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

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

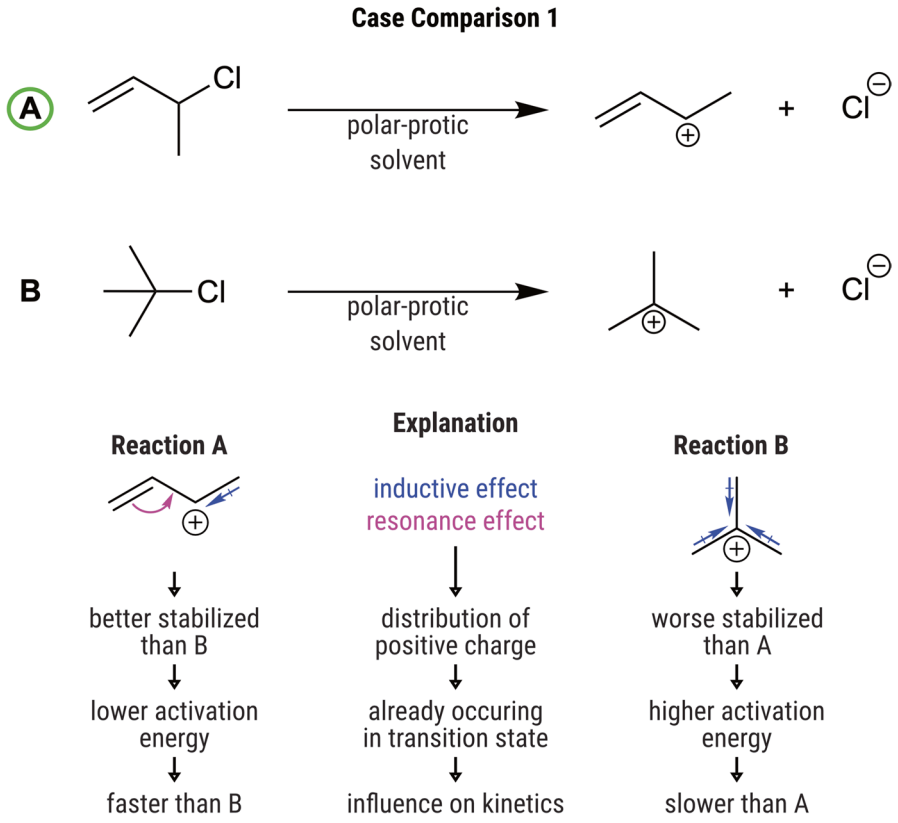

Fig. 1 Case comparison 1 with a sample solution included: The column on the left argues for reaction A, while the column on the right argues for reaction B. The column in the middle illustrates the general reasoning path. The faster reaction is highlighted in green

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.

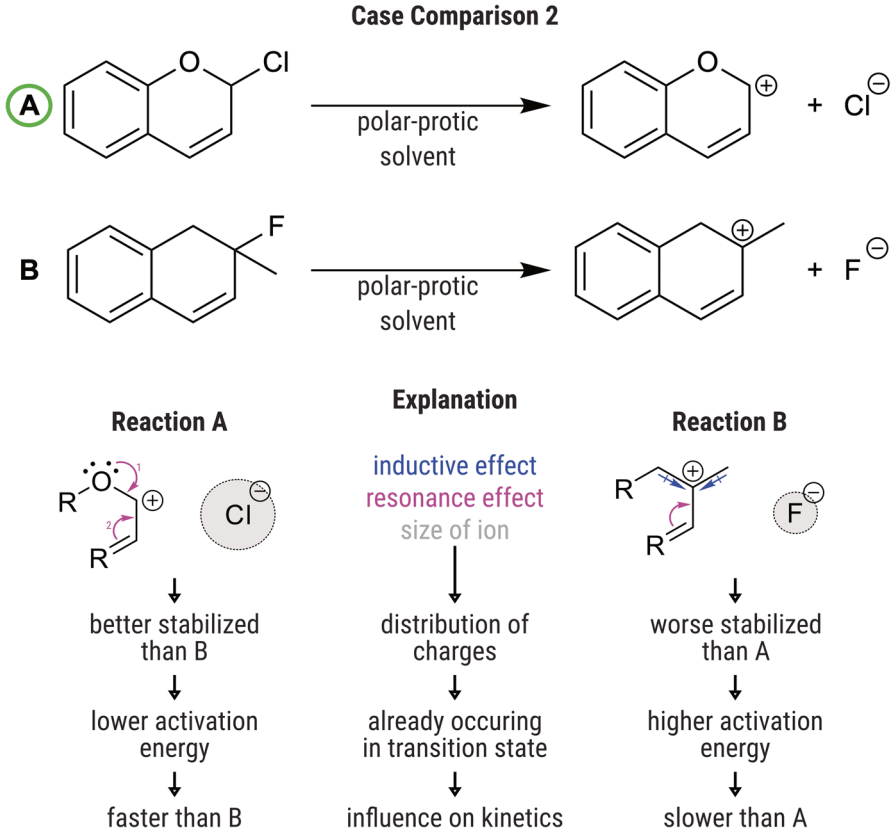

Fig. 2 Case comparison 2 with a sample solution included: The column on the left argues for reaction A, while the column on the right argues for reaction B. The column in the middle illustrates the general reasoning path. The faster reaction is highlighted in green

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).

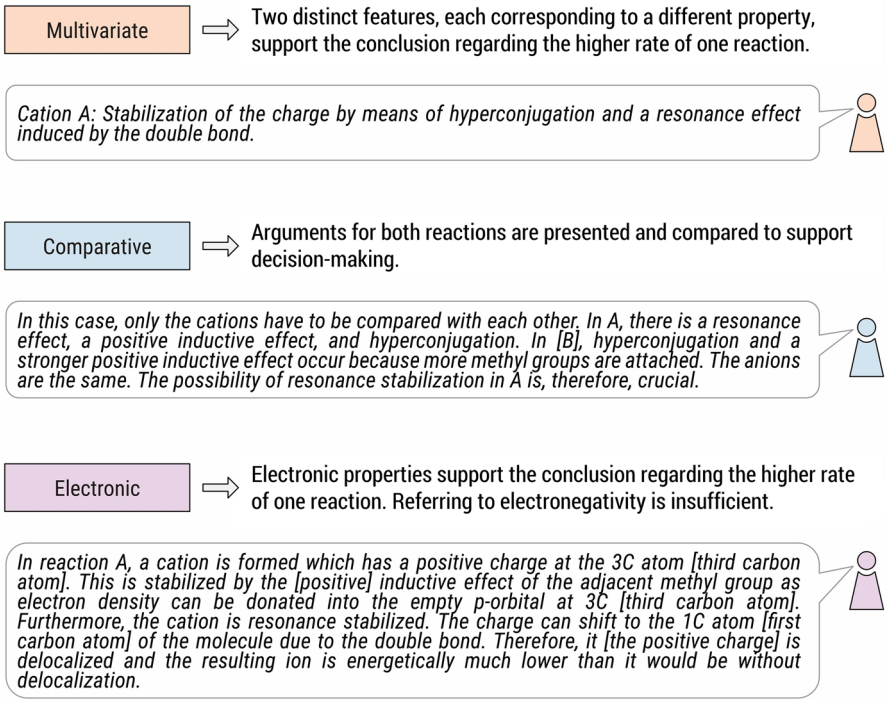

Fig. 3 Research-informed reasoning types with definitions and student-written examples included

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

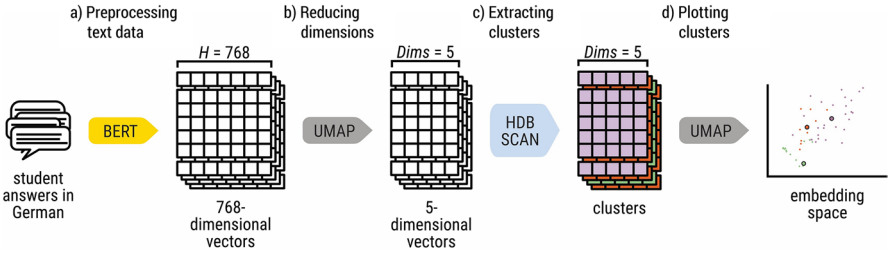

Fig. 4 Analysis steps for the development of the ML-based rubric

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).

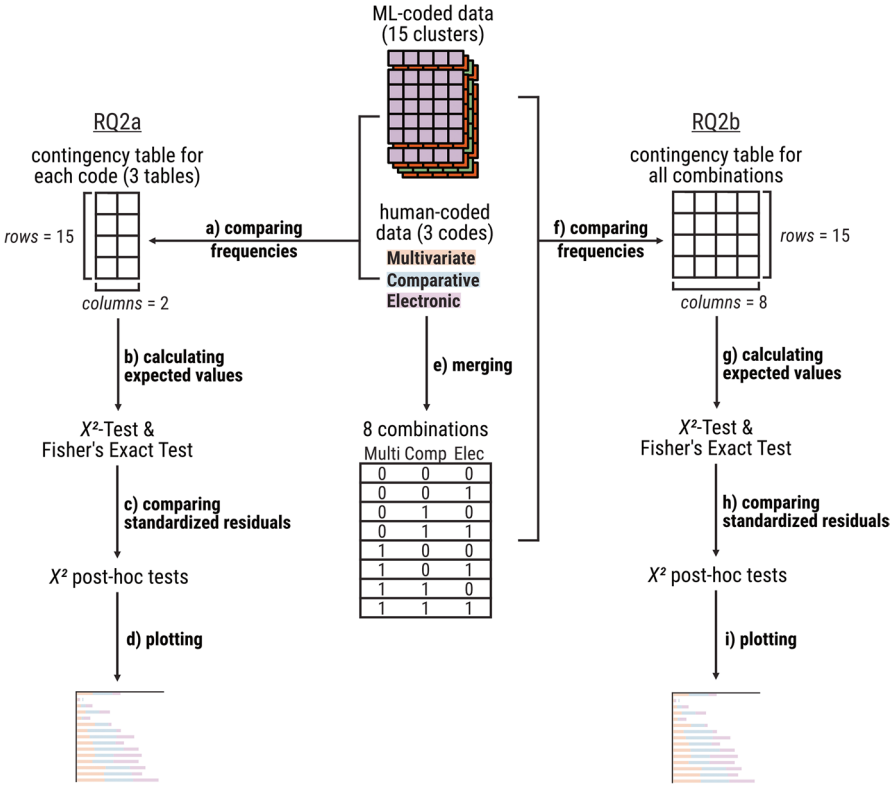

Fig. 5 Steps of the quantitative analysis. Letters a) – i) represent the sequence of the applied analysis steps

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.

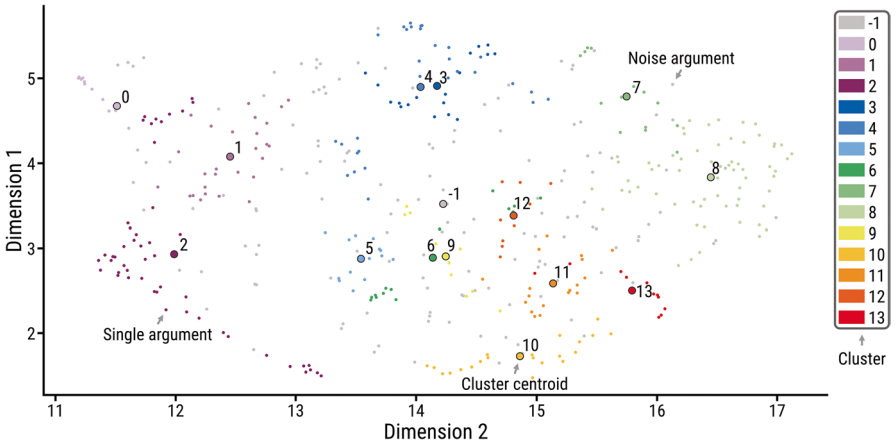

Fig. 6 Embedding space of the extracted clusters. *Note*: Each data point in the embedding space represents a student's written response. Different colors visualize different clusters. The gray points scattered throughout the embedding space show noise arguments that cannot be distinctly assigned to 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 ($n_{expected} = 10.86$, $n_{observed} = 1$, $p < .001$)

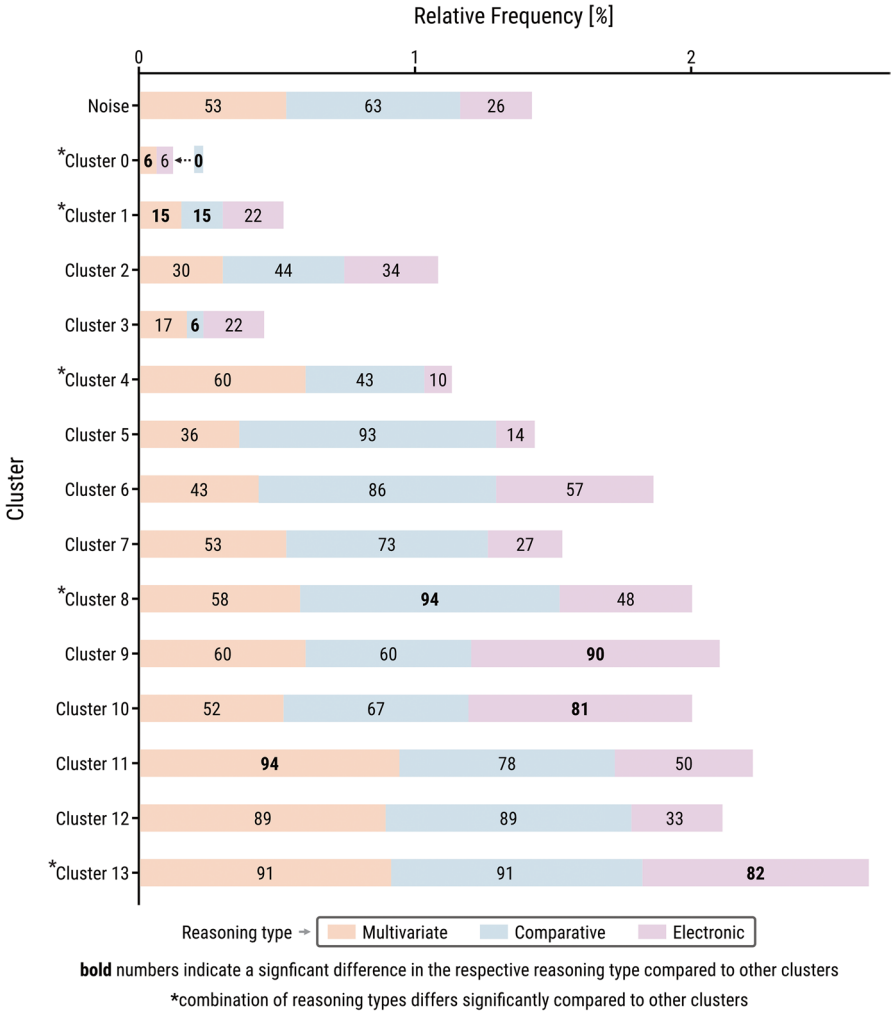

Fig. 7 Relative frequencies of the reasoning types in each cluster

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.

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].

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