracing trajectories: Using machine learning to map how chemistry students' mechanistic reasoning develops over time

Chapter 9 presents a research manuscript by Martin, Kubsch, Yik, Burlingham, and Graulich that is currently in preparation.



RESEARCH ARTICLE

Tracing trajectories: Using machine learning to map how chemistry students' mechanistic reasoning develops over time

Paul P. Martin[1], Marcus Kubsch[2], Brandon J. Yik[3], Benjamin T. Burlingham[4], Nicole Graulich[1,*]

Abstract

Mechanistic reasoning—identifying, describing, or analyzing the underlying mechanisms driving natural phenomena—is fundamental across science domains. However, students often face challenges with characterizing unobservable entities and their activities, which contributes to context-dependent variations in how students demonstrate mechanistic reasoning within and across cohorts. Prior research offered rather fragmented insights into the longitudinal development of students' mechanistic reasoning, providing limited guidance for supporting mechanistic reasoning over time. Meanwhile, machine learning (ML) models can automatically analyze students' mechanistic reasoning, enabling a more detailed analysis. This quantitative study leverages an ML model to automatically assess and monitor the development of undergraduate students' mechanistic reasoning throughout an organic chemistry course. This approach allowed us to map the trajectories of students' mechanistic reasoning and identify key factors influencing students' success. Our results show that students' mechanistic reasoning is context-dependent but tends to improve over time with adaptive support. The ability to apply disciplinary core ideas related to energetics emerged as a key predictor of more complex mechanistic reasoning, along with high chemistry-specific concept knowledge and confidence in solving exercises. Our findings inform evidence-based approaches to supporting students' mechanistic reasoning in undergraduate science courses over time.

KEYWORDS mechanistic reasoning, longitudinal assessment, learning progression, organic chemistry, machine learning

Affiliation

[1]Institute of Chemistry Education, Justus-Liebig-University, Heinrich-Buff-Ring 17, 35392 Giessen, Germany
[2]Department of Physics, Freie Universität Berlin, Arnimallee 14, 14195 Berlin, Germany
[3]Department of Chemistry, University of Georgia, 302 East Campus Road, Athens, GA 30602, United States
[4]Department of Chemistry, Indiana University, 800 E. Kirkwood Ave., Bloomington, IN 47405-7102, United States
*E-Mail corresponding author: Nicole.Graulich@didaktik.chemie.uni-giessen.de

1 | Introduction

Science practices, such as constructing explanations, building arguments, and developing models, enable students to analyze the natural and engineered world (NRC, 2012; OECD, 2023). Several of these science practices require mechanistic reasoning, which includes identifying, describing, or analyzing the underlying and often unobservable mechanisms of natural phenomena (Krist et al., 2019; Russ et al., 2008). Mechanistic reasoning is essential throughout science domains as it allows students to understand how and why phenomena occur (Glennan & Illari, 2017; Illari & Williamson, 2012; Machamer et al., 2000). Explaining, predicting, or controlling scientific phenomena is challenging without mechanistic reasoning since mechanistic reasoning decodes the sequence and causality of underlying mechanisms (Haskel-Ittah, 2023). Thus, developing mechanistic reasoning is crucial for students (e.g., Shtechman et al., 2024; Talanquer, 2018a, 2018b; van Mil et al., 2016).

Mechanistic reasoning is a complex practice (Craver, 2007; Illari & Williamson, 2012). Students need to describe the target phenomenon, identify the entities and their properties that contribute to that phenomenon, and recognize the occurring activities (Russ et al., 2008). Given these complexities, students frequently encounter challenges in mechanistic reasoning across science domains, e.g., when students assign properties to entities, explain their multifaceted interactions, or determine the appropriate number of entities required for a causal explanation (Bachtiar et al., 2022). For instance, students' description of entities is often based on surface-level properties, rather than on an analysis of implicit properties and interactions (e.g., Bhattacharyya & Bodner, 2005; Ferguson & Bodner, 2008; Graulich et al., 2019). Consequently, students often view mechanisms as a linear sequence of unrelated steps and neglect the multifaceted interactions of the entities in dynamic activities (e.g., Kraft et al., 2010; Weinrich & Sevian, 2017). Yan and Talanquer (2015) also found that students face challenges in constructing causal explanations, where activities result from dynamic random interactions among entities. Due to these challenges, a recent review of mechanistic reasoning in science education emphasizes the need to explicitly support students' mechanistic reasoning to help them explain how and why phenomena occur (Bachtiar et al., 2022).

However, students' mechanistic reasoning is highly context-dependent and varies widely in complexity at the same educational level, showcasing considerable heterogeneity within and between students (e.g., Braun & Graulich, 2024; Crandell et al., 2019; Deng & Flynn, 2021; Moreira et al., 2019; Southard et al., 2017, Weinrich & Talanquer, 2016). Prior research, for example, in chemistry education, examined this heterogeneity in students' mechanistic reasoning but offered limited insights into how mechanistic reasoning develops over time due to several constraints: Some studies assessed students' mechanistic reasoning longitudinally across multiple semesters but only in the context of specific topics, such as acid–base reactions (Crandell et al., 2019), intermolecular forces (Noyes & Cooper, 2019), or nucleophilic substitutions (Crandell et al., 2020). Other studies were restricted by a small number of assessments (Lieber et al., 2022a), small sample sizes (Watts et al., 2021; Weinrich & Sevian, 2017), or cross-sectional approaches (Franovic et al., 2023; Shiroda et al., 2024).

Collectively, prior research has provided valuable but fragmented insights into the longitudinal development of students' mechanistic reasoning. As a result, our understanding of the development of students' mechanistic reasoning and its interactions with other variables remains limited. Given the central role of mechanistic reasoning in understanding mechanisms (e.g., Grove et al., 2012), a longitudinal study of students' mechanistic reasoning is essential because it can offer guidance for supporting mechanistic reasoning over time. In this work,

we started to address this gap by analyzing how students' mechanistic reasoning develops longitudinally across contexts in undergraduate organic chemistry to examine the dynamic interplay of multiple components of students' mechanistic reasoning over time.

2 | CONCEPTUAL BACKGROUND: MECHANISTIC REASONING IN CHEMISTRY

Modeling the development of students' mechanistic reasoning necessitates hypotheses about how their mechanistic reasoning may progress over time and how it is influenced by other learning-related variables. Thus, we propose a model that offers a holistic view of mechanistic reasoning and its interdependencies with other variables to examine its development in undergraduate organic chemistry (Figure 1). Generally, the proposed model considers domain-specific phenomena as starting points for students' mechanistic reasoning (Russ et al., 2008). These phenomena are driven by underlying mechanisms that provide a phenomenon's sequence of steps (Glennan & Illari, 2017). Mechanistic reasoning enables students to identify, describe, or analyze the underlying mechanisms to generate claims that explain, predict, or control the target phenomenon. We build upon the *essential epistemic heuristics for mechanistic reasoning* to characterize students' mechanistic reasoning given their broad applicability across science domains (Krist et al., 2019).

The *essential epistemic heuristics for mechanistic reasoning* provide a structured approach to categorize the steps students may follow in their

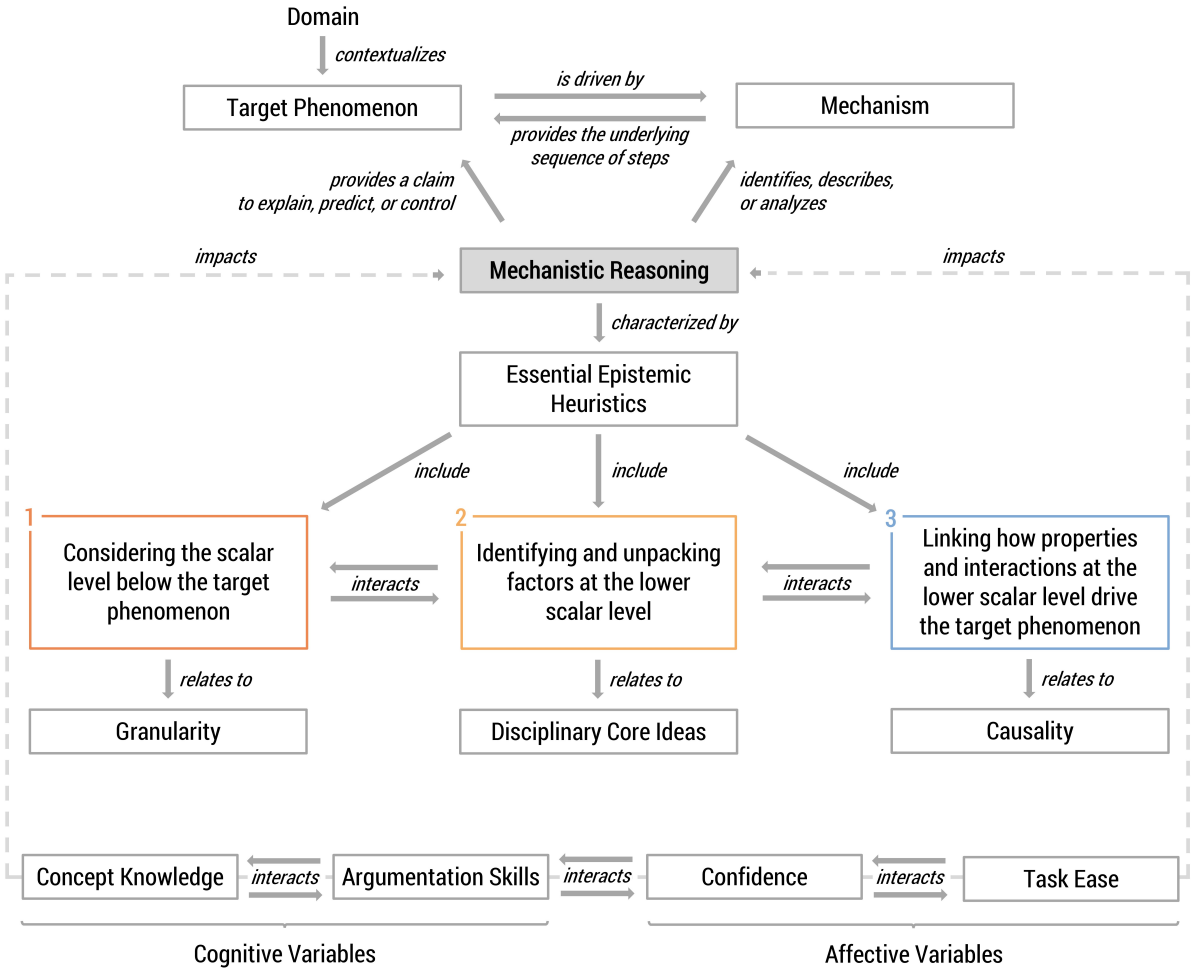


FIGURE 1 Model detailing a holistic view of mechanistic reasoning, its interdependencies, and operationalization.

mechanistic reasoning (Krist et al., 2019). The first *essential epistemic heuristic* emphasizes reasoning across scalar levels, particularly analyzing entities, activities, and their properties at a scalar level below the target phenomenon (Krist et al., 2019). The specific scalar level below the target phenomenon varies across science domains, so the concept of granularity can be used to identify the scalar levels at which a phenomenon is analyzed (Darden, 2002). At the lower scalar level, students must identify and unpack the entities, activities, and their related properties—the second *essential epistemic heuristic* (Krist et al., 2019). Identifying and unpacking these entities, activities, and properties requires a solid understanding of disciplinary core ideas since the target phenomenon is always contextualized within a specific domain (Russ et al., 2008). Finally, the unpacked properties must be arranged to establish a causal sequence of events over time or space. For this reason, the third *essential epistemic heuristic* describes to what extent the entities, activities, and properties of the proposed mechanism explain the observed phenomenon causally, which involves linking the identified entities and activities back to the original scalar level and assessing whether the unpacked properties provide a causal explanation (Krist et al., 2019).

Our model incorporates the influence of cognitive variables, such as students' chemical concept knowledge and domain-general argumentation skills on tasks that do not necessarily prompt mechanistic reasoning, and affective variables, such as their perceived confidence and task ease. These variables may be crucial for developing mechanistic reasoning. The next sections unpack the components of our model and situate them within the chemistry-specific context of this study.

2.1 First *essential epistemic heuristic*: Reasoning across scalar levels

Reasoning at a scalar level below the target phenomenon allows students to identify the underlying mechanism (Krist et al., 2019). The specific scalar level below the target phenomenon varies depending on the domain, context, and task, and the idea of granularity can be used to assess the scalar levels addressed in students' mechanistic reasoning (Darden, 2002; Deng & Flynn, 2021; Southard et al., 2017; van Mil et al., 2013). For instance, students must navigate between the scalar levels of reactions, molecules, atoms, and electrons in chemistry contexts. In molecular biology contexts, students should consider biological systems, chemical reactions, and molecules. Other phenomena may necessitate macroscopic perspectives, shifting between global and organismal levels, while some may require microscopic perspectives that detail subatomic structures (Darden, 2002).

In chemical contexts, variations in students' use of granularity—the scalar levels at which students analyze a phenomenon—are apparent even within the same task (Deng & Flynn, 2021). For example, most students addressed an acid–base reaction on the scalar levels of reactions and molecules, incorporating measurable properties such as equilibrium direction, pK_a values, and conjugate acid strength (Deng & Flynn, 2021). Reactions and molecules often represent the target phenomenon in chemistry, so this type of mechanistic reasoning can be considered as having a *phenomenological* granularity. Conversely, other students included *electronic* properties like electronegativity or electron delocalization in their mechanistic reasoning, which address the scalar levels of atoms and electrons and, thus, an *electronic* level of granularity (Deng & Flynn, 2021). Overall, Deng and Flynn's (2021) analysis illustrates students' heterogeneous use of granularity, suggesting that examining how granularity develops within comparable contexts could be

valuable for adapting instruction and learning environments.

2.2 Second *essential epistemic heuristic*: Identifying and unpacking relevant factors

The second *essential epistemic heuristic* includes identifying and unpacking relevant entities, activities, and their corresponding properties at the lower scalar level (Krist et al., 2019). Disciplinary core ideas can be used to capture students' mechanistic reasoning through the second *essential epistemic heuristic* since entities and activities are always embedded within a domain-specific context (Cooper et al., 2017; Russ et al., 2008). The appropriate disciplinary core ideas for mechanistic reasoning in a context emerge from examining the properties of entities and activities (Haskel-Ittah et al., 2020).

In chemical contexts, *structural* features mostly shape the properties of entities (Goodwin, 2008). These *structural* features include atom types, connectivity, molecular shape, bonding patterns, and functional groups. Disciplinary core ideas derived from *structural* features might comprise *charged species and substrates*, *nucleophiles and electrophiles*, or *acids and bases* (Martin et al., 2024). In contrast to *structural* features, *energetic* features govern the properties of activities (Goodwin, 2008). For example, *energetic* features like equilibrium direction or reaction rate impact the activities of chemical, biological, and physical mechanisms. Disciplinary core ideas derived from *energetic* features could encompass *thermodynamics and kinetics* (Martin et al., 2024). Little is known about which disciplinary core ideas support students' mechanistic reasoning. Identifying the disciplinary core ideas that advance students' mechanistic reasoning could be a first step in providing targeted support that helps students identify and unpack entities, activities, and their properties (Haskel-Ittah et al., 2020).

2.3 Third *essential epistemic heuristic*: Linking how properties drive the phenomenon

Once relevant entities, activities, and properties are identified and unpacked, these factors must be organized to establish a causal sequence of events over time or space (Krist et al., 2019). In other words, the third *essential epistemic heuristic* captures how students integrate the first two heuristics—reasoning across scalar levels and identifying factors at the lower scalar level—to connect the unpacked entities, activities, and properties back to the target phenomenon. Following this, the third e*ssential epistemic heuristic* searches for the underlying causes of emerging mechanisms, which is why the idea of causality can be used to assess how students connect the properties and their interactions back to the target phenomenon (Sevian & Talanquer, 2014).

How students integrate the idea of causality into their mechanistic reasoning can vary across science domains (Grotzer, 2003; Russ et al., 2008). Students' use of causality ranges from simply listing properties of entities and activities without establishing connections, to identifying cause-and-effect relations between all relevant properties (e.g., Bodé et al., 2019; Deng & Flynn, 2021; Moon et al., 2019; Moreira et al., 2019; Sevian & Talanquer, 2014; Weinrich & Talanquer, 2016). A longitudinal investigation of students' use of causality would reveal whether students can transfer or advance their causality across contexts or whether they encounter unique challenges within each new context.

Overall, students' mechanistic reasoning can vary in complexity along these heuristics. For instance, some students may identify properties at the level of the target phenomenon while others may analyze properties at the scalar level below the target phenomenon. Similarly, some students may list unpacked properties while others may establish cause-and-effect relations between those properties. Thus, our domain-specific operationalization of the three *essential*

epistemic heuristics represents a continuum for assessing the complexity of students' mechanistic reasoning regarding the use of granularity, disciplinary core ideas, and causality. We use the term *reasoning complexity* to quantify students' mechanistic reasoning through the lens of the *essential epistemic heuristics*.

2.4 Mechanistic reasoning as a tool to make claims about phenomena

Drawing from this operationalization, little is known about how the three *essential epistemic heuristics* interact with one another. For example, prior studies analyzed the granularity and causality of students' mechanistic reasoning separately (Bodé et al., 2019; Deng & Flynn, 2021). Likewise, little is known about which disciplinary core ideas advance students' use of granularity or causality. As a result, it remains unclear whether supporting students' mechanistic reasoning in one heuristic can promote complexity gains in another.

Ultimately, mechanistic reasoning can serve as a tool to make claims about the target phenomenon (Russ et al., 2008). Here, we hypothesize that students who demonstrate more complex mechanistic reasoning are more likely to make correct claims about the target phenomenon. However, evidence concerning the relationship between the complexity of students' mechanistic reasoning and the correctness of their claims remains inconclusive. For instance, Bodé et al. (2019) observed that students' reasoning complexity was comparable among those who made correct and incorrect claims about the plausibility of two reactions in undergraduate organic chemistry. Conversely, Deng and Flynn (2021) found that students exhibiting more complex mechanistic reasoning about acid–base equilibria demonstrated a statistically higher frequency of correct claims. Qualitative research suggests that students who make correct claims can link entities, activities, and their properties at a more complex level (Cruz-Ramírez de Arellano & Towns, 2014; Watts, Park, et al., 2022). Given the inconclusive relationship between reasoning complexity and claim correctness, our proposed model examines this connection (Figure 1). A longitudinal study investigating the evolving relationship between reasoning complexity and claim correctness may clarify the relationship between these skills.

2.5 Analyzing students' mechanistic reasoning within a broader system: Including further variables

Several variables may influence the complexity of students' mechanistic reasoning. For instance, students' prior knowledge influences their science learning and may be linked to their reasoning complexity (e.g., Glasson, 1989; Johnson & Lawson, 1998). Furthermore, research indicates that students who apply mechanistic reasoning demonstrate high concept knowledge in tasks that involve knowledge transfer, such as predicting the products of organic chemistry mechanisms (Grove et al., 2012). This finding suggests that complex mechanistic reasoning may enhance students' understanding of chemical concepts even in tasks that do not explicitly require mechanistic reasoning—and perhaps the reverse may also hold. However, students who did not engage in mechanistic reasoning performed equally to their peers who used mechanistic approaches on easier problems, where problem-solving was possible based on factual recall (Grove et al., 2012). This inconsistency underscores the inconclusive relationship between students' concept knowledge and their mechanistic reasoning.

Alongside domain-specific concept knowledge, some studies hypothesize that domain-general skills, such as scientific argumentation, can advance students' mechanistic reasoning (e.g., Choi et al., 2013; Lieber & Graulich, 2022; Lieber et al., 2022a; Moon et al., 2016; Pabuccu &

Erduran, 2017). However, little is known about whether these rather domain-general argumentation skills suffice for complex mechanistic reasoning. Hence, longitudinally examining the multifaceted relationships between students' mechanistic reasoning, concept knowledge, and argumentation skills would provide valuable insights to advance mechanistic reasoning support.

Besides domain-general argumentation skills or domain-specific concept knowledge, evidence suggests that a positive self-concept—the belief in one's ability to succeed in a subject—is crucial for developing a meaningful understanding of science concepts in high school (Nieswandt, 2007). Additionally, students with correct and more complex mechanistic reasoning reported higher confidence levels in undergraduate organic chemistry (Gao et al., 2024). Modeling the relationship between students' mechanistic reasoning and their affective perceptions can illuminate the interplay of various variables impacting students' mechanistic reasoning within a holistic system.

3 | Research questions

Mechanistic reasoning is essential across science domains because it helps students explain, predict, or control phenomena. Nevertheless, the intricate nature of mechanistic reasoning suggests that students encounter context-dependent challenges when reasoning about mechanisms, which leads to substantial heterogeneity, particularly over time. Given this heterogeneity, explicitly supporting students' mechanistic reasoning over an extended period is crucial. However, the development of students' mechanistic reasoning depends on multiple variables, whose interplay remains unclear (Figure 1). Consequently, how to best design longitudinal support for students' mechanistic reasoning is uncertain. A deeper understanding of how students' mechanistic reasoning evolves and how it interacts with other variables may inform instructional approaches that support mechanistic reasoning, which is why we address the following research questions (RQs):

1. What mechanistic reasoning trajectories do undergraduate organic chemistry students follow over the course length?
2. How does students' claim correctness change over time, and what is the relationship between reasoning complexity and claim correctness?
3. Which *essential epistemic heuristics* predict students' mechanistic reasoning trajectories?
    a. To what extent do granularity and causality develop concurrently over the course length?
    b. How do disciplinary core ideas in students' mechanistic reasoning vary across trajectories?
4. Which cognitive and affective variables predict students' mechanistic reasoning trajectories?

4 | Methods

This study automatically assessed students' mechanistic reasoning throughout one semester of an undergraduate organic chemistry course using a machine learning (ML) model, which can analyze students' mechanistic reasoning in large-scale longitudinal assessments (Martin & Graulich, 2023). We used the classifications generated by this ML model to provide adaptive support at four points during the semester-long course. Such adaptive support promises to be more effective than uniform approaches, especially in undergraduate organic chemistry (Asmussen et al., 2024; Dood et al., 2020; Hedtrich & Graulich, 2018a, 2018b; Lieber et al., 2022a, 2022b). Further, the continuous ML output allowed us to monitor students' reasoning complexity throughout the semester, model their mechanistic reasoning trajectories (hereafter referred to as *reasoning trajectories*),

and identify the variables contributing to more productive trajectories.

4.1 Study context

This study provided students enrolled in an Organic Chemistry II course at a public university in the Midwestern United States during the Winter 2023 semester with ML-assisted adaptive support to advance their mechanistic reasoning and model their reasoning trajectories. The Institutional Review Board for the Social and Behavioral Sciences of a university in the mid-Atlantic region of the United States approved this study under Protocol #5937. Informed consent was obtained from all participants whose data were analyzed. At the beginning of the semester, 314 undergraduate students enrolled in the Organic Chemistry II course agreed to participate in this longitudinal study. To encourage participation, students received credit for completing study activities as a course assignment, where 80% of the credit was given for completing the assignments and 20% for the correctness of some closed-response questions. By the end of the semester, 266 students completed all four required assignments and constructed a total of 9,221 open-ended responses. Data from students who did not complete all assignments were excluded from the analysis.

Of the 266 students whose data were included, 19 chose not to provide demographic information. Among the remaining students, 148 identified as women, 93 as men, and six as other gender identities. The participants' ages ranged from 18 to 24 years while they were, on average, in their second year of undergraduate studies. The students majored in various disciplines, such as biochemistry, biology, biotechnology, chemistry, and neuroscience. Students identified their ethnic and racial background as follows: 19.7% Asian or Asian-American; 4.2% Black, Afro-Caribbean, or African American; 0.4% Indigenous American or Alaska Native; 5.4% Latino/a/e/x or Hispanic American; 3.3% Middle Eastern, North African, or Arab American; 59.0% Non-Hispanic White or Euro-American; and 8.0% multiracial.

4.2 Study design

We developed and implemented four online scaffolds on the survey platform Qualtrics to support students' mechanistic reasoning. The scaffolds prompted students to judge the plausibility of alternative reaction products for four reaction mechanisms throughout one semester: Williamson ether synthesis, enamine hydrolysis, aldol condensation, and Claisen condensation (Figure 2) (Lieber et al., 2022a, 2022b; Martin & Graulich 2024a). We used an ML model to evaluate students' mechanistic reasoning, tailor scaffolds to meet their needs, and track their reasoning complexity (Figure 2).

The first scaffold included diagnostic tasks to assess students' argumentation skills and chemical concept knowledge (Figure 2). Here, closed-response tasks guided students in structuring domain-general and organic chemistry-specific arguments. These tasks prompted students, among others, to identify claims, evidence, and reasoning in given arguments, distinguish between non-electronic and electronic reasoning statements, classify the elaborateness and interconnectedness of arguments, and build chemically sound arguments. Furthermore, open-ended tasks were designed to help students activate their chemical concept knowledge for four disciplinary core ideas. These tasks prompted students, for instance, to judge the leaving group quality of charged species, identify the strongest nucleophile/electrophile or acid/base in a reaction mixture, estimate the energetic level of molecules and ions, and compare the rate of two contrasted reactions. Lieber et al. (2022b) provide a comprehensive overview of the diagnostic tasks. After completing these diagnostic tasks, students judged the plausibility of alternative reaction products for the ether

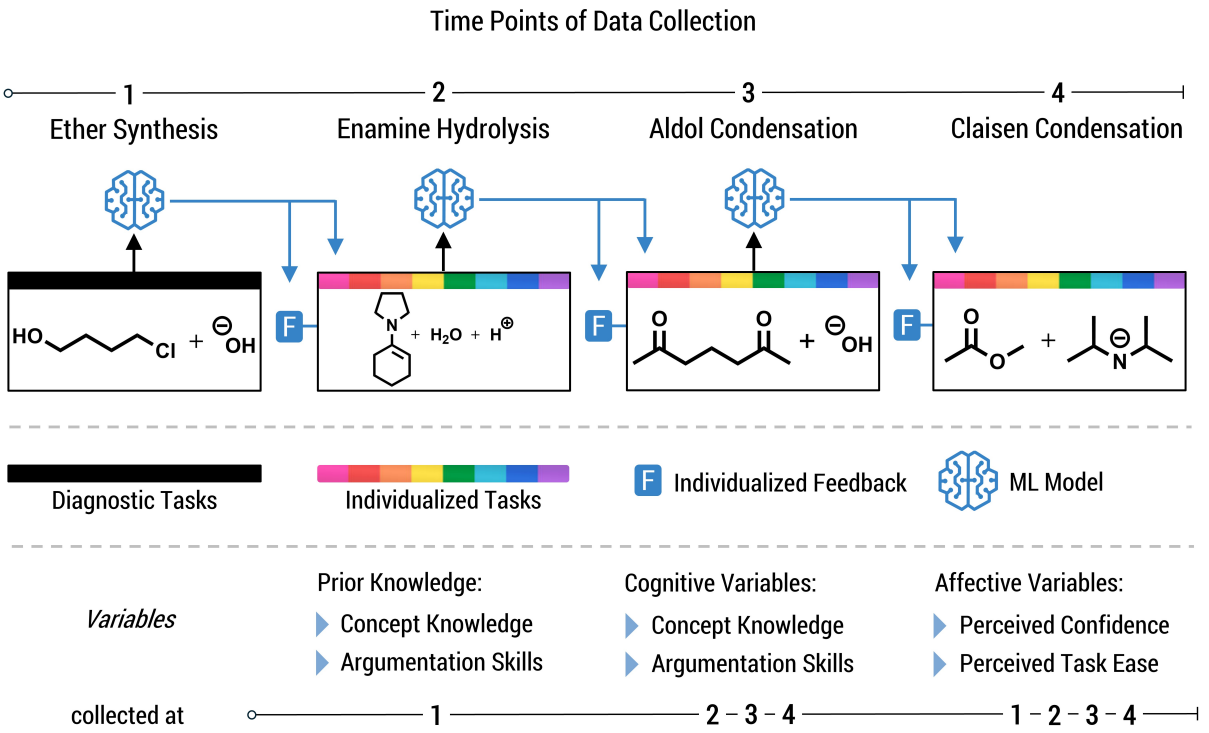


FIGURE 2 Using a machine learning model to support Organic Chemistry II students' mechanistic reasoning adaptively throughout one semester.

synthesis mechanism through open-ended items. The ML model automatically evaluated their responses about each product's plausibility (Martin & Graulich, 2024b; Martin et al., 2024). This enabled us to individualize the second scaffold to each student's current reasoning complexity.

The second scaffold comprised three parts: First, an individualized formative assessment featuring approximately 15 closed-response tasks, with task difficulties adjusted to a student's reasoning complexity, supported each student in structuring scientific arguments and activating chemical concepts. Moreover, the second scaffold provided individualized feedback before students judged the alternative reaction products in the enamine hydrolysis. To individualize feedback, students were assigned to one of seven groups based on their reasoning complexity in previous responses, with all members of the same group receiving identical feedback. The feedback highlighted students' current reasoning complexity and pinpointed areas for improvement. Afterward, students judged the plausibility of alternative reaction products for the enamine hydrolysis through open-ended tasks, which were again automatically evaluated by the ML model. Based on the ML model's output, the third and subsequently fourth scaffolds were adapted to each student's reasoning complexity.

To scaffold students' mechanistic reasoning about the plausibility of alternative reaction products, we employed a simplified version of Toulmin's argumentation pattern: the claim-evidence-reasoning model (McNeill & Krajcik, 2011; McNeill et al., 2006). A *claim* is a statement that requires a causal explanation. This explanation is provided by domain-specific information or scientific data, referred to as *evidence*, which draws on various sources to substantiate the claim. The *reasoning* then links the claim and evidence by establishing a cause-effect relationship. In our context, students were prompted to claim whether the displayed product is plausible or implausible, provide evidence to explain their judgment, and connect the claim and evidence in reasoning statements

by considering electronic, steric, or energetic effects (Figure 3). This task format was previously used and evaluated in a non-ML-assisted setting to encourage students to weigh alternatives (Lieber & Graulich, 2020, 2022). Examples of students' responses are available in Figure 3 and elsewhere (Lieber et al., 2022a, 2022b; Martin & Graulich, 2024b; Martin et al., 2024). To assess the correctness of students' claims in this setting, we evaluated whether students selected the correct option in the single-choice item that provided the possibilities of each product being either plausible or implausible (Figure 3).

The fourth author of this study was the course instructor. During instruction, students were introduced to several reaction mechanisms, including the ether synthesis, enamine hydrolysis, aldol condensation, and Claisen condensation. Scaffolds were assigned as a homework activity after these mechanisms had been taught in class. Scaffolds were implemented in weeks 6, 8, 12, and 13 of the semester, with each taking approximately 75–90 minutes to complete. Since prior research found that supporting students' mechanistic reasoning with the claim-evidence-reasoning model activates more ideas in their mechanistic reasoning (Lieber & Graulich, 2022), the instructor adopted this scaffolding to support students' mechanistic reasoning throughout the entire organic chemistry course.

4.3 Additional variables

In addition to evaluating students' reasoning complexity and claim correctness, we assessed four additional variables (Figure 1). These variables were assessed using items developed and validated by Lieber et al. (2022b). First, we measured students' chemical concept knowledge—initially as prior knowledge before the first data collection point and subsequently

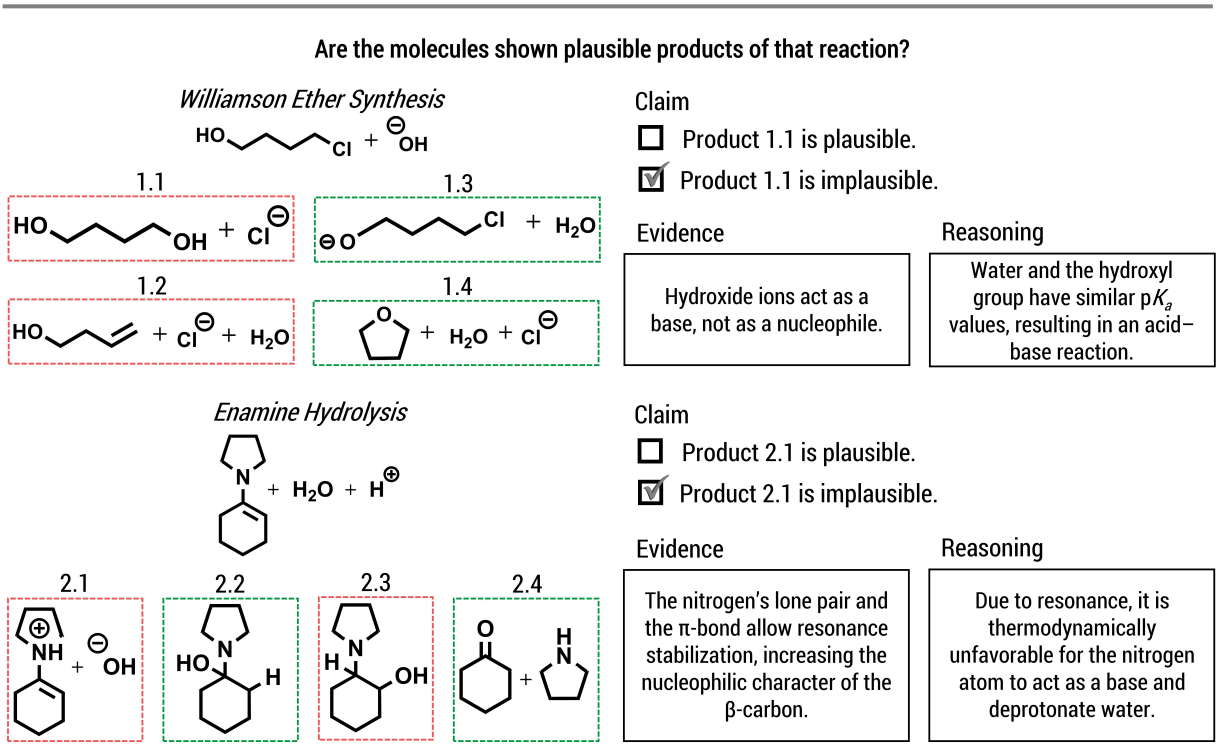


FIGURE 3 Students judged the plausibility of four alternative products for four mechanisms, including a Williamson ether synthesis and enamine hydrolysis. During this process, students made claims about the plausibility of each product, supported their claims with evidence, and connected the claims and evidence through reasoning (Lieber et al., 2022a). Students could construct more than one response for every alternative product. *Note:* Implausible products (1.1, 1.2, 2.1, 2.3) are highlighted in red while plausible ones (1.3, 1.4, 2.2, 2.4) are marked in green.

at time points two, three, and four of the study (Figure 2). Prior chemical concept knowledge was assessed using seven open-ended items. For example, students were asked to explain, based on electronic properties, which of two molecules is more likely to act as a nucleophile in a given reaction (Section 4.2). Research assistants manually scored these open-ended responses using the rubric applied by Lieber et al. (2022b). This rubric awards up to 2 points per item: 1 point for selecting the correct molecule and 1 point for providing a correct explanation. At time points two, three, and four, chemical concept knowledge was assessed with closed-response items, including single-choice, multiple-choice, and ranking tasks, often involving contrasting cases as introduced by Graulich and Schween (2018). For instance, students ranked four molecules by decreasing nucleophilicity. Points were awarded as follows: Students received 1 point for correct answers in single-choice tasks, 1 point per correct selection in multiple-choice tasks, and 1 point per correctly ordered molecule in ranking tasks.

Similarly, students' argumentation skills were assessed with closed-response items both before the first time point of the study and at subsequent time points during the semester (Figure 2). Tasks included assigning given statements to claim, evidence, or reasoning components of a scientific argument (Section 4.2). Scoring followed the same approach as for the closed-response tasks assessing chemical concept knowledge. Students' scores for chemical concept knowledge and argumentation skills assessed at time points two, three, and four of the study were first averaged to analyze general trends across the semester but then disaggregated by time point to account for context-dependent relationships with mechanistic reasoning. Scores are reported as percentages of correctly solved tasks.

Besides these cognitive variables, we assessed two affective variables: students' perceived confidence and perceived ease when engaging with organic chemistry tasks (Figure 2). These variables were measured using four-point Likert scales for each of the four alternative reaction products in every mechanism, resulting in four items assessing perceived confidence and task ease for each of the four time points. For example, students rated their confidence in solving the tasks with the options *very unconfident*, *unconfident*, *confident*, and *very confident*. Responses were converted to numerical values and summed for each mechanism, yielding scores ranging from 0 (low confidence, low ease) to 12 (high confidence, high ease). Scores were then averaged across the four mechanisms to identify general trends but also analyzed separately by time point to account for context-dependent relationships between mechanistic reasoning, perceived confidence, and perceived task ease.

4.4 ML model training

Students' open-ended responses were automatically evaluated using an ML model. This ML model is similar to the one reported by Martin et al. (2024). The ML model includes a pre-trained, transformer-based deep neural network for sequence classification to analyze students' responses based on the *PyTorch* deep learning framework implemented in Python (Paszke et al., 2019). Further technological details are available in Martin and Graulich (2024b) and Martin et al. (2024). The hyperparameters of the best-performing model are a learning rate of 5e-6, a batch size of 32, and trained over 51 epochs. The machine-human score agreements are as follows: accuracy = 0.9231, macro precision = 0.9149, macro recall = 0.9067, weighted F_1 score = 0.9208, macro F_1 score = 0.9072, Matthews Correlation Coefficient = 0.9155, and Cohen's κ = 0.9152.

The ML model was trained on a dataset comprising 3,430 student responses. This dataset included 1,450 responses reported in Lieber et al. (2022a, 2022b) judging a product's

plausibility for Williamson ether synthesis and Claisen condensation, 536 responses from a university in Western Canada judging a product's plausibility for enamine hydrolysis and aldol condensation, 196 expert responses written by chemistry education researchers, 848 responses from Watts, Dood, et al. (2022) providing explanations for a base-free Wittig reaction and thalidomide hydrolysis, and 400 responses generated by ChatGPT (Martin & Graulich, 2024b). Student responses related to the base-free Wittig reaction and thalidomide hydrolysis were included in the training dataset to improve ML model reliability (Martin, Dood, et al., under review). Further, ChatGPT-generated responses were incorporated because integrating authentic student responses with chatbot-generated ones yielded the highest machine-human score agreements across different validation scenarios (Martin & Graulich, 2024b). Additional information regarding the ML model development is available in Martin et al. (2024).

4.5 Rubric for machine coding

To evaluate the complexity of students' mechanistic reasoning, we used a previously developed rubric for machine coding drawing from the *essential epistemic heuristics* (Figure 4) (Krist et al., 2019). This rubric comprises the *levels of granularity* as an operationalization of students' use of granularity (Bodé et al., 2019; Deng et al., 2022; Deng & Flynn, 2021; Soo et al., 2019), disciplinary core ideas relevant to mechanistic reasoning (Cooper et al., 2017), and the *modes of reasoning* as an operationalization of students' use of causality (Sevian & Talanquer, 2014).

In the rubric, two *levels of granularity* classify the first *essential epistemic heuristic*. A *phenomenological granularity* indicates reasoning at the level of the target phenomenon, whereas an *electronic granularity* considers the scalar level below the target phenomenon (Bodé et al., 2019; Deng et al., 2022; Deng & Flynn, 2021). *Phenomenological* reasoning addresses explicit or measurable properties of entities or activities, such as pK_a values, the direction of chemical equilibrium, or reaction rate, without considering electron distribution. By contrast, *electronic* reasoning details the distribution and movement of electrons in chemical compounds, considering electronic properties like resonance, hyperconjugation, or inductive effects.

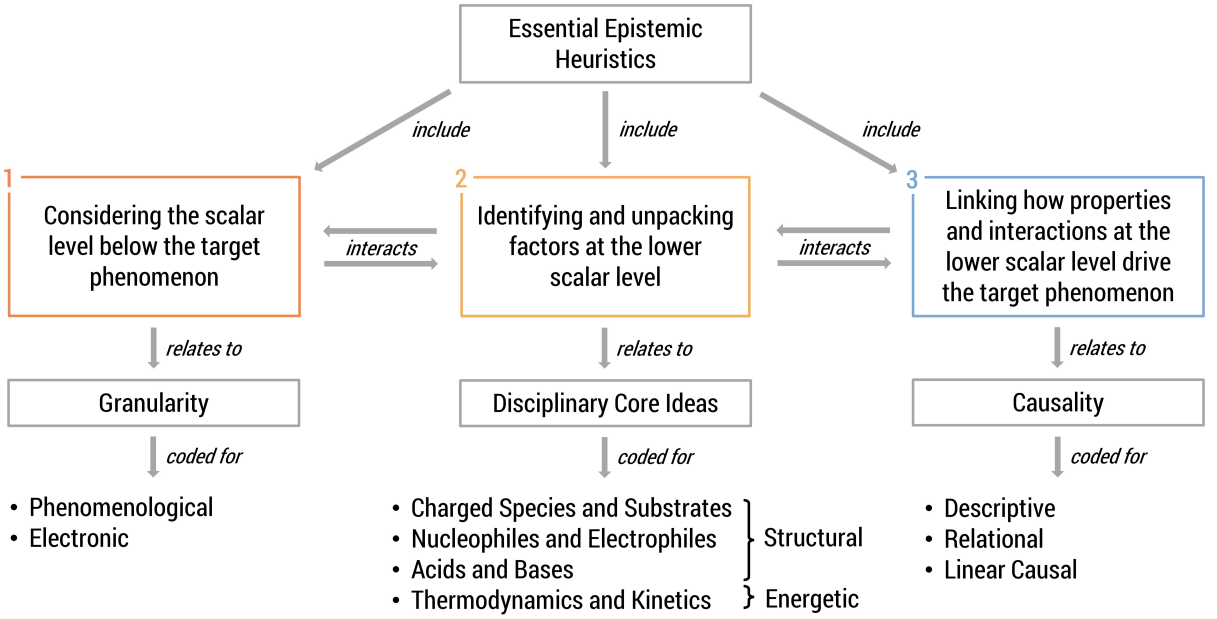


FIGURE 4 Operationalization of the *essential epistemic heuristics* (Krist et al., 2019) to classify students' mechanistic reasoning in organic chemistry using granularity, disciplinary core ideas, and causality.

Therefore, the *electronic* level offers a more in-depth reasoning of a mechanism's steps and causes.

Disciplinary core ideas related to *structural* and *energetic* features determine the entities, activities, and properties students identified and unpacked, reflecting the second *essential epistemic heuristic* in the rubric (Bodé et al., 2019; Cooper et al., 2017; Deng & Flynn, 2021). *Structural* reasoning focuses on the properties of entities. To categorize the entities and properties students used in their mechanistic reasoning, we subdivided this *structural* level into three disciplinary core ideas based on the properties students assigned to the entities: *charged species and substrates*, *nucleophiles and electrophiles*, and *acids and bases* (Martin et al., 2024). Conversely, *energetic* reasoning examines the properties of activities. In chemical contexts, thermodynamic and kinetic changes during chemical reactions, such as the impact of factors like enthalpy and entropy on a reaction's favorability, represent properties of activities. We introduced the disciplinary core idea of *thermodynamics and kinetics* to capture students' *energetic* reasoning (Martin et al., 2024). Responses that do not address these disciplinary core ideas are categorized as *absent*.

To evaluate students' use of causality as conceptualized by the third *essential epistemic heuristic*, we used three levels: *descriptive*, *relational*, and *linear causal* (Sevian & Talanquer, 2014). *Descriptive* reasoning recognizes explicit properties of salient entities without connecting them. *Relational* reasoning goes further by identifying explicit and implicit properties of entities or activities and correlating spatial or temporal links among them. Finally, *linear causal* reasoning invokes cause-effect relations between explicit and implicit properties of entities or activities to explain phenomena through the static interplay of these properties.

In sum, our rubric measures students' reasoning complexity by considering the granularity at which phenomena are explained, the detail with which disciplinary core ideas are unpacked, and the causality used to establish links between scalar levels. By doing so, the rubric assesses each response holistically by categorizing it into one of 21 mechanistic reasoning patterns. The ML model coded the *levels of granularity* (absent = 0, phenomenological = 1, electronic = 2) and *modes of reasoning* (absent = 0, descriptive = 1, relational = 2, linear causal = 3) for increasing complexity, which resulted in a score ranging from zero to five for each student response (Figure 5). The scores were assigned regardless of the applied disciplinary core ideas: *charged*

		Structural						Energetic	
Disciplinary Core Ideas		Charged Species and Substrates		Nucleophiles and Electrophiles		Acids and Bases		Thermodynamics and Kinetics	
Granularity		Phen.	Elec.	Phen.	Elec.	Phen.	Elec.	Phen.	Elec.
Causality	Absent	0							
	Descriptive	2	3*	2	3	2	3	2	3*
	Relational	3	4	3	4	3	4	3	4
	Linear Causal		5	4		4		4	5

FIGURE 5 Rubric to classify students' mechanistic reasoning with numbers indicating the complexity scores. *Note:* "Phen." denotes *phenomenological* reasoning, and "Elec." denotes *electronic* reasoning. Gray rectangles visualize categories that could not be automatically evaluated. The asterisk (*) symbolizes that these categories were combined for automated coding since they reflect the same reasoning pattern.

species and substrates, *nucleophiles and electrophiles*, *acids and bases*, or *thermodynamics and kinetics*. Further details on rubric development and challenges encountered in automatically evaluating specific rubric categories are documented by Martin et al. (2024).

4.6 Data analysis

This longitudinal study explored students' reasoning complexity at four time points during one semester, repeatedly measuring the same variables in the same cohort. Vermunt (2010) outlined a three-step process for such an analysis: modeling students' trajectories, clustering these trajectories, and using a multinomial logistic regression model to link the clustered trajectories to explanatory variables. Based on this approach, we investigated how students' reasoning trajectories varied and which cognitive and affective variables contributed to the observed differences. Due to the exploratory nature of the study, students' reasoning trajectories were inductively derived from the data after the study was completed. Thus, students were unaware of their assigned trajectory.

In this study, the four reaction mechanisms represent the time points of data collection (Figure 2). The four alternative products corresponding to each reaction mechanism were treated as subtasks and aggregated for this analysis. Since students could provide multiple responses to a given alternative product, we calculated the mean reasoning complexity for each student for each reaction mechanism. Consequently, the analysis is performed at the student level separated by reaction mechanisms. All statistical analyses were conducted with a significance level of $\alpha < .05$ unless noted otherwise, performed in R (version 4.3.2) (R Core Team, 2023) with RStudio as an integrated development environment (IDE) (RStudio Team, 2023) using base R packages unless specified otherwise.

4.6.1 Clustering students' reasoning complexity to model trajectories (RQ1)

To cluster students' reasoning complexity into trajectories, we applied the *kml* package using a *k*-means algorithm tailored to exploratory longitudinal data (Genolini et al., 2015; Genolini & Falissard, 2010, 2011). The *kml* package applies the *k*-means algorithm by Hartigan and Wong (Hartigan & Wong, 1979) with the Gower-adjusted Euclidean distance metric (Gower, 1966) to model piecewise linear trajectories. The benefits of this approach are numerous (Genolini & Falissard, 2011): it requires no assumptions about the trajectory's shape, it does not necessitate normality or parametric assumptions within trajectories, and it is robust concerning numerical convergence. However, the exploratory nature of this technique has certain drawbacks (Genolini & Falissard, 2011): it is impossible to test the fit between the identified partition and a theoretical model, the starting conditions are set randomly, and the algorithm may converge to a local optimum. To address these limitations, we reran longitudinal *k*-means analyses with cluster numbers ranging from two to ten using 100 varying starting values for each predefined cluster number, resulting in 900 cluster solutions. This approach increased the likelihood of finding a global maximum (Genolini et al., 2015; Genolini & Falissard, 2011). An alternative approach to modeling trajectories is detailed in Supporting Information A.

We integrated students' use of granularity and causality to model their reasoning trajectories, rather than analyzing these heuristics separately. The significant medium correlation between both heuristics justifies this approach ($r = 0.478$, 95% CI [0.430, 0.523], $p < .001$) (Cohen, 1988). We selected the time points of data collection as the metric for modeling the trajectories, resulting in equidistant trajectories. Because the

number of weeks between the time points varied, using weeks as a time metric would have resulted in unevenly spaced trajectories. Adopting the time points of data collection as a time metric likely simplifies students' actual reasoning trajectories but increases their interpretability.

4.6.2 Determining the optimal cluster solution (RQ1)

Estimating the optimal number of clusters in *k*-means clustering is challenging due to the need to predefine the number of clusters. Also, clusters are often not well-separated, making it difficult to assign students without error (Den Teuling et al., 2023). Running the algorithm multiple times with different cluster numbers and selecting the best solution based on statistical criteria that account for within-trajectory compactness and between-trajectory separation is useful (Genolini et al., 2015; Genolini & Falissard, 2011; Liu et al., 2010). Computing multiple statistical criteria can further strengthen the robustness of the findings.

To determine the optimal number of trajectories, we used a two-step approach: First, we tried to interpret the trajectories. Second, we quantitatively evaluated the trajectory solutions using parametric and non-parametric quality criteria. Our analysis of the quality criteria revealed a disagreement on the optimal number of trajectories: five criteria supported a ten-trajectory solution, three favored a two-trajectory solution, one recommended a four-trajectory solution, and one suggested a five-trajectory solution. After careful consideration, we chose the five-trajectory solution for the following reasons: Qualitatively, the five trajectories provided interpretable and distinct reasoning trajectories without inappropriately merging them. More trajectories did not offer additional insights, fewer missed important nuances. Additionally, all quality criteria supported the five-trajectory solution with acceptable values, and the Bayesian Information Criterion (BIC)—a conservative and widely used metric for cluster selection—also favored this solution. Detailed information on our selection of the five-trajectory solution, the quality criteria, and alternative solutions are available as Supporting Information A. Furthermore, we analyze the longitudinal differences across the five trajectories using a hierarchical linear model (HLM) with random effects, detailed in Supporting Information B.

4.6.3 Performing a mixed ANOVA to analyze students' claim correctness (RQ2)

To monitor students' ability to make a correct claim about a product's plausibility, a skill we call *claim correctness,* we applied a mixed analysis of variance (ANOVA) as a general linear model. Broadly, a mixed ANOVA tests for differences in means across repeated measures and between-subject variables (Twisk, 2019). Our study examined how the percentage of correct claims evolved, using the time points of data collection as a repeated measures variable and students' reasoning trajectories as a between-subject variable. These variables were dummy-coded because of their categorical nature by setting orthogonal contrasts (Table 1). For example, contrast C1 compares the Williamson ether synthesis—the first time point of data collection—against the combined effects of the other three mechanisms.

Table 2 demonstrates that the model incorporating students' reasoning trajectories, the time points of data collection, and their interaction provides the best fit. We offer a detailed discussion of this mixed ANOVA model in Section 5.2. We applied the *nlme* package for linear modeling (Pinheiro et al., 2011).

TABLE 1 Orthogonal contrasts for dummy-coding the categorical variables to estimate and report the models.

Variable	Contrast	Reference Group	Target Group
Context	C1	Ether Synthesis	Enamine Hydrolysis + Aldol Condensation + Claisen Condensation
Context	C2	Enamine Hydrolysis	Aldol Condensation + Claisen Condensation
Context	C3	Aldol Condensation	Claisen Condensation
Trajectory	T1	Trajectories A + B	Trajectories C + D + E
Trajectory	T2	Trajectory C	Trajectories D + E
Trajectory	T3	Trajectory A	Trajectory B
Trajectory	T4	Trajectory D	Trajectory E

TABLE 2 Model fit statistics of the mixed ANOVA as a general linear model.

ID	Model	*df*	AIC	BIC	Deviance	Test	*L*.Ratio	*p*-value	η_p^2
1	Baseline	4	209	229	201	-	-	-	-
2	Context	7	−78	−43	−92	1 vs 2	293	< .001	0.355
3	Trajectory	11	−98	−44	−120	2 vs 3	29	< .001	0.445
4	Context × Trajectory	23	−97	18	−143	3 vs 4	22	.036	0.029

Note: Models build upon each other; for instance, model 3 includes all effects of model 2 with *Trajectory* as an additional fixed effect. *df* = degrees of freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; *L*.Ratio = likelihood ratio; η_p^2 = measure of effect size ($\eta_p^2 > 0.01$ indicates a small effect, $\eta_p^2 > 0.06$ indicates a medium effect, $\eta_p^2 > 0.14$ indicates a large effect (Cohen, 1988)).

Generally, mixed ANOVAs have many assumptions concerning the within- and between-subject elements. First, data are expected to be normally distributed at each level. Our data showed no evidence of skew or kurtosis since all values were below 2.228 for skew and 4.213 for kurtosis (Kline, 2023). However, results of the Shapiro-Wilk normality test, applied to the residuals from the factorial ANOVA model, suggested that model residuals differed significantly from a normal distribution (W = 0.964, p < .001). A visual inspection of boxplots for each reaction mechanism highlighted no outliers. We also assessed the homogeneity of variance assumption. Levene's test revealed a violation of this assumption for time point three (F(4,261) = 2.42, p = .049) and four (F(4,261) = 7.04, p < .001). Conversely, homogeneity of variance was satisfied for time point one (F(4,261) = 0.35, p = .846) and two (F(4,261) = 1.24, p = .294). Further, Box's M-test indicated no violation of the homogeneity of covariance matrices (M = 1.82, p = .768). Since some assumptions were violated, we also conducted a robust mixed ANOVA using the *WRS* package (Wilcox & Schönbrodt, 2019). The results from this analysis were consistent with those of the traditional approach, so further details of the robust analysis are not discussed.

Initially, we aimed to analyze the development of students' claim correctness using an HLM with random effects, similar to the approach described in Supporting Information B. To determine the need for an HLM, we compared a baseline model with a random intercept model using a likelihood ratio test. The test showed no significant improvement in fit for the more complex model (χ^2(1) = 2.068, p = 0.150), and the adjusted intraclass correlation (ICC) was 0.037. Since an ICC below 0.100 suggests that an HLM is

unsuitable, we applied a mixed ANOVA instead as described above.

4.6.4 Performing MANOVA and discriminant function analysis to disaggregate students' reasoning complexity (RQ3a)

To estimate the impact of granularity and causality on students' reasoning trajectories, we implemented a multivariate analysis of variance (MANOVA). In general, MANOVA examines the influence of a predictor variable on more than one outcome variable (Bray & Maxwell, 1985). In our analysis, we performed an omnibus MANOVA and four MANOVAs separated by time points to test whether differences among trajectories still exist if students' use of granularity and causality are not aggregated. The MANOVAs performed for each time point, considered as post hoc analyses of the omnibus MANOVA, applied a Bonferroni-corrected significance level of $p < .013$. Given that all test statistics (e.g., Pillai-Bartlett's Trace, Hotelling's Trace, Wilks' Lambda, and Roy's Largest Root) yielded consistent results, we chose to report Pillai-Bartlett's Trace (Bray & Maxwell, 1985). Statistically significant results prompted us to conduct discriminant function analyses as post hoc tests to quantify the unique contributions of students' use of granularity and causality to differences among trajectories. We used the *car* package for the MANOVA analyses (Fox et al., 2007) and the *MASS* package for discriminant function analyses (Ripley et al., 2013).

In our analysis, multivariate normality, assessed via multivariate Shapiro-Wilk tests, and homogeneity of covariance matrices, measured via comparisons of variances and covariances across matrices, were not fully met across all trajectories and time points. Considering these findings, we also conducted a robust MANOVA using the *WRS2* package (Mair & Wilcox, 2020). Since the findings of this analysis were consistent with the non-robust analysis, we do not report the robust MANOVA in further detail.

4.6.5 Performing a χ^2 test to identify trajectory differences in students' use of disciplinary core ideas (RQ3b)

To determine differences across trajectories in students' use of disciplinary core ideas, we conducted a χ^2 test. A χ^2 test assesses whether a significant association between categorical variables exists based on observed and expected frequencies (Agresti, 2012). In addition to the four disciplinary core ideas outlined in Section 4.5, the *absent* code was also incorporated in this analysis. Besides, students who addressed multiple disciplinary core ideas simultaneously (n(ether synthesis) = 60; n(enamine hydrolysis) = 49; n(aldol condensation) = 72; n(Claisen condensation) = 48) were included in a separate bin.

In our analysis, we first computed a contingency table showing the distribution of disciplinary core ideas across reasoning trajectories. Every student contributed to only one cell in the contingency table, so data independence was satisfied. Then, we conducted a χ^2 test to assess the association between trajectories and the applied disciplinary core ideas (Agresti, 2012; McHugh, 2013). Due to some expected frequencies being less than five, we also conducted a Fisher's exact test using simulated data to confirm results (Patefield, 1981). Finally, we performed a post hoc analysis of the standardized residuals to pinpoint specific differences among trajectories based on the frequencies of disciplinary core ideas per trajectory (Agresti, 2012; Beasley & Schumacker, 1995). We repeated this analysis separately for each of the four time points. The χ^2 tests for each time point, treated as post hoc analyses of the omnibus χ^2 test, used a Bonferroni-corrected significance level of $p < .013$. The post hoc analyses of the standardized residuals applied a Šidák-adjusted significance level of $p < .003$ (Šidák, 1967) as recommended by Beasley and Schumacker (1995). We used the *gmodels* package for this χ^2 analysis (Warnes et al., 2018).

4.6.6 Performing multinomial logistic regression to explain trajectory differences (RQ4)

As a last analysis step, we implemented a multinomial logistic regression to identify the cognitive and affective variables differing among reasoning trajectories. Generally, multinomial logistic regression is a statistical method used to model the relationship between a categorical variable with more than two possible outcomes and one or more predictor variables (Menard, 2002). To predict students' trajectories, we established a regression model based on the four variables described above (Figure 1). In Supporting Information C, we provide the standardized regression coefficients, standard errors, and odds ratios of the regression model.

Except for students' prior knowledge, the other variables were continuously assessed throughout the semester but averaged for this analysis step. Accordingly, we also conducted multinomial logistic regressions separated by these variables as post hoc analyses to identify the specific reaction mechanisms in which the trajectories differed. The multinomial logistic regressions performed for the four continuously assessed variables are considered as post hoc analyses of the omnibus multinomial logistic regression, so we used a Bonferroni-corrected significance level of $p < .013$. Supporting Information C provides detailed information on these post hoc analyses.

Some assumptions must be satisfied to perform a multinomial logistic regression (Field et al., 2012): To evaluate multicollinearity, we calculated variance inflation factors (VIFs), where values exceeding 10 are problematic; in our analysis, the VIF values ranged from 1.100 to 3.271, which is not concerning. The assumption of linearity of the logit was confirmed by the non-significant interactions between the predictor variables and their logits in a regression model.

5 | Results

5.1 RQ1: Modeling students' reasoning trajectories

We monitored students' reasoning complexity longitudinally (Figure 6a) and clustered their reasoning trajectories into more generalizable patterns. The five-trajectory solution provided the best statistical fit (Supporting Information A), offering a nuanced view of students' reasoning complexity over time (Figure 6b). Trajectory A (yellow) represents the highest reasoning complexity with a starting score of around 3.5, minor fluctuations, and a slightly increasing trend throughout the semester. Trajectory B (red), indicative of moderate-to-high reasoning complexity, begins at a score of 3.0 and remains relatively steady with a slight increase by the end of the term. Trajectories C (blue) and D (green) characterize low-to-moderate reasoning complexity with a starting score of 2.5. Despite similar complexity at the beginning of the study, the trajectories diverge significantly over time. Lastly, trajectory E (purple) represents the lowest reasoning complexity, beginning with a score near 2.5 but experiencing two drops during enamine hydrolysis and aldol condensation. This trajectory remains low with an ending score of 2.0. Overall, trajectories A and B are more stable and highlight higher reasoning complexity; trajectories C, D, and especially E exhibit greater challenges in maintaining or improving reasoning complexity over time.

In addition to modeling students' reasoning trajectories, we calculated their gains in reasoning complexity from the first to the last assignment (Table 3). To facilitate comparisons between trajectories, we report Cohen's d_z to analyze each trajectory's gain in reasoning complexity. Overall, the adaptive support significantly improved students' reasoning complexity with a small effect size as indicated by a Cohen's d_z of 0.36. However, the impact of the adaptive support varied across trajectories. The high- and moderate-to-high-achieving

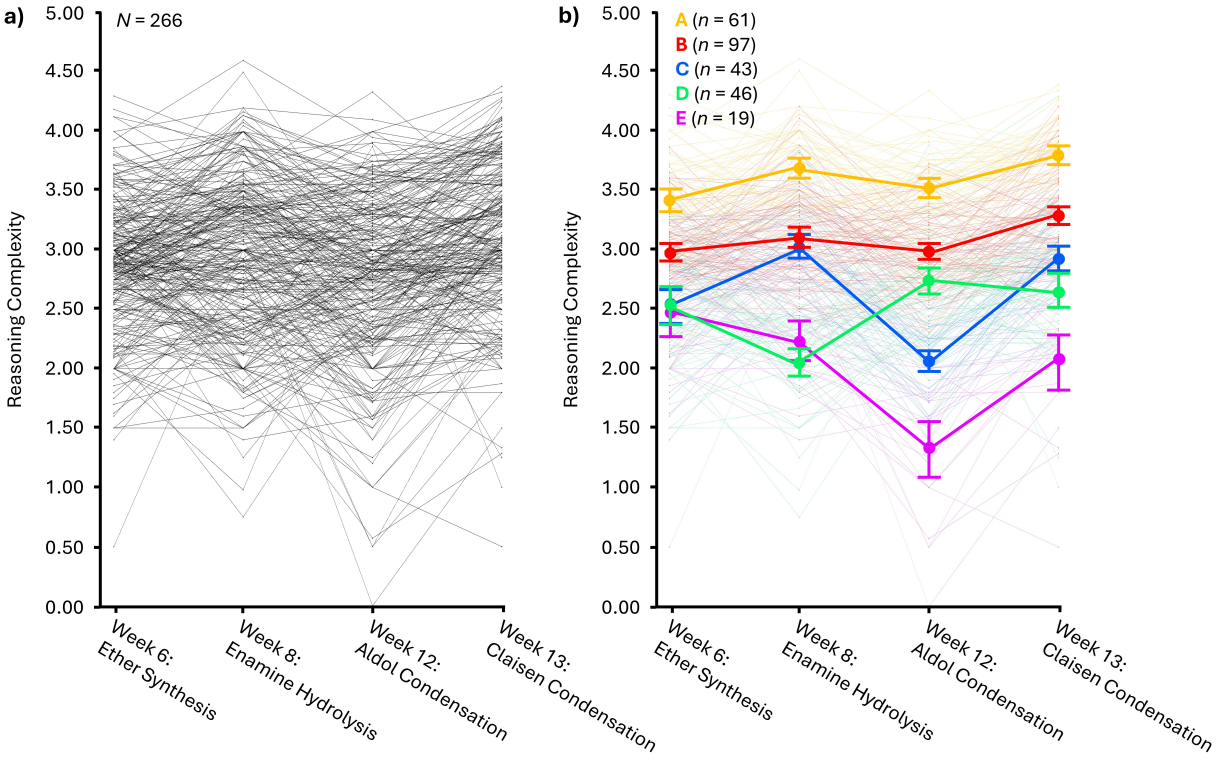


FIGURE 6 a) Students' reasoning trajectories. b) Classification of students' reasoning trajectories into five generalizable patterns using a longitudinal clustering technique. *Note:* Error bars represent 95% confidence intervals adjusted for longitudinal data.

trajectories A and B demonstrated medium gains throughout the semester with Cohen's d_z of 0.64 and 0.55, respectively. The low-to-moderate-achieving trajectories C and D displayed varied gains: trajectory C improved with a Cohen's d_z of 0.54, whereas trajectory D showed a negligible gain in Cohen's d_z of 0.14. Notably, trajectory E experienced a decline in reasoning complexity with a medium effect of 0.65 in Cohen's d_z.

In sum, the first step of the analysis reveals five distinct reasoning trajectories, with medium increases in reasoning complexity over time for trajectories A, B, and C—including approximately 76% of all students. Trajectories D and E showed greater challenges in mechanistic reasoning, particularly trajectory E declined over time. Differences between higher- and lower-achieving were maintained.

TABLE 3 Reasoning complexity scores of trajectories separated by reaction mechanisms.

Trajectory (Students)	Ether Synthesis		Enamine Hydrolysis		Aldol Condensation		Claisen Condensation		Gain: Start-End
	$\bar{x}_1$	σ_1	$\bar{x}_2$	σ_2	$\bar{x}_3$	σ_3	$\bar{x}_4$	σ_4	$\bar{x}_4 - \bar{x}_1$ (Cohen's d_z)
A (n = 61)	3.41	0.41	3.67	0.37	3.48	0.36	3.75	0.32	0.34 (0.64***)
B (n = 97)	2.96	0.34	3.09	0.42	2.95	0.34	3.26	0.41	0.30 (0.55***)
C (n = 43)	2.52	0.57	2.98	0.39	2.02	0.31	2.88	0.39	0.36 (0.54***)
D (n = 46)	2.51	0.51	2.04	0.41	2.72	0.40	2.62	0.58	0.11 (0.14)
E (n = 19)	2.45	0.53	2.19	0.43	1.29	0.58	2.03	0.58	–0.42 (0.65*)
Combined	2.88	0.56	2.96	0.69	2.76	0.72	3.11	0.65	0.23 (0.36***)

Note: $\bar{x}$ indicates means and σ indicates standard deviations. Cohen's d_z = measure of effect size (Cohen's d_z > 0.20 indicates a small effect, Cohen's d_z > 0.50 indicates a medium effect, Cohen's d_z > 0.80 indicates a large effect (Cohen, 1988)). *p < .05; **p < .01; ***p < .001.

5.2 RQ2: Changes in claim correctness over time and its relation to reasoning complexity

Similar to monitoring students' reasoning complexity, we also tracked their claim correctness—that is, how students' ability to make a correct claim improved over time. Conceptually, reasoning complexity and claim correctness are different skills: Reasoning complexity refers to the granularity and causality of students' mechanistic reasoning while claim correctness assesses whether students correctly judged the plausibility of the provided alternative products in single-choice items. We explored the extent to which reasoning complexity and claim correctness are correlated.

Figure 7 illustrates the trajectories of students' claim correctness across the organic chemistry reaction mechanisms. Consistent with Figure 6b, the five trajectories include the same students as the reasoning trajectories. Trajectory A (yellow) exhibits the highest claim correctness. This trajectory starts below 60% correctness for ether synthesis and steadily improves to approximately 95% correctness for Claisen condensation. Trajectories B (red), C (blue), and D (green) follow a similar pattern: Each trajectory starts at approximately 60%, just above trajectory A, and demonstrates consistent improvement during enamine hydrolysis and aldol condensation. Trajectory B reaches about 90% correctness for Claisen condensation while trajectories C and D experience a slight decline in correctness within this mechanism. Lastly, trajectory E (purple) begins with the lowest percentage of correct claims. Despite a slight decrease in claim correctness for the enamine hydrolysis, this trajectory improves during the condensation reactions, achieving around 70% correctness by the final mechanism. Overall, Figure 7 highlights that although students from different trajectories exhibited varying claim correctness, the average claim correctness improved across all trajectories.

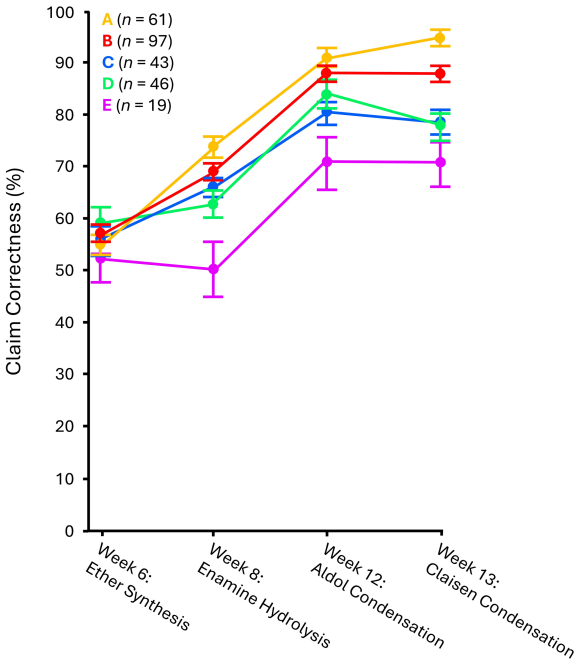


FIGURE 7 Changes in students' claim correctness over time. *Note:* The letters correspond to the reasoning trajectories presented in Figure 6b. Error bars represent 95% confidence intervals adjusted for longitudinal data.

A mixed ANOVA performed as a general linear model, which examined the effects of the reaction mechanisms and students' reasoning trajectories on their claim correctness, supports these visual observations: Several main effects related to the reaction mechanisms are statistically significant, implying that students' claim correctness varied by mechanism. Specifically, students' claim correctness was significantly lower in the ether synthesis compared to the other three mechanisms (Table 4: ID 4.1), lower in the enamine hydrolysis compared to the condensation reactions (Table 4: ID 4.2), but comparable for the two condensation reactions (Table 4: ID 4.3). Likewise, several main effects related to students' reasoning trajectories were statistically significant, suggesting that students of different reasoning trajectories demonstrated varying levels of claim correctness. This finding emphasizes that classifying students into five trajectories—based on their reasoning complexity—also reflects their claim correctness. This is especially true for students in

TABLE 4 Fixed effects included in the best-fitting mixed ANOVA model (Table 2: ID 4) analyzing the impact of reaction mechanisms and reasoning trajectories on students' claim correctness.

ID	Fixed Effects	Contrast	*b*	*SE b*	*df*	*t*-value	*p*-value	Cohen's f^2
4.0	Intercept	–	0.712	0.010	783	75.263	< .001	7.234
4.1	**Context**	**C1**	**0.051**	**0.004**	**783**	**11.620**	**< .001**	**0.176**
4.2	**Context**	**C2**	**0.063**	**0.006**	**783**	**10.051**	**< .001**	**0.129**
4.3	Context	C3	−0.005	0.011	783	−0.427	.669	< 0.001
4.4	**Trajectory**	**T1**	**−0.019**	**0.004**	**261**	**−5.289**	**< .001**	**0.107**
4.5	Trajectory	T2	−0.015	0.009	261	−1.590	.113	0.010
4.6	Trajectory	T3	−0.014	0.011	261	−1.319	.188	0.007
4.7	**Trajectory**	**T4**	**−0.048**	**0.018**	**261**	**−2.625**	**.009**	**0.026**

Note: Contrasts C1−C3 and T1−T4 are defined in Table 1. *b* = regression coefficient; *SE b* = standard error of regression coefficient; *df* = degrees of freedom; Cohen's f^2 = measure of effect size (Cohen's f^2 > 0.02 indicates a small effect, Cohen's f^2 > 0.15 indicates a medium effect, Cohen's f^2 > 0.35 indicates a large effect (Cohen, 1988)). Significant effects are bolded.

trajectories A and B, who showed significantly higher claim correctness than students in trajectories C, D, and E (Table 4: ID 4.4), with students in trajectory D outperforming those in trajectory E (Table 4: ID 4.7). However, trajectory C showed similar claim correctness compared to the combination of trajectories D and E (Table 4: ID 4.5) while claim correctness was also similar across trajectories A and B (Table 4: ID 4.6). Finally, only one of twelve interactions between reaction mechanisms and trajectories was statistically significant, indicating that the progression rate in claim correctness was relatively consistent across trajectories. Due to the insignificant test statistics, further results are not reported.

Collectively, this analysis reveals that students' claim correctness improved across all reasoning trajectories with the higher-achieving trajectories A and B outperforming the lower-achieving trajectories C, D, and E. The main effects including *Trajectory* imply that reasoning complexity and claim correctness are associated since trajectories modeled on reasoning complexity also differ in their claim correctness. A small but significant correlation between students' reasoning complexity and claim correctness substantiates their association (r = 0.178, 95% CI [0.120, 0.236], p < .001).

5.3 RQ3a: Concurrent development of granularity and causality

To explain the emergence of different reasoning trajectories, we analyzed whether any trajectory showed disproportionately high or low complexity specifically in students' use of granularity or causality. Figure 8 visualizes students' use of granularity and causality across the four reaction mechanisms. Again, the colors correspond to the reasoning trajectories depicted in Figure 6b. The trajectories of students' use of granularity and causality share many commonalities. Both figures present similar overall trends across the four reaction mechanisms and the relative ranking of the trajectories remains stable: trajectory A (yellow) consistently outperforms the others, followed by trajectory B (red), while trajectory E (purple) shows the lowest complexity. Moreover, most trajectories demonstrate a notable complexity decrease during aldol condensation, followed by an increase during Claisen condensation. Nonetheless, some variations arise between students' use of granularity and causality. For example, the causality is more linear and less variable in trajectories A and B than their granularity; thus, granularity might be more sensitive to context-specific mechanistic challenges for some students. Moreover,

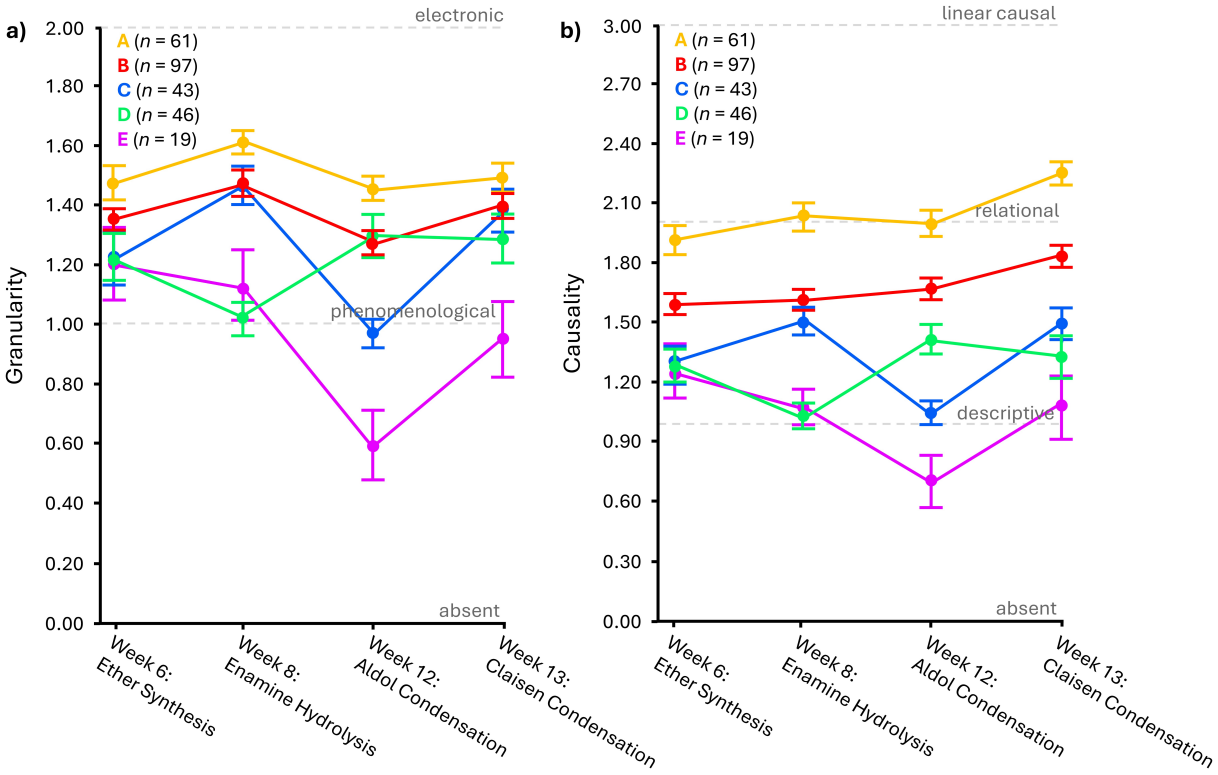


FIGURE 8 Trajectories of students' use of a) granularity and b) causality. *Note:* The letters correspond to the reasoning trajectories presented in Figure 6b. Error bars represent 95% confidence intervals adjusted for longitudinal data.

students' use of causality rarely exceeded a value of 2.00 representing *relational* reasoning, showcasing their challenges in achieving the highest causality: *linear causal* reasoning—a challenge not observed for students' use of granularity. However, the similarities between the trajectories outweigh their differences.

To quantitatively assess whether granularity and causality are interrelated, we conducted four MANOVAs separated by reaction mechanisms, where students' use of granularity and causality was predicted from their trajectory. All MANOVAs were statistically significant (Table 5), indicating that the five reasoning trajectories differ when granularity and causality are analyzed separately, rather than aggregated. We conducted discriminant function analyses as post hoc tests by calculating linear discriminants (LDs) to maximize the separation between trajectories and determine whether granularity or causality better differentiates them. Only the first linear discriminant (LD1) was statistically significant (Table 6), suggesting that the reasoning trajectories can be differentiated along LD1. For the ether synthesis, LD1 explains 99.93% of the variance in students' use of granularity and causality. The coefficients of LD1 suggest that this discriminant has a greater impact on causality than on granularity (Table 6), which reflects that causality better differentiates between trajectories on LD1. Both coefficients have a positive relationship with LD1, implying that increases in granularity and causality both positively influence LD1. Hence, this discriminant does not separate granularity and causality, rather LD1 is a discriminant that equally influences both heuristics. Therefore, the combined effect of granularity and causality explains trajectory differences. Analyzing LD1 for the other reaction mechanisms yields similar conclusions; additional LDs are not reported as they did not significantly differentiate between trajectories.

TABLE 5 Test statistics of the MANOVAs to investigate the association between granularity and causality.

Reaction Mechanism (Week of Assignment)	*df1*	*df2*	*F*	*p*-value	*V*	η_p^2
Ether Synthesis (6)	8	522	19.64	< .001	0.410	0.205
Enamine Hydrolysis (8)	8	522	36.72	< .001	0.720	0.360
Aldol Condensation (12)	8	522	44.95	< .001	0.815	0.408
Claisen Condensation (13)	8	522	31.53	< .001	0.651	0.325
Combined	8	2118	95.40	< .001	0.530	0.265

Note: df = degrees of freedom, *F* = test statistic, *V* = Pillai-Bartlett's Trace, η_p^2 = measure of effect size (η_p^2 > 0.01 indicates a small effect, η_p^2 > 0.06 indicates a medium effect, η_p^2 > 0.14 indicates a large effect (Cohen, 1988)).

TABLE 6 Results of the linear discriminant function analyses for the first linear discriminant.

Reaction Mechanism (Week of Assignment)	Explained Variance	β(causal.)	β(granular.)	*df*	*F*	*p*-value
Ether Synthesis (6)	0.9993	1.127	0.359	4	45.24	< .001
Enamine Hydrolysis (8)	0.9694	1.217	0.669	4	127.80	< .001
Aldol Condensation (12)	0.9708	1.380	0.767	4	183.20	< .001
Claisen Condensation (13)	0.9491	1.393	0.588	4	90.66	< .001
Combined	0.9825	1.236	0.336	4	277.30	< .001

Note: β represents the standardized coefficients of the linear discriminants. "Causal." denotes causality, and "granular." denotes granularity. *df* = degrees of freedom, *F* = test statistic.

Together, our analysis shows that a single discriminant function that positively relates to granularity and causality can explain an average of 98% of the variance in students' use of granularity and causality (Table 6). For this reason, the trajectory differences in Figure 6b do not result from a disproportionately high or low complexity of some trajectories in only one heuristic. In other words, high granularity is associated with high causality, and vice versa. The medium-to-large correlation between the two heuristics supports this conclusion (Table 7). Consequently, students' use of granularity, reflecting the first *essential epistemic heuristic*, and causality, representing the third *essential epistemic heuristic*, are interrelated.

TABLE 7 Correlation between granularity and causality.

Reaction Mechanism (Week of Assignment)	*df*	*t*	*p*-value	*r*	95% CI	
					LL	*UL*
Ether Synthesis (6)	264	5.736	< .001	0.333	0.222	0.436
Enamine Hydrolysis (8)	264	12.363	< .001	0.606	0.523	0.677
Aldol Condensation (12)	264	11.979	< .001	0.593	0.510	0.666
Claisen Condensation (13)	264	6.309	< .001	0.362	0.253	0.462
Combined	1062	17.741	< .001	0.478	0.430	0.523

Note: df = degrees of freedom, *t* = test statistic, *r* = correlation coefficient, *LL* = lower limit, *UL* = upper limit.

5.4 RQ3b: Differences in students' use of disciplinary core ideas by trajectories

As outlined above, variations in students' use of granularity and causality alone do not account for trajectory differences. To better understand the differences in trajectories, we analyzed the disciplinary core ideas students used to justify their claims: *charged species and substrates*, *nucleophiles and electrophiles*, *acids and bases*, and *thermodynamics and kinetics*. A χ^2 test revealed that students of different trajectories applied distinct disciplinary core ideas in their mechanistic reasoning (Table 8).

To identify the specific differences among the trajectories, we analyzed the standardized residuals of the χ^2 test based on the disciplinary core ideas students applied more or less frequently than statistically expected. In detail, the moderate-to-high-achieving trajectory B showed significantly fewer *absent* responses than statistically expected ($n_{expected}$ = 5.835; $n_{observed}$ = 0; p = .001) while the lowest-achieving trajectory E produced significantly more *absent* responses ($n_{expected}$ = 1.143; $n_{observed}$ = 9; p < .001). Moreover, trajectory B reasoned about *energetic* ideas more than statistically expected ($n_{expected}$ = 87.883; $n_{observed}$ = 120; p = .002) while trajectories D ($n_{expected}$ = 41.677; $n_{observed}$ = 21; p = .002) and E ($n_{expected}$ = 17.214; $n_{observed}$ = 4; p = .001) reasoned about them significantly less.

Subsequently, we examined differences in students' use of disciplinary core ideas separated by reaction mechanisms as a post hoc analysis to account for the context-dependency of their reasoning complexity. We found significant differences in students' use of disciplinary core ideas for enamine hydrolysis, aldol condensation, and Claisen condensation (Table 8).

Specifically, the low-to-moderate-achieving trajectory D showed more *absent* responses in enamine hydrolysis than statistically expected ($n_{expected}$ = 0.519; $n_{observed}$ = 3; p = .001). Similarly, the lowest-achieving trajectory E produced more *absent* responses than statistically expected in aldol condensation ($n_{expected}$ = 0.571; $n_{observed}$ = 7; p < .001) and Claisen condensation ($n_{expected}$ = 0.143; $n_{observed}$ = 2; p < .001). These findings suggest that students of trajectories D and E struggled to integrate disciplinary core ideas into their mechanistic reasoning, which may have contributed to their lower reasoning complexity. Similarly, the lowest-achieving trajectory E

TABLE 8 Results of χ^2 tests to identify differences in students' use of disciplinary core ideas.

Reaction Mechanism (Week of Assignment)	χ^2 test					Fischer's exact test
	df	χ^2	*p*-value	Cramer's *V*	Cohen's *W*	*p*-value
Mult	14	26.92	.137	0.159	0.318	.152
Enamine Hydrolysis (8)	14	73.26	< .001	0.262	0.525	< .001
Aldol Condensation (12)	14	101.35	< .001	0.309	0.617	< .001
Claisen Condensation (13)	98	54.56	< .001	0.226	0.453	.005
Combined	12	111.11	< .001	0.187	0.323	< .001

Note: The omnibus χ^2 test in "Combined" compares only the frequencies between *absent*, *structural*, and *energetic* reasoning to identify general associations between students' reasoning trajectories and their use of disciplinary core ideas. The post hoc χ^2 tests further explore these associations by breaking down *structural* granularity into its three corresponding disciplinary core ideas. Cramer's *V* = measure of effect size (*V* > 0.06 indicates a small effect, *V* > 0.17 indicates a medium effect, *V* > 0.29 indicates a large effect (Cohen, 1988)); Cohen's *W* = measure of effect size (*W* > 0.10 indicates a small effect, *W* > 0.30 indicates a medium effect, *W* > 0.50 indicates a large effect (Cohen, 1988)).

engaged less with *thermodynamics and kinetics* across aldol condensation ($n_{expected}$ = 4.571; $n_{observed}$ = 0; p = .001) and Claisen condensation ($n_{expected}$ = 5.707; $n_{observed}$ = 0; p < .001). In contrast, the moderate-to-high achieving trajectory B applied *thermodynamics and kinetics* more frequently in enamine hydrolysis ($n_{expected}$ = 35.737; $n_{observed}$ = 51; p = .002).

In sum, the analysis reveals that the higher-achieving students of trajectory B integrated thermodynamic and kinetic ideas more frequently into their mechanistic reasoning; the lower-achieving students of trajectories D and E struggled with this integration. This emphasizes that using thermodynamic and kinetic ideas is associated with higher reasoning complexity, which also shows that the *second essential epistemic heuristic* is connected to the other two heuristics.

5.5 RQ4: Predicting students' reasoning trajectories based on cognitive and affective variables

We performed a multinomial logistic regression to determine which cognitive and affective variables predict trajectory differences over the semester-long course. As cognitive variables, we analyzed students' chemical concept knowledge and argumentation skills, categorized as prior knowledge and within-study knowledge (Figure 2). As affective variables, we examined students' perceived confidence and ease when engaging with organic chemistry tasks. The χ^2 test for the regression incorporating these variables was statistically significant ($\chi^2(24) = 141.82$; p < .001; R^2 = 0.183), indicating that the assessed variables are associated with students' reasoning trajectories. Supporting Information C (Table S4) provides the standardized regression coefficients, standard errors, and odds ratios of the regression model.

First, students' prior chemical concept knowledge differed significantly across several trajectories (Figure 9a). Here, trajectory A significantly outperformed all other trajectories, and trajectory B showed higher prior chemical concept knowledge than trajectories C, D, and E. Likewise, students' within-study chemical concept knowledge varied significantly across most trajectories, except for trajectories C and D as well as C and E (Figure 9b). We also identified differences in students' prior argumentation skills between some trajectories (Figure 9c). Differences indicate that the high-achieving trajectory A had significantly higher argumentation skills before the study began while the other trajectories showed consistent argumentation skills. By contrast, students' within-study argumentation skills differed only between trajectories A and D (Figure 9d). The statistical similarity of the trajectory's within-study argumentation skills shows that students' mechanistic reasoning in undergraduate organic chemistry may be less associated with domain-general argumentation skills than with chemistry-specific concept knowledge. Chemical concept knowledge is more associated with students' reasoning complexity than their argumentation skills.

Analogously, we examined the extent to which the two affective variables—students' perceived confidence and perceived task ease—predict trajectory differences. Students' confidence while solving organic chemistry tasks varied significantly between many trajectories (Figure 9e) with the higher-achieving students feeling more confident. On the contrary, the perceived task ease differed only between three trajectories (Figure 9f), where the high-achieving trajectory A rated tasks as easier than most other trajectories. Consequently, perceived confidence is more associated with students' reasoning complexity than their perceived task ease.

To account for the context-dependency of students' reasoning complexity, we examined differences in the assessed cognitive and affective variables categorized by reaction mechanisms as a post hoc analysis. First, we

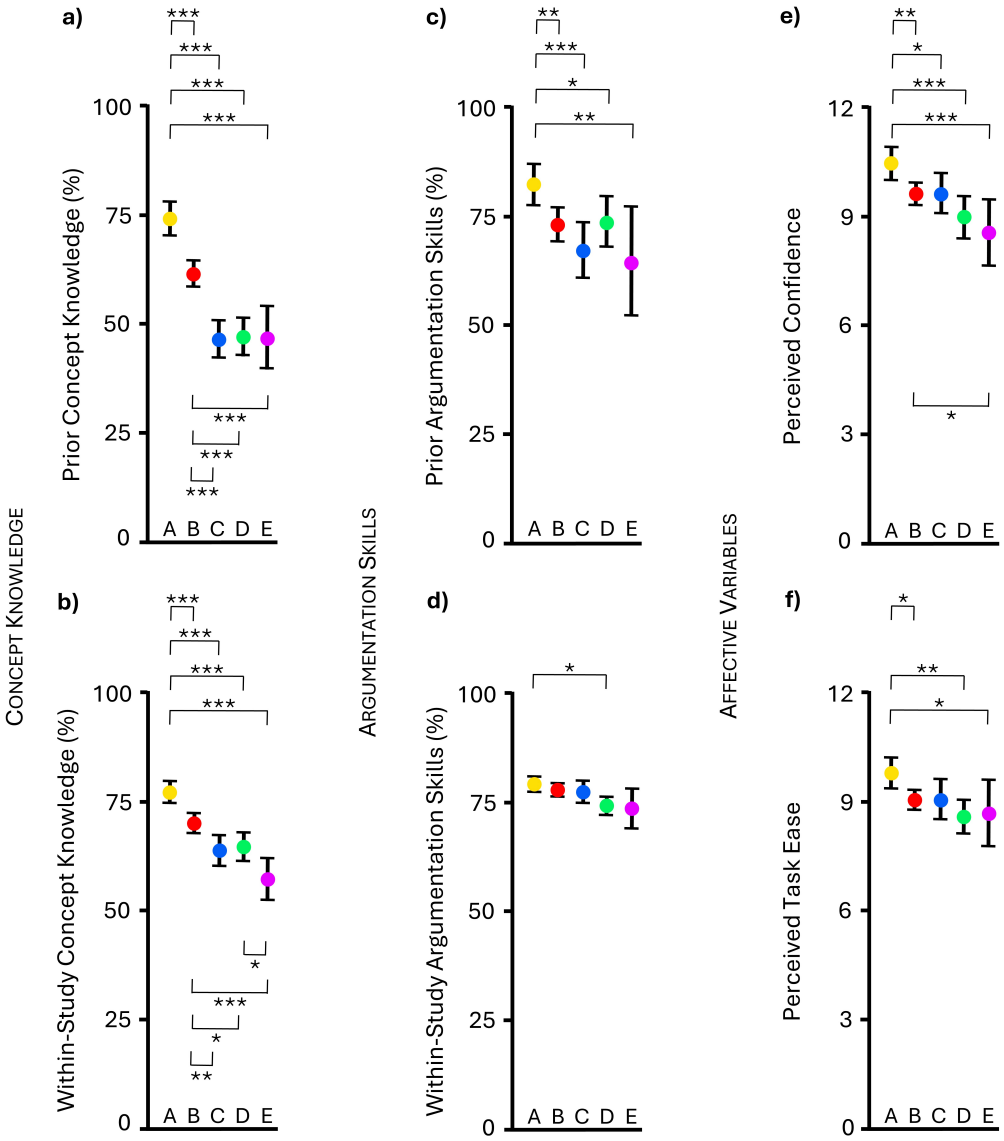


FIGURE 9 Distribution of concept knowledge, argumentation skills, and affective variables across trajectories. *Note:* The letters correspond to the reasoning trajectories presented in Figure 6b. Error bars represent 95% confidence intervals adjusted for longitudinal data. *$p < .05$; **$p < .01$; ***$p < .001$ (no p-value correction applied).

analyzed variations in students' within-study chemical concept knowledge and argumentation skills across time points two to four of the study. For instance, the logistic regression assessing students' within-study chemical concept knowledge used their concept knowledge from i) enamine hydrolysis (time point two), ii) aldol condensation (time point three), and iii) Claisen condensation (time point four) as three variables to predict students' reasoning trajectories. Supporting Information C (Tables S5–S7) presents the standardized regression coefficients, standard errors, and odds ratios.

The χ^2 test for the multinomial logistic regression analyzing students' within-study chemical concept knowledge was statistically significant ($\chi^2(12) = 64.29$; $p < .001$; $R^2 = 0.083$). Figure 10a presents that students' chemical concept knowledge varied considerably across trajectories. Likewise, the χ^2 test for the multinomial logistic regression for students' within-study argumentation skills was also statistically significant ($\chi^2(12) = 34.54$; $p = .001$; $R^2 = 0.045$), although fewer trajectory differences were observed (Figure 10b). This underscores that students' chemical concept knowledge is

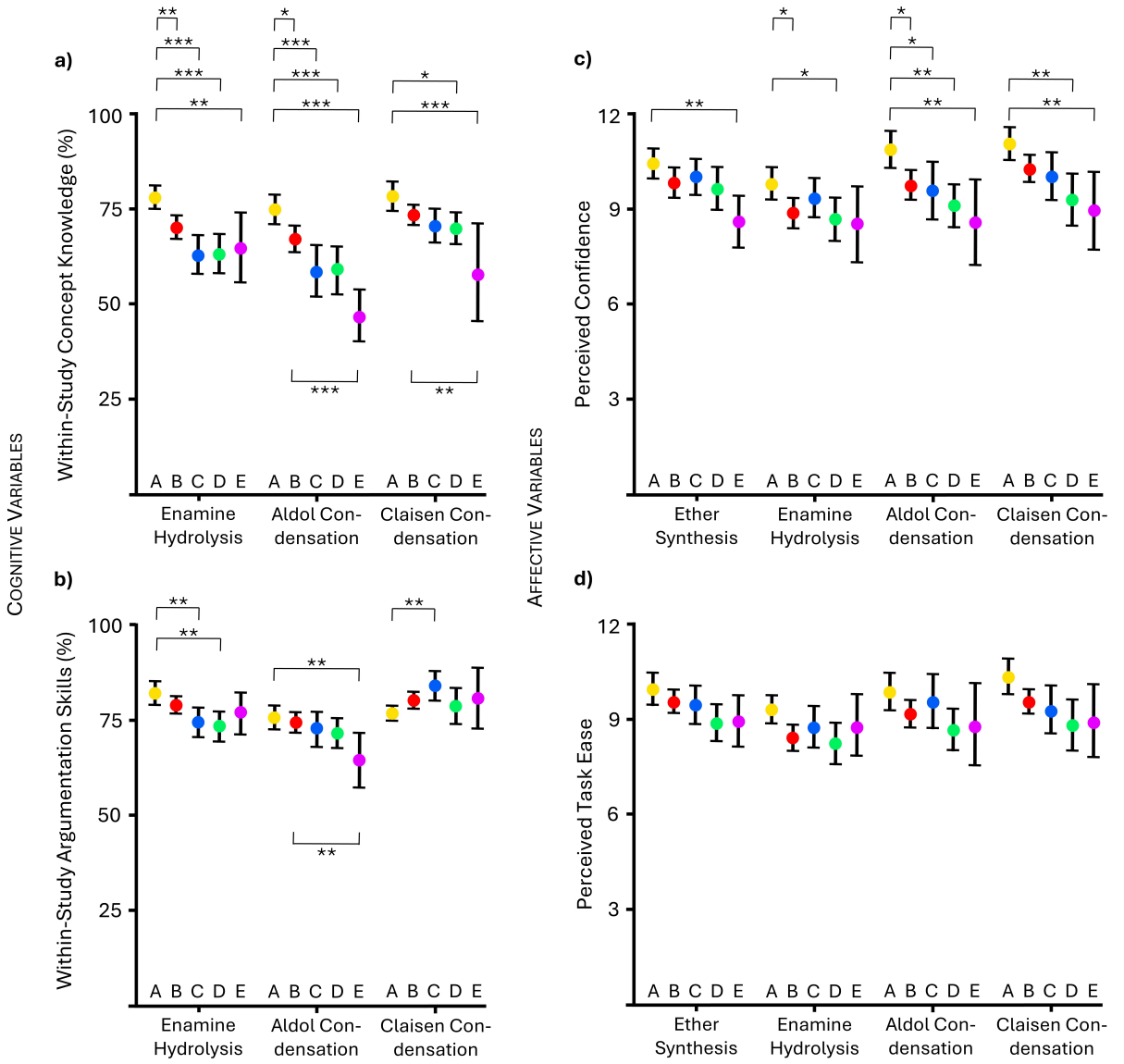


FIGURE 10 Distribution of cognitive and affective variables across trajectories separated by reaction mechanisms. *Note:* The letters correspond to the reasoning trajectories presented in Figure 6b. Error bars represent 95% confidence intervals adjusted for longitudinal data. *p < .013; **p < .003; ***p < .0003 (Bonferroni correction as outlined in Section 4.6.7).

more strongly associated with their reasoning complexity than with their argumentation skills. Additionally, students' within-study chemical concept knowledge and argumentation skills seem to develop differently. Achievement differences between trajectories persisted and, in some cases, were even exacerbated for domain-specific skills, such as chemical concept knowledge. Conversely, achievement differences between students tend to be minimized for domain-general skills like argumentation. This finding suggests that students may be more adept at transferring domain-general skills from one reaction mechanism to another. Domain-specific skills may rely more heavily on the context.

Ultimately, we conducted post hoc analyses to examine variations in students' affective perceptions across reaction mechanisms. The χ^2 test for the multinomial logistic regression for students' confidence was statistically significant ($\chi^2(16) = 32.02$; $p = .010$; $R^2 = 0.041$). Figure 10c presents the specific trajectory differences, which demonstrate the high confidence of the higher-achieving students. Lastly, the χ^2 test for

the multinomial logistic regression for students' perceived task ease was insignificant ($\chi^2(16) = 21.28$; $p = .168$; $R^2 = 0.027$). Figure 10d illustrates the consistent ratings across reaction mechanisms. This showcases that perceived confidence is more strongly associated with students' reasoning complexity than perceived task ease.

Together, the multinomial logistic regression reveals that students' reasoning trajectories are, among others, associated with their prior chemical concept knowledge, within-study chemical concept knowledge, and perceived confidence. In particular, chemical concept knowledge showed stronger associations with reasoning complexity than domain-general argumentation skills.

6 | Discussion

This study used an ML model to longitudinally monitor undergraduate organic chemistry students' mechanistic reasoning as operationalized by the *essential epistemic heuristics* (Section 2). We mapped students' reasoning trajectories and identified key variables contributing to an increased reasoning complexity.

6.1 RQ1: Modeling students' reasoning trajectories

First, our analysis shows that students followed five distinct reasoning trajectories over the semester. None of these trajectories exhibit a strictly linear progression across all mechanisms—particularly the lower-achieving trajectories demonstrate several fluctuations throughout the semester. Consequently, students' mechanistic reasoning seems to follow a non-linear sequence without clearly defined levels of progression. This finding confirms Southard et al.'s (2017) observation in molecular and cellular biology that students' mechanistic reasoning is highly context-dependent. Thus, the development of students' mechanistic reasoning cannot be based on a model of stable-developing cognition. Students' reasoning complexity in one mechanism does not reliably reflect their reasoning complexity in another.

One reason for the variability in students' mechanistic reasoning across mechanisms is that each reaction mechanism in organic chemistry poses unique challenges for students (Deng & Flynn, 2021). Some reaction mechanisms are inherently more complex and require a deeper analysis of multiple steps or the involvement of more advanced chemical concepts (Flynn & Featherstone, 2017). Especially lower-achieving students struggle to determine which chemical concepts to apply within a given reaction mechanism due to challenges in recognizing the underlying connections between those concepts (e.g., Bhattacharyya & Bodner, 2005; Ferguson & Bodner, 2008; Grove et al., 2012). As a result, these students perceive the activation of chemical concepts as entirely different from one reaction mechanism to the next, which leads to persistent contextual challenges.

6.2 RQ2: Changes in claim correctness over time and its relation to reasoning complexity

Based on the reasoning trajectories, we monitored students' claim correctness and found that students improved their ability to make correct claims over the semester. Students with higher reasoning complexity made more correct claims. The identified correlation of reasoning complexity and claim correctness aligns with a qualitative study by Cruz-Ramírez de Arellano and Towns (2014), which demonstrates that undergraduate organic chemistry students who made correct claims about alkyl halide reactions were better at estimating the basicity and nucleophilicity of substances, analyzing intermediates and transition states, and linking entities, activities, and properties.

A more granular analysis of the relationship between reasoning complexity and claim correctness in our context showed that students' mechanistic reasoning was more complex with correct claims in the enamine hydrolysis, aldol condensation, and Claisen condensation; however, this trend did not hold for the ether synthesis (Martin, Yik, et al., under review). This observation may explain the conflicting evidence by Deng and Flynn (2021) and Bodé et al. (2019): Deng and Flynn (2021) found that students' use of causality was significantly higher for students who made correct claims compared to those with incorrect claims in an acid–base context. Conversely, Bodé et al. (2019) found that students' use of causality did not differ significantly between students with correct and incorrect claims when evaluating the plausibility of two unimolecular nucleophilic substitutions. Given that the reactions analyzed by Bodé et al. (2019) and the ether synthesis studied herein involve nucleophilic substitution (S_N) mechanisms, it is unsurprising that these reactions deviate from the observed trend. Students' familiarity with the S_N mechanisms—which are introduced early in undergraduate organic chemistry and follow a clear reaction pattern—may have minimized the influence of claim-making in their decision-making. Rather than engaging in mechanistic reasoning, students may have relied on factual recall to reach conclusions without a time-consuming, complex mechanistic reasoning process. Grove et al. (2012) identified a similar mechanistic reasoning pattern in undergraduate organic chemistry, where students successfully solved simpler problems by relying on factual recall, rather than by applying mechanistic reasoning.

Our findings suggest that instructional strategies aimed at supporting reasoning complexity should also address students' claim correctness—something we did not implement in this study—to help them achieve higher reasoning complexity. In doing so, instruction should guide students in constructing and justifying claims more consciously (McNeill & Pimentel, 2010). Besides, our findings indicate that various skills, such as reasoning complexity and claim correctness, interact differently over time, highlighting the importance of longitudinally examining how cognitive variables interrelate in science learning.

6.3 RQ3: Predicting students' reasoning trajectories based on the *essential epistemic heuristics*

To better understand the structure of mechanistic reasoning in organic chemistry, we subsequently examined the interconnections among the *essential epistemic heuristics* (Krist et al., 2019). Our findings indicate a strong relationship between students' use of granularity and causality. Additionally, integrating thermodynamic and kinetic ideas increased students' reasoning complexity, which suggests an interdependence among the *essential epistemic heuristics*. These results expand previous research that often distinguished between domain-general practices (Machamer et al., 2000; Russ et al., 2008) and domain-specific knowledge (Duncan, 2007) in mechanistic reasoning. Specifically, our findings indicate that domain-general heuristics, as reflected in granularity and causality, are closely linked to domain-specific disciplinary core ideas. This aligns with Haskel-Ittah et al. (2020), who found that applying the heuristic of unpacking factors at the lower scalar level relies on sufficient domain-specific knowledge, suggesting that this heuristic may not be activated without such knowledge. We can even identify which domain-specific knowledge is critical for high reasoning complexity in organic chemistry: thermodynamic and kinetic ideas.

Our findings contribute to existing literature on students' challenges in reasoning about energetics across science domains. For instance, high school chemistry students rarely considered energetic changes during chemical reactions and

seldom referred to energy minimization to explain the driving force of a reaction (Pölloth et al., 2023). Comparably, high school biology students often overlooked energy production in their mechanistic explanations of cellular respiration and struggled to connect energetic processes with mechanistic representations (Ariely et al., 2024). These findings highlight that energy concepts are often weakly connected to other disciplinary core ideas, so students face several challenges in reasoning about energetics (e.g., Podschuweit & Bernholt, 2020). This observation is concerning because energy is a fundamental concept across the sciences, and students who successfully apply energy concepts in one science domain are more likely to transfer their understanding to another science domain (Park & Liu, 2016). For this reason, Kohn et al. (2018) call for a stronger integration of energy concepts across the sciences. Instructors should engage in interdisciplinary dialogue to develop cohesive teaching strategies. In addition, instruction should address thermodynamics and kinetics as central factors initiating natural phenomena, rather than mentioning them as an afterthought.

6.4 RQ4: Predicting students' reasoning trajectories based on cognitive and affective variables

Furthermore, we explored the cognitive and affective variables influencing trajectory differences. Here, students' argumentation skills were almost consistent across trajectories. Domain-general skills such as argumentation are valuable; however, it seems that they do not suffice for complex reasoning about mechanisms in organic chemistry. By contrast, significant variations were found for higher- and lower-achieving students' concept knowledge, which confirms its importance in achieving complex mechanistic reasoning (Kranz et al., 2023). This finding aligns with Grove et al.'s (2012) observation that students who apply mechanistic reasoning can better solve organic chemistry problems that require knowledge transfer. Consequently, emphasizing domain-specific concepts is more decisive in fostering mechanistic reasoning than supporting domain-general argumentation skills, which favors a science learning approach where domain-specific concepts are prioritized over domain-general skills (Duncan, 2007). This approach could enhance future interventions aimed at improving students' mechanistic reasoning. For example, Lieber et al. (2022a, 2022b) provided adaptive support based on students' needs in structuring scientific arguments or applying chemical concepts. However, aiming to deepen students' concept knowledge in advance may be more beneficial than simultaneously supporting scientific argumentation *and* concept knowledge.

Finally, students' confidence in solving organic chemistry tasks correlated with more complex mechanistic reasoning. Gao et al. (2024) found the same relationship in undergraduate organic chemistry. The statistically insignificant χ^2 test for the multinomial logistic regression analyzing students' perceived task ease suggests that students did not feel overwhelmed by the organic chemistry tasks, which indicates that their cognitive load remained manageable.

7 | Implications

Our study provided adaptive support for students' mechanistic reasoning, adapted to their reasoning complexity in the previous mechanism. However, our analysis suggests that students' reasoning complexity in one mechanism does not consistently predict their reasoning complexity in another. Following this, it may be beneficial to support students' mechanistic reasoning first within a single mechanism without immediately moving to the next since a new mechanism can introduce new student challenges. As the curriculum advances and instructors provide adaptive support,

assessing students' mechanistic reasoning based on the cumulative sequence of prior mechanisms, rather than solely on the most recent one, may better address students' needs.

So far, we only provided concept-strategic feedback related to the three *essential epistemic heuristics*—such as guidance on addressing electronic properties, identifying the most nucleophilic center in a molecule, or linking various chemical properties—to increase students' reasoning complexity. Other feedback types, including motivational, corrective, or conceptual feedback (Zech, 1998), could address students' reasoning complexity *and* claim correctness. For example, corrective feedback could prevent the reinforcement of incorrect mechanistic reasoning and may activate more productive ideas in advance. In addition, conceptual feedback—such as definitions of nucleophilicity and electrophilicity—could assist students in revisiting and strengthening key organic chemistry concepts (e.g., Asmussen et al., 2024). Hence, future research could investigate how integrating motivational, corrective, or conceptual feedback influences students' reasoning complexity and claim correctness.

Moreover, our analysis demonstrates that mechanistic reasoning is a multifaceted construct encompassing various interconnected heuristics. To date, we only investigated the association between students' use of granularity, disciplinary core ideas, and causality. Nevertheless, students' mechanistic reasoning may also vary across other dimensions. For example, future studies could examine how students link *structural* and *energetic* properties in their mechanistic reasoning over time. Furthermore, we did not consider reasoning differences for each of the four alternative products in a given mechanism. A more granular analysis of differences in mechanistic reasoning between alternative products could reveal which mechanisms are more accessible to students, which can guide a learning progression that introduces increasingly complex reaction mechanisms.

Finally, we found that variables like students' chemistry-specific concept knowledge and confidence influence their mechanistic reasoning. Consequently, students' mechanistic reasoning should be examined within a broader system of cognitive and affective variables (Figure 1). For example, supporting students' attitudes toward science may impact their mechanistic reasoning (e.g., Collini et al., 2023, 2024). Similarly, the learning objectives in science classes as well as instructors' and students' epistemic beliefs likely shape how students engage with and develop mechanistic reasoning (DeGlopper & Stowe, 2024). Understanding these influences can provide a more holistic view of students' mechanistic reasoning and inform instructional practices that better support mechanistic reasoning in complex science contexts.

8 | Limitations

Several limitations should be considered when interpreting the findings. First, we excluded data from students who did not complete all assignments, which may have introduced a selection bias. Most students who withdrew from the course did so either at the beginning or near the end of the semester. It is possible that these students faced the greatest challenges in the organic chemistry course. As a result, the findings of this study may not fully capture the reasoning trajectories of students who faced the most challenges.

Furthermore, students' reasoning trajectories are based on data collected at only four time points. While these trajectories enabled a detailed analysis of students' mechanistic reasoning over time, future studies could incorporate additional time points and a broader range of reaction mechanisms.

9 | Conclusion

In this article, we operationalized the *essential epistemic heuristics for mechanistic reasoning* (Krist et al., 2019) to examine the development of students' mechanistic reasoning in undergraduate organic chemistry and its connections to other cognitive and affective variables (Figure 1). Our analysis contributes to a deeper understanding of how students' mechanistic reasoning develops and how it interacts with other variables, which may help design instructional approaches that support mechanistic reasoning over an extended period. Our longitudinal analysis revealed five distinct reasoning trajectories among undergraduate organic chemistry students, which showcases the non-linear, context-dependent nature of how students' mechanistic reasoning develops. Importantly, students with higher reasoning complexity were more likely to make correct claims about the target phenomenon. Furthermore, our findings underscore the interconnectedness of granularity and causality—reflecting the first and third heuristics. We also observed challenges among lower-achieving students with thermodynamic and kinetic ideas—linked to the second heuristic. Analyzing students' mechanistic reasoning within a broader system suggests that domain-general skills like argumentation are valuable; however, domain-specific concept knowledge is more important to support complex mechanistic reasoning.

By tracing students' reasoning trajectories, our analysis provides several recommendations for supporting students' mechanistic reasoning. First, support should consider the correctness of students' claims regarding the plausibility of mechanisms since reasoning complexity and claim correctness are interconnected. Further, instruction should address granularity *or* causality but does not need to separate these heuristics given their interrelation. Instruction should also help lower-achieving students connect the disciplinary core idea of *thermodynamics and kinetics* to mechanisms. As chemical concept knowledge plays a more significant role in complex mechanistic reasoning than domain-general argumentation skills, students' chemical concept knowledge should be continuously assessed and supported in formative assessments throughout the semester. Lastly, students' confidence should be fostered by consistently providing feedback that highlights their strengths and areas for future improvement.

Acknowledgments

We thank Peter Wulff and David Kranz for their assistance in implementing the machine learning analysis. Moreover, we thank Leonie Lieber, Ira Caspari-Gnann, and Krenare Ibraj for sharing their research instrument and data. Similarly, we thank Amber Dood, Field Watts, Ginger Shultz, and their research team for providing access to their Writing-to-Learn data to train our machine learning model. We also thank Brett McCollum and Nathan Ackroyd for allowing us to pilot our learning resources and gather data at their institution. Our further appreciation goes to Lukas Neuser, Constantin Stremmer, Marie Sahba, and Felix Blödtner for their assistance in implementing the study, as well as to all members of the Graulich group for productive discussions. Finally, Paul P. Martin thanks the German Chemical Industry Association (Verband der Chemischen Industrie) for supporting him with the Kekulé Fellowship. This publication is part of the first author's doctoral thesis (Dr. rer. nat.) at the Faculty of Biology and Chemistry, Justus-Liebig-University Giessen, Germany.

Conflict of interest statement

The authors Paul P. Martin, Marcus Kubsch, Brandon J. Yik, Benjamin T. Burlingham, and Nicole Graulich declare that they have no conflicts of interest.

REFERENCES

Agresti, A. (2012). *Categorical data analysis* (2nd ed.). Wiley.

Ariely, M., Nazaretsky, T., & Alexandron, G. (2024). Causal-mechanical explanations in biology: Applying automated assessment for personalized learning in the science classroom. *Journal of Research in Science Teaching*, *61*(8), 1858-1889. https://doi.org/10.1002/tea.21929

Asmussen, G., Rodemer, M., & Bernholt, S. (2024). Stepping stones to success: A qualitative investigation of the effectiveness of adaptive stepped supporting tools for problem-solving in organic chemistry to design an intelligent tutoring system. *International Journal of Science Education*, *ASAP Article*, 1-23. https://doi.org/10.1080/09500693.2024.2361933

Bachtiar, R. W., Meulenbroeks, R. F. G., & van Joolingen, W. R. (2022). Mechanistic reasoning in science education: A literature review. *EURASIA Journal of Mathematics, Science and Technology Education*, *18*(11), em2178. https://doi.org/10.29333/ejmste/12512

Beasley, T. M., & Schumacker, R. E. (1995). Multiple regression approach to analyzing contingency tables: Post hoc and planned comparison procedures. *The Journal of Experimental Education*, *64*(1), 79-93. https://doi.org/10.1080/00220973.1995.9943797

Bhattacharyya, G., & Bodner, G. M. (2005). "It gets me to the product": How students propose organic mechanisms. *Journal of Chemical Education*, *82*(9), 1402-1407. https://doi.org/10.1021/ed082p1402

Bodé, N. E., Deng, J. M., & Flynn, A. B. (2019). Getting past the rules and to the WHY: Causal mechanistic arguments when judging the plausibility of organic reaction mechanisms. *Journal of Chemical Education*, *96*(6), 1068-1082. https://doi.org/10.1021/acs.jchemed.8b00719

Braun, I., & Graulich, N. (2024). Exploring diversity: student's (un-)productive use of resonance in organic chemistry tasks through the lens of the coordination class theory. *Chemistry Education Research and Practice*, *25*(3), 643-671. https://doi.org/10.1039/D3RP00298E

Bray, J. H., & Maxwell, S. E. (1985). *Multivariate analysis of variance* (1st ed.). Sage.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience* (1st ed.). Clarendon Press.

Choi, A., Hand, B., & Greenbowe, T. (2013). Students' written arguments in general chemistry laboratory investigations. *Research in Science Education*, *43*, 1763-1783. https://doi.org/10.1007/s11165-012-9330-1

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) Routledge.

Collini, M. A., Miguel, K., Weber, R., & Atkinson, M. B. (2024). Investigating changes in students' attitudes towards organic chemistry: a longitudinal study. *Chemistry Education Research and Practice*, *25*(3), 613-624. https://doi.org/10.1039/D3RP00228D

Collini, M. A., Rocha, L. A., Ford, J. E., Weber, R., & Atkinson, M. B. (2023). Characterizing and identifying influences on undergraduates' attitudes towards organic chemistry. *Chemistry Education Research and Practice*, *24*(2), 723-739. https://doi.org/10.1039/D2RP00256F

Cooper, M. M., Posey, L. A., & Underwood, S. M. (2017). Core ideas and topics: Building up or drilling down? *Journal of Chemical Education*, *94*(5), 541-548. https://doi.org/10.1021/acs.jchemed.6b00900

Crandell, O. M., Kouyoumdjian, H., Underwood, S. M., & Cooper, M. M. (2019). Reasoning about reactions in organic chemistry: Starting it in general chemistry. *Journal of Chemical Education*, *96*(2), 213-226. https://doi.org/10.1021/acs.jchemed.8b00784

Crandell, O. M., Lockhart, M. A., & Cooper, M. M. (2020). Arrows on the page are not a good gauge: Evidence for the importance of causal mechanistic explanations about nucleophilic substitution in organic chemistry. *Journal of Chemical Education*, *97*(2), 313-327. https://doi.org/10.1021/acs.jchemed.9b00815

Cruz-Ramírez de Arellano, D., & Towns, M. H. (2014). Students' understanding of alkyl halide reactions in undergraduate organic chemistry. *Chemistry Education Research and Practice*, *15*(4), 501-515. https://doi.org/10.1039/C3RP00089C

Darden, L. (2002). Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward/backward chaining.

Philosophy of Science, *69*(S3), S354–S365. https://doi.org/10.1086/341858

DeGlopper, K. S., & Stowe, R. L. (2024). Modeling students' epistemic cognition in undergraduate chemistry courses: a review. *Chemistry Education Research and Practice*, *25*(3), 594-612. https://doi.org/10.1039/D3RP00348E

Den Teuling, N. G. P., Pauws, S. C., & van den Heuvel, E. R. (2023). A comparison of methods for clustering longitudinal data with slowly changing trends. *Communications in Statistics–Simulation and Computation*, *52*(3), 621-648. https://doi.org/10.1080/03610918.2020.1861464

Deng, J. M., Carle, M. S., & Flynn, A. B. (2022). Students' reasoning in chemistry arguments and designing resources using constructive alignment. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (1st ed., pp. 74-89). The Royal Society of Chemistry. https://doi.org/10.1039/9781839167782-00074

Deng, J. M., & Flynn, A. B. (2021). Reasoning, granularity, and comparisons in students' arguments on two organic chemistry items. *Chemistry Education Research and Practice*, *22*(3), 749-771. https://doi.org/10.1039/D0RP00320D

Dood, A. J., Dood, J. C., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2020). Using the research literature to develop an adaptive intervention to improve student explanations of an S_N1 reaction mechanism. *Journal of Chemical Education*, *97*(10), 3551-3562. https://doi.org/10.1021/acs.jchemed.0c00569

Duncan, R. G. (2007). The role of domain-specific knowledge in generative reasoning about complicated multileveled phenomena. *Cognition and Instruction*, *25*(4), 271-336. https://doi.org/10.1080/07370000701632355

Ferguson, R., & Bodner, G. M. (2008). Making sense of the arrow-pushing formalism among chemistry majors enrolled in organic chemistry. *Chemistry Education Research and Practice*, *9*(2), 102-113. https://doi.org/10.1039/B806225K

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R* (1st ed.). Sage.

Flynn, A. B., & Featherstone, R. B. (2017). Language of mechanisms: Exam analysis reveals students' strengths, strategies, and errors when using the electron-pushing formalism (curved arrows) in new reactions. *Chemistry Education Research and Practice*, *18*(1), 64-77. https://doi.org/10.1039/C6RP00126B

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., & Graves, S. (2007). *Package 'car'*. [Computer Program].

Franovic, C. G. C., Noyes, K., Stoltzfus, J. R., Schwarz, C. V., Long, T. M., & Cooper, M. M. (2023). Undergraduate chemistry and biology students' use of causal mechanistic reasoning to explain and predict preferential protein-ligand binding activity. *Journal of Chemical Education*, *100*(5), 1716-1727. https://doi.org/10.1021/acs.jchemed.2c00737

Gao, S., Outlaw, T. C., Liang-Lin, J. G., Feng, A., Shimomura, R., Roizen, J. L., & Cox, C. T. (2024). Analysis of resources applied to rationalize elimination mechanisms. *Chemistry Education Research and Practice*, *25*(1), 62-78. https://doi.org/10.1039/D3RP00031A

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d: R packages to cluster longitudinal data. *Journal of statistical software*, *65*(4), 1-34. https://doi.org/10.18637/jss.v065.i04

Genolini, C., & Falissard, B. (2010). KmL: k-means for longitudinal data. *Computational Statistics*, *25*(2), 317-328. https://doi.org/10.1007/s00180-009-0178-4

Genolini, C., & Falissard, B. (2011). Kml: A package to cluster longitudinal data. *Computer Methods and Programs in Biomedicine*, *104*(3), e112-e121. https://doi.org/10.1016/j.cmpb.2011.05.008

Glasson, G. E. (1989). The effects of hands-on and teacher demonstration laboratory methods on science achievement in relation to reasoning ability and prior knowledge. *Journal of Research in Science Teaching*, *26*(2), 121-131. https://doi.org/10.1002/tea.3660260204

Glennan, S., & Illari, P. (2017). Varieties of mechanisms. In S. Glennan & P. Illari (Eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (1st ed., pp. 91–103). Routledge.

Goodwin, W. M. (2008). Structural formulas and explanation in organic chemistry. *Foundations of*

Chemistry, *10*, 117-127. https://doi.org/10.1007/s10698-007-9033-2

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, *53*(3-4), 325-338. https://doi.org/10.1093/biomet/53.3-4.325

Graulich, N., Hedtrich, S., & Harzenetter, R. (2019). Explicit versus implicit similarity—Exploring relational conceptual understanding in organic chemistry. *Chemistry Education Research and Practice*, *20*(4), 924-936. https://doi.org/10.1039/C9RP00054B

Graulich, N., & Schween, M. (2018). Concept-oriented task design: Making purposeful case comparisons in organic chemistry. *Journal of Chemical Education*, *95*(3), 376-383. https://doi.org/10.1021/acs.jchemed.7b00672

Grotzer, T. A. (2003). Learning to understand the forms of causality implicit in scientifically accepted explanations. *Studies in Science Education*, *39*(1), 1-74. https://doi.org/10.1080/03057260308560195

Grove, N. P., Cooper, M. M., & Cox, E. L. (2012). Does mechanistic thinking improve student success in organic chemistry? *Journal of Chemical Education*, *89*(7), 850-853. https://doi.org/10.1021/ed200394d

Haas, D. B., Watts, F. M., Dood, A. J., & Shultz, G. V. (2024). Analysis of organic chemistry students' developing reasoning elicited by a scaffolded case comparison activity. *Chemistry Education Research and Practice*, *25*(3), 742-759. https://doi.org/10.1039/D4RP00021H

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society*, *28*(1), 100-108. https://doi.org/10.2307/2346830

Haskel-Ittah, M. (2023). Explanatory black boxes and mechanistic reasoning. *Journal of Research in Science Teaching*, *60*(4), 915-933. https://doi.org/10.1002/tea.21817

Haskel-Ittah, M., Duncan, R. G., Vázquez-Ben, L., & Yarden, A. (2020). Reasoning about genetic mechanisms: Affordances and constraints for learning. *Journal of Research in Science Teaching*, *57*(3), 342-367. https://doi.org/10.1002/tea.21595

Hedtrich, S., & Graulich, N. (2018a). Translating numbers into feedback: Providing students with automatically generated feedback. *American Journal of Educational Research*, *6*(2), 108-116. https://doi.org/10.12691/education-6-2-3

Hedtrich, S., & Graulich, N. (2018b). Using software tools to provide students in large classes with individualized formative feedback. *Journal of Chemical Education*, *95*(12), 2263-2267. https://doi.org/10.1021/acs.jchemed.8b00173

Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*, 119-135. https://doi.org/10.1007/s13194-011-0038-2

Johnson, M. A., & Lawson, A. E. (1998). What are the relative effects of reasoning ability and prior knowledge on biology achievement in expository and inquiry classes? *Journal of Research in Science Teaching*, *35*(1), 89-103. https://doi.org/10.1002/(SICI)1098-2736(199801)35:1<89::AID-TEA6>3.0.CO;2-J

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford publications.

Kohn, K. P., Underwood, S. M., & Cooper, M. M. (2018). Energy connections and misconnections across chemistry and biology. *CBE—Life Sciences Education*, *17*(1), ar3,1-ar3,17. https://doi.org/10.1187/cbe.17-08-0169

Kraft, A., Strickland, A. M., & Bhattacharyya, G. (2010). Reasonable reasoning: multi-variate problem-solving in organic chemistry. *Chemistry Education Research and Practice*, *11*(4), 281-292. https://doi.org/10.1039/C0RP90003F

Kranz, D., Schween, M., & Graulich, N. (2023). Patterns of reasoning—Exploring the interplay of students' work with a scaffold and their conceptual knowledge in organic chemistry. *Chemistry Education Research and Practice*, *24*(2), 453-477. https://doi.org/10.1039/D2RP00132B

Krist, C., Schwarz, C. V., & Reiser, B. J. (2019). Identifying essential epistemic heuristics for guiding mechanistic reasoning in science learning. *Journal of the Learning Sciences*, *28*(2), 160-205. https://doi.org/10.1080/10508406.2018.1510404

Lieber, L. S., & Graulich, N. (2020). Thinking in alternatives—A task design for challenging students' problem-solving approaches in organic chemistry. *Journal of Chemical Education*,

97(10), 3731-3738. https://doi.org/10.1021/acs.jchemed.0c00248

Lieber, L. S., & Graulich, N. (2022). Investigating students' argumentation when judging the plausibility of alternative reaction pathways in organic chemistry. *Chemistry Education Research and Practice*, *23*(1), 38-53. https://doi.org/10.1039/D1RP00145K

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022a). Closing the gap of organic chemistry students' performance with an adaptive scaffold for argumentation patterns. *Chemistry Education Research and Practice*, *23*(4), 811-828. https://doi.org/10.1039/D2RP00016D

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022b). Students' individual needs matter: A training to adaptively address students' argumentation skills in organic chemistry. *Journal of Chemical Education*, *99*(7), 2754-2761. https://doi.org/10.1021/acs.jchemed.2c00213

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of internal clustering validation measures. In G. I. Webb, B. Liu, C. Zhang, D. Gunopulos & X. Wu (Eds.), *2010 IEEE International Conference on Data Mining* (10th ed., pp. 911-916). IEEE. https://doi.org/10.1109/ICDM.2010.35

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1-25. https://doi.org/10.1086/392759

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior research methods*, *52*, 464-488. https://doi.org/10.3758/s13428-019-01246-w

Martin, P. P., Dood, A. J., Watts, F. M., Shultz, G., & Graulich, N. (under review). To what extent can we trust a machine? Establishing trustworthiness guidelines to evaluate machine learning models across chemistry contexts. *Journal of Research in Science Teaching*.

Martin, P. P., & Graulich, N. (2023). When a machine detects student reasoning: a review of machine learning-based formative assessment of mechanistic reasoning. *Chemistry Education Research and Practice*, *24*(2), 407-427. https://doi.org/10.1039/D2RP00287F

Martin, P. P., & Graulich, N. (2024a). Lehre in der organischen Chemie individualisieren [Individualized teaching in organic chemistry]. *Nachrichten aus der Chemie*, *72*(3), 8-11. https://doi.org/10.1002/nadc.20244141003

Martin, P. P., & Graulich, N. (2024b). Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence. *Computers and Education: Artificial Intelligence*, *7*, Article 100265. https://doi.org/10.1016/j.caeai.2024.100265

Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2024). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*, *61*(8), 1757-1792. https://doi.org/10.1002/tea.21903

Martin, P. P., Yik, B. J., Burlingham, B. T., & Graulich, N. (under review). Reasoning about alternatives: How students reason about plausible and implausible mechanisms in chemistry. In R. W. Bachtiar, M. Haskel-Ittah, & N. Graulich (Eds.), *Mechanistic Reasoning in Science Education*.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia Medica*, *23*(2), 143-149. https://doi.org/10.11613/bm.2013.018

McNeill, K. L., & Krajcik, J. (2011). *Supporting grade 5-8 students in constructing explanations in science: The claim, evidence, and reasoning framework for talk and writing*. Pearson Allyn & Bacon.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences*, *15*(2), 153-191. https://doi.org/10.1207/s15327809jls1502_1

McNeill, K. L., & Pimentel, D. S. (2010). Scientific discourse in three urban classrooms: The role of the teacher in engaging high school students in argumentation. *Science Education*, *94*(2), 203-229. https://doi.org/10.1002/sce.20364

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Sage.

Moon, A., Moeller, R., Gere, A. R., & Shultz, G. V. (2019). Application and testing of a framework for characterizing the quality of scientific reasoning in chemistry students' writing on ocean acidification. *Chemistry Education Research and Practice*, *20*(3), 484-494. https://doi.org/10.1039/C9RP00005D

Moon, A., Stanford, C., Cole, R., & Towns, M. (2016). The nature of students' chemical reasoning employed in scientific argumentation in physical chemistry. *Chemistry Education Research and Practice*, *17*(2), 353-364. https://doi.org/10.1039/C5RP00207A

Moreira, P., Marzabal, A., & Talanquer, V. (2019). Using a mechanistic framework to characterise chemistry students' reasoning in written explanations. *Chemistry Education Research and Practice*, *20*(1), 120-131. https://doi.org/10.1039/C8RP00159F

National Research Council [NRC]. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

Nieswandt, M. (2007). Student affect and conceptual understanding in learning chemistry. *Journal of Research in Science Teaching*, *44*(7), 908-937. https://doi.org/10.1002/tea.20169

Noyes, K., & Cooper, M. M. (2019). Investigating student understanding of London dispersion forces: A longitudinal study. *Journal of Chemical Education*, *96*(9), 1821-1832. https://doi.org/10.1021/acs.jchemed.9b00455

Organization for Economic Co-operation and Development [OECD]. (2023). *PISA 2025 science framework.* Oxford University Press.

Pabuccu, A., & Erduran, S. (2017). Beyond rote learning in organic chemistry: The infusion and impact of argumentation in tertiary education. *International Journal of Science Education*, *39*(9), 1154-1172. https://doi.org/10.1080/09500693.2017.1319988

Park, M., & Liu, X. (2016). Assessing understanding of the energy concept in different science disciplines. *Science Education*, *100*(3), 483-516. https://doi.org/10.1002/sce.21211

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (32nd ed., pp. 8024-8035). Curran Associates Inc. https://doi.org/10.48550/arXiv.1912.01703

Patefield, W. (1981). Algorithm AS 159: An efficient method of generating random R×C tables with given row and column totals. *Journal of the Royal Statistical Society*, *30*(1), 91-97. https://doi.org/10.2307/2346669

Pinheiro, J., Bates, D., & R Core Team (2011). *Package 'nlme: Linear and nonlinear mixed effects models'*. [Computer Program].

Podschuweit, S., & Bernholt, S. (2020). Investigating network coherence to assesss students' conceptual understanding of energy. *Education Sciences*, *10*(4), Article 103. https://doi.org/10.3390/educsci10040103

Pölloth, B., Diekemper, D., & Schwarzer, S. (2023). What resources do high school students activate to link energetic and structural changes in chemical reactions?—A qualitative study. *Chemistry Education Research and Practice, 24*(4), 1153-1173. https://doi.org/10.1039/D3RP00068K

R Core Team (2023). *R: A language and environment for statistical computing.* [Computer Program].

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). *Package 'mass'*. [Computer Program].

RStudio Team (2023). *RStudio: Integrated development environment for R*. [Computer Program].

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science education*, *92*(3), 499-525. https://doi.org/10.1002/sce.20264

Sevian, H., & Talanquer, V. (2014). Rethinking chemistry: a learning progression on chemical thinking. *Chemistry Education Research and Practice*, *15*(1), 10-23. https://doi.org/10.1039/C3RP00111C

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*(318), 626-633. https://doi.org/10.1080/01621459.1967.10482935

Shiroda, M., Franovic, C. G.-C., de Lima, J., Noyes, K., Babi, D., Beltran-Flores, E., Kesh, J., McKay, R. L., Persson-Gordon, E., Cooper, M. M., Long, T. M.,

Schwarz, C. V., & Stoltzfus, J. R. (2024). Examining and supporting mechanistic explanations across chemistry and biology courses. *CBE—Life Sciences Education*, *23*(3), ar38,1-ar38,21. https://doi.org/10.1187/cbe.23-08-0157

Shtechman, Y., Ergazaki, M., & Haskel-Ittah, M. (2024). Mechanistic reasoning in biology among elementary school children: Epistemic criteria for evaluation and construction. *Journal of the Learning Sciences*, 1-36. https://doi.org/10.1080/10508406.2024.2432682

Southard, K. M., Espindola, M. R., Zaepfel, S. D., & Bolger, M. S. (2017). Generative mechanistic explanation building in undergraduate molecular and cellular biology. *International Journal of Science Education*, *39*(13), 1795-1829. https://doi.org/10.1080/09500693.2017.1353713

Talanquer, V. (2018a). Exploring mechanistic reasoning in chemistry. In J. Yeo, T. W. Teo, & K.-S. Tang (Eds.), *Science education research and practice in Asia-Pacific and beyond* (1st ed., pp. 39-52). Springer Singapore. https://doi.org/10.1007/978-981-10-5149-4_3

Talanquer, V. (2018b). Importance of understanding fundamental chemical mechanisms. *Journal of Chemical Education*, *95*(11), 1905-1911. https://doi.org/10.1021/acs.jchemed.8b00508

Twisk, J. W. R. (2019). *Applied mixed model analysis: a practical guide* (2nd ed.). Cambridge University Press.

van Mil, M. H., Boerwinkel, D. J., & Waarlo, A. J. (2013). Modelling molecular mechanisms: A framework of scientific reasoning to construct molecular-level explanations for cellular behaviour. *Science & Education*, *22*, 93-118. https:/doi.org/10.1007/s11191-011-9379-7

van Mil, M. H., Postma, P. A., Boerwinkel, D. J., Klaassen, K., & Waarlo, A. J. (2016). Molecular mechanistic reasoning: Toward bridging the gap between the molecular and cellular levels in life science education. *Science Education*, *100*(3), 517-585. https://doi.org/10.1002/sce.21215

Vermunt, J. K. (2010). Latent class modeling with variables: Two improved three-step approaches. *Political Analysis*, *18*(4), 450-469. https://doi.org/10.1093/pan/mpq025

Warnes, G. R., Bolker, B., Lumley, T., Warnes, M. G. R., & Imports, M. (2018). *Package 'gmodels'*. [Computer Program].

Watts, F. M., Dood, A. J., & Shultz, G. V. (2022). Developing machine learning models for automated analysis of organic chemistry students' written descriptions of organic reaction mechanisms. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (1st ed., pp. 285-303). The Royal Society of Chemistry. https://doi.org/10.1039/9781839167782-00285

Watts, F. M., Park, G. Y., Petterson, M. N., & Shultz, G. V. (2022). Considering alternative reaction mechanisms: Students' use of multiple representations to reason about mechanisms for a writing-to-learn assignment. *Chemistry Education Research and Practice*, *23*(2), 486-507. https://doi.org/10.1039/D1RP00301A

Watts, F. M., Zaimi, I., Kranz, D., Graulich, N., & Shultz, G. V. (2021). Investigating students' reasoning over time for case comparisons of acyl transfer reaction mechanisms. *Chemistry Education Research and Practice*, *22*(2), 364-381. https://doi.org/10.1039/D0RP00298D

Weinrich, M. L., & Sevian, H. (2017). Capturing students' abstraction while solving organic reaction mechanism problems across a semester. *Chemistry Education Research and Practice*, *18*(1), 169-190. https://doi.org/10.1039/C6RP00120C

Weinrich, M. L., & Talanquer, V. (2016). Mapping students' modes of reasoning when thinking about chemical reactions used to make a desired product. *Chemistry Education Research and Practice*, *17*(2), 394-406. https://doi.org/10.1039/C5RP00208G

Wilcox, R. R., & Schönbrodt, F. (2019). *Package 'WRS: A package of R. R. Wilcox' robust statistics functions'*. [Computer Program].

Yan, F., & Talanquer, V. (2015). Students' ideas about how and why chemical reactions happen: Mapping the conceptual landscape. *International Journal of Science Education*, *37*(18), 3066-3092. https://doi.org/10.1080/09500693.2015.1121414

Zech, F. (1998). *Grundkurs Mathematikdidaktik: Theoretische und praktische Anleitungen für das*

Lehren und Lernen von Mathematik [Basics in math education: Theoretical and practical instructions for teaching and learning math] (1st ed.). Julius Beltz. KG.

SUPPORTING INFORMATION A: DETAILS ON LONGITUDINAL CLUSTER ANALYSIS

Evaluating the output of longitudinal cluster analyses involves several critical decisions that can significantly affect the results and potentially lead to varying conclusions. To enhance the transparency of our findings, we report the specific choices made throughout the analysis. In doing so, we adhere to the Guidelines for Reporting on Latent Trajectory Studies by van de Schoot et al. (2017), which aim to standardize the reporting of latent trajectory studies for increased consistency.

DISTRIBUTION OF THE OBSERVED VARIABLES

The outcome variable in longitudinal cluster analyses can take various forms, so evaluating its statistical properties is essential. In the following sections, we examine students' reasoning complexity concerning (1) data independence, (2) measurement scale, (3) normal distribution, and (4) homogeneity of variance and covariance matrices.

Data independence: In our longitudinal study, data from different students is considered independent because each student's reasoning complexity is assumed to be unrelated to any other student's reasoning complexity. However, as students are measured multiple times in our repeated-measures design, scores within a single student are dependent.

Measurement scale: We performed the statistical analysis at the student level. By doing so, we calculated the mean scores of students' reasoning complexity separated by reaction mechanisms. For instance, a student could have constructed eight responses with complexity scores of 2, 2, 3, 3, 4, 4, 4, and 5, resulting in a total reasoning complexity of 3.375. Consequently, our data is measured at the interval level.

Normal distribution: Data should be normally distributed at each level of the design. No skewness or kurtosis was found since all values were below the thresholds of 1.156 and 2.294, respectively (Kline, 2023). However, the Shapiro-Wilk test applied to the residuals from the factorial ANOVA model indicated a significant deviation from normality ($W = 0.991$, $p < .001$). Boxplots for each reaction mechanism revealed no outliers.

Homogeneity of variance and covariance matrices: Given the between-subjects nature of the design, we assessed the homogeneity of variance assumption using Levene's test, which identified violations for the ether synthesis ($F(4,261) = 4.77$, $p = .001$), aldol condensation ($F(4,261) = 3.79$, $p = .005$), and Claisen condensation ($F(4,261) = 3.42$, $p = .010$) but not for the enamine hydrolysis ($F(4,261) = 0.26$, $p = .904$). Finally, Box's M-test showed a violation of the homogeneity of the covariance matrix assumption ($M = 82.61$, $p < .001$). Given these violations, we also performed robust tests.

ALTERNATIVE SPECIFICATIONS OF TRAJECTORIES

As an alternative approach to modeling reasoning trajectories, we also applied a latent class growth analysis and growth mixture modeling using the R-package *lcmm* (Proust-Lima et al., 2017). Both approaches estimate latent classes with specific trajectories and account for the probability of class membership and variability within classes. A two-class growth mixture model with random intercept, fixed slope, and a cubic trajectory fitted the data best as indicated by the lowest Bayesian Information Criterion (BIC). This model categorized students into a high-achieving (94.7%) and a low-achieving trajectory (5.3%). However, we

decided not to use this model for further analyses for two reasons: First, the assumption that students' mechanistic reasoning follows cubic trajectories is probably not easily transferable across organic chemistry contexts. Second, the large proportion of students in the high-achieving trajectory covers individual differences, so this model is rather useless for analyzing nuanced variations in students' reasoning complexity as envisioned in this article. Given these limitations, we applied the longitudinal *k*-means analysis to model students' reasoning trajectories. This approach models piecewise linear trajectories, which offers better transferability and a more nuanced analysis.

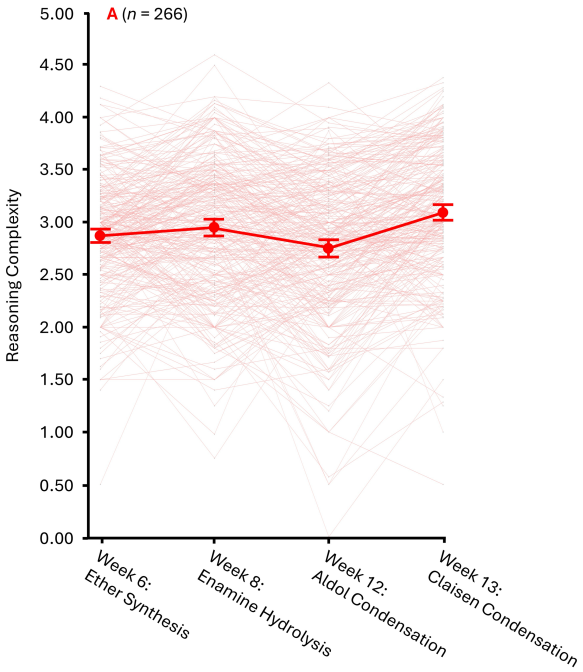


FIGURE S1 Reasoning trajectory aggregated for the entire student cohort.

ALTERNATIVE TRAJECTORY SOLUTIONS

A goal of longitudinal cluster analyses is to determine the optimal number of trajectories. An approach that begins with a one-trajectory solution and incrementally adds trajectories to determine which model best fits the data is often recommended (van de Schoot et al., 2017). Accordingly, we calculated a one-trajectory solution and repeated the longitudinal *k*-means analysis with cluster numbers ranging from two to ten, using 100 different starting points for each of the nine cluster numbers. Subsequently, we present the visualizations of students' reasoning trajectories for an increasing number.

<u>One-trajectory solution:</u>

The one-trajectory solution represents the aggregated reasoning trajectory across all students over one semester, measured across four reaction mechanisms: ether synthesis, enamine hydrolysis, aldol condensation, and Claisen condensation (Figure S1). Table 3 of the main article details descriptive statistics of this trajectory. The trajectory reveals that while students' reasoning complexity varies by reaction mechanisms, it tends to improve over time. However, aldol condensation emerged as the most challenging mechanism for most students despite adaptive support.

<u>Two-trajectory solution:</u>

The two-trajectory solution distinguishes between two groups (Figure S2). Trajectory A (red) starts at a complexity level of approximately 3.1, maintaining relative stability with only minor fluctuations and an increase toward Claisen condensation. In contrast, the lower-achieving trajectory B (yellow) starts at a complexity level of 2.5, shows two drops during enamine hydrolysis and aldol condensation but demonstrates substantial improvement by the end of the semester.

<u>Three-trajectory solution:</u>

Figure S3 illustrates the three-trajectory solution. Trajectory A (green) represents the highest-achieving students. This trajectory's reasoning complexity starts at approximately 3.4 and remains consistently high across all four mechanisms. Students assigned to this trajectory improve their reasoning complexity throughout the semester. Trajectory B (red) exhibits

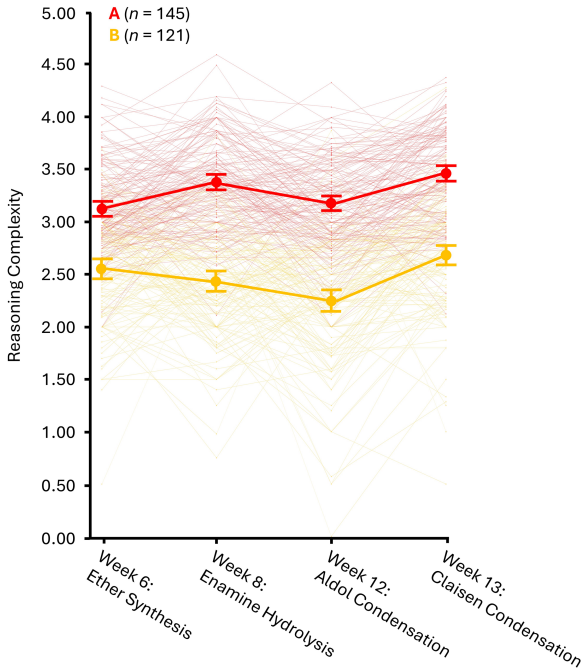


FIGURE S2 Reasoning trajectories of the two-trajectory solution.

moderate reasoning complexity, which begins around 2.9, decreases during aldol condensation, and increases during Claisen Condensation. Students assigned to this trajectory show high variation in reasoning complexity. Finally, the lowest-achieving trajectory C (yellow) corresponds roughly to the yellow trajectory B of the two-trajectory solution (Figure S2). Both trajectories start with a reasoning complexity score of around 2.5, demonstrate small declines during enamine hydrolysis and aldol condensation, but increase their reasoning complexity toward Claisen condensation.

Four-trajectory solution:

The trajectories visualized in Figure S4 categorize the students into four trajectories. Trajectory A (yellow) of the four-trajectory solution closely aligns with the green trajectory A of the three-trajectory solution while the red trajectories in both solutions match each other. Thus, the four-trajectory solution reveals that the yellow trajectory C of the three-trajectory solution contains two distinct patterns: Trajectory C (green) in the four-trajectory solution starts with a reasoning complexity of around 2.5. This trajectory shows a decline in complexity during enamine hydrolysis, where scores reach their lowest point, followed by an increase during the subsequent condensation reactions. In turn,

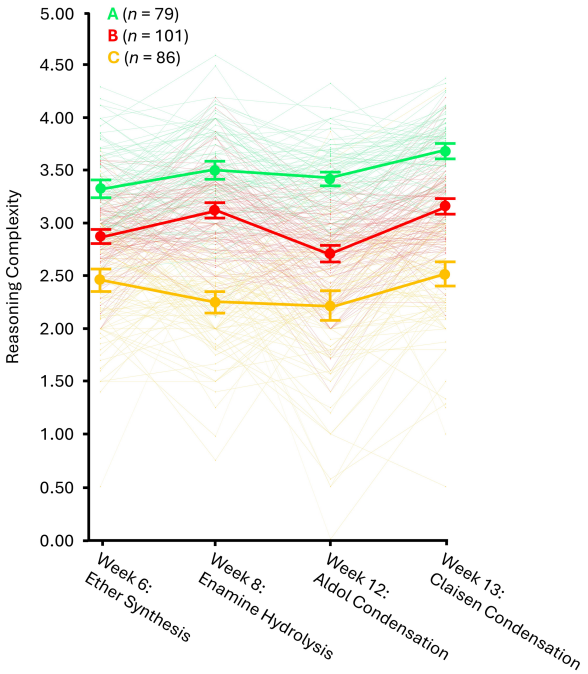


FIGURE S3 Reasoning trajectories of the three-trajectory solution.

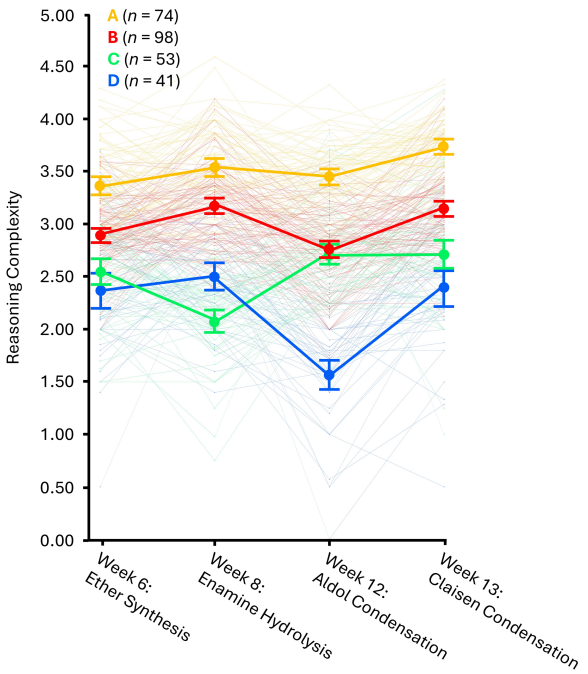


FIGURE S4 Reasoning trajectories of the four-trajectory solution.

trajectory D (blue), which begins with the lowest initial reasoning complexity, displays a downward trend during aldol condensation, before showing a complexity increase during Claisen condensation.

Five-trajectory solution:

The five-trajectory solution is visualized and discussed in the main article (Figure 6b) and in Figure S5. The yellow, red, and green trajectories are robust across the four- and five-trajectory solutions, where the red trajectory of the five-trajectory solution is more consistent than its red counterpart in the four-trajectory solution. The more granular classification of students' reasoning complexity leads to the blue and purple trajectories.

Six-trajectory solution:

An analysis of the six-trajectory solution indicates that most trajectories identified in the five-trajectory solution are robust when the cluster number is increased (Figure S6): the yellow, blue, green, and purple trajectories in the five-trajectory solution correspond to the green,

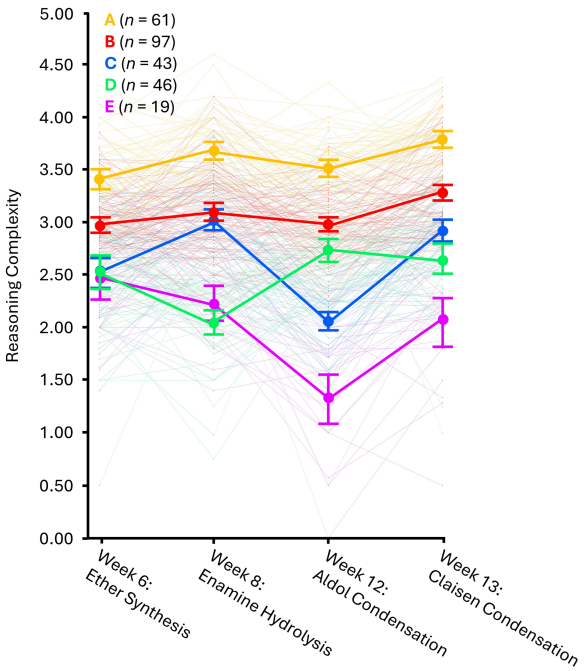


FIGURE S5 Reasoning trajectories of the five-trajectory solution.

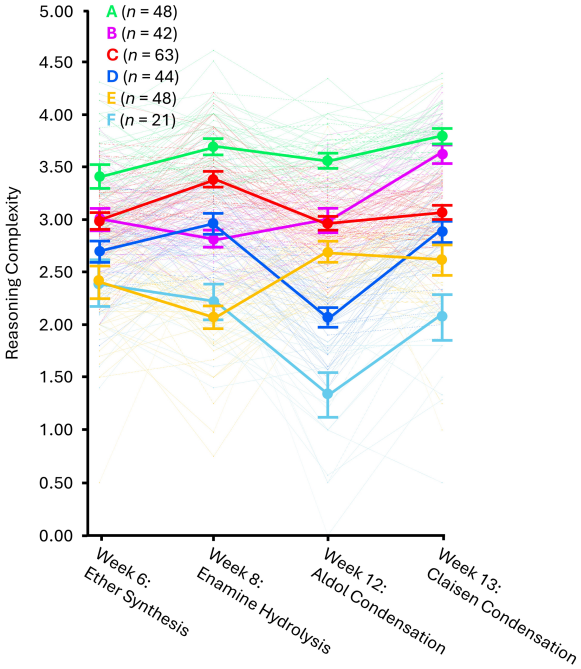


FIGURE S6 Reasoning trajectories of the six-trajectory solution.

blue, yellow, and light blue trajectories in the six-trajectory solution. The only change from the five- to the six-trajectory solution is the disaggregation of the red trajectory, represented by the red and purple trajectories in the six-trajectory solution. The consistency of four trajectories across both solutions demonstrates the robustness of the five-trajectory solution, which is why introducing additional trajectories does not reveal substantial new insights into students' reasoning complexity.

QUALITY CRITERIA TO COMPARE TRAJECTORY FIT

We applied several statistical criteria to identify the optimal trajectory number. These quality criteria characterize a "good" solution as one that models trajectories that are compact and well-separated from one another. Thus, most criteria measure within-trajectory compactness as well as between-trajectory separation and calculate their ratio. We applied five non-parametric and five parametric quality criteria to identify the best-fitting trajectory solution

(Genolini et al., 2015): the Calinski-Harabasz criterion, the Calinski-Harabasz-Kryszczuk variant, the Calinski-Harabasz-Genolini variant, the Ray-Turi criterion, the Davies-Bouldin criterion, two variants of the BIC assuming different numbers of independent observations, and three variants of the Akaike Information Criterion (AIC) also based on different assumptions about the number of independent observations. Among these criteria, some need to be maximized, others should be minimized. However, the *kml* package calculates the criteria that should be maximized directly and provides the inverse of those that should be minimized. Consequently, all criteria are treated as if they should be maximized, facilitating their comparison (Figures S7 and S8).

The quality criteria disagree on the optimal number of trajectories (Figures S6 and S7). Five criteria support a ten-trajectory solution, three favor a two-trajectory solution, one suggests a four-trajectory solution, and one recommends a five-trajectory solution. Despite this disagreement, none of the criteria indicates a poor fit of the five-trajectory solution. The acceptable statistical adequacy of the five-trajectory solution aligns with its high interpretability, specificity, and robustness as discussed above, which is why we decided to proceed with the five-trajectory solution in subsequent analyses.

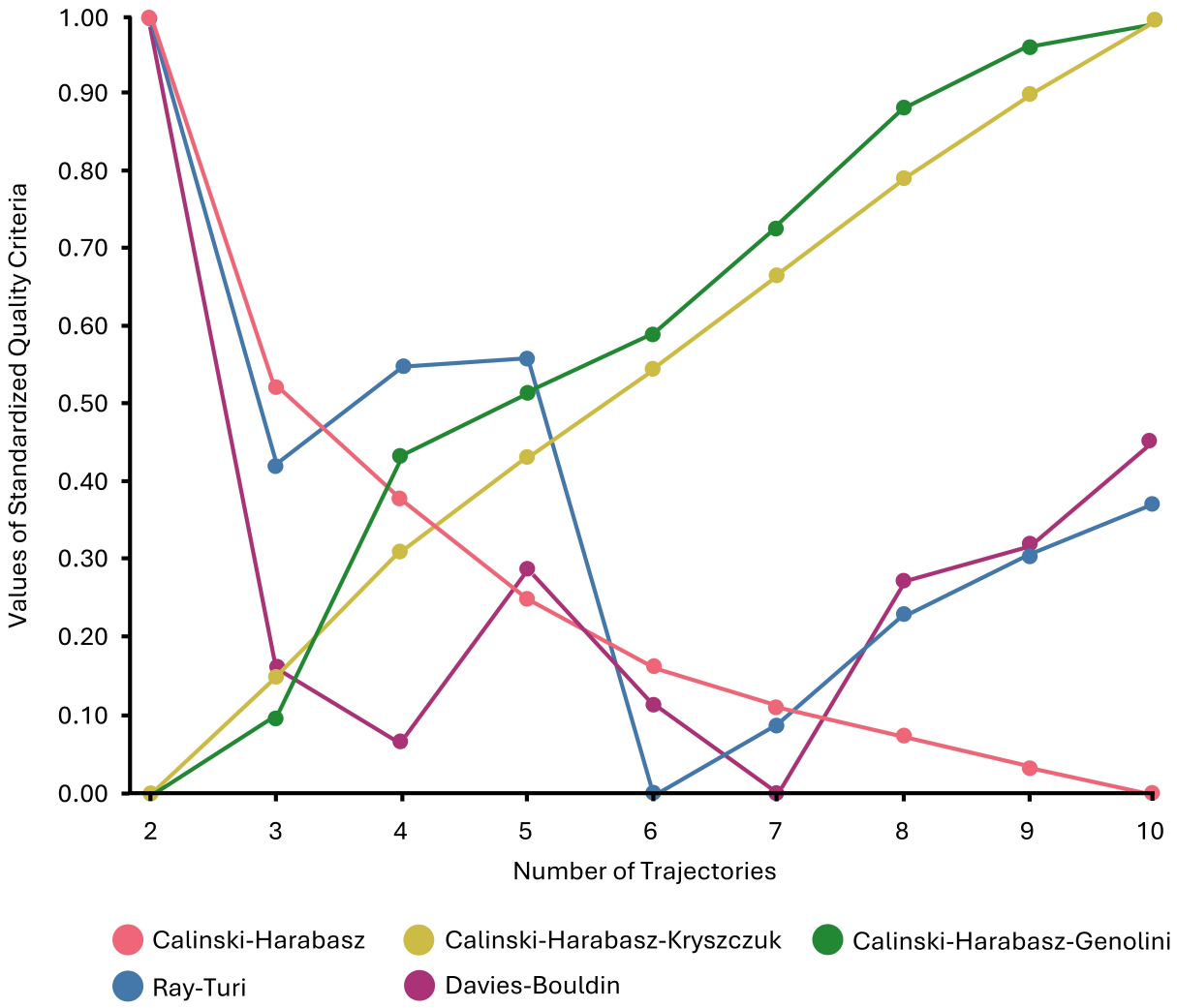


FIGURE S7 Values of the five non-parametric standardized quality criteria evaluating the fit of solutions with two to ten trajectories.

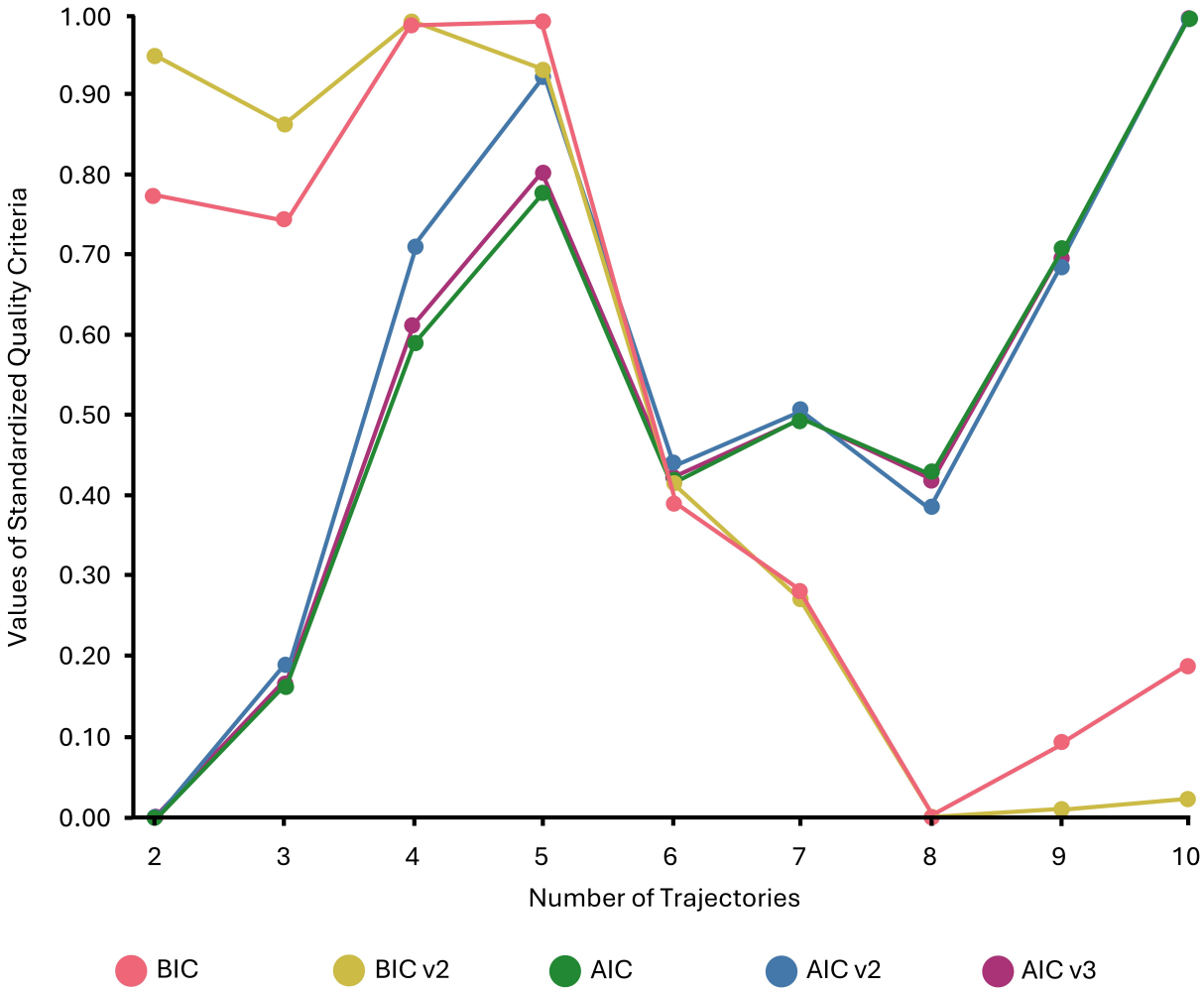


FIGURE S8 Values of the five parametric standardized quality criteria evaluating the fit of solutions with two to ten trajectories.

REFERENCES

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d: R packages to cluster longitudinal data. *Journal of statistical software*, *65*(4), 1-34. https://doi.org/10.18637/jss.v065.i04

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford publications.

Proust-Lima, C., Philipps, V., Diakite, A., & Liquet, B. (2017). *Package 'lcmm: Extended mixed models using latent classes and latent processes'*. [Computer Program].

van de Schoot, R., Sijbrandij, M., Winter, S. D., Depaoli, S., & Vermunt, J. K. (2017). The GRoLTS-Checklist: Guidelines for reporting on latent trajectory studies. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(3), 451-467. https://doi.org/10.1080/10705511.2016.1247646

SUPPORTING INFORMATION B: HIERARCHICAL LINEAR MODELING TO ANALYZE STUDENTS' REASONING TRAJECTORIES

We applied hierarchical linear models (HLMs) with random effects to investigate the trajectories of students' reasoning complexity. HLMs facilitate flexible analysis of hierarchically structured repeated-measures data (Hox et al., 2017; Raudenbush & Bryk, 2002). In our study, students judged the plausibility of alternative reaction products in four assignments throughout a semester. The analysis was conducted at the student level with repeated measures from the 266 students (level 1)

grouped into the five reasoning trajectories (level 2). We used a two-level model with students nested within trajectories and time as a predictor to reflect the longitudinal nature of the data.

We categorized students' reasoning complexity according to two categorical predictor variables: the mechanism students reasoned about (dummy-coded using the orthogonal contrasts C1−C3 in Table 1), and the trajectory to which they belong (dummy-coded using the orthogonal contrasts T1−T4 in Table 1). We examined how these variables influenced students' reasoning complexity as defined by the rubric outlined in Sections 2 and 4.5. Normalization or standardization was not required because of the dummy coding. Due to the large sample size, we used the maximum-likelihood estimator. We applied the *nlme* package to set up the HLMs (Pinheiro et al., 2011).

STEP-BY-STEP ANALYSIS WITH HLMS

We established HLMs to analyze changes in students' reasoning complexity over time. Analyzing HLMs involves several steps: evaluating the need for an HLM, adding fixed effects, introducing random slopes, and incorporating interaction terms (Field et al., 2012; Hox et al., 2017; Raudenbush & Bryk, 2002). To determine the necessity of an HLM, we first fitted a baseline model that included only an intercept and then allowed intercepts to vary across students, which is a random intercept model (Table S1: ID 1 and 2). We compared the model fits using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) with smaller values indicating better-fitting models. Additionally, we performed a likelihood ratio test, comparing the deviance of the simpler model to the more complex one using a χ^2 distribution. Comparisons revealed that the more complex model fitted the data better (Table S1: ID 2). Further, the adjusted intraclass correlation (ICC) was 0.428, indicating that 42.8% of the data's variation is attributable to differences between trajectories. An ICC above 0.100 suggests that an HLM is preferred over a non-hierarchical approach, confirming the appropriateness of using an HLM in our context.

Given our focus on adaptively supporting students' mechanistic reasoning, we hypothesized that students' reasoning complexity would improve over time. Analogously, we expected that students' reasoning complexity would differ depending on their reasoning trajectories as we explicitly modeled these trajectories to identify group differences. To test these hypotheses, we incorporated the reaction mechanisms and

TABLE S1 Model fit statistics of HLMs.

ID	Model	*df*	AIC	BIC	Deviance	Test	*L.*Ratio	*p*-value	Effect size
1	Intercept-Only	2	2163	2173	2159	-	-	-	-
2	Random Intercept	3	1939	1954	1933	1 vs 2	226	< .001	-
3	Context	6	1874	1904	1862	2 vs 3	71	< .001	0.084
4	Trajectory	10	1379	1428	1359	3 vs 4	504	< .001	0.026
5	Random Slope	19	1356	1451	1318	4 vs 5	40	< .001	0.999
6	Context × Trajectory	31	1165	1319	1103	5 vs 6	251	< .001	0.236

Note: Models build upon each other; for example, model 4 includes all effects of model 3 with *Trajectory* as an additional fixed effect. *df* = degrees of freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; *L.*Ratio = likelihood ratio.

students' reasoning trajectories as fixed effects in the random-intercept model. Including these variables improved model fit (Table S1: IDs 3 and 4).

We then investigated whether adding a random slope further improved model fit. In line with standard practices for longitudinal data, we allowed the effect of the reaction mechanism—the time point of data collection—to vary across students (Field et al., 2012). AIC and BIC values showed that the random-slopes model better fitted the data (Table S1: ID 5), suggesting that students' reasoning complexity evolved differently depending on their reasoning trajectory.

Afterward, we introduced the interaction term between the reaction mechanisms and students' reasoning trajectories into the random-slopes model. This interaction improved model fit (Table S1: ID 6), indicating that the combination of both variables influenced students' reasoning complexity. Table S2 presents the random effects of all HLMs. We provide a detailed discussion of the best-fitting final model below, using orthogonal contrasts for comparisons (Table 1).

ASSUMPTIONS CHECK: HLM ANALYSIS OF STUDENTS' REASONING COMPLEXITY

We evaluated whether our model met the necessary assumptions (Hox et al., 2017). Initially, we assessed the assumptions related to the least-squares residuals resulting from a linear regression separated by students. A histogram confirmed that these residuals were normally distributed; however, a scatterplot indicated a lack of homogeneity of variance. The residuals of the random intercepts were not normally distributed as evidenced by a significant Shapiro-Wilk test (p = .025). The residuals of the random slopes were partly normally distributed (p = .084 for contrast C1; p = .017 for contrast C2; p = .001 for contrast C3).

We also conducted a Rainbow test, which supported the appropriateness of applying linear models in our context due to the insignificant test statistic (F(532, 524) = 0.96, p = .662). To assess multicollinearity, we calculated variance inflation factors (VIFs). Values above 10 are problematic; in our case, the values ranged from 1.014 to 1.150, which is not concerning. Further examination of studentized residuals, Cook's distance, hat values, and the covariance ratio revealed no substantial irregularities.

Since not all assumptions were met, we also performed a robust HLM analysis using the *robustlmm* package (Koller, 2016). The results from this analysis were consistent with those obtained from the traditional HLM, so further details are not discussed.

RESULTS OF THE HLM ANALYSIS

To examine how students' reasoning complexity varied across different reaction mechanisms and reasoning trajectories, we interpreted the best-

TABLE S2 Random effects of HLMs.

ID	Model	σ^2	τ_{00}	τ_{11} (C1)	τ_{11} (C2)	τ_{11} (C3)
2	Random Intercept	0.255	0.191	-	-	-
3	Context	0.233	0.196	-	-	-
4	Trajectory	0.210	< 0.001	-	-	-
5	Random Slope	0.025	0.029	0.016	0.039	0.098
6	Context × Trajectory	0.021	0.030	0.014	0.025	0.078

Note: σ^2 represents the variance of students' reasoning complexity, τ_{00} denotes the random intercept residual, and τ_{11} indicates the random slopes residual. Contrasts C1−C3 are defined in Table 1.

fitting HLM described above (Table S1: ID 6). This model reveals significant two-way interactions, indicating that the interplay of reaction mechanisms and reasoning trajectories influenced students' reasoning complexity (Table S3). Due to these two-way interactions, the main effects should be interpreted cautiously.

Two contrasts involving the variable *Context* as a fixed effect are significant (Table S3: IDs 6.2 and 6.3), indicating that students' reasoning complexity is strongly context-dependent. In detail, students demonstrated slightly lower reasoning complexity in the two condensation reactions compared to enamine hydrolysis (Table S3: ID 6.2). Additionally, students' reasoning complexity was higher for Claisen condensation than for aldol condensation (Table S3: ID 6.3), suggesting that the students encountered substantial challenges in reasoning about aldol condensation. These challenges explain the lack of a clear upward trend in reasoning complexity over the semester.

Furthermore, significant variations in reasoning complexity were observed across trajectories with all contrasts being statistically significant (Table S3: IDs 6.4 to 6.7). The combination of the higher-achieving trajectories A and B significantly outperformed the combination of the lower-achieving trajectories C, D, and E (Table S3: ID 6.4). Comparably, trajectory C outperformed trajectories D and E (Table S3: ID 6.5), trajectory A outperformed trajectory B (Table S3: ID 6.6),

TABLE S3 Fixed effects included in the best-fitting HLM (Table S1: ID 6) analyzing the impact of the reaction mechanisms and reasoning trajectories on students' reasoning complexity.

ID	Fixed Effects	Contrast	*b*	*SE b*	*df*	*t*-value	*p*-value	Cohen's f^2
6.0	Intercept	–	2.740	0.013	783	205.589	< .001	53.981
6.1	Context	C1	−0.010	0.009	783	−1.106	.269	0.002
6.2	**Context**	**C2**	**−0.031**	**0.012**	**783**	**−2.595**	**.010**	**0.009**
6.3	**Context**	**C3**	**0.209**	**0.021**	**783**	**9.916**	**< .001**	**0.126**
6.4	**Trajectory**	**T1**	**−0.193**	**0.005**	**261**	**−38.609**	**< .001**	**5.711**
6.5	**Trajectory**	**T2**	**−0.123**	**0.013**	**261**	**−9.526**	**< .001**	**0.348**
6.6	**Trajectory**	**T3**	**−0.257**	**0.015**	**261**	**−16.622**	**< .001**	**1.059**
6.7	**Trajectory**	**T4**	**−0.240**	**0.026**	**261**	**−9.313**	**< .001**	**0.332**
6.8	**Context × Trajectory**	**C1 × T1**	**−0.019**	**0.003**	**783**	**−5.603**	**< .001**	**0.040**
6.9	Context × Trajectory	C2 × T1	−0.009	0.005	783	−1.867	.062	0.005
6.10	**Context × Trajectory**	**C3 × T1**	**0.021**	**0.008**	**783**	**2.685**	**.007**	**0.009**
6.11	**Context × Trajectory**	**C1 × T2**	**−0.036**	**0.009**	**783**	**−4.198**	**< .001**	**0.023**
6.12	**Context × Trajectory**	**C2 × T2**	**0.064**	**0.012**	**783**	**5.500**	**< .001**	**0.039**
6.13	**Context × Trajectory**	**C3 × T2**	**−0.090**	**0.020**	**783**	**−4.408**	**< .001**	**0.025**
6.14	Context × Trajectory	C1 × T3	−0.010	0.010	783	−0.979	.328	0.001
6.15	Context × Trajectory	C2 × T3	0.012	0.014	783	0.835	.404	0.001
6.16	Context × Trajectory	C3 × T3	0.006	0.024	783	0.238	.812	< 0.001
6.17	**Context × Trajectory**	**C1 × T4**	**−0.070**	**0.017**	**783**	**−4.122**	**< .001**	**0.022**
6.18	**Context × Trajectory**	**C2 × T4**	**−0.193**	**0.023**	**783**	**−8.265**	**< .001**	**0.087**
6.19	**Context × Trajectory**	**C3 × T4**	**0.210**	**0.041**	**783**	**5.142**	**< .001**	**0.034**

Note: The definitions of contrasts C1−C3 and T1−T4 are provided in Table 1. *b* = regression coefficient; *SE b* = standard error of regression coefficient; *df* = degrees of freedom; Cohen's f^2 = measure of effect size (Cohen's $f^2 > 0.02$ indicates a small effect, Cohen's $f^2 > 0.15$ indicates a medium effect, Cohen's $f^2 > 0.35$ indicates a large effect (Cohen, 1988)). Significant effects are bolded.

and trajectory D outperformed trajectory E (Table S3: ID 6.7).

Ultimately, significant interactions were detected, suggesting that the gains in reasoning complexity varied across trajectories, as shown in Figure 6b. For instance, the significant contrast C1 × T1 indicates that the increase in reasoning complexity from the Williamson ether synthesis to the other three mechanisms was statistically different for trajectories A and B compared to trajectories C, D, and E (Table S3: ID 6.8). Most other interactions were also statistically significant and can be interpreted equally. Only trajectories A and B exhibited identical gains in reasoning complexity over time (Table S3: IDs 6.14 to 6.16). Nonetheless, the other significant interactions demonstrate not just trajectory differences but also differences in how these trajectories change over time. This finding validates the applied clustering technique since it effectively identified varying growth patterns.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) Routledge.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R* (1st ed.). Sage.

Hox, J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge.

Koller, M. (2016). robustlmm: an R package for robust estimation of linear mixed-effects models. *Journal of statistical software*, *75*(6), 1-24. https://doi.org/10.18637/jss.v075.i06

Pinheiro, J., Bates, D., & R Core Team (2011). *Package 'nlme: Linear and nonlinear mixed effects models'*. [Computer Program].

Raudenbush, S. W., & Bryk A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.

Supporting Information C: Details on multinomial logistic regression to analyze trajectory differences

To evidence the statistical differences shown in Figure 9 of the main article, we present the standardized regression coefficients, standard errors, and odds ratios of the multinomial logistic regression model analyzing the impact of the cognitive and affective variables on students' reasoning trajectories.

The odds ratios can be interpreted as follows: In Table S4, the odds ratio of 2.49 for students' prior chemical concept knowledge when comparing trajectories A and B suggests that for a standard deviation increase in the prior chemical concept knowledge score, there is a 2.49 times increase in the odds of being a member of trajectory A compared to being a member of trajectory B when all other variables are held constant.

To address the context-dependency of students' mechanistic reasoning, we also conducted post hoc analyses to examine how the assessed cognitive and affective variables predict differences between trajectories across reaction mechanisms. The multinomial logistic regressions detailed in Tables S5, S6, and S7 correspond to Figures 10a, 10b, and 10c of the main article. Further details from the multinomial logistic regression analyzing students' perceived task ease (Figure 10d) are not reported due to the insignificant test statistic.

TABLE S4 Multinomial logistic regression to identify variables that explain trajectory differences.

	Concept Knowledge			**Argumentation Skills**			**Affective Variables**		
	Prior Concept Knowledge			**Prior Argumentation Skills**			**Perceived Confidence**		
	β	**SE *β***	**OR**	***β***	**SE *β***	**OR**	***β***	**SE *β***	**OR**
A vs. B	**0.912*****	**0.221**	**2.49**	**0.496****	**0.191**	**1.64**	**0.510****	**0.185**	**1.67**
A vs. C	**2.048*****	**0.304**	**7.75**	**0.802*****	**0.223**	**2.23**	**0.498***	**0.222**	**1.64**
A vs. D	**2.020*****	**0.296**	**7.54**	**0.447***	**0.220**	**1.56**	**0.773*****	**0.217**	**1.64**
A vs. E	**2.025*****	**0.366**	**7.57**	**0.864****	**0.276**	**2.37**	**1.025*****	**0.282**	**2.79**
B vs. C	**1.136*****	**0.248**	**3.11**	0.306	0.183	1.36	−0.013	0.194	0.99
B vs. D	**1.108*****	**0.239**	**3.03**	−0.049	0.183	0.95	0.263	0.184	1.30
B vs. E	**1.112*****	**0.321**	**3.04**	0.368	0.244	1.45	**0.514***	**0.254**	**1.67**
C vs. D	−0.028	0.262	0.97	−0.035	0.214	0.70	0.276	0.221	1.32
C vs. E	−0.024	0.338	0.98	0.062	0.267	1.06	0.527	0.282	1.69
D vs. E	0.004	0.332	1.00	0.417	0.268	1.52	0.251	0.271	1.29
	Within-Study Concept Knowledge			**Within-Study Argumentation Skills**			**Perceived Task Ease**		
	β	**SE *β***	**OR**	***β***	**SE *β***	**OR**	***β***	**SE *β***	**OR**
A vs. B	**0.760*****	**0.212**	**2.14**	0.081	0.169	1.08	**0.383***	**0.173**	**1.47**
A vs. C	**1.327*****	**0.258**	**3.77**	0.134	0.210	1.14	0.384	0.211	1.47
A vs. D	**1.251*****	**0.250**	**3.49**	**0.411***	**0.202**	**1.51**	**0.661****	**0.210**	**1.94**
A vs. E	**1.871*****	**0.330**	**6.49**	0.480	0.269	1.62	**0.543***	**0.275**	**1.72**
B vs. C	**0.568****	**0.206**	**1.76**	0.053	0.190	1.05	0.001	0.192	1.00
B vs. D	**0.491***	**0.198**	**1.63**	0.330	0.183	1.39	0.278	0.187	1.32
B vs. E	**1.111*****	**0.286**	**3.04**	0.400	0.255	1.49	0.160	0.259	1.17
C vs. D	−0.077	0.227	0.93	0.276	0.218	1.32	0.278	0.222	1.32
C vs. E	0.544	0.299	1.72	0.347	0.281	1.41	0.160	0.285	1.17
D vs. E	**0.620***	**0.295**	**1.86**	0.070	0.274	1.07	−0.118	0.281	0.89

Note: β = standardized regression coefficient, *SE β* = standard error of regression coefficient, *OR* = odds ratio. Significant effects are bolded. *p < .05; **p < .01; ***p < .001.

TABLE S5 Multinomial logistic regression analyzing differences in students' within-study chemical concept knowledge separated by mechanisms.

	Enamine Hydrolysis			**Aldol Condensation**			**Claisen Condensation**		
	β	**SE *β***	**OR**	***β***	**SE *β***	**OR**	***β***	**SE *β***	**OR**
A vs. B	**0.616****	**0.201**	**1.85**	**0.521***	**0.207**	**1.68**	0.379	0.188	1.46
A vs. C	**1.086*****	**0.239**	**2.96**	**0.974*****	**0.242**	**2.65**	0.532	0.225	1.70
A vs. D	**1.067*****	**0.234**	**2.91**	**0.970*****	**0.237**	**2.64**	**0.590***	**0.218**	**1.80**
A vs. E	**0.949****	**0.292**	**2.58**	**1.492*****	**0.295**	**4.44**	**1.176*****	**0.275**	**3.24**
B vs. C	0.470	0.194	1.60	0.453	0.197	1.57	0.154	0.195	1.17
B vs. D	0.451	0.188	1.57	0.449	0.191	1.57	0.212	0.187	1.24
B vs. E	0.334	0.259	1.40	**0.971*****	**0.255**	**2.64**	**0.797****	**0.245**	**2.22**
C vs. D	−0.019	0.211	0.98	−0.004	0.214	1.00	0.058	0.220	1.06
C vs. E	−0.136	0.278	0.87	0.518	0.266	1.68	0.644	0.268	1.90
D vs. E	−0.118	0.274	0.89	0.522	0.261	1.69	0.586	0.260	1.80

Note: *β* = standardized regression coefficient, *SE β* = standard error of regression coefficient, *OR* = odds ratio. Significant effects are bolded. *p < .013; **p < .003; ***p < .0003.

TABLE S6 Multinomial logistic regression analyzing differences in students' within-study argumentation skills separated by mechanisms.

	Enamine Hydrolysis			Aldol Condensation			Claisen Condensation		
	β	SE β	OR	β	SE β	OR	β	SE β	OR
A vs. B	0.354	0.190	1.42	0.114	0.175	1.12	−0.236	0.168	0.79
A vs. C	**0.670****	**0.221**	**1.95**	0.235	0.211	1.27	**−0.598****	**0.215**	**0.55**
A vs. D	**0.725****	**0.215**	**2.07**	0.305	0.204	1.36	−0.168	0.199	0.85
A vs. E	0.471	0.281	1.60	**0.766****	**0.262**	**2.15**	−0.295	0.269	0.74
B vs. C	0.316	0.185	1.37	0.122	0.192	1.13	−0.362	0.196	0.70
B vs. D	0.372	0.178	1.45	0.191	0.183	1.21	0.068	0.182	1.07
B vs. E	0.117	0.256	1.12	**0.653****	**0.245**	**1.92**	−0.059	0.256	0.94
C vs. D	0.056	0.203	1.06	0.070	0.217	1.07	0.430	0.224	1.54
C vs. E	−0.199	0.276	0.82	0.531	0.269	1.70	0.303	0.286	1.35
D vs. E	−0.255	0.271	0.78	0.461	0.262	1.59	−0.127	0.278	0.88

Note: β = standardized regression coefficient, *SE* β = standard error of regression coefficient, *OR* = odds ratio. Significant effects are bolded. *p < .013; **p < .003; ***p < .0003 (Bonferroni correction as outlined in Section 4.6.7).

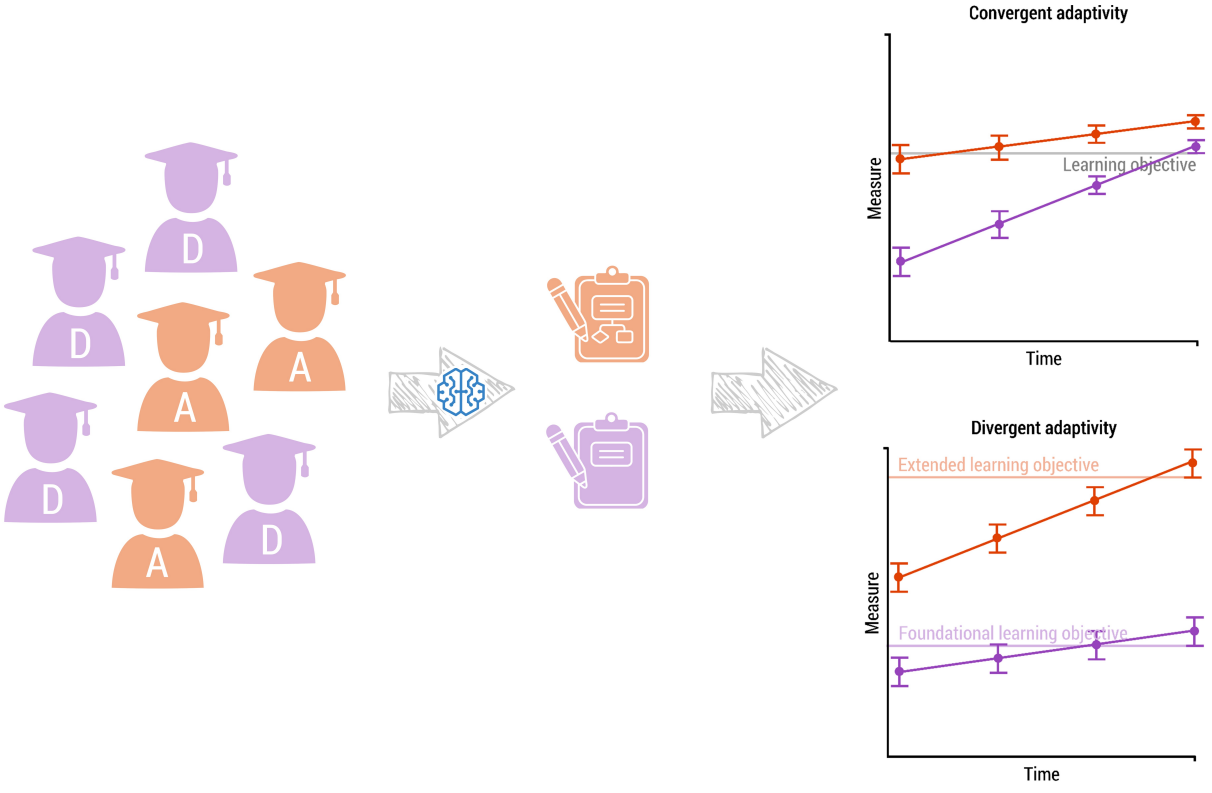
TABLE S7 Multinomial logistic regression analyzing differences in students' confidence separated by mechanisms.

	Ether Synthesis			Enamine Hydrolysis			Aldol Condensation			Claisen Condensation		
	β	SE β	OR	β	SE β	OR	β	SE β	OR	β	SE β	OR
A vs. B	0.271	0.172	1.31	**0.448***	**0.179**	**1.56**	**0.525***	**0.187**	**1.69**	0.365	0.181	1.44
A vs. C	0.250	0.210	1.28	0.239	0.216	1.27	**0.583***	**0.223**	**1.79**	0.485	0.218	1.62
A vs. D	0.364	0.204	1.44	**0.527***	**0.208**	**1.69**	**0.779****	**0.218**	**2.18**	**0.762****	**0.214**	**2.14**
A vs. E	**0.838****	**0.275**	**2.31**	0.579	0.270	1.78	**0.972****	**0.281**	**2.64**	**0.891****	**0.278**	**2.44**
B vs. C	−0.022	0.191	0.98	−0.209	0.192	0.81	0.057	0.191	1.06	0.120	0.193	1.13
B vs. D	0.093	0.183	1.10	0.080	0.178	1.08	0.253	0.182	1.29	0.397	0.185	1.49
B vs. E	0.567	0.257	1.76	0.131	0.247	1.14	0.446	0.252	1.56	0.526	0.255	1.69
C vs. D	0.115	0.219	1.12	0.289	0.219	1.33	0.196	0.217	1.22	0.277	0.219	1.32
C vs. E	0.588	0.284	1.80	0.340	0.278	1.41	0.389	0.278	1.48	0.406	0.280	1.50
D vs. E	0.474	0.277	1.61	0.052	0.268	1.05	0.192	0.270	1.21	0.129	0.270	1.14

Note: β = standardized regression coefficient, *SE* β = standard error of regression coefficient, *OR* = odds ratio. Significant effects are bolded. *p < .013; **p < .003; ***p < .0003 (Bonferroni correction as outlined in Section 4.6.7).

10 | To what extent is adaptive support equitable? Exploring the impact of machine learning-assisted adaptive support on students' mechanistic reasoning

Chapter 10 presents a research manuscript by Martin, Kubsch, Yik, Burlingham, and Graulich that is currently in preparation.



RESEARCH ARTICLE

To what extent is adaptive support equitable? Exploring the impact of machine learning-assisted adaptive support on students' mechanistic reasoning

Paul P. Martin[1], Marcus Kubsch[2], Brandon J. Yik[3], Benjamin T. Burlingham[4], Nicole Graulich[1,*]

Abstract

Education research of the past decades shows that students vary in their cognitive abilities and affective characteristics, resulting in considerable student heterogeneity. This heterogeneity challenges educators to create instructional settings that account for each student's unique needs. Emerging research on adaptive learning methodologies claims that artificial intelligence (AI) techniques like machine learning (ML) can help reduce student heterogeneity and mitigate inequities in learning due to their potential to provide adaptive support. However, the implementation of adaptive learning reflects a critical tension between *convergent adaptivity*, which seeks to mitigate differences in students' learning outcomes, and *divergent adaptivity*, which enables all students to progress at their own pace. In this article, we quantitatively examine the extent to which divergent adaptivity, facilitated through ML, can achieve its hypothesized aim of enabling all students to make progress in their learning. Specifically, we illustrate a semester-long study conducted within an undergraduate organic chemistry course that applied an ML model to adaptively support 266 chemistry students through four interventions as they reasoned about reaction mechanisms. We identified five distinct learning trajectories throughout these adaptive interventions. Our findings show that while students with higher prior knowledge progressed throughout the semester, students with lower prior knowledge faced challenges maintaining their progress. This indicates that the adaptive support maintained and partly exacerbated differences in students' learning outcomes. Consequently, adaptivity may foster individual growth; however, it may also favor students with higher prior knowledge and specific intersectional identities unless carefully tailored. Our study contributes to the existing literature by connecting AI, adaptive support, and equity considerations. Furthermore, it highlights the importance of a fine-grained analysis of students' prior knowledge and learning gains to identify who truly benefits from adaptive support.

KEYWORDS student heterogeneity, educational debt, adaptivity, personalization, machine learning

Affiliation

[1]Institute of Chemistry Education, Justus-Liebig-University, Heinrich-Buff-Ring 17, 35392 Giessen, Germany
[2]Department of Physics, Freie Universität Berlin, Arnimallee 14, 14195 Berlin, Germany
[3]Department of Chemistry, University of Georgia, 302 East Campus Road, Athens, GA 30602, United States
[4]Department of Chemistry, Indiana University, 800 E. Kirkwood Ave., Bloomington, IN 47405-7102, United States
*E-Mail corresponding author: Nicole.Graulich@didaktik.chemie.uni-giessen.de

1 | Introduction

Due to structural, systemic, and institutional barriers, students enter formal education with diverse cognitive abilities, affective-emotional traits, and socioeconomic backgrounds (Castle et al., 2024). Addressing this heterogeneity challenges educators across domains to design instructional settings that meet diverse student needs (Dumont & Ready, 2023). Adaptive learning, often enabled by digital technology, offers a promising approach to tailoring support to students' needs (Xie et al., 2019). Adaptive learning can be implemented using various methods, such as adaptive prompting, scaffolding, adaptive feedback, intelligent tutoring systems (ITSs), or formative assessments driven by artificial intelligence (AI) (Plass & Pawar, 2020b). In contrast to adaptive support, a one-size-fits-all approach may pressure students with lower prior knowledge to solve tasks that are too challenging while it prompts students with higher prior knowledge to practice skills they have already mastered (Tomlinson et al., 2003).

Educational research has increasingly highlighted the potential of adaptive support to reduce student heterogeneity and mitigate inequities in students' learning, especially given the recent advancements in machine learning (ML) to personalize instruction (e.g., Li et al., 2021; Roshanaei et al., 2023; Viberg et al., 2024). For instance, AI-driven ITSs are anticipated to provide personalized learning experiences comparable to one-on-one human tutoring (e.g., Kizilcec et al., 2017; Ma et al., 2014; Steenbergen-Hu & Cooper, 2014; VanLehn, 2011), potentially exceeding human capabilities in providing adaptive support (Kulik & Fletcher, 2016). Such expectations arise from AI's presumed capacity to fine-tune when and how instructional support needs to be delivered, offer content-specific, adaptive feedback, and prioritize students most in need (Roshanaei et al., 2023; Viberg et al., 2024). Numerous reviews and meta-analyses have already explored the effects of AI-driven adaptive learning systems like ITSs for implementing personalized instruction (e.g., Kulik & Fletcher, 2016; Ma et al., 2014; Slavin et al., 2009; Steenbergen-Hu & Cooper, 2013, 2014; VanLehn, 2011). While some reviews and meta-analyses found no significant improvement in students' learning outcomes from using ITSs compared to traditional, non-adaptive instruction (Slavin et al., 2009; Steenbergen-Hu & Cooper, 2013), others reported mostly medium improvement in students' learning outcomes (Kulik & Fletcher, 2016; Ma et al., 2014; Steenbergen-Hu & Cooper, 2014; VanLehn, 2011).

Given these rather inconclusive findings, it becomes evident that the effectiveness of ITSs depends on various factors. For instance, Kulik and Fletcher (2016) found that the type of post-test was associated with ITS effectiveness, with locally developed tests measuring greater improvements in students' learning outcomes than standardized ones. Furthermore, ITSs were more effective when instructors were experienced in using such a system, which allowed them to provide extra adaptive support while students interacted with the ITS (Kulik & Fletcher, 2016). This implies that ITSs can be used as a diagnostic tool to help instructors offer targeted assistance for students who need it most and challenge students with higher prior knowledge with more advanced tasks (Holstein et al., 2018). However, Steenbergen-Hu and Cooper (2013) found negative effects of ITSs in K-12 mathematics learning for students with lower prior knowledge, which indicates that the impact of ITSs may vary across students.

In sum, much evidence suggests that adaptive support may improve students' learning outcomes. The degree of improvement varies based on the applied instructional strategies, the educational level, the subject domain, students' peer interactions, and how well these approaches are integrated with other teaching, learning, and assessment practices (Deunk et al., 2018; Smale-Jacobse et al., 2019). However, as

shown by Steenbergen-Hu and Cooper's (2013) review, effectively addressing the diverse needs of students with varying prior knowledge remains a significant challenge in adaptive learning. Therefore, whether all students benefit equally from adaptive support is uncertain.

2 | Theoretical Background on Adaptivity in Education

At its core, adaptive learning aims to scale personalized education, so instruction can better focus on a student's prior knowledge, problem-solving approaches, affect, motivation, self-regulated learning strategies, or metacognition (Aleven et al., 2016). Adaptivity is a spectrum rather than a binary characteristic. Aleven et al. (2016) defined three properties to describe the degree of a system's adaptivity: (1) using data on domain-specific student challenges, (2) varying pedagogical decisions depending on psychological measures of individual students, and (3) responding interactively to student actions. This definition demonstrates that adaptivity is a process in which learning systems assess current student attributes and modify the instructional setting to improve learning (Plass & Pawar, 2020a; Shute & Zapata-Rivera, 2012).

2.1 Analyzing adaptivity in education through the enhanced educational debt framework

In recent years, the necessity for supporting students with diverse systemic predispositions through adaptive learning has gained prominence across domains (e.g., Dumont & Ready, 2023). Sbeglia and Nehm (2024) introduced the *enhanced educational debt framework* to examine (1) how learning outcomes vary after instruction and (2) identify who is successfully improving their learning outcomes (Figure 1). In doing so, the term *educational debt* acknowledges the accumulated structural, systemic, and institutional barriers that have historically restricted access to learning opportunities for students from systemically marginalized groups (Ladson-Billings, 2006). Accordingly, educational debt shifts the focus from viewing differences in achievement as a gap that needs to be closed to understanding these differences as an accumulative disadvantage that society owes to students from systemically marginalized groups (Shukla et al., 2022). The term achievement gap is intentionally avoided as it implies a deficit perspective on student ability or effort, potentially placing undue pressure on students with lower prior knowledge to close the achievement gap between themselves and their peers (Gutiérrez, 2008; Sbeglia & Nehm, 2024).

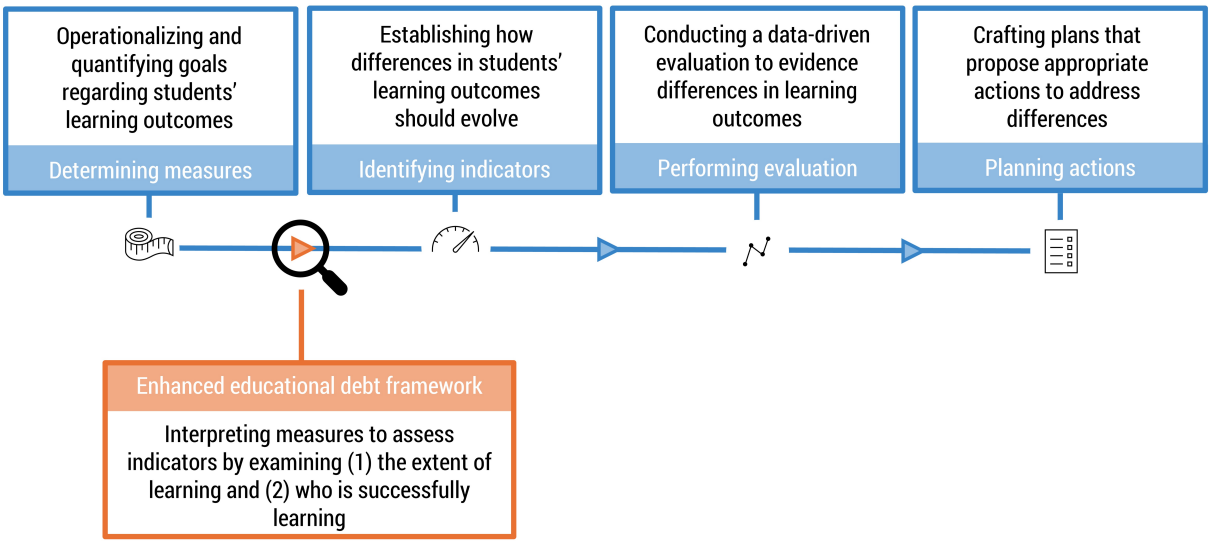


FIGURE 1 Interpreting measures to make data-driven claims about indicators and enable the planning of future actions using the enhanced educational debt framework (Sbeglia & Nehm, 2024).

As conceptualized in the enhanced educational debt framework, instructional strategies can impact educational debt in three ways: they may *mitigate*, *maintain*, or *exacerbate* the debt (Van Dusen et al., 2022). To identify the impact of instructional strategies on educational debt, data-driven evaluations based on quantifiable measures of learning, such as course grades, exam scores, pass rates, or class participation, are conducted. These data-driven evaluations help identify differences in students' engagement or learning outcomes. Evidence-based indicators can ultimately serve as a foundation for creating targeted action plans to improve future instruction (Figure 1).

While many differences in learning outcomes are likely associated with educational debt, it is important to distinguish between differences in learning outcomes and educational debt (Sbeglia & Nehm, 2024). First, educational debt may not necessarily cause identifiable differences, especially in the later stages of education. For example, structural, systemic, and institutional barriers may prevent students from systemically marginalized groups from accessing undergraduate studies, which is why certain differences in learning outcomes may no longer be visible in undergraduate education—although they stem from an educational debt. In contrast, some differences may be linked to other factors unrelated to educational debt. For instance, a biology major taking an organic chemistry course may have less interest in the subject than a chemistry major, resulting in reduced engagement with the course material. While these variations in student engagement may result in different learning outcomes, these differences arise from personal choices based on individual interests and career aspirations, rather than from structural, systemic, or institutional barriers. Thus, factors like personal interests can also result in learning-related differences.

2.2 Aims of adaptivity in education

Adaptivity in education can serve two fundamentally different purposes: mitigating differences in students' learning outcomes through *convergent adaptivity* or allowing all students to progress at their own pace through *divergent adaptivity*, which may maintain or even exacerbate differences (Deunk et al., 2015, 2018; Smale-Jacobse et al., 2019). Convergent adaptivity establishes a learning objective that is roughly aligned with the prior knowledge of students with higher prior knowledge (Figure 2a). This learning objective remains the same for all. Students who do not meet the defined learning objective receive adaptive support until they reach it. Conversely, students who already meet the objective are provided with additional resources for enrichment, which may extend their learning but are not directly related to the primary learning objective. Following this, convergent adaptivity intends to ensure that all students reach a similar learning outcome, which is why this practice allocates additional instructional time and resources to students with lower prior knowledge—even if it reduces time for their peers. Accordingly, convergent adaptivity may help mitigate differences in learning outcomes (Figure 2a).

In contrast to convergent adaptivity, divergent adaptivity allows learning objectives to vary between students (Figure 2b), which means these learning objectives are aligned with external criteria that correspond to each student's individual prior knowledge (Heckhausen, 1974). Students with lower prior knowledge should at least reach a foundational learning objective, exemplified in Figure 2b by the objective for these students to attain a measure of 2.00. Adaptivity encourages students with lower prior knowledge to gradually improve their skills until they attain this foundational learning objective. Students with higher prior knowledge are simultaneously challenged with more advanced learning opportunities that promote a higher-order learning objective. To

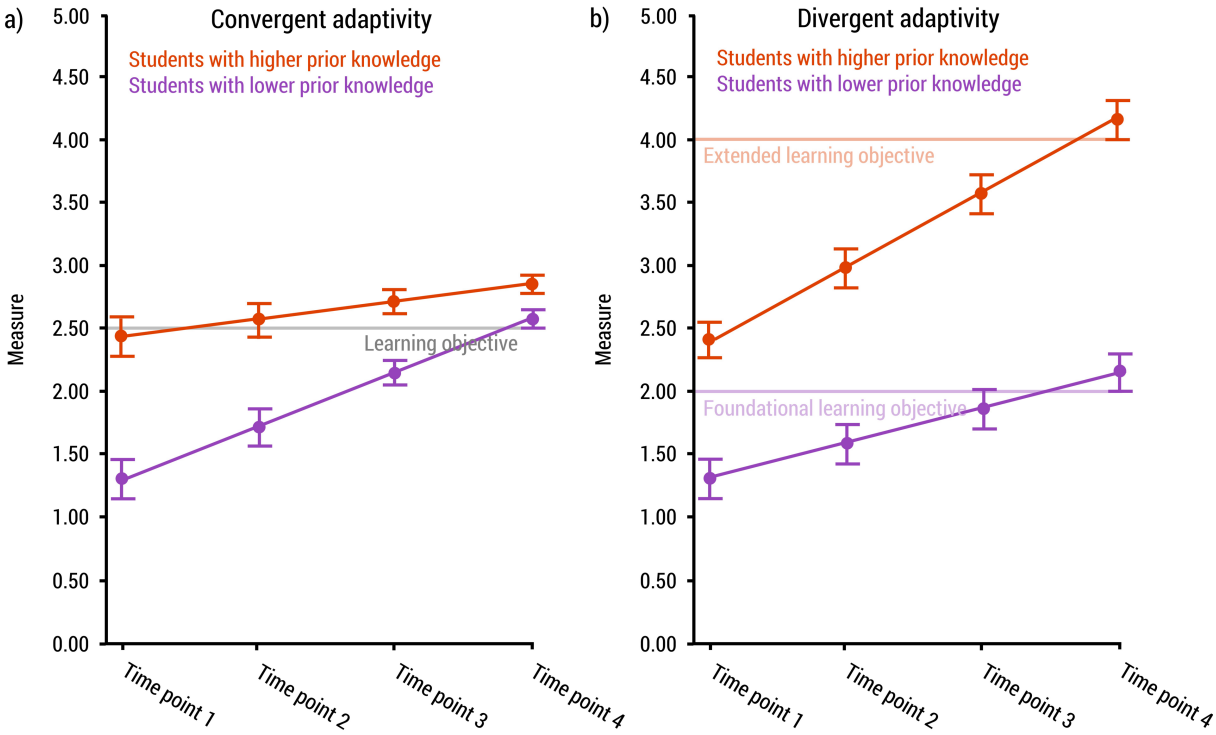


FIGURE 2 Prototypical long-term impact of a) convergent adaptivity and b) divergent adaptivity on a fictitious measure, separated for groups of students with higher and lower prior knowledge.

ensure that every student meets their personalized learning objective, instructional time is distributed evenly among all students. Although this may lead to inequitable outcomes, it enables all students to reach the foundational learning objective *and* leverage their full learning potential—even if it may maintain or exacerbate learning-related differences between students with lower and higher prior knowledge (Figure 2b) (Dumont & Ready, 2023). Maintained or exacerbated differences in students' learning outcomes are often observed in adaptive instructional settings; this phenomenon is subsumed under the term *Matthew effect*. This effect describes how students with higher prior knowledge progress more significantly in adaptive instructional settings due to greater access to learning opportunities that address higher-order cognitive operations while students with lower prior knowledge struggle to "catch up" (Stanovich, 1986).

Both approaches to adaptivity have valid arguments (Figure 3). Convergent adaptivity aligns with an equity conceptualization of *equality of outcomes*, aiming to mitigate differences in students' learning outcomes by providing additional support to students with lower prior knowledge (Dumont & Ready, 2023; Sbeglia & Nehm, 2024). Thus, adaptive support is primarily directed toward students who need more support and stops once they reach outcomes comparable to their peers. This adaptivity approach may be necessary in certain situations. For instance, providing additional support to students with lower prior knowledge is indispensable for essential skills like reading and writing. Since these skills are essential for future participation in society, convergent adaptivity is one approach to ensure that students are not excluded from future societal engagement. A similar situation arises when differences in students' learning outcomes are directly related to educational debt—structural, systemic, and institutional barriers that have historically limited learning opportunities for students from systemically marginalized groups. Here, convergent adaptivity may mitigate the effects of educational debt. By doing so, it is, however, important to ensure that addressing

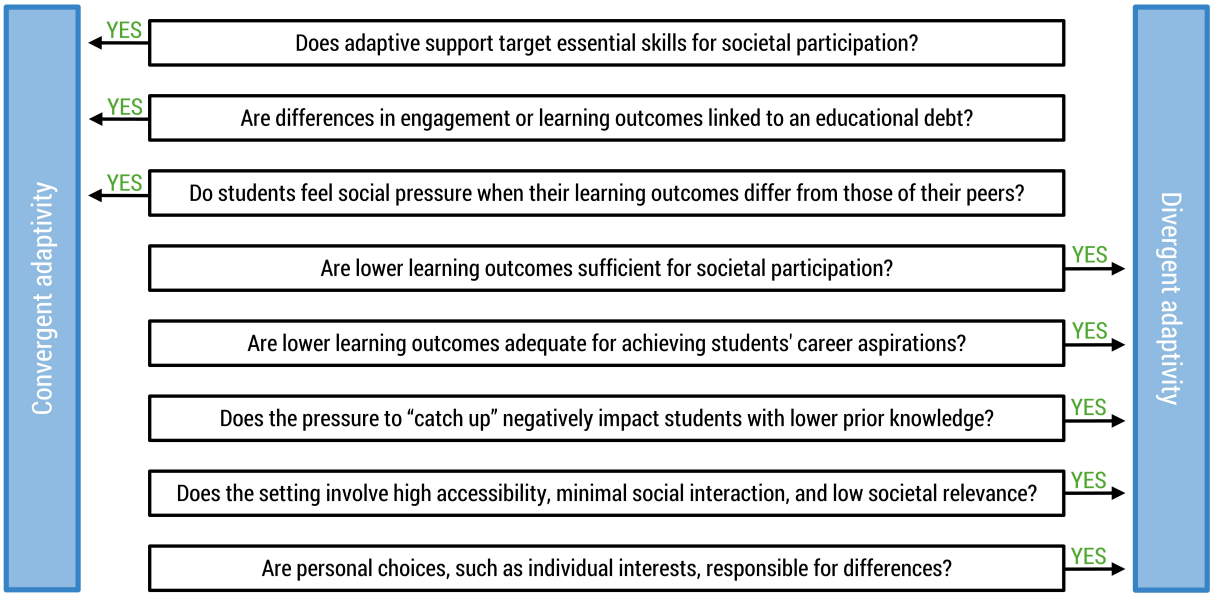


FIGURE 3 Guiding questions to determine whether convergent or divergent adaptivity should be applied.

structural, systemic, or institutional issues contributing to educational debt is not overlooked (Gutiérrez, 2008). Besides, convergent adaptivity may be necessary when students experience social pressure if they do not meet the same learning outcomes as their peers.

By contrast, an *adequacy*-oriented conceptualization of equity provides valid arguments for divergent adaptivity (Dumont & Ready, 2023; Holstein & Doroudi, 2022). According to this conceptualization, divergent adaptivity is appropriate when meeting a foundational learning objective is sufficient for participation in society or future career aspirations, and higher-order learning objectives—while significant within a specific domain—may not have broader societal relevance. For example, determining whether a nucleophilic substitution or elimination reaction is more likely to occur is a key learning objective in undergraduate organic chemistry with limited societal relevance. Instruction may help all students assess the plausibility of these two chemical mechanisms for less structurally complex species while some students may be supported in extending this skill to more complex species. Besides, the expectation that students with lower prior knowledge demonstrate higher learning gains than their peers can lead to excessive social pressure on these students. Divergent adaptivity may alleviate the pressure on these students to perform at the same level as their peers. Likewise, it may ensure that convergent adaptivity does not hinder the progress of students with higher prior knowledge. Moreover, divergent adaptivity may be suitable in highly accessible instructional settings where social interaction is limited and the learning objective lacks societal relevance, as may occur in personalized online learning. Also, divergent adaptivity may be appropriate when differences in engagement or learning outcomes stem from personal choices driven by individual interests or future career aspirations, rather than from structural or institutional barriers.

Together, these different aims underscore an adaptivity dilemma of how to allocate limited instructional time to students with varying needs (Deunk et al., 2018; Holstein et al., 2018). Convergent adaptivity primarily supports students with lower prior knowledge, which may reduce instructional time for their peers. Conversely, divergent adaptivity equally allocates instructional time among all students, acknowledging that differences in learning outcomes may continue to exist.

Determining which adaptivity aim to prioritize is also a sociopolitical issue; however, there is no prevailing consensus on which aim to pursue in a given context. For example, Blok (2004) and Bosker (2005) outline that governmental documents, such as the American Success for All (Slavin, 1996), the Australian Early Reading Literacy Project (Crévola & Hill, 1998), or its Dutch variant (Veen et al., 2000), use the term adaptivity almost exclusively for addressing the educational needs of students with accessibility needs. In this way, adaptivity is viewed as a compensatory strategy to meet the needs of students with lower prior knowledge. However, although these governmental documents emphasize convergent adaptivity, they are frequently operationalized as divergent adaptivity in domain-specific curricular frameworks or learning resources (Blok, 2004; Bosker, 2005). Moreover, the inconclusive evidence on the long-term impact of convergent or divergent adaptivity on students' learning outcomes complicates making informed decisions about when to apply specific adaptivity strategies and for what aims (Deunk et al., 2015). Consequently, whether an adaptive learning strategy can successfully achieve its hypothesized aim for students with different prior knowledge remains largely uncertain. In this article, we start to address this challenge by illustrating a study that examines the long-term impact of divergent adaptivity, implemented through ML, on students' learning.

3 | Conducting a data-driven evaluation as an illustrative case: Using divergent adaptivity to support chemistry students' mechanistic reasoning

In the following sections, we illustrate a study that used divergent adaptivity to support undergraduate organic chemistry students in reasoning about reaction mechanisms over one semester. We applied divergent adaptivity because students enrolled in the organic chemistry course pursued different majors, including biochemistry, biology, biotechnology, chemistry, and neuroscience, among others, and we do not expect a biology major, for instance, to have or need the same level of expertise in reasoning about reaction mechanisms as a chemistry major. Therefore, applying divergent adaptivity may align with students' interests and career aspirations, so the implemented divergent adaptivity may have several benefits. For example, it may allow students with a stronger interest in organic chemistry to reach their full learning potential, and it may also equip students with less interest with the skills necessary for understanding mechanisms in everyday contexts (Stowe et al., in press).

Within the illustrated case, we aim to explore which students benefit most from the implemented divergent adaptive support. In doing so, we use the complexity of students' reasoning about reaction mechanisms as a measure to perform a data-driven evaluation that assesses differences in this measure over time (Figure 1). Our evaluation is guided by five questions: (1) What differences in students' reasoning about reaction mechanisms are present at the beginning of our implementation of adaptive support? (2) How effective is our adaptive support overall? (3) How does our adaptive support impact differences in students' reasoning about reaction mechanisms over time? (4) How does our adaptive support impact student heterogeneity in mechanistic reasoning complexity over time? (5) Which students benefit the most from adaptive support?

3.1 The need to adaptively support undergraduate chemistry students' mechanistic reasoning

A key objective of science education is to enable students to understand the natural and engineered world by engaging them in science practices, such as developing models, constructing explanations, or formulating

arguments (National Research Council, 2012). Mechanistic reasoning, which is required in several of these science practices, enables students to analyze the underlying mechanisms of observed phenomena (Glennan, 2002; Illari & Williamson, 2012), thereby providing students with a tool to understand the natural and engineered world. Mechanistic reasoning is defined as the ability to identify, describe, or analyze the underlying and often unobservable mechanisms of scientific phenomena to explain, predict, or control their outcomes (Krist et al., 2019; Russ et al., 2008). Students' mechanistic reasoning can vary in complexity due to variations in their skills to make predictions and explain causal relationships (Russ et al., 2008).

Students face heterogeneous challenges in demonstrating mechanistic reasoning across science domains (Bachtiar et al., 2022). For example, Crandell et al. (2019) highlight six dimensions of student heterogeneity in mechanistic reasoning. They found that the complexity of students' mechanistic reasoning about acid–base reactions varied across (1) chemistry course types, (2) chemistry curricula, and (3) chemical contexts. Additionally, student heterogeneity was evident (4) within a single cohort, (5) between cohorts, and (6) over time. Other studies further evidence students' heterogeneity in mechanistic reasoning (e.g., Crandell et al., 2020; Crowder et al., 2024; Deng & Flynn, 2021; Frost et al., 2023; Noyes & Cooper, 2019; Southard et al., 2017; Watts et al., 2021; Yik et al., 2023).

Given this heterogeneity, numerous studies have implemented adaptive support to guide students' mechanistic reasoning (e.g., Ariely et al., 2024; Asmussen et al., 2024; Dood et al., 2020a, 2020b; Hedtrich & Graulich, 2018a, 2018b; Lieber et al., 2022a, 2022b; Young et al., 2021). For instance, Ariely et al. (2024) and Dood et al. (2020b) demonstrated that applying an ML model to evaluate students' mechanistic reasoning allowed them to implement divergent adaptive support. Students who received this adaptive support advanced their mechanistic reasoning significantly compared to a control group. Accordingly, adaptive support appears to improve students' mechanistic reasoning. Nonetheless, while both of these studies compared their adaptive support to a control group activity or exercise, they did not investigate which students benefited most from their adaptive support and how students' learning outcomes changed over time. Hence, the impact of adaptive support on students' heterogeneous challenges in mechanistic reasoning remains largely unexplored.

3.2 Context of the illustrated study

To investigate the impact of adaptive support on students' mechanistic reasoning, we illustrate a study that provided 266 undergraduate students enrolled in an Organic Chemistry II course at a large, public, research-intensive institution in the Midwestern United States with divergent adaptive support during the Winter 2023 semester. Of the 266 students whose data were included, 19 chose not to provide demographic information. Among the remaining students, 148 self-identified as women (59.9%), 93 as men (37.7%), and six as other gender identities (2.4%). The students majored in various disciplines, such as biochemistry (*n* = 31; 12.6%), biology (*n* = 91; 36.8%), biotechnology (*n* = 8; 3.2%), chemistry (*n* = 21; 8.5%), human biology (*n* = 15; 6.1%), and neuroscience (*n* = 57; 23.1%), among others. For further analysis, we only distinguished between chemistry/biochemistry majors and non-chemistry/biochemistry majors. Students identified their ethnic and racial background as follows: Asian or Asian-American (*n* = 48; 19.4%); Black, Afro-Caribbean, or African American (*n* = 10; 4.0%); Indigenous American or Alaska Native (*n* = 1; < 1.0%); Latino/a/e/x or Hispanic American (*n* = 13; 5.3%); Middle Eastern, North African, or Arab American (*n* = 10; 4.0%); Non-Hispanic White or Euro-American (*n* = 143; 57.9%); and multiracial (*n* = 14; 5.7%). Eight students chose

not to disclose their ethnic or racial background. For further analysis, we distinguished between students who identify with an ethnicity that is federally considered of an underrepresented minority (URM) status (i.e., students who identify as Black, Afro-Caribbean, or African American; Indigenous American or Alaska Native; Latino/a/e/x or Hispanic American; or multiracial including at least one of these racial/ethnic categories) or those who do not belong to a URM (i.e., students who identify as Asian or Asian American; Middle Eastern, North African, or Arab American; Non-Hispanic White or Euro-American).

We developed and implemented four online scaffolds using the Qualtrics survey platform. These scaffolds guided students' mechanistic reasoning in judging the plausibility of alternative reaction products across four organic chemistry reaction mechanisms throughout the semester-long course: ether synthesis, enamine hydrolysis, aldol condensation, and Claisen condensation (Lieber et al., 2022a, 2022b; Martin & Graulich, 2024b). We used a previously developed ML model to automatically code students' mechanistic reasoning, adapt the scaffolds to the mechanistic reasoning complexity a student demonstrated in the prior scaffold, and track how their mechanistic reasoning complexity developed over the study's four time points (Figure 4). Details on ML model development can be found elsewhere (Martin & Graulich, 2024a, 2024c; Martin et al., 2024).

Our study was intentionally designed to align more closely with divergent adaptivity to allow all students to progress at their own pace. Specifically, tasks were adapted to support students in demonstrating increasingly complex mechanistic reasoning; students with higher mechanistic reasoning complexity were not restricted to advancing toward higher-order learning objectives. These objectives varied among students depending on their mechanistic reasoning complexity in the preceding tasks. Martin et al. (in preparation) report this study in detail.

3.3 Data analysis

The analyses were performed in R (version 4.3.2) (R Core Team, 2023) using RStudio as the integrated development environment (IDE) (RStudio Team, 2023), primarily relying on base R

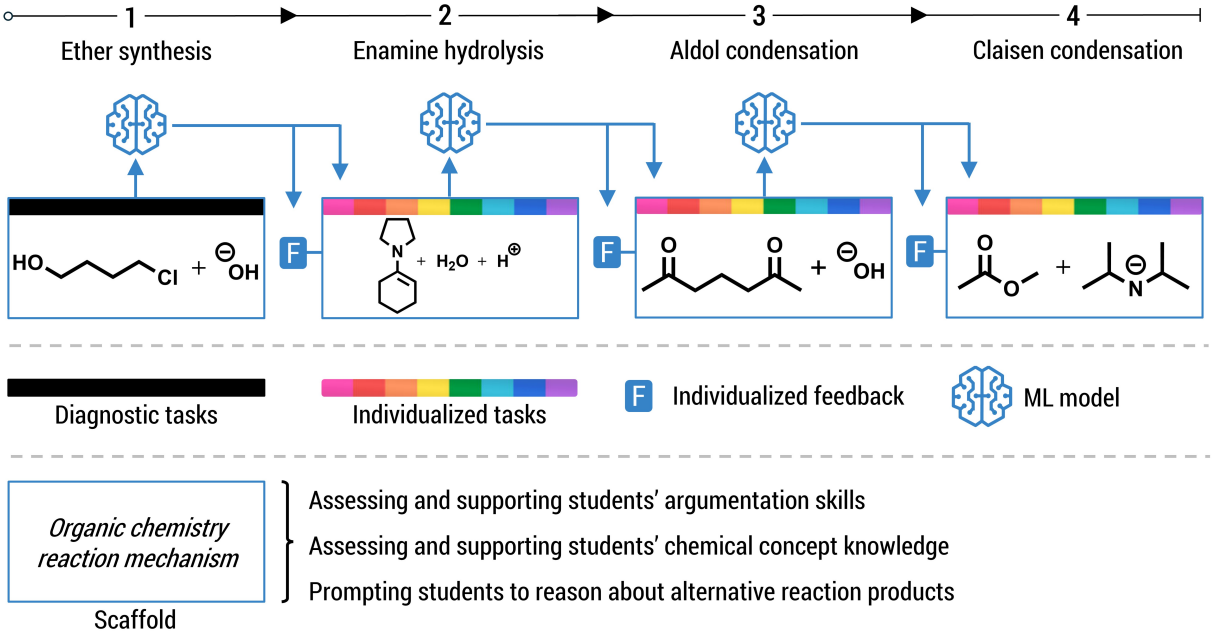


FIGURE 4 Using machine learning to support Organic Chemistry II students' mechanistic reasoning adaptively at four time points throughout one semester.

packages unless specified otherwise. Unless stated otherwise, all statistical analyses were conducted at a significance level of $\alpha < .05$.

3.3.1 Modeling students' mechanistic reasoning trajectories

To examine how students' mechanistic reasoning complexity evolved under divergent adaptive support, we aggregated each student's mechanistic reasoning complexity across multiple responses within every time point. This aggregate—ranging from 0.00 for low mechanistic reasoning complexity to 5.00 for high mechanistic reasoning complexity—was used to model students' mechanistic reasoning trajectories, which reflect different growth patterns. We modeled students' mechanistic reasoning trajectories inductively from the data after the study. To classify students into different trajectories based on their mechanistic reasoning complexity, we used the *kml* package, which applies a *k*-means algorithm designed for exploratory longitudinal data to model piecewise linear trajectories (Genolini et al., 2015; Genolini & Falissard, 2010, 2011). Further details on the analysis, such as how we determined the optimal number of trajectories and assessed their robustness, are available in Martin et al. (in preparation).

3.3.2 Investigating the gains in students' mechanistic reasoning over time

We conducted a dependent *t*-test to examine the gains in students' mechanistic reasoning complexity from the first to the final time point of the study. Dependent *t*-tests were performed both across all trajectories to assess the overall effectiveness of the adaptive support and separately for each trajectory to determine whether all trajectories benefited equally from the support and whether differences were mitigated, maintained, or exacerbated. We report Cohen's d_z as a measure of effect size.

Next, we performed a Grambsch test (Grambsch, 1994)—a robust alternative to the *F*-test for repeated-measures data—to assess the equality of variances in students' mechanistic reasoning complexity between the first and final time points of the study. A statistically significant Grambsch test indicates that the variances of the residuals change over time. The test was performed using the *PairedData* package (Champely, 2013).

3.3.3 Identifying the students who are improving their mechanistic reasoning

To investigate which students benefited from the adaptive support, we used an intersectional approach (Crenshaw 1989, 1991). Traditional analyses of singular identities (e.g., analyzing gender and race/ethnicity separately) overlook the complex intersections of these identities and their impact on students' educational experiences (e.g., Bauer et al., 2021; Crenshaw, 1991; Eddy et al., 2015; Riegle-Crumb et al., 2019; Wright et al., 2016). We categorized students based on three binary variables: binary gender (e.g., women and men), racial/ethnic background (e.g., URM and non-URM), and students' major (e.g., non-chemistry/biochemistry and chemistry/biochemistry). We included the distinction between chemistry/biochemistry majors (hereafter referenced as "majors") and non-chemistry/biochemistry majors (hereafter referenced as "non-majors") since the latter group may face institutional inequities in their undergraduate chemistry studies due to fewer opportunities to engage in discipline-specific learning communities. The intersections of these three identity categories resulted in eight distinct combinations of intersectional identities (Table 1). The six students with other gender identities and 27 students who chose not to report all demographic data were excluded from this analysis. Our analysis is limited to these three social identities due to data availability and ethical considerations related to the collection and use of sensitive demographic information.

TABLE 1 Distribution of students' intersectional identities.

	Chemistry/biochemistry major		Non-chemistry/biochemistry major	
	Men	**Women**	**Men**	**Women**
URM	3	3	10	22
Non-URM	18	24	56	97

We used a χ^2 test to examine associations between students' mechanistic reasoning trajectories and their intersectional identities (Agresti, 2013). Since some expected frequencies in the contingency table fell below five, we also conducted a Fisher's exact test using simulated data with 10,000 replications to ensure the robustness of our findings (Patefield, 1981). Finally, we performed a post hoc analysis of the standardized residuals from the χ^2 test to identify specific associations between students' mechanistic reasoning trajectories and their intersectional identities (Agresti, 2013; Beasly & Schumacker, 1995), highlighting combinations that occurred more or less frequently than statistically expected. We applied the *gmodels* package to perform the χ^2 test and Fisher's exact test (Warnes et al., 2005).

3.4 Findings

On average, our longitudinal adaptive support significantly improved students' mechanistic reasoning complexity from the study's first time point ($\bar{x}_1$ = 2.88, σ_1 = 0.56) to the last time point ($\bar{x}_4$ = 3.11, σ_4 = 0.65) with a small effect size: $t(265) = -5.909$, $p < .001$, $d_z = 0.36$. At first glance, it appears that the divergent adaptive support benefited all students' mechanistic reasoning. However, a more detailed analysis revealed that the effectiveness of the adaptive support varied significantly across students.

3.4.1 Modeling students' mechanistic reasoning trajectories

We monitored students' mechanistic reasoning across the semester-long course (Figure 5a) and used a longitudinal cluster technique to group students into five trajectories (Figure 5b), each linked to varying mechanistic reasoning complexity across the four time points. Students with the highest mechanistic reasoning complexity, illustrated by trajectory A (yellow), started at a complexity level of about 3.5 and followed a slightly upward trend throughout the semester with minor fluctuations. Students with moderate-to-high mechanistic reasoning complexity, visualized by trajectory B (red), began at a complexity of 3.0 and showed stable mechanistic reasoning complexity with an increase by the term's end. Students with low-to-moderate mechanistic reasoning complexity, represented by trajectories C (blue) and D (green), started at a similar complexity of around 2.5 but diverged significantly, showing considerable heterogeneity in mechanistic reasoning complexity over time. Lastly, students with the lowest mechanistic reasoning complexity, reflected by trajectory E (purple), began with a score near 2.5 but experienced declines during the third and fourth time points, ultimately ending with a complexity of 2.0.

To reiterate, the support was adapted to students' mechanistic reasoning complexity at the preceding time point of the study. For example, the adaptive support at the second time point was based on students' mechanistic reasoning complexity at the first time point. This means that, for instance, students in trajectory A consistently received support targeting more complex mechanistic reasoning while students in trajectory E received support addressing less complex mechanistic reasoning.

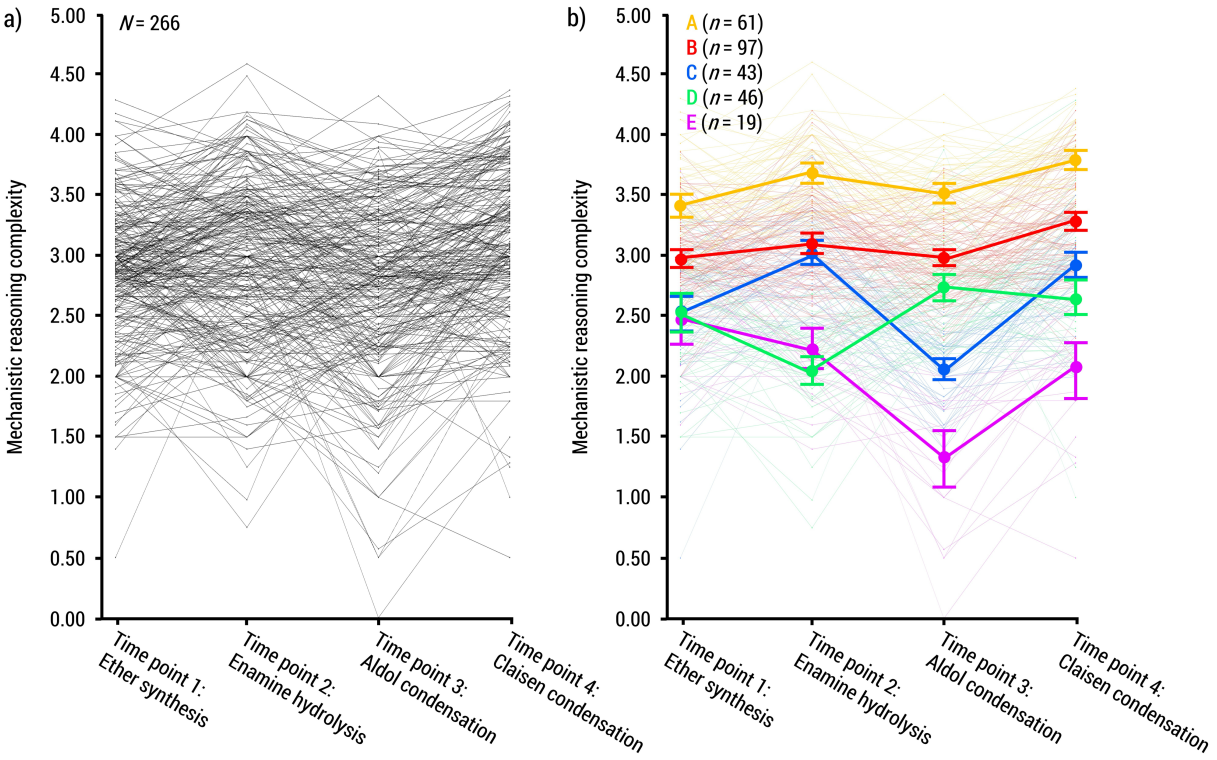


FIGURE 5 a) Students' mechanistic reasoning trajectories. b) Classifying students' mechanistic reasoning trajectories into five generalizable patterns through longitudinal clustering. *Note:* Error bars visualize 95% confidence intervals adjusted for longitudinal data.

3.4.2 Investigating the gains in students' mechanistic reasoning over time

In addition to modeling students' mechanistic reasoning trajectories, we calculated their gains in mechanistic reasoning complexity from the first to the last time point (Table 2). We report Cohen's d_z to statistically compare the gains in mechanistic reasoning complexity across trajectories. Trajectories A and B showed medium gains in mechanistic reasoning complexity reflected by a Cohen's d_z of 0.64 and 0.55, respectively. Trajectory C also exhibited a medium gain in mechanistic reasoning complexity with a Cohen's d_z of 0.54. Trajectory D displayed a negligible gain with a Cohen's d_z of 0.14. Finally, trajectory E experienced a decline in mechanistic reasoning complexity with a medium effect as demonstrated by a Cohen's d_z of 0.65. These results indicate that the adaptive support had a differential impact on students across trajectories.

Table 2 further reveals that the standard deviations tended to increase over time due to the adaptive support. Specifically, the standard deviation of students' mechanistic reasoning complexity increased significantly from the first time point (σ_1 = 0.56) to the final time point (σ_4 = 0.65) of the study (Z = –2.188, p = 0.029). This suggests that the adaptive support contributed to greater heterogeneity in students' mechanistic reasoning.

3.4.3 Identifying the students who are improving their mechanistic reasoning

The χ^2 test revealed no significant association between students' mechanistic reasoning trajectories and their intersectional identities (Table 1) ($\chi^2(32)$ = 38.30, p = 0.233; Cohen's W = 0.377; Cramer's V = 0.188). This result was further confirmed by Fisher's exact test (p = 0.143). However, a possible trend was identified when examining the standardized residuals. Non-URM men who are majors were more frequently assigned to trajectory A than statistically expected ($n_{expected}$ = 4.094, $n_{observed}$ = 8, p = 0.054). In contrast, URM women who are non-majors were less frequently assigned to trajectory A

TABLE 2 Mechanistic reasoning complexity of trajectories separated by time points.

Trajectory (Students)	Time Point 1		Time Point 2		Time Point 3		Time Point 4		Overall Gain
	$\bar{x}_1$	σ_1	$\bar{x}_2$	σ_2	$\bar{x}_3$	σ_3	$\bar{x}_4$	σ_4	$\bar{x}_4 - \bar{x}_1$ (Cohen's d_z)
A (n = 61)	3.41	0.41	3.67	0.37	3.48	0.36	3.75	0.32	0.34 (0.64***)
B (n = 97)	2.96	0.34	3.09	0.42	2.95	0.34	3.26	0.41	0.30 (0.55***)
C (n = 43)	2.52	0.57	2.98	0.39	2.02	0.31	2.88	0.39	0.36 (0.54***)
D (n = 46)	2.51	0.51	2.04	0.41	2.72	0.40	2.62	0.58	0.11 (0.14)
E (n = 19)	2.45	0.53	2.19	0.43	1.29	0.58	2.03	0.58	–0.42 (0.65*)
Combined	2.88	0.56	2.96	0.69	2.76	0.72	3.11	0.65	0.23 (0.36***)

Note: $\bar{x}$ indicates means and σ indicates standard deviations. Cohen's d_z = measure of effect size (Cohen's d_z > 0.20 indicates a small effect, Cohen's d_z > 0.50 indicates a medium effect, Cohen's d_z > 0.80 indicates a large effect (Cohen, 1988)). *p < .05; **p < .01; ***p < .001.

than statistically expected ($n_{expected}$ = 5.004, $n_{observed}$ = 1, p = 0.074); these students were more often assigned to trajectory C ($n_{expected}$ = 3.871, $n_{observed}$ = 8, p = 0.036). All other standardized residuals showed no notable deviations.

In summary, our analysis identified five distinct mechanistic reasoning trajectories. These trajectories revealed significant differences in students' mechanistic reasoning complexity at the beginning of the study. In particular, the mechanistic reasoning complexity of trajectory A was 0.5 points higher than that of trajectory B, and trajectory B scored 0.5 points higher than trajectories C, D, and E. The adaptive support improved students' mechanistic reasoning complexity across all trajectories with an overall gain of 0.23 points, indicating a small gain in mechanistic reasoning complexity over time. However, trajectories A, B, and C exhibited medium gains, trajectory D showed a negligible gain, and trajectory E demonstrated a medium decline. Thus, differences in mechanistic reasoning complexity were maintained between trajectories A, B, and C; differences between these trajectories and trajectories D and E were even exacerbated. The adaptive support did not reduce the heterogeneity in students' mechanistic reasoning complexity; instead, student heterogeneity in mechanistic reasoning complexity tended to increase. Beyond that, non-URM men who are majors seemed to benefit more from the adaptive support than URM women who are non-majors

4 | Discussion of findings

In the illustrated study, we implemented ML-assisted divergent adaptivity to support undergraduate organic chemistry students' mechanistic reasoning. Our findings on students' prior knowledge and their learning gains align partially with the studies by Koedinger et al. (2023) and Sbeglia and Nehm (2024). Koedinger et al. (2023) conducted a meta-analysis across 27 datasets of student interactions with online practice systems from elementary to postsecondary courses in math, science, and language arts. They revealed that students typically entered courses with moderate prior knowledge, evidenced by approximately 65% of correct responses across these datasets. Similarly, Sbeglia and Nehm (2024) investigated the effects of non-adaptive instruction in undergraduate evolutionary biology courses. They found that students generally exhibited low evolution knowledge, ranging from 46% to 59%, and held numerous non-normative conceptions. In our study, students began with a mechanistic reasoning complexity of 2.88 out of 5.00—equivalent to 58% of the total score—mirroring the findings from the other two studies. More importantly, all studies indicate that students'

knowledge at course entry varied significantly, indicating substantial differences in prior knowledge. Both Koedinger et al.'s (2023) and our analysis suggest that scores of students with higher and lower prior knowledge differed by approximately 20% at course entry. Non-URM men who are majors tended to exhibit the highest prior knowledge in our study. Sbeglia and Nehm (2023) reported smaller differences in prior knowledge, ranging from 3% to 9%, with URM students or those lacking prior undergraduate biology courses scoring lower than their peers.

Regarding students' learning gains, Sbeglia and Nehm (2024) observed significant gains in evolution knowledge and a reduction in non-normative conceptions over time, predominantly with large effect sizes. Differences in evolution knowledge and non-normative conceptions between students with varying binary genders, racial/ethnic backgrounds, or prior undergraduate biology courses were maintained but not exacerbated. Koedinger et al. (2023) also found that students with lower and higher prior knowledge progressed at similar rates, leading to maintained differences. As a result, these studies conclude that differences in learning outcomes are not attributable to inherent variations in students' learning rates. Instead, differences in students' learning outcomes arise from inequities in environmental factors, such as access to high-quality learning opportunities (Gutiérrez, 2008; Koedinger et al., 2023). Here, seven learning opportunities are, on average, required to master a knowledge component—almost irrespective of a student's learning rate (Koedinger et al., 2023). However, students with lower prior knowledge need about 13 learning opportunities to master a knowledge component while those with higher prior knowledge require only four. To reiterate, these differences stem from environmental factors, such as fewer learning opportunities leading to lower prior knowledge, not from slower learning rates. The similarity in students' learning rates underscores that enhanced access to high-quality learning opportunities can reduce differences in learning outcomes. All students can master higher-order cognitive operations—provided they have access to high-quality learning opportunities and are willing to invest effort in them (Koedinger et al., 2023).

Koedinger et al.'s (2023) and Sbeglia and Nehm's (2024) findings on students' learning gains partly contrast with our findings. The gains in mechanistic reasoning complexity in our study had a medium effect, and significant improvements were only observed for trajectories A, B, and C—encompassing approximately 76% of all students. This indicates that only trajectories A, B, and C maintained their differences in mechanistic reasoning complexity; conversely, differences between these trajectories and trajectories D and E were exacerbated. This suggests that the adaptivity employed in our study may still benefit students with higher prior knowledge. Other research assumed that adaptivity can exacerbate differences in learning outcomes because more challenging learning paths often provide learning opportunities that address higher-order cognitive operations, and students with lower prior knowledge are less frequently assigned to them (Terrin & Triventi, 2023; Van de Werfhorst & Mijs, 2010). Following that, students with higher prior knowledge can benefit disproportionately from these advanced learning opportunities, exacerbating the differences between themselves and their peers (Adams, 1990; Kempe et al., 2011)—a phenomenon known as the Matthew effect (Stanovich, 1986). In their review, Steenbergen-Hu and Cooper (2013) identified exacerbating learning-related differences in K-12 mathematics classrooms using ITSs, where students with higher prior knowledge benefitted more from adaptive support than those with lower prior knowledge. These negative effects of ITSs on students with lower prior knowledge are also observed in other domains (e.g., Doroudi & Brunskill, 2019; Hansen

& Reich, 2015; Rau, 2015). Thus, relying solely on ITSs to implement adaptivity appears insufficient to reduce student heterogeneity and mitigate inequities in students' learning.

These findings may explain why some students benefited more from the adaptive support than others in our instructional setting. In our setting, students with lower mechanistic reasoning complexity were provided with tasks that targeted less complex mechanistic reasoning. As a result, these students had fewer opportunities to practice more complex mechanistic reasoning, even though seven high-quality learning opportunities are necessary to master a knowledge component (Koedinger et al., 2023). Consequently, our adaptive instructional setting may have restricted gains in mechanistic reasoning complexity for these students by design. Alternatively, overly challenging tasks might have also hindered gains in mechanistic reasoning complexity for these students. Therefore, implementing divergent adaptivity through ML is not straightforward and may have a detrimental impact on students with lower prior knowledge. This presents a new adaptivity dilemma. While ML-assisted adaptive support can extend instructional time, determining the optimal task difficulty for every student to improve learning outcomes is not straightforward. Both too easy and too challenging tasks can hinder students' progress, highlighting the complexity of providing effective adaptive support. In addition to our divergent adaptivity, elements of convergent adaptivity may further support student learning.

Overall, convergent and divergent adaptivity offer valuable benefits. As instructors and education researchers, it is essential to determine the learning objectives within a course at the individual and cohort levels to decide which adaptivity aim to prioritize, particularly when adaptivity is implemented with digital technologies (Deunk et al., 2018; Holstein & Doroudi, 2022). Convergent adaptivity may help mitigate differences in students' learning outcomes, divergent adaptivity allows each student to progress independently. Nevertheless, both approaches are likely not implemented distinctly in everyday instruction (Deunk et al., 2015). For instance, educational technologies can combine convergent and divergent adaptivity by keeping students' learning outcomes within a curriculum synchronized to a certain extent (Holstein & Doroudi, 2022). This means that students with higher prior knowledge are permitted to progress at their own pace; however, once they reach key milestones, they are provided with opportunities for enrichment to deepen their understanding (Deunk et al., 2015). These opportunities allow students with higher prior knowledge to explore specific areas in greater depth without significantly outpacing their peers in the curriculum.

5 | Crafting action plans: Reflecting on our ML-assisted divergent adaptivity to improve future implementations

The findings of the illustrated study suggest that ML-assisted divergent adaptivity alone is insufficient to reduce student heterogeneity in mechanistic reasoning complexity and mitigate inequities in students' learning. Building on the steps proposed in the enhanced educational debt framework (Figure 1), we have developed an action plan to improve future implementations of our ML-assisted divergent adaptivity. Above all, integrating elements of convergent adaptivity may provide additional support for student learning. For example, when using an ML model to implement divergent adaptivity, it may be beneficial to add elements of human-implemented adaptivity that align with convergent aims to increase the effectiveness of the adaptivity for all students. Therefore, incorporating instructional strategies into our ML-assisted divergent adaptivity that follow convergent aims may be valuable in future implementations. However, the context

dependency of the trajectories highlights the importance of not ending convergent adaptivity too early. Achieving the convergence of students' learning outcomes at one time point does not guarantee similar learning outcomes at subsequent times. Therefore, efforts to converge students' learning outcomes should extend across a sequence of time points, rather than being limited to a single time point.

To converge students' learning outcomes over time, identifying students with lower mechanistic reasoning complexity at the beginning of the semester could allow instructors to provide additional human tutoring to these students. This human tutoring could include pre-teaching, extended instruction, or supporting instruction for review, offered on a voluntary basis (Smets & Struyven, 2018). For example, instructors can provide extra guidance to students like those assigned to trajectory E in this illustrated study. Since trajectory E comprises only 7.1% of students, offering additional human tutoring could be feasible, even in large-enrollment courses. Additionally, incorporating peer learning assistants in class or encouraging peer tutoring outside of class could be effective strategies for implementing convergent adaptivity where students with higher prior knowledge assist their peers (Holstein et al., 2018; Holstein & Doroudi, 2022; Sbeglia & Nehm, 2024). Supporting students with lower prior knowledge must be approached with sensitivity and avoid labeling them as low-achievers.

Identifying students with lower mechanistic reasoning complexity may be more reliable when their trajectory across multiple tasks is known since context-dependent variations in their trajectories make it difficult to adapt future formative assessments based on a student's mechanistic reasoning complexity in a single prior context. For this reason, it would be more appropriate to implement a series of formative assessments early in the semester and identify students who may benefit from more support based on their mechanistic reasoning complexity across multiple tasks and contexts. Incorporating a greater number of formative assessments may increase the effectiveness of adaptive support and provide a more detailed understanding of students' learning trajectories under adaptive support. Following that, future implementations could increase the frequency of adaptive formative assessments to better support students and offer more fine-grained insights into when and to what extent students' mechanistic reasoning complexity tends to converge or diverge. However, increasing the number of formative assessments may also disadvantage students whose time for coursework is constrained by other academic duties, familial responsibilities, or work obligations, among other responsibilities in their already busy daily lives.

An instructor dashboard could facilitate monitoring and supporting students with lower mechanistic reasoning complexity. This interactive digital tool provides educators with real-time data on student learning outcomes to pinpoint a student's challenges, monitor student progress, identify inequities, and adapt instruction (e.g., Bayer et al., 2024; Karademir et al., 2024; Knoop-van Campen et al., 2023). Evidence suggests that integrating a dashboard into traditional teaching practices may benefit student learning across educational levels (Karademir et al., 2024; Martinez-Maldonado et al., 2015). Furthermore, these dashboards appear to help instructors use a wider range of pedagogical content knowledge, offer more useful process-focused feedback, and dedicate significantly more time to supporting students with lower prior knowledge (Holstein et al., 2018; Karademir et al., 2024; Knoop-van Campen & Molenaar, 2020; Knoop-van Campen et al., 2023; Molenaar & Knoop-van Campen, 2019). Conversely, instructors without a dashboard provided less overall feedback, gave performance feedback over useful process feedback to students with lower prior knowledge, and dedicated instructional time

equivalently among all students (Holstein et al., 2018; Knoop-van Campen et al., 2023). Accordingly, teacher dashboards may promote convergent adaptivity by prompting instructors to prioritize students most in need (Holstein et al., 2018; Karademir et al., 2024).

Besides linking elements of divergent and convergent adaptivity, the extent to which adaptive learning is integrated with other teaching, learning, and assessment practices also appears important for their implementation success (Deunk et al., 2018; Smale-Jacobse et al., 2019). Consequently, in addition to the adaptive support methodology, numerous factors, including the type of course, the epistemological beliefs of instructors and learners, student motivation, and other contextual elements, must be considered for effective adaptive learning. Therefore, future implementations of ML-assisted divergent adaptivity should thoroughly integrate this support into traditional coursework, encompassing lectures, exercises, and homework. Relying solely on adaptive systems as supplements to conventional teaching practices without fostering dynamic interactions between instructors and students as well as properly aligning adaptive teaching, learning, and assessment methods is likely ineffective (Kulik & Fletcher, 2016).

Discussing divergent adaptivity through an adequacy-oriented conceptualization of equity prompts further assessment-related questions. For example, whether all students should take the same assessment after receiving adaptive support is debatable. Uniform assessments seem to conflict with divergent adaptivity since they do not account for each student's progress or educational goals. Accordingly, future instruction could implement alternative assessment approaches in adaptive learning environments (Yik et al., 2024). For instance, Clark and Talbert (2023) summarized four key features of alternative grading practices, which can inform the design of formative and summative assessments in classrooms that apply divergent adaptivity. First, they emphasize that learning objectives must be clear and measurable so students can demonstrate their assessment progress. Second, formative feedback must provide students with helpful evaluative information on their work to progress toward the learning objectives. Third, assignment marks should signal a student's progress toward meeting the defined objectives, rather than solely mirroring a number of reached points; for example, these marks can be expressed as words or short phrases that clarify a student's understanding and progress, such as "satisfactory" and "try again". Fourth, students should be allowed to reattempt tasks and resubmit their solutions for evaluative feedback without facing grade penalties. In this context, specifications grading (Nilson, 2014) and ungrading (Blum, 2020), among alternative grading practices, have become increasingly popular ways of grading, especially across science domains (Ferguson & Bonner, 2024; Hackerson et al., 2024).

6 | Conclusions

Students' diverse cognitive abilities and affective characteristics challenge educators to design adaptive instructional settings. Adaptive support promises to provide personalized learning experiences and advance students' understanding. By doing so, adaptive support can pursue two fundamentally different aims: the *convergence* or *divergence* of students' learning outcomes. This study offers initial insights into how divergent adaptivity can impact student progress across various levels of prior knowledge. Our approach allowed students to advance at their own pace, which benefited mostly students with moderate to high prior knowledge, so the adaptive support maintained and partly exacerbated differences in learning outcomes.

Our study highlights the importance of evaluating adaptivity in education through an

equity framework to ensure that adaptive support benefits all students and does not leave some behind. In doing so, our study provides neither evidence that ML-assisted divergent adaptivity reduces student heterogeneity in mechanistic reasoning complexity nor that it mitigates learning-related inequities. This means that the implemented adaptivity does not automatically ensure learning gains for all students. (e.g., Roberts-Mahoney et al., 2016). Following this, simply implementing ML-assisted divergent adaptivity does not necessarily benefit all students. To ensure more equitable outcomes, incorporating elements of convergent adaptivity, such as dashboards to identify students with lower prior knowledge or additional human tutoring through pre-teaching, extended instruction, or supplementary instruction, could help students with lower prior knowledge better progress within our adaptive setting. These considerations showcase that finding an optimal balance in supporting progress for all students remains a persistent challenge.

Beyond that, the learning process, for example, related to students' mechanistic reasoning in organic chemistry, seems to be non-linear, further complicating a straightforward implementation of ML-assisted divergent adaptivity to guide students' progress. For this reason, we conclude that introducing divergent adaptivity through ML should be complemented with human-implemented convergent adaptivity to support all students. Leveraging digital technologies like ML to enable adaptivity presents new complexities, such as determining whether the adaptivity aim should be to converge or diverge students' learning outcomes.

Acknowledgments

This publication is part of the first author's doctoral thesis (Dr. rer. nat.) at the Faculty of Biology and Chemistry, Justus-Liebig-University Giessen, Germany. We thank Peter Wulff and David Kranz for their assistance in implementing the machine learning analysis. Moreover, we thank Leonie Lieber, Ira Caspari-Gnann, and Krenare Ibraj for sharing their research instrument and data. Similarly, we thank Amber Dood, Field Watts, Ginger Shultz, and their research team for providing access to their Writing-to-Learn data to train our machine learning model. We also thank Brett McCollum and Nathan Ackroyd for allowing us to pilot our learning resources and gather data at their institution. Our further appreciation goes to Lukas Neuser, Constantin Stremmer, Marie Sahba, and Felix Blödtner for their assistance in implementing the study, as well as to all members of the Graulich group for productive discussions. Finally, Paul P. Martin thanks the German Chemical Industry Association (Verband der Chemischen Industrie) for supporting him with the Kekulé Fellowship. We acknowledge that ChatGPT was used to translate the Dutch-language references Blok (2004), Bosker (2005), and Veen et al. (2000) into English.

Conflict of interest statement

The authors Paul P. Martin, Marcus Kubsch, Brandon J. Yik, Benjamin T. Burlingham, and Nicole Graulich declare that they have no conflicts of interest.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print* (1st ed.). MIT Press.

Agresti, A. (2013). *Categorical data analysis* (1st ed.). Wiley.

Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2016). Instruction based on adaptive learning technologies. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (2nd ed., pp. 522-560). Routledge.

Ariely, M., Nazaretsky, T., & Alexandron, G. (2024). Causal-mechanical explanations in biology: Applying automated assessment for personalized

learning in the science classroom. *Journal of Research in Science Teaching*, *61*(8), 1858-1889. https://doi.org/10.1002/tea.21929

Asmussen, G., Rodemer, M., & Bernholt, S. (2024). Stepping stones to success: A qualitative investigation of the effectiveness of adaptive stepped supporting tools for problem-solving in organic chemistry to design an intelligent tutoring system. *International Journal of Science Education*, 1-23. https://doi.org/10.1080/09500693.2024.2361933

Bachtiar, R. W., Meulenbroeks, R. F. G., & van Joolingen, W. R. (2022). Mechanistic reasoning in science education: A literature review. *EURASIA Journal of Mathematics, Science and Technology Education*, *18*(11), Article em2178. https://doi.org/10.29333/ejmste/12512

Bauer, G. R., Churchill, S. M., Mahendran, M., Walwyn, C., Lizotte, D., & Villa-Rueda, A. A. (2021). Intersectionality in quantitative research: A systematic review of its emergence and applications of theory and methods. *SSM - Population Health*, *14*, Article 100798. https://doi.org/10.1016/j.ssmph.2021.100798

Bayer, V., Mulholland, P., Hlosta, M., Farrell, T., Herodotou, C., & Fernandez, M. (2024). Co-creating an equality diversity and inclusion learning analytics dashboard for addressing awarding gaps in higher education. *British Journal of Educational Technology*, *55*(5), 2058-2074. https://doi.org/10.1111/bjet.13509

Beasley, T. M., & Schumacker, R. E. (1995). Multiple regression approach to analyzing contingency tables: Post hoc and planned comparison procedures. *The Journal of Experimental Education*, *64*(1), 79–93. https://doi.org/10.1080/00220973.1995.9943797

Blum, S. D. (2020). *Ungrading: Why rating students undermines learning (and what to do instead)* (1st ed.). West Virginia University Press.

Blok, H. (2004). Adaptief onderwijs: Betekenis en effectiviteit [Adaptive Education: Meaning and Effectiveness]. *Pedagogische Studiën*, *81*(1). https://pedagogischestudien.nl/article/view/14617

Bosker, R. J. (2005). *De grenzen van gedifferentieerd onderwijs [The limits of differentiated education]* (1st ed.). Rijksuniversiteit Groningen.

Castle, S. D., Byrd, W. C., Koester, B. P., Pearson, M. I., Bonem, E., Caporale, N., ... & Matz, R. L. (2024). Systemic advantage has a meaningful relationship with grade outcomes in students' early STEM courses at six research universities. *International Journal of STEM Education*, *11*(1), Article 14. https://doi.org/10.1186/s40594-024-00474-7

Champely, S. (2013). *PairedData* (Version 1.1.1) [Computer software]. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/package=PairedData

Clark, D., & Talbert, R. (2023). *Grading for growth: A guide to alternative grading practices that promote authentic learning and student engagement in higher education* (1st ed.). Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Crandell, O. M., Kouyoumdjian, H., Underwood, S. M., & Cooper, M. M. (2019). Reasoning about reactions in organic chemistry: Starting it in general chemistry. *Journal of Chemical Education*, *96*(2), 213-226. https://doi.org/10.1021/acs.jchemed.8b00784

Crandell, O. M., Lockhart, M. A., & Cooper, M. M. (2020). Arrows on the page are not a good gauge: Evidence for the importance of causal mechanistic explanations about nucleophilic substitution in organic chemistry. *Journal of Chemical Education*, *97*(2), 313-327. https://doi.org/10.1021/acs.jchemed.9b00815

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics. *University of Chicago Legal Forum, 1989*(1), Article 8. http://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*(6), 1241-1299. https://doi.org/10.2307/1229039

Crévola, C. A., & Hill, P. W. (1998). Evaluation of a whole-school approach to prevention and intervention in early literacy. *Journal of Education for Students Placed at Risk*, *3*(2), 133-157. https://doi.org/10.1207/s15327671espr0302_4

Crowder, C. J., Yik, B. J., Frost, S. J. H., Cruz-Ramírez de Arellano, D., & Raker, J. R. (2024). Impact of prompt cueing on level of explanation sophistication for organic reaction mechanisms. *Journal of Chemical Education*, *101*(2), 398-410. https://doi.org/10.1021/acs.jchemed.3c00710

Deng, J. M., & Flynn, A. B. (2021). Reasoning, granularity, and comparisons in students' arguments on two organic chemistry items. *Chemistry Education Research and Practice*, *22*(3), 749-771. https://doi.org/10.1039/D0RP00320D

Deunk, M. I., Doolaard, S., Smale-Jacobse, A. E., & Bosker, R. J. (2015). *Differentiation within and across classrooms: A systematic review of studies into the cognitive effects of differentiation practices* (1st ed.). RUG/GION.

Deunk, M. I., Smale-Jacobse, A. E., de Boer, H., Doolaard, S., & Bosker, R. J. (2018). Effective differentiation practices: A systematic review and meta-analysis of studies on the cognitive effects of differentiation practices in primary education. *Educational Research Review*, *24*, 31-54. https://doi.org/10.1016/j.edurev.2018.02.002

Dood, A. J., Dood, J. C., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2020a). Analyzing explanations of substitution reactions using lexical analysis and logistic regression techniques. *Chemistry Education Research and Practice*, *21*(1), 267-286. https://doi.org/10.1039/C9RP00148D

Dood, A. J., Dood, J. C., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2020b). Using the research literature to develop an adaptive intervention to improve student explanations of an S_N1 reaction mechanism. *Journal of Chemical Education*, *97*(10), 3551-3562. https://doi.org/10.1021/acs.jchemed.0c00569

Doroudi, S., & Brunskill, E. (2019). Fairer but not fair enough: On the equitability of knowledge tracing. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (1st ed., pp. 335-339). Association for Computing Machinery. https://doi.org/10.1145/3303772.3303838

Dumont, H., & Ready, D. D. (2023). On the promise of personalized learning for educational equity. *npj Science of Learning*, *8*(1), Article 26. https://doi.org/10.1038/s41539-023-00174-x

Eddy, S. L., Brownell, S. E., Thummaphan, P., Lan, M.-C., & Wenderoth, M. P. (2015). Caution, student experience may vary: Social identities impact a student's experience in peer discussions. *CBE—Life Sciences Education*, *14*(4), Article ar45. https://doi.org/10.1187/cbe.15-05-0108

Ferguson, J. H., & Bonner, L. A. (2024). Ungrading in organic chemistry: Students assessing themselves and reflecting on their learning. *Frontiers in Education*, *9,* Article 1394042. https://doi.org/10.3389/feduc.2024.1394042

Frost, S. J. H., Yik, B. J., Dood, A. J., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2023). Evaluating electrophile and nucleophile understanding: A large-scale study of learners' explanations of reaction mechanisms. *Chemistry Education Research and Practice*, *24*(2), 706-722. https://doi.org/10.1039/D2RP00327A

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software*, *65*(4), 1-34. https://doi.org/10.18637/jss.v065.i04

Genolini, C., & Falissard, B. (2010). Kml: k-means for longitudinal data. *Computational Statistics*, *25*(2), 317-328. https://doi.org/10.1007/s00180-009-0178-4

Genolini, C., & Falissard, B. (2011). Kml: A package to cluster longitudinal data. *Computer Methods and Programs in Biomedicine*, *104*(3), e112-e121. https://doi.org/10.1016/j.cmpb.2011.05.008

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, *69*(S3), S342-S353. https://doi.org/10.1086/341857

Grambsch, P. M. (1994). Simple robust tests for scale differences in paired data. *Biometrika*, *81*(2), 359-372. https://doi.org/10.1093/biomet/81.2.359

Gutiérrez, R. (2008). A gap-gazing fetish in mathematics education? Problematizing research on the achievement gap. *Journal for Research in Mathematics Education*, *39*(4), 357-364. https://doi.org/10.5951/jresematheduc.39.4.0357

Hackerson, E. L., Slominski, T., Johnson, N., Buncher, J. B., Ismael, S., Singelmann, L., ... & Momsen, J. L. (2024). Alternative grading practices in undergraduate STEM education: A scoping review. *Disciplinary and Interdisciplinary Science*

Education Research, *6*, Article 15. https://doi.org/10.1186/s43031-024-00106-8

Hansen, J. D., & Reich, J. (2015). Democratizing education? Examining access and usage patterns in massive open online courses. *Science*, *350*(6265), 1245-1248. https://doi.org/10.1126/science.aab3782

Heckhausen, H. (1974). *Leistung und Chancengleichheit [Performance and equity of opportunity]* (1st ed.). Hogrefe.

Hedtrich, S., & Graulich, N. (2018a). Translating numbers into feedback: Providing students with automatically generated feedback. *American Journal of Educational Research*, *6*(2), 108-116. https://doi.org/10.12691/education-6-2-3

Hedtrich, S., & Graulich, N. (2018b). Using software tools to provide students in large classes with individualized formative feedback. *Journal of Chemical Education*, *95*(12), 2263-2267. https://doi.org/10.1021/acs.jchemed.8b00173

Holstein, K., & Doroudi, S. (2022). Equity and artificial intelligence in education. In W. Holmes & K. Porayska-Pomsta (Eds.), *The ethics of artificial intelligence in education: Practices, challenges, and debates* (1st ed., pp. 151-173). Routledge.

Holstein, K., Hong, G., Tegene, M., McLaren, B. M., & Aleven, V. (2018). The classroom as a dashboard: Co-designing wearable cognitive augmentation for K-12 teachers. In *Proceedings of the 8th International Conference on Learning Analytics & Knowledge* (1st ed., pp. 79-88). Association for Computing Machinery. https://doi.org/10.1145/3170358.3170377

Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*(1), 119-135. https://doi.org/10.1007/s13194-011-0038-2

Karademir, O., Borgards, L., Di Mitri, D., Strauß, S., Kubsch, M., Brobeil, M., ... & Drachsler, H. (2024). Following the impact chain of the LA cockpit: An intervention study investigating a teacher dashboard's effect on student learning. *Journal of Learning Analytics*, *11*(2), 215-228. https://doi.org/10.18608/jla.2024.8399

Kempe, C., Eriksson-Gustavsson, A. L., & Samuelsson, S. (2011). Are there any Matthew effects in literacy and cognitive development? *Scandinavian Journal of Educational Research*, *55*(2), 181-196. https://doi.org/10.1080/00313831.2011.554699

Kizilcec, R. F., Saltarelli, A. J., Reich, J., & Cohen, G. L. (2017). Closing global achievement gaps in MOOCs. *Science*, *355*(6322), 251-252. https://doi.org/10.1126/science.aag2063

Knoop-van Campen, C. A. N., & Molenaar, I. (2020). How teachers integrate dashboards into their feedback practices. *Frontline Learning Research*, *8*(4), 37-51. https://doi.org/10.14786/flr.v8i4.641

Knoop-van Campen, C. A. N., Wise, A., & Molenaar, I. (2023). The equalizing effect of teacher dashboards on feedback in K-12 classrooms. *Interactive Learning Environments*, *31*(6), 3447-3463. https://doi.org/10.1080/10494820.2021.1931346

Koedinger, K. R., Carvalho, P. F., Liu, R., & McLaughlin, E. A. (2023). An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences*, *120*(13), Article e2221311120. https://doi.org/10.1073/pnas.2221311120

Krist, C., Schwarz, C. V., & Reiser, B. J. (2019). Identifying essential epistemic heuristics for guiding mechanistic reasoning in science learning. *Journal of the Learning Sciences*, *28*(2), 160-205. https://doi.org/10.1080/10508406.2018.1510404

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, *86*(1), 42-78. https://doi.org/10.3102/0034654315581420

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in US schools. *Educational Researcher*, *35*(7), 3-12. https://doi.org/10.3102/0013189X035007003

Li, F., He, Y., & Xue, Q. (2021). Progress, challenges and countermeasures of adaptive learning: A systematic review. *Educational Technology & Society*, *24*(3), 238-255. https://www.jstor.org/stable/27032868

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022a). Closing the gap of organic chemistry students' performance with an adaptive scaffold for argumentation patterns. *Chemistry Education*

Research and Practice, *23*(4), 811-828. https://doi.org/10.1039/D2RP00016D

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022b). Students' individual needs matter: A training to adaptively address students' argumentation skills in organic chemistry. *Journal of Chemical Education*, *99*(7), 2754-2761. https://doi.org/10.1021/acs.jchemed.2c00213

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, *106*(4), 901-918. https://doi.org/10.1037/a0037123

Martin, P. P., & Graulich, N. (2024a). Beyond language barriers: Allowing multiple languages in postsecondary chemistry classes through multilingual machine learning. *Journal of Science Education and Technology*, *33*(3), 333-348. https://doi.org/10.1007/s10956-023-10087-4

Martin, P. P., & Graulich, N. (2024b). Lehre in der organischen Chemie individualisieren [Individualized teaching in organic chemistry]. *Nachrichten aus der Chemie*, *72*(3), 8-11. https://doi.org/10.1002/nadc.20244141003

Martin, P. P., & Graulich, N. (2024c). Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence. *Computers and Education: Artificial Intelligence*, *7*, Article 100265. https://doi.org/10.1016/j.caeai.2024.100265

Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2024). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*, *61*(8), 1757-1792. https://doi.org/10.1002/tea.21903

Martin, P. P., Kubsch, M., Yik, B. J., Burlingham, B. T., & Graulich, N. (in preparation). Tracing trajectories: Using machine learning to map how chemistry students' mechanistic reasoning develops over time.

Martinez-Maldonado, R., Clayphan, A., Yacef, K., & Kay, J. (2015). MTFeedback: Providing notifications to enhance teacher awareness of small group work in the classroom. *IEEE Transactions on Learning Technologies*, *8*(2), 187-200. https://doi.org/10.1109/TLT.2014.2365027

Molenaar, I., & Knoop-van Campen, C. A. N. (2019). How teachers make dashboard information actionable. *IEEE Transactions on Learning Technologies*, *12*(3), 347-355. https://doi.org/10.1109/TLT.2018.2851585

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academic Press.

Nilson, L. B. (2014). *Specifications grading: Restoring rigor, motivating students, and saving faculty time* (1st ed.). Routledge.

Noyes, K., & Cooper, M. M. (2019). Investigating student understanding of London dispersion forces: A longitudinal study. *Journal of Chemical Education*, *96*(9), 1821-1832. https://doi.org/10.1021/acs.jchemed.9b00455

Patefield, W. (1981). Algorithm AS 159: An efficient method of generating random R×C tables with given row and column totals. *Journal of the Royal Statistical Society*, *30*(1), 91-97. https://doi.org/10.2307/2346669

Plass, J. L., & Pawar, S. (2020a). Adaptivity and personalization in games for learning. In J. L. Plass, R. E. Mayer, & B. D. Homer (Eds.), *Handbook of game-based learning* (1st ed., pp. 263-281). MIT Press.

Plass, J. L., & Pawar, S. (2020b). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education*, *52*(3), 275-300. https://doi.org/10.1080/15391523.2020.1719943

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.2) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org

Rau, M. (2015). Why do the rich get richer? A structural equation model to test how spatial skills affect learning with representations. In O. C. Santos, J. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, ... & M. C. Desmarais (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (1st ed., pp. 350-357). International Educational Data Mining Society.

Riegle-Crumb, C., King, B., & Irizarry, Y. (2019). Does stem stand out? Examining racial/ethnic gaps in persistence across postsecondary fields. *Educational Researcher*, *48*(3), 133-144. https://doi.org/10.3102/0013189X19831006

Roberts-Mahoney, H., Means, A. J., & Garrison, M. J. (2016). Netflixing human capital development:

Personalized learning technology and the corporatization of K-12 education. *Journal of Education Policy*, *31*(4), 405-420. https://doi.org/10.1080/02680939.2015.1132774

Roshanaei, M., Olivares, H., & Lopez, R. R. (2023). Harnessing AI to foster equity in education: Opportunities, challenges, and emerging strategies. *Journal of Intelligent Learning Systems and Applications*, *15*(4), 123-143. https://doi.org/10.4236/jilsa.2023.154009

RStudio Team. (2023). *RStudio: Integrated development environment for R* (Version 2023.9.1.494) [Computer software]. RStudio. https://www.rstudio.com

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education*, *92*(3), 499-525. https://doi.org/10.1002/sce.20264

Sbeglia, G. C., & Nehm, R. H. (2024). Building conceptual and methodological bridges between SSE's diversity, equity, and inclusion statement and educational actions in evolutionary biology. *Evolution*, *78*(5), 809-820. https://doi.org/10.1093/evolut/qpae026

Shukla, S. Y., Theobald, E. J., Abraham, J. K., & Price, R. M. (2022). Reframing learning outcomes: Moving beyond achievement gaps. *CBE—Life Sciences Education*, *21*(2), Article es2. https://doi.org/10.1187/cbe.21-05-0130

Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. J. Durlach & A. M. Lesgold (Eds.), *Adaptive technologies for training and education* (1st ed., pp. 7-27). Cambridge University Press. https://doi.org/10.1017/CBO9781139049580.004

Slavin, R. E. (1996). *Education for all* (1st ed.). CRC Press.

Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research*, *79*(2), 839-911. https://doi.org/10.3102/0034654308330968

Smale-Jacobse, A. E., Meijer, A., Helms-Lorenz, M., & Maulana, R. (2019). Differentiated instruction in secondary education: A systematic review of research evidence. *Frontiers in Psychology*, *10*, Article 2366. https://doi.org/10.3389/fpsyg.2019.02366

Smets, W., & Struyven, K. (2018). Realist review of literature on catering for different instructional needs with preteaching and extended instruction. *Education Sciences*, *8*(3), Article 113. https://doi.org/10.3390/educsci8030113

Southard, K. M., Espindola, M. R., Zaepfel, S. D., & Bolger, M. S. (2017). Generative mechanistic explanation building in undergraduate molecular and cellular biology. *International Journal of Science Education*, *39*(13), 1795-1829. https://doi.org/10.1080/09500693.2017.1353713

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360-407. https://www.jstor.org/stable/747612

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K-12 students' mathematical learning. *Journal of Educational Psychology*, *105*(4), 970-987. https://doi.org/10.1037/a0032447

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, *106*(2), 331-347. https://doi.org/10.1037/a0034752

Stowe, R. L., Russ, R. S., & Wells, L. B. (in press). Mechanistic reasoning in-school vs. mechanistic reasoning in-life. In R. W. Bachtiar, M. Haskel-Ittah, & N. Graulich (Eds.), *Mechanistic reasoning in science education*. Springer.

Terrin, É., & Triventi, M. (2023). The effect of school tracking on student achievement and inequality: A meta-analysis. *Review of Educational Research*, *93*(2), 236-274. https://doi.org/10.3102/00346543221100850

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., ... & Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted*, *27*(2-3), 119-145. https://doi.org/10.1177/016235320302700203

Van de Werfhorst, H. G., & Mijs, J. J. B. (2010). Achievement inequality and the institutional structure of educational systems: A comparative perspective. *Annual Review of Sociology*, *36*, 407-428. https://doi.org/10.1146/annurev.soc.012809.102538

Van Dusen, B., Nissen, J., Talbot, R. M., Huvard, H., & Shultz, M. (2022). A QuantCrit investigation of society's educational debts due to racism and sexism in chemistry student learning. *Journal of Chemical Education*, *99*(1), 25-34. https://doi.org/10.1021/acs.jchemed.1c00352

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197-221. https://doi.org/10.1080/00461520.2011.611369

Veen, A. M., Roeleveld, J., & Leseman, P. P. M. (2000). *Evaluatie van Kaleidoscoop en Piramide: Eindrapportage [Evaluation of Kaleidoscope and Pyramid: Final report]* (1st ed.). SCO-Kohnstamm Instituut.

Viberg, O., Kizilcec, R. F., Wise, A. F., Jivet, I., & Nixon, N. (2024). Advancing equity and inclusion in educational practices with AI-powered educational decision support systems (AI-EDSS). *British Journal of Educational Technology*, *55*(5), 1974-1981. https://doi.org/10.1111/bjet.13507

Warnes, G. R., Bolker, B., Lumley, T., & Johnson, R. C. (2023). *gmodels: Various R programming tools for model fitting* (Version 2.19.1) [Computer software]. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/package=gmodels

Watts, F. M., Zaimi, I., Kranz, D., Graulich, N., & Shultz, G. V. (2021). Investigating students' reasoning over time for case comparisons of acyl transfer reaction mechanisms. *Chemistry Education Research and Practice*, *22*(2), 364-381. https://doi.org/10.1039/D0RP00298D

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE—Life Sciences Education*, *15*(2), Article ar23. https://doi.org/10.1187/cbe.15-12-0246

Xie, H., Chu, H.-C., Hwang, G.-J., & Wang, C.-C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Computers & Education*, *140*, Article 103599. https://doi.org/10.1016/j.compedu.2019.103599

Yik, B. J., Dood, A. J., Frost, S. J. H., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2023). Generalized rubric for level of explanation sophistication for nucleophiles in organic chemistry reaction mechanisms. *Chemistry Education Research and Practice*, *24*(1), 263-282. https://doi.org/10.1039/D2RP00184E

Yik, B. J., Machost, H., Streifer, A. C., Palmer, M. S., Morkowchuk, L., & Stains, M. (2024). Students' perceptions of specifications grading: Development and evaluation of the perceptions of grading schemes (PGS) instrument. *Journal of Chemical Education*, *101*(9), 3723-3738. https://doi.org/10.1021/acs.jchemed.4c00698

Young, K. R., Schaffer, H. E., James, J. B., & Gallardo-Williams, M. T. (2021). Tired of failing students? Improving student learning using detailed and automated individualized feedback in a large introductory science course. *Innovative Higher Education*, *46*(2), 133-151. https://doi.org/10.1007/s10755-020-09527-5

11 | Appendix

All appendices can be accessed digitally on the online platform of the *Open Science Framework* (OSF), a repository for open-access research and educational resources.

- For the **informed consent agreement form**, please refer to osf.io/nfx6k, file "Scaffold1_Ether-Synthesis.PDF," pages 1-4.
- For the **demographic questionnaire**, refer to osf.io/nfx6k, file "Scaffold4_Claisen-Condensation.PDF," pages 108-110.
- For the **adaptive scaffolds** designed and used with *Qualtrics*, please refer to osf.io/nfx6k, files "Scaffold1_Ether-Synthesis.PDF," "Scaffold2_Enamine-Hydrolysis.PDF," "Scaffold3_Aldol-Condensation.PDF," and "Scaffold4_Claisen-Condensation.PDF."
- For the **tasks on judging a product's plausibility**, please see Figure 5 (Chapter 1) and refer to osf.io/nfx6k, files "Scaffold1_Ether-Synthesis.PDF," "Scaffold2_Enamine-Hydrolysis.PDF," "Scaffold3_Aldol-Condensation.PDF," and "Scaffold4_Claisen-Condensation.PDF."
- For an extensive overview of **sample solutions**, refer to osf.io/nfx6k, files "SampleSolution_Ether-Synthesis.PDF," "SampleSolution_Enamine-Hydrolysis.PDF," "SampleSolution_Aldol-Condensation.PDF," and "SampleSolution_Claisen-Condensation.PDF."

Scan this QR code to access all appendices at: https://doi.org/10.17605/OSF.IO/NFX6K

Martin, P. P., Yik, B. J., Burlingham, B. T., & Graulich, N. (2025). Scaffolds to support students' mechanistic reasoning adaptively in organic chemistry. *Open Science Framework*. https://doi.org/10.17605/OSF.IO/NFX6K

The educational resources are published under the terms of a Creative Commons Attribution 4.0 International License (CC BY 4.0) (https://creativecommons.org/licenses/by/4.0/).

12 | Acknowledgments

I sincerely thank everyone who has supported me throughout my doctoral journey. Your expertise, guidance, and encouragement have been invaluable in shaping my professional and personal growth. The past three years have been an enriching experience, including numerous learning opportunities, significant collaborations, and wonderful moments in our research group. I am grateful for all the experiences I have gained in academia and beyond.

Prof. Dr. Nicole Graulich

First and foremost, I sincerely thank Nicole for her exceptional supervision and support. Her guidance—at any time, even on weekends and holidays—has allowed me to learn and grow immensely. Nicole has placed great trust in me from the beginning of this project. She allowed me to explore my research interests and guided the conceptualization of all our research manuscripts. Her thoughtful project administration and encouragement ensured a structured and productive research process. Her investigation and conceptualization of key ideas provided a strong foundation for my research. The scientific discussions with Nicole have been instrumental in shaping this work. I am grateful for her many theoretical, conceptual, methodological, and analytical ideas. Her expertise and insightful feedback throughout the planning, conceptualization, writing, review, and editing process have been indispensable in refining and synthesizing the research presented in this dissertation. Without her support, this dissertation would not have been possible. I am thankful that she made the past three years an incredibly enriching experience. Her dedication and mentorship have not only fostered my growth as a researcher but have also had a profound impact on my personal development. I sincerely appreciate everything she has done to support and guide me.

Prof. Dr. Marcus Kubsch

I sincerely thank Marcus for his continuous support and profound insights throughout my doctoral journey. Even before I began this journey, Marcus was willing to offer his advice. He shaped the project with his conceptual and methodological support. His contribution has been remarkable, including his willingness to review two of my manuscripts, “When a machine detects student reasoning” and “To what extent can we trust a machine.” Later, Marcus became an important collaborator in analyzing the collected data, discussing conceptual frameworks, refining research objectives, and preparing manuscripts. With his expertise in learning progression analytics, he greatly advanced the depth and quality of our work as outlined in the manuscripts “Tracing trajectories” and “To what extent is adaptive support equitable.” I am grateful for his willingness to guide and review my work, including this dissertation as a member of the examination committee. I truly appreciate his expertise, which has been a guiding force throughout my research.

Prof. Dr. Katja Lengnink

I sincerely thank Katja for her critical perspective on the application of artificial intelligence in society. I have learned a lot from her concept of algorithmic literacy, which played a significant role in shaping my work. Her ideas motivated me to write the manuscript “To what extent can we trust a machine.” Her thought-provoking questions have been instrumental in advancing my work. I also appreciate her willingness to be part of my examination committee.

Prof. Dr. Richard Göttlich

I sincerely thank Richard for his expertise in organic chemistry, which has greatly informed this project. I also thank Richard for his willingness to be part of the examination committee.

Prof. Dr. Brandon Yik

I sincerely thank Brandon for our collaboration during my doctoral journey. In the early stages of this project, Brandon's methodological expertise was key to my growth as a researcher. His support continued throughout the entirety of my doctoral journey. I am especially grateful for his dedication during the implementation of our study. Brandon went above and beyond to make the study possible: reaching out to numerous organic chemistry instructors, managing the IRB approval, and revising the learning resources—all of which required countless hours that I deeply appreciate. Brandon was also instrumental in the conceptualization of three manuscripts: "Mechanistic reasoning about alternatives," "Tracing trajectories," and "To what extent is adaptive support equitable." His methodological and conceptual advice significantly advanced the work. I am truly thankful for the effort he put into our joint work. Additionally, I sincerely appreciate his thoughtful review of this dissertation.

Prof. Dr. Peter Wulff

I sincerely thank Peter for immediately offering his support with the machine learning analysis. I am grateful for his generosity in sharing the Python scripts for the supervised and unsupervised machine learning analyses and his continuous methodological support throughout the implementation of this analysis. He always had an answer whenever I had technical questions, and his assistance was invaluable. I am especially thankful for his help in conceptualizing the manuscript "Exploring new depths"—without him, this manuscript would not have been possible. His methodological and analytical advice was crucial in helping me make rapid progress in applying machine learning methodology.

Prof. Dr. Ben Burlingham

I sincerely thank Ben for being willing to implement our study as an organic chemistry instructor. His careful implementation of our study was vital to its success, and his dedication ensured everything went smoothly. I also appreciate his genuine interest in supporting his students optimally and his enthusiasm for mechanistic reasoning. Ben's revision of our learning resources was key to improving them. I am grateful for his availability and our seamless collaboration.

Dr. Amber Dood

I sincerely thank Amber for our scientific exchange in the early months of my doctoral journey. She shared her methodological expertise, which made the start of my research much smoother. Her support continued throughout my entire journey. I am especially grateful that we could co-write the manuscript "To what extent can we trust a machine." I deeply appreciate her conceptualization of "AI trustworthiness," which she made available for our joint work. Also, I sincerely thank Amber for sharing her data with me, which I could use for machine learning model training in other projects.

Dr. Field Watts

I sincerely thank Field for his important methodological input at the start of my doctoral journey, where he shared his expertise and research ideas with me. Additionally, he contributed significantly to the manuscript “To what extent can we trust a machine” by providing thorough feedback.

Prof. Dr. Ginger Shultz

I sincerely thank Ginger for her input and support on the manuscript “To what extent can we trust a machine.” I am also grateful for her permission to use the data collected at her institution for machine learning model training.

Prof. Dr. Brett McCollum

I sincerely thank Brett for organizing the pilot study to test our learning resources. His support helped us refine our resources.

Prof. Dr. Nathan Ackroyd

I sincerely thank Nathan for his willingness to implement our learning resources in his organic chemistry course, allowing us to test and refine them.

Dr. Dennis Gerbig

I sincerely thank Dennis for providing me access to the machine learning infrastructure at the Department of Biology and Chemistry at Justus-Liebig-University during the initial phases of this project. His support in applying machine learning methodologies was instrumental to the progress of my research.

Dr. Irina Braun

I sincerely thank Irina for being such an important support during my doctoral journey. Above all, I enjoyed sitting across from her and spending hours together in front of the computer screen—often late into the evening. Even without many words, we always understood each other. We often shared similar views on many things. I also had many wonderful moments with Irina during my teaching. Early on in my work, we had the opportunity to co-lead the laboratory sessions of the Chemistry Education 2 module, and we had an amazing time doing so. I’m also grateful that we could co-author a publication towards the end of my doctoral journey. Additionally, I would like to sincerely thank Irina for her thorough review of this dissertation and assistance in formatting it.

Dr. Leonie Lieber

I sincerely thank Leonie for allowing me to take on her research project. Her thorough introduction to the subject matter enabled me to grasp the foundations of her qualitative and quantitative research. Leonie also provided me with all her materials. She allowed me to build upon her task format on alternative reaction products and to use her data for machine learning model training. The articles “Exploring new depths,” “Beyond language barriers,” and “Navigating the data frontier” would not have been possible without these data. Furthermore, I sincerely thank Leonie for the trust she placed in me as a researcher. Leonie recognized potential in me long before I saw it in myself—already during my studies. Leonie supported me and opened many doors for me. Without her, I probably would never

have started my work in chemistry education research. Finally, I would like to thank Leonie for her careful review of this dissertation.

David Kranz

I sincerely thank David for his interest in contributing to the development of our automated scoring model. I appreciate that he was immediately ready to assist me and did everything necessary to push the analysis forward. His expertise in programming and his willingness to improve my Python scripts took the technical aspect of our work to the next level. This includes his involvement in the work "Exploring new depths." I also thank David for the opportunity to co-author the paper "Revealing rubric relations." His methodological and conceptual advice was key to shaping this publication. Beyond these joint projects during my doctoral journey, I had the opportunity to collaborate with David as a research assistant on his research project during my studies. His introduction to the foundations of chemistry education research greatly supported my doctoral journey. I am also grateful for the chance to contribute to the manuscript "Should we scaffold it." During our collaborations, we shared many great moments, and I deeply respect David as a human being and a researcher.

Axel Langner

I sincerely thank Axel for creating such a positive atmosphere in our office. It was enjoyable working with him. Besides, I appreciate all the organizational tasks Axel took on in our research group.

Elias Heinrich

I sincerely thank Elias for critical and profound discussions about science and other topics. I always value his perspective. Additionally, his insights into the practical aspects of teaching chemistry have helped me better position my work within the field.

Sarah Saupe

I sincerely thank Sarah for all the fun moments we shared. Sarah's humor made everything more enjoyable.

Dr. Annette Geuther

I sincerely thank Dr. Geuther, with whom I have already enjoyed working during my studies. I learned a lot from her chemistry education courses. Our conversations and her guidance throughout my doctoral journey enriched my personal and professional development.

Bettina Romberg

I sincerely thank Bettina for all the fun and encouraging conversations. She supported me throughout my doctoral journey, and we always had something to laugh about. Bettina provided invaluable empathy to me. Also, I would like to thank her for her input in our laboratory work.

HEIKO BARTH

I sincerely thank Heiko, who was always someone I could rely on for technical or laboratory-related questions. Without Heiko, the laboratory sessions in our teaching would not have been possible. We all greatly benefit from his expertise. I would also like to thank Heiko for his support in formatting this dissertation.

SILKE WAGNER

I sincerely thank Silke, who always puts a smile on everyone's face. Her relaxed and cheerful attitude made it a pleasure to prepare our chemistry education courses together and discuss chemistry education topics, which often led me to reflect on my research.

DR. JANINE FELLER

I sincerely thank Janine for the insightful discussions on chemistry education, especially regarding adaptive support in high school chemistry. Our conversations prompted me to reflect on my research.

UNDERGRADUATE RESEARCH ASSISTANTS: FELIX BLÖDTNER, CONSTANTIN STREMMER, LUKAS NEUSER, MARIE SAHBA

I sincerely thank my undergraduate research assistants, Felix Blödtner, Constantin Stremmer, Lukas Neuser, and Marie Sahba. I am grateful to Felix for his assistance in coding students' mechanistic reasoning. I also thank him for translating students' mechanistic reasoning from English into German. Furthermore, I thank Felix and Lukas for paraphrasing numerous student responses using chatbots or generating new ones. Additionally, I am thankful to Constantin for implementing a network analysis of students' mechanistic reasoning. His analysis was useful for understanding the connections within students' responses. Finally, I am grateful to my research assistants for their efforts in implementing the study. All research assistants coded closed-response items, assigned adaptive learning pathways, and systematically collected the data—sometimes even on weekends. Without their help, it would not have been possible to carry out our study.

TO THE STUDENTS WHO PARTICIPATED IN OUR STUDIES

I sincerely thank all the students who willingly participated in the studies and allowed us to collect their data to improve future organic chemistry teaching. Your contribution will play a key role in advancing educational practices.

VERBAND DER CHEMISCHEN INDUSTRIE (GERMAN CHEMICAL INDUSTRY ASSOCIATION)

I sincerely thank the Verband der Chemischen Industrie (German Chemical Industry Association) for supporting me with the Kekulé Fellowship during my doctoral journey.

TO MY FAMILY AND FRIENDS

I sincerely thank my parents and friends for their constant support throughout my doctoral journey. They always had my back and provided the support and space to dedicate myself fully to my research. You are amazing!

Ich erkläre: Ich habe die vorgelegte Dissertation selbstständig und ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht. Ich stimme einer evtl. Überprüfung meiner Dissertation durch eine Antiplagiat-Software zu. Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ niedergelegt sind, eingehalten.