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

DOI: 10.1002/tea.21903

RESEARCH ARTICLE

JRST | WILEY

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

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

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

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

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

Abstract

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

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

KEYWORDS

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

1 | INTRODUCTION

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

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

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

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

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

2 | METHODOLOGICAL CONSIDERATIONS

2.1 | ML and NLP basics

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

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

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

2.2 | Literature overview on capturing students' argumentation with ML

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

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

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

3 | STUDY OBJECTIVES

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

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

3.1 | Research questions

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

Pedagogical focus:

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

Methodological focus:

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

4 | METHODS

4.1 | Setting of data collection

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

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

4.2 | Research instrument

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

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

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

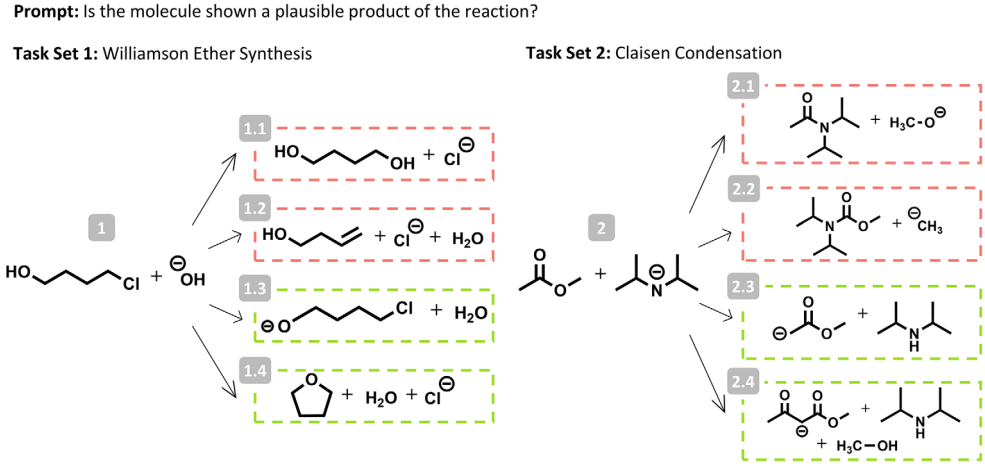
To scaffold the argumentation about competing chemical reactions, Lieber et al. (2022a, b) utilized a simplified version of Toulmin's argumentation pattern: the frequently applied three-component CER model based on the terms *claim*, *evidence (data)*, and *reasoning (warrant)* (McNeill et al., 2006).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

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

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

4.3 | Computational grounded theory

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



Sample solution for product 1.1

Claim: The product is not plausible.

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

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

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

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

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

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

4.3.1 | Pattern detection step

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

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

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

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

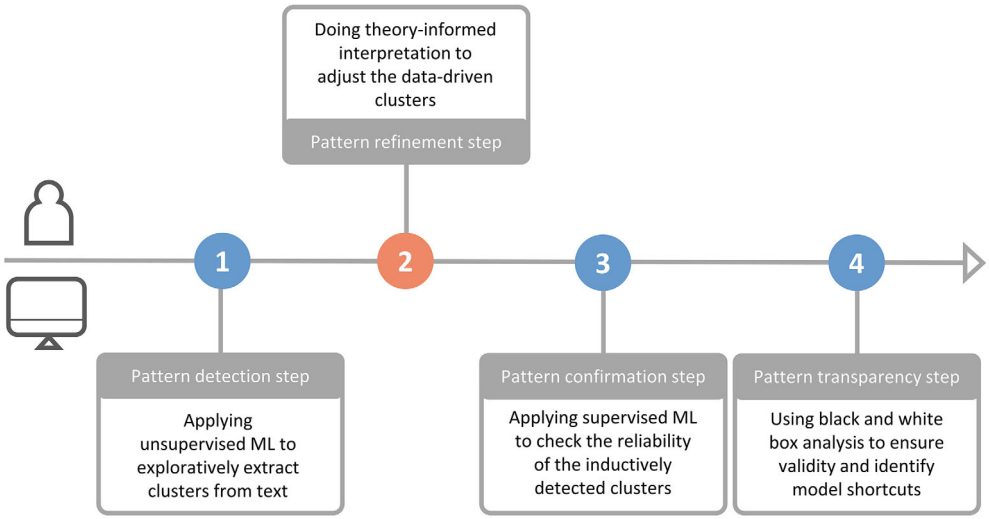


FIGURE 2 Revisited four-step computational grounded theory framework.

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

4.3.2 | Pattern refinement step

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

4.3.3 | Pattern confirmation step

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

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

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

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

4.3.4 | Pattern transparency step

We expanded *computational grounded theory* in our analysis by a fourth step, the pattern transparency step (Figure 2). During this step, methods for explaining the internal and external workings of an ML algorithm such as black and white box analysis or external validation can be applied. Black box analysis helps understand how an algorithm operates solely based on its observable input–output behavior. As the internal workings of ML models can be difficult to interpret, black box analysis is a time-efficient method for determining which features are driving a model's output. For implementing black box analysis, we modified four input sentences with a low level of argument complexity by systematically adding or removing words or phrases according to predefined criteria of our scoring rubric so that these sentences just reached the next level of the rubric (see sections 5.2 and 5.4). Subsequently, we checked whether the changed input led to the desired algorithmic classification.

However, black box analysis does not investigate the internal workings of an algorithm, which has led to the development of complex techniques for white boxing an algorithm. White box analysis aims to interpret the decision-making process of ML models by identifying the feature importance of individual words via post-hoc explanations (cf., Gombert et al., 2022). For white box analysis, we applied a technique called SHapley Additive exPlanations (SHAP) to calculate the importance scores of individual words for each rubric category (Lundberg &

Lee, 2017). This technique extracts keywords associated with positive or negative classifications in a rubric category and identifies critical model shortcuts (cf., Geirhos et al., 2020).

In sum, the pattern transparency step is essential for ensuring that an algorithm makes unbiased decisions. Transparently explaining the results of an ML algorithm is especially helpful when its decisions have a significant impact on human lives since white boxing an algorithm can build trust in ML methodology (e.g., Caruana & Nori, 2022).

5 | RESULTS AND DISCUSSION

5.1 | RQ1: Argumentation patterns uncovered by integrating unsupervised deep learning with human interpretation

To answer RQ1, we used HDBSCAN to extract 22 distinct clusters, an additional noise cluster, their 10 most representative words, and sample arguments (Table 1). By qualitatively evaluating the most representative words as well as sample responses, we derived descriptions of each topic. Figure 3 presents the two-dimensional embedding space of the extracted clusters. Clusters that are similar in content also appear close to each other in the two-dimensional embedding space: For example, arguments in clusters 2 and 3 apply acid–base theories on an electronic level, with minor differences in the interconnectedness of applied concepts. Arguments in clusters 4 and 5 correlate the thermodynamic state of the reagents to implicit structural properties, such as bond energies or resonance. In clusters 8, 9, 10, 11, and 12, arguments focus on the energetic aspects of a reaction in various degrees of elaborateness, in contrast to structural features addressed in many other clusters. Arguments in clusters 14, 15, and 16 apply acid–base theories on a rather phenomenological level, e.g., by citing pKa-values, or connect acid–base chemistry to thermodynamics or kinetics, e.g., by quantifying chemical equilibrium or estimating reaction rate. Arguments in clusters 17, 20, and 21 use steric hindrance in conjunction with hybridization or leaving group quality, while arguments in clusters 18 and 19 connect nucleophilicity and electrophilicity to basicity or electronic effects.

Clustering students' written arguments uncovered various argumentation patterns (Table 1) and revealed the degree of interconnectedness, elaborateness, and specificity of the applied concepts. The clusters comprised varying chemical topics and different levels of sophistication, indicating that the clustering approach captured the content *and* complexity of students' reasoning. Since students were explicitly prompted to include different chemical concepts in their argumentation, it is not surprising that these two aspects are also reflected in the clusters. However, clustering was exclusively data-driven, meaning that it groups data points based on the distance in the embedding space. While clustering helped us identify patterns across many responses, this approach did not provide a classification of the identified clusters influenced by theoretical frameworks, instructional goals, or learning objectives. Including research-informed classifications of students' argumentation grounded on subject matter expertise may, nonetheless, extend the validity of the analysis and ease the subsequent process of designing effective personalized instruction (Carlsen & Ralund, 2022; Kubsch et al., 2023; Nelson, 2020; Sherin, 2013). Hence, we combined the clustering approach with research-informed considerations to ensure that instructional decisions are data- *and* theory-driven.

TABLE 1 Overview of the extracted clusters.

ID	Cluster size	Description	Top words	Student example
−1	255	Redescribing the displayed reaction without referring to chemical concepts	Reaction, carbon, group, forms, S_N2, attacks, molecule, stabilizes, occurs, product	"Because the hydroxide ion can attack the carbon in an S_N2 fashion."
0	51	Estimating entropic changes based on the number of reactants and products	Entropy, reactant, product, favorable, increases, molecule, reaction, changes, unfavorable, forms	"This will increase the entropy of the system. There are more molecules in the products than in the reactants, which is a favorable transformation for entropy."
1	122	Estimating leaving group quality based on base strength, electronegativity, or polarizability	Leaves, good, group, poor, solvent, octet, electronegative, base, weak, stabilizes	"Chloride is a good leaving group. Chloride is polarizable and electronegative, making it well able to stabilize a negative charge effectively."
2	13	Invoking the Lewis acid–base theory	Acid, donates, base, proton, electron, acts, Lewis, accepts, carbon, pair	"The hydroxide anion will function as a Lewis base because it will donate an electron pair to the alpha carbon which forms the bond between the alpha carbon and hydroxide."
3	19	Applying electronic effects to infer acid strength	Carbon, electron, reacts, molecule, near, attacks, withdrawing, pair, forms, unstable	"The hydrogen atom, that is removed from the first molecule, is the most acidic one since the inductive effects from the oxygens in the molecule withdraw electron density, lowering its pKa."
4	15	Identifying resonance stabilization in carbanions	Carbanion, stabilizes, resonance, carbocation, results, product, forms, ketone, enolate, unstable	"The carbanion produced is stable because resonance forms are present that stabilize it."
5	21	Referring to bond enthalpies	Unfavorable, bond, wants, carbon, group, takes, molecule, satisfies, forms, stabilizes	"The reaction is energetically unfavorable. The breaking of the C-O bond is more expensive than the energy released from forming a C–N bond."
6	16	Identifying electrophiles based on electron density	Electrophile, density, carbon, electron, atom, withdrawing, attacks, takes, electronegative, makes	"Chlorine is much more electronegative than carbon; thus, it withdraws electron density and makes carbon an excellent electrophile."
7	12	Identifying acids and bases	Weak, acid, base, water, strong, proton, donates, molecule, atom, small	"The amine ion is a strong base that can deprotonate the ester. The amine ion acts as a base and the ester acts as an acid."

(Continues)

TABLE 1 (Continued)

ID	Cluster size	Description	Top words	Student example
8	78	Applying electronegativity to estimate the stability of the products	Electronegative, negative, charge, stabilizes, atom, carbon, better, effects, unstable, higher	"The negative charge is on a more electronegative atom. Oxygen is more electronegative than nitrogen since it is closer to completing its octet and so pulls electrons toward its nucleus. Products that have the negative charge on the more electronegative atom are more stable and so are more likely to be products."
9	48	Locating resonance stabilization	Resonance, stabilizes, bond, negative, charge, delocalized, electron, ring, pushes, pair	"The negative charge in the product is stabilized. There is resonance, so the electrons can form a double bond and then shift to the oxygen atom."
10	15	Focusing on charges as a surface-level indicator of stability	Negative, charge, positive, partial, makes, bond, reforms, prefers, carbon, likely	"The product is negatively charged. Molecules are most stable when limited charges are present."
11	98	Attributing energy levels to molecules	Bond, product, enthalpy, energy, stabilizes, reactant, favorable, reaction, forms, breaks	"The products are not very stable since they are not the lowest energy of this reaction."
12	15	Identifying resonance stabilization	Aldol, enolate, resonance, stabilizes, intermediate, energy, low, product, reactant, able	"The product participates in resonance, allowing it to be stabilized."
13	13	Weighing nucleophilic and basic properties by considering steric effects	Amine, bulky, alkoxy, acts, amide, nucleophile, approaches, easy, molecule, tertiary	"The amine is more likely to act as a base than as a nucleophile. The amine's bulkiness makes it difficult to approach the molecule, but it is much easier to remove a proton."
14	14	Citing pKa-values for identifying the strongest acid	pKa, low, favorable, higher, product, small, reversible, reaction, large, water	"The product has a higher pKa. With a higher pKa, the reaction is more likely to happen."
15	54	Describing proton transfer or comparing its reaction rate	Deprotonates, base, alcohol, strong, acid, proton, group, pKa, water, anion	"Acid–base reactions occur faster. Thus, this would not be a nucleophilic attack."

TABLE 1 (Continued)

ID	Cluster size	Description	Top words	Student example
16	22	Quantifying chemical equilibrium based on pKa-values	Ester, pKa, acid, base, amine, conjugate, low, deprotonates, weak, equilibrium	"Water, the conjugate acid of hydroxide, and alcohols have a very similar pKa, so there will be an equilibrium between the deprotonation of the alcohol by hydroxide and the protonation of the alkoxide by water."
17	37	Identifying leaving groups	Alkyl, primary, eliminates, group, substitution, reaction, better, favorable, S_N2, leaves	"Primary alkyl halide is more likely to undergo S_N2 since chloride is a good leaving group."
18	106	Identifying electrophiles and nucleophiles based on charges	Nucleophile, electrophile, attacks, electron, acts, good, lone, pair, charge, positive	"Hydroxide ion is a nucleophile. It bears a negative charge."
19	48	Identifying nucleophiles or weighing nucleophilic and basic properties	S_N2, nucleophile, reaction, E2, E1, primary, base, attacks, good, strong	"LDA is acting like a strong base, not like a nucleophile in an S_N2 reaction."
20	22	Connecting steric hindrance to hybridization	Steric, hindrance, S_N2, reaction, primary, interference, bulky, prevents, reaches, hybridization	"S_N2 reactions can occur. The carbon center attached to the chlorine is primary and sp3 hybridized, so there is no steric hindrance or orbital interference."
21	14	Connecting steric hindrance to the number of substituents	Hindrance, steric, substituent, bulky, attacks, effects, tertiary, likely, primary, spatial	"Chlorine is primary. There is very little steric hindrance."

5.2 | RQ2: Analysis of the *modes of reasoning* and *levels of granularity* applied in students' argumentation

As proposed by *computational grounded theory*, we looked for a research-informed framework to establish links between the identified clusters so that the ML output can be used to guide students' learning progression. Looking for a theoretical framework was led by two requirements: On the one hand, the theory-driven considerations needed to differentiate students' argumentation skills and their chemical concept knowledge in as much detail as possible. On the other hand, these considerations must allow for using ML in an instructional setting to accompany students' learning progression over time. A theoretical framework consisting of the *modes of reasoning* (Sevian & Talanquer, 2014) and the *levels of granularity* (Bodé et al., 2019; Darden, 2002; Deng et al., 2023; Deng & Flynn, 2021; Luisi, 2002; Machamer et al., 2000; Southard et al., 2016; Talanquer, 2018; van Mil et al., 2013) best reflected the cluster analysis findings.

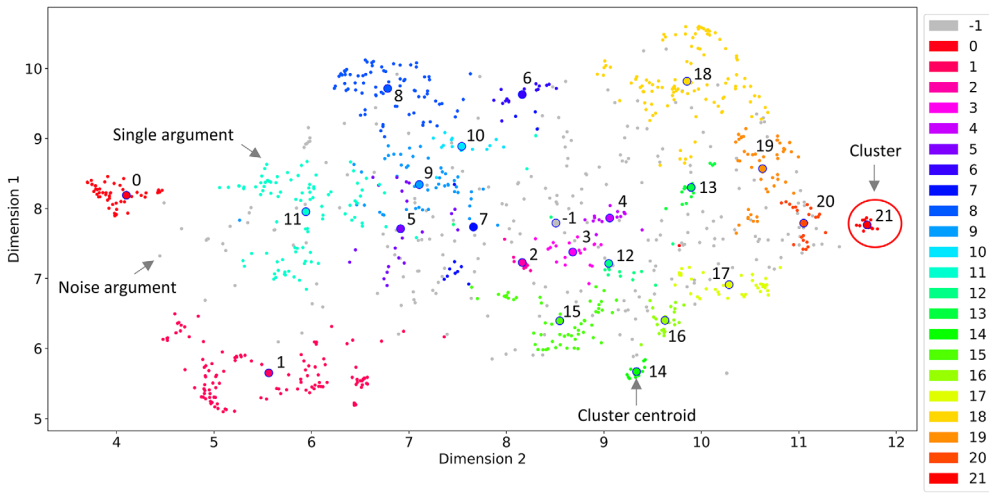


FIGURE 3 Two-dimensional embedding space of the extracted clusters. Each dot in the visualization corresponds to a single argument and different clusters are indicated by different colors. Noisy arguments, which are arguments that are too vague or include multiple topics, are represented by gray dots. Larger dots visualize cluster centroids. Dimensions 1 and 2 indicate some sort of principal components.

To evaluate the elaborateness and interconnectedness of applied concepts in students' scientific argumentation, Sevian and Talanquer (2014) proposed four *modes of reasoning*, namely, *descriptive*, *relational*, *linear causal*, and *multicomponent causal* (Table 2). These *modes of reasoning* discuss the complexity of novice or expert reasoning regarding their abilities to connect various concepts, provide sound justifications, and develop sophisticated explanations (Russ et al., 2008; Sevian & Talanquer, 2014). A major commitment underpinning these *modes of reasoning* is that students' understanding of science subjects cannot only be examined by assessing their content knowledge but also by evaluating how they integrate new information into their existing cognitive network (Sevian & Talanquer, 2014). So far, the *modes of reasoning* have been applied in various research contexts to analyze student reasoning in chemistry (e.g., Bodé et al., 2019; Carle et al., 2021; Deng et al., 2022, 2023; Deng & Flynn, 2021; Moon et al., 2019; Moreira et al., 2019; Sevian & Talanquer, 2014; Weinrich & Talanquer, 2016).

Arguments about scientific phenomena cannot only be generated in different *modes of reasoning* but also at different *levels of granularity* (Bodé et al., 2019; Deng et al., 2023; Deng & Flynn, 2021; Soo, 2019). The concept of *granularity* is characterized by the level of specificity to which an argument refers, assuming that different tasks require different grain sizes to explain the underlying processes. In general, the *level of granularity* varies depending on the context at hand but often ranges from global to submicroscopic perspectives (Darden, 2002). In our study, we adopt four *levels of granularity*, namely, *structural*, *energetic*, *phenomenological*, and *electronic* (Table 3), that were already applied in Deng and Flynn (2021) and Deng et al. (2023). These four levels are in alignment with Krist et al.'s (2019) *epistemic heuristics* which can be applied to characterize students' mechanistic reasoning across scientific content areas. The *epistemic heuristics* include identifying all factors of the phenomenon under investigation, considering the scalar level below, decomposing the factors at this lower scalar level, and linking those factors back to the target phenomenon. The target phenomenon of Krist et al. (2019) corresponds to *phenomenological* descriptions, which are characterized by argumentation about

TABLE 2 *Modes of reasoning* to characterize the elaborateness and interconnectedness of students' argumentation.

Mode of reasoning	Description	Student example arguing about reaction 1.3 (Figure 1)
Descriptive	Redescribing a phenomenon by recognizing explicit properties of single salient components without providing any connections or cause-and-effect relationships between these components	"This reaction is plausible. The alcohol acts as an acid and hydroxide ions act as a base." (Cluster 7)
Relational	Discussing the relationships between different components as well as their explicit and implicit properties in a correlative fashion or providing surface-level justifications of the established relations constrained by a reduction of variables	"This reaction is plausible because the pKa value of water is like the pKa value of an alcohol. So, the acidic alcohol is easily deprotonated by the basic hydroxide ions to form alkoxide." (Cluster 14)
Linear causal	Identifying cause-and-effect interactions between explicit and implicit properties of most components or explaining mechanisms based on the static interplay of different concepts	"This reaction is plausible. Alkoxide and hydroxyl exist in equilibrium. Hydroxide ions act as a base and deprotonate the acidic alcohol. The two involved OH groups have very similar pKa values and are, thus, expected to interchange protons easily." (Cluster 16)
Multi-component causal	Perceiving phenomena as the dynamic interplay of several interrelated concepts, explaining systematically how different properties affect the reactivity of the components, and weighing how different concepts relate to each other	*Not observed in the data*

Note: Descriptions of the *modes of reasoning* are adapted from Sevian and Talanquer (2014), student examples are chosen from the data analyzed in this study. Cluster IDs can be found in Table 1.

explicit or measurable properties of the given structures. The lower scalar level arises, in turn, from the interplay of the *energetic*, *structural*, and *electronic* properties of the involved components.

Based on these theoretical considerations, we qualitatively analyzed students' arguments again. A *mode of reasoning* and *level of granularity* was assigned to every argument, which is in the following illustrated by the student examples listed in Table 2. These examples address the concepts of acidity and basicity while arguing about reaction 1.3 (Figure 1).

In the *descriptive* example, the student claimed that the reaction is plausible; however, it is only described that the alcohol molecule can act as an acid and the hydroxide ion as a base, lacking any cause-and-effect relationship. In the *relational* argument, the student compared the pKa values of water and alcohol molecules, which are essential chemical properties, to support the claim. Highlighting the similarity in pKa values demonstrates a stronger understanding of the acid–base concept, resulting in an increased elaborateness of the argument. In the *linear causal* example, the student explains, based on pKa values, that alkoxide and hydroxyl exist in equilibrium. This detailed explanation of why in a chemical equilibrium a fast exchange of

TABLE 3 *Levels of granularity* proposed by Deng and Flynn (2021) to classify the grain size of students' argumentation.

Level of granularity	Description
Structural	Focusing on chemical compounds as well as their properties, which are derived from structural features such as the connectivity of atoms, atom size, the presence of charges, and salient centers
Energetic	Focusing on thermodynamic and kinetic changes during chemical reactions, which are whether a reaction is reactant- or product-favored, whether any entropic effects impact the reaction process, and at which rate a reaction occurs
Phenomenological	Reasoning about measurable properties of components such as pKa values, the direction of chemical equilibrium, or reaction rate without including electron distribution
Electronic	Reasoning about the distribution and movement of electrons in chemical compounds by including electronic effects like resonance, hyperconjugation, or inductive effects to explain how electron density impacts the properties of a molecule

protons can occur strengthens the causal link in the argument and provides a comprehensive understanding of the factors contributing to the plausibility of acid–base reactions. The *modes of reasoning* helped us classify the degree of causality in each argument.

Despite different levels of causality, the three arguments all address structural properties related to acid–base reaction 1.3 (Figure 1) by relying on the identification of acidic and basic functional groups and the influence of their pKa values on the reaction outcome. In contrast, other students argued about acid–base reaction 1.3 (Figure 1) from a kinetic point of view by simply claiming that acid–base reactions occur faster than any other reaction type. Such an argumentation does not include any structural properties; as a result, we distinguished, inspired by the *levels of granularity*, between a *structural* and *energetic* argumentation.

In addition, none of the three examples above included electronic properties such as electron density or orbital interactions, which is why we classified these examples as *phenomenological*. Nonetheless, other students integrated electronic properties in their argumentation about acid–base reaction 1.3 (Figure 1). For example, one student argued that hydroxide ions are strong bases due to their small atom size and low polarizability, leading to a localization of the negative charge. This argument highlights electronic properties related to the concept of basicity, showing the necessity to differentiate between *phenomenological* and *electronic* arguments as proposed in the *levels of granularity*.

With this qualitative analysis in mind, we established a two-dimensional holistic scoring rubric by adapting the *modes of reasoning* and *levels of granularity* (Figure 4). This scoring rubric is nominally scaled. The number of clusters detected with HDBSCAN guided the process of rubric design as this number served as a reference for estimating how many categories could be clearly distinguished in the scoring rubric. For the *levels of granularity*, we found by interpreting the detected clusters that classifying every argument into two of the four proposed levels described the responses in our sample best. Accordingly, every argument has one of four codes *structural-phenomenological*, *structural-electronic*, *energetic-phenomenological*, or *energetic-electronic*. Additionally, the *structural level of granularity* was divided into three sublevels because HDBSCAN detected topics with a *structural* focus for three different concepts.

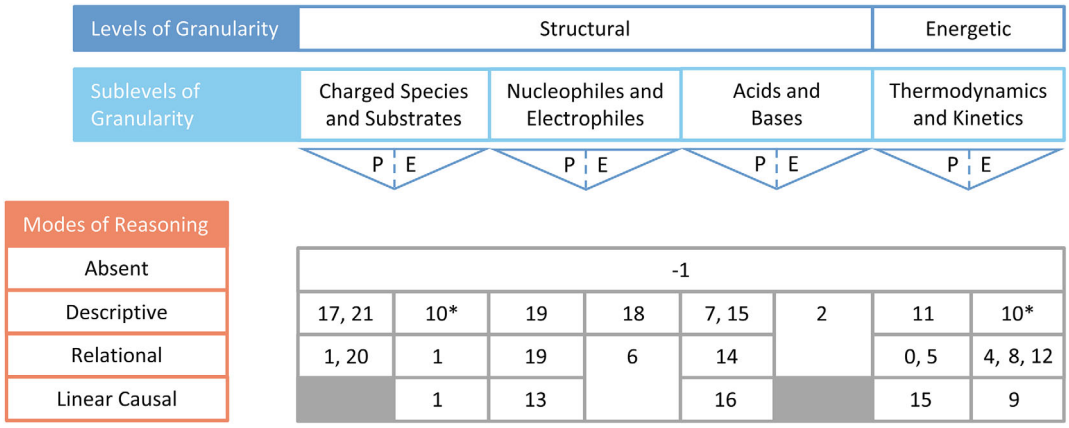


FIGURE 4 Classification of the clusters identified through HDBSCAN according to the two-dimensional scoring rubric. For instance, cluster −1 mostly contains *absent* responses. Numbers indicate the cluster ID listed in Table 1. P stands for *phenomenological*, E for *electronic*. The asterisk shows that clusters belong together.

Here, we introduced the *structural sublevels* of *charged species and substrates*, *nucleophiles and electrophiles*, and *acids and bases*. This distinction helped us consider the chemical entities students included in their argumentation. To clarify what we mean by *energetic*, we renamed this level to *thermodynamics and kinetics*. For the *modes of reasoning*, we introduced the mode *absent* for off-task responses. Since instances of *multicomponent causal reasoning* were not prompted, we did not include this *mode* in our coding scheme.

After including research-informed considerations, we examined if the theory- and data-driven analysis of students' argumentation matched by assigning every cluster to the *mode of reasoning* and *level of granularity* that was most frequently applied across all student arguments in that cluster (Figure 4). In doing so, we found that the clusters partially represented the theoretical framework—and vice versa. Consequently, we iteratively modified the classification of HDBSCAN by dividing some large clusters into further subclusters and aggregating clusters that contained similar ideas (Table 4). For example, cluster 19 could be divided into two further subclusters since this cluster incorporated two different *modes of reasoning* with enough sample responses to precisely differentiate between these two *modes*. By contrast, clusters 17 and 21 were aggregated because both referred to a *descriptive*, *phenomenological* argumentation about *charged species and substrates*. Aligning the results of the data-driven cluster analysis with theoretical considerations contributed to establishing well-distinguishable coding categories (cf., Carlsen & Ralund, 2022). However, because of ambiguous wording in students' argumentation as well as the small number of arguments in clusters 2, 6, and 10, it was not useful for subsequent supervised automation to further divide these clusters as conceptualized in the theoretical framework (Figure 4). Besides, *structural-phenomenological* reasoning with a focus on *charged species and substrates* and a *linear causal mode of reasoning* was not present in our dataset at all, which is why we excluded this category from our analysis, leading to a total of 21 categories (Figure 4).

HDBSCAN facilitated rubric design as it detected patterns across many arguments, which enhanced the validity of our scoring rubric as it is grounded in data. Furthermore, HDBSCAN identified distinct topics students reasoned about. In particular, this approach assisted us in differentiating between well-interpretable and vague categories. Integrating research-informed

TABLE 4 Final modifications of the clusters identified through HDBSCAN guided by the *modes of reasoning* and *levels of granularity*.

Modification	Justification
Dividing a cluster into subclusters	
Dividing cluster 1 into three subclusters	Responses in cluster 1 explain why a leaving group is good or bad by referring to base strength, electronegativity, or polarizability. The high number of responses indicates the heterogeneity and coarse level of data differentiation. The cluster was divided into three subclusters which attribute leaving group quality to (i) low base strength, (ii) high electronegativity, and (iii) high polarizability to better differentiate the data.
Dividing cluster 15 into two subclusters	Responses in cluster 15 justify the occurrence of proton transfer by identifying acids and bases or consider that acid–base reactions occur faster than substitution reactions. Despite the similar wording, these two ideas are different in quality when judging the plausibility of chemical reactions. Hence, the cluster was divided into two subclusters to distinguish between describing proton transfer and considering kinetics.
Dividing cluster 19 into two subclusters	Responses in cluster 19 identify nucleophiles and electrophiles or weigh nucleophilic properties against basic properties. Since combining nucleophilic and basic properties is considered a higher *mode of reasoning*, these arguments were separated from rather *descriptive* ones by dividing this cluster into two subclusters.
Aggregating various clusters	
Aggregating clusters 0 and 5	Responses in cluster 0 reason about entropic changes by referring to the number of reactants and products, while responses in cluster 5 reason about enthalpic changes by referring to bond energies. Although these clusters are distinct and easily interpretable, they address, from a technical point of view, a system's thermodynamic changes. Due to the technical comparability, the clusters were unified.
Aggregating clusters 4, 8, and 12	Responses in cluster 4 identify resonance stabilization in carbanions, while responses in cluster 12 identify resonance stabilization in general, which is why they were aggregated. Similarly, responses in cluster 8 apply electronegativity, which is another electronic property, to estimate the energetic level of molecules. So, these clusters address the thermodynamic stability of molecules on an electronic level, leading to their combination.
Aggregating clusters 17 and 21	Clusters 17 and 21 are *phenomenological* and *descriptive* in nature. Responses in both clusters only recognize and describe salient properties of the displayed molecules, which are properties that are explicitly shown by Lewis structures. Consequently, these clusters were aggregated.

considerations finally showed where clusters needed to be further differentiated or, in contrast, be aggregated (cf., Carlsen & Ralund, 2022). Taken together, refining the detected clusters influenced by subject matter expertise uncovered new argumentation patterns and laid the foundation for the subsequent process of pattern confirmation.

TABLE 5 Metrics to measure score agreements.

Accuracy	Cohen's κ	Weighted F_1 score	Macro F_1 score
Percentage of agreement between two raters	Agreement between two raters beyond chance agreement	Agreement metrics for imbalanced datasets accounting for the number and type of prediction errors as the harmonic mean of precision and recall	
		Weighted mean of all per-class F_1 scores	Arithmetic, i.e., unweighted mean of all per-class F_1 scores

Based on the new scoring rubric, a student research assistant and the first author scored the entire dataset holistically by assigning one of 20 mutually exclusive scores to answer RQ2. Here, we needed to exclude cluster 3 from analysis (Figure 4) since it contained only a few arguments, which could also be classified into one of the other clusters. Across all rubric categories, we achieved *almost perfect* inter-rater reliability (cf., Landis & Koch, 1977) as shown by an accuracy of 84%, Cohen's κ of 0.83, a weighted F_1 score of 0.85, and a macro F_1 score of 0.78 (for a definition of these metrics, see Table 5). After that, codes were discussed until full agreement was reached. The distribution of the arguments across the 20 categories is shown in Table 6. Along with other researchers, we found that students at the same educational level applied different *modes of reasoning* to solve the same exercise (Bodé et al., 2019; Deng & Flynn, 2021; Moreira et al., 2019; Weinrich & Talanquer, 2016). The students in this study constructed mostly *descriptive* or *relational* arguments, which is also a common observation in other studies (cf., Carle et al., 2021; Deng et al., 2022; Deng & Flynn, 2021). Moreover, when arguing about Williamson ether synthesis, students applied predominantly a *phenomenological level of granularity*.

5.3 | RQ3: Level of accuracy of the deep learning model

After human scoring, we checked if a deep neural network can accurately classify students' arguments according to the holistic, 20-category rubric. For this reason, we compared the accuracy of different large language models on the validation set. *BERT base uncased*, *RoBERTa base, RoBERTa large*, and *SciBERT scivocab uncased* reached a maximum macro F_1 score of 0.88 across all hyperparameter configurations, while *BERT large uncased* performed best with a maximum macro F_1 score of 0.89. *BERT large uncased* achieved the highest macro F_1 score with 38 epochs, a batch size of 8, and a learning rate of 1e-5. We then applied this best hyperparameter configuration to measure the accuracy of the model on the unseen test set.

Across all 20 categories, our multi-class model achieved *almost perfect* machine-human score agreement as indicated by an accuracy of 87%, Cohen's κ of 0.86, a weighted F_1 score of 0.86, and a macro F_1 score of 0.80. Machine-human score agreements were slightly higher than human-human score agreements. To be noted, we used the initial human scores, which may contain errors as a rater may have lacked a complete understanding of the rubric categories, to calculate human-human score agreements while training the model on consensus scores. As presented in Table 7, the macro F_1 scores per category ranged from 0.50 to 1.00, showing that the applied deep learning model identified some categories more accurately than others. Predictably, the algorithm struggled to detect categories with fewer sample arguments because ML is based on an inductive learning procedure that is dependent on diverse datasets.

TABLE 6 Number of arguments per rubric category and task.

	Structural						Energetic	
	Charged species and substrates		Nucleophiles and electrophiles		Acids and bases		Thermodynamics and kinetics	
	P	E	P	E	P	E	P	E
Intramolecular Williamson ether synthesis								
Absent	6							
Descriptive	78	24*	27	13	63	35	37	24*
Relational	23	20	48	36	34		40	19
Linear causal	0	21	7		11	0	14	7
Claisen condensation								
Absent	17							
Descriptive	49	46*	20	18	27	35	10	46*
Relational	15	12	16	38	24		42	104
Linear causal	0	15	15		9	0	8	25

Note: The asterisk shows that clusters belong together.
Abbreviations: E, *electronic*; P, *phenomenological*.

TABLE 7 F_1 scores listed per rubric category.

	Structural						Energetic	
	Charged species and substrates		Nucleophiles and electrophiles		Acids and bases		Thermodynamics and kinetics	
	P	E	P	E	P	E	P	E
Absent	0.57							
Descriptive	0.89	0.92*	0.96	0.86	0.91	0.88	0.56	0.92*
Relational	0.60	0.57	0.81	0.95	1.00		0.94	0.98
Linear causal	n.e.	0.67	0.57		1.00	n.e.	0.50	0.82

Note: The asterisk shows that clusters belong together.
Abbreviations: E, *electronic*; P, *phenomenological*; n.e., non-existent.

To subsequently perform a baseline comparison, we also tried to automate the holistic scoring with a support vector machine. A support vector machine is a more traditional, so-called shallow supervised ML algorithm that seeks to find the best possible boundary between different classes in a dataset. We chose a support vector machine since it performed well in a classification task on students' use of the Lewis acid–base theory across different assessment items (Yik et al., 2021). However, by adopting the *modes of reasoning* and *levels of granularity*, the support vector machine achieved only *fair* machine-human score agreement with an accuracy of 31%, Cohen's κ of 0.25, a weighted F_1 score of 0.25, and a macro F_1 score of 0.21. This demonstrates that applying a deep neural network was necessary to solve the holistic scoring problem, which is in line with other findings demonstrating that deep neural networks perform better than more traditional ML models in educational contexts

(e.g., Gombert et al., 2022; Wulff, Mientus, et al., 2022). However, the accuracy of ML methods strongly depends on the complexity of the task at hand. For instance, Küchemann et al. (2020) found that more traditional ML models outperformed deep neural networks for simpler scoring tasks.

In sum, we built a deep learning model with *almost perfect* accuracy capable of automatically evaluating the nature of students' arguments across different chemical reactions and concepts. This model can be used in further formative assessments to analyze students' argumentation irrespective of the displayed chemical reaction.

5.4 | RQ4: Consensus between human and deep learning scoring guidelines

To answer RQ4, we validated the accuracy of our deep learning architecture in three different ways. First, we used *ChatGPT*, a large language model that can analyze and produce human-like language, to generate 20 fictitious student arguments (cf., Dai et al., 2023), one per category, based on the cluster descriptions (Table 1). We then tested whether our deep learning architecture can classify these *ChatGPT*-generated arguments into the correct category, which helped us estimate if the 20 categories are distinct from each other and if the cluster descriptions are valid. Our deep learning architecture classified 17 fictitious student arguments into the correct category. This implies that the cluster descriptions are comprehensible, distinguishable, and valid for generating additional fictitious arguments. In addition, this approach provides further support for the validity of the clustering method used to define the categories.

For black box analysis, we selected a *phenomenological*, *descriptive* argument from each of the four rubric sublevels *charged species and substrates*, *nucleophiles and electrophiles*, *acids and bases*, and *thermodynamics and kinetics*. Afterward, we added word groups listed in our scoring guidelines to these *phenomenological*, *descriptive* arguments so that they reached the next category of the scoring rubric in the respective sublevel. For example, we added the phrase "as they are a weak base" to a *phenomenological*, *descriptive* argument in the sublevel *charged species and substrates* and checked whether this additional phrase was sufficient for the algorithm to classify the argument as *phenomenological*, *relational*. With this approach, we simulated if students' learning progression can be accompanied by the deep learning architecture over time. The algorithm correctly classified 18 out of 20 arguments, which highlights the potential of the algorithm in monitoring students' learning progression even over a longer period.

Furthermore, the model explainer SHAP, a technique for white boxing an algorithm (Lundberg & Lee, 2017), was used to calculate the importance scores of individual words for each rubric category, that is, how individual words impact algorithmic output. In doing so, we formed the arithmetic mean of the importance scores across all arguments. For clarity, we present only the output for four exemplary rubric categories herein (Figure 5). Positive scores have a positive impact on the model output in the respective category and increase, thus, the probability of a positive classification, while negative scores have a negative impact on the model output in the respective category and decrease, thus, the probability of a positive classification. In general, our algorithm utilized similar keywords than human raters in categories with higher F_1 scores (Figure 5a, b). Words like *resonance*, *electron*, *octet*, and *dipole* relate, for instance, the thermodynamic stability of the reagents to electron distribution, which is why these words correspond to the human scoring guidelines in the category *thermodynamics and kinetics, electronic*, *relational* (Figure 5a). Furthermore, the words *electron*, *electronegative*, *inductive*,

FIGURE 5 Importance scores of the 10 words that impact algorithmic output most significantly for four sample categories.

resonance, and *withdrawing* can be used to relate nucleophilic and electrophilic properties of reactive centers to their electron density (Figure 5b). In one category, we identified that the algorithmic decision relied almost exclusively on a single word, which is a critical model shortcut (Figure 5c). Although the word *equilibrium* is key to argumentation about *acids and bases*, the algorithm neglected other words like *pKa*, *acidic*, or *basic* in this category, which also mirrors human shortcuts in the scoring procedure. In the category *charged species and substrate*, *electronic*, *phenomenological*, a category with a lower F_1 score, the algorithm relied partly on expectable words like *polarizable*, *leaving*, and *electron*, but the impact of the other words on the model output was not explainable based on human scoring guidelines (Figure 5d).

Collectively, the implemented validation techniques demonstrated that the deep learning model adopts human scoring guidelines to a great extent. In particular, the deep learning architecture achieved an accuracy of 85% when using *ChatGPT* for external validation and an accuracy of 90% when conducting black box analysis. These metrics are in alignment with the overall model accuracy of 87%. Unsurprisingly, arguments from categories with lower F_1 scores (Table 7) were more difficult to classify. Furthermore, white box analysis determined the most significant words associated with positive or negative classifications in each category, which helped estimate if the words consulted by the deep learning model correspond to human scoring guidelines. Expectedly, the words used to classify arguments in categories with higher F_1 scores were comparable to keywords human raters looked for, while discrepancies arose in categories with lower F_1 scores. Overall, these validation methods provided additional performance metrics, helped us gain a better understanding of the model generalizability, and may increase confidence in decisions derived from the algorithm reported herein.

6 | LIMITATIONS

From a technological point of view, large language models such as BERT are trained on uncleaned language data, among others, from the Internet or Wikipedia. For this reason, large language models might acquire stereotyped biases (e.g., Caliskan et al., 2017), which can lead to the perpetuation of discrimination against racially marginalized groups in science assessment (Cheuk, 2021; Cheuk et al., 2019), such as second language learners (Deng et al., 2022; Deng & Flynn, 2023). Reducing bias can include a more controlled training process or applying subject-sensitive, situation-specific standards for examining bias along with creating countermeasures against discrimination (Grimm, Steegh, Çolakoğlu, et al., 2023; Grimm, Steegh, Kubsch, & Neumann, 2023).

Moreover, we did not investigate the difficulty of detecting the clusters without deep learning. We only showed via post-hoc explanations that the features consulted by the deep learning model correspond to human scoring guidelines. Specifically, the deep learning model relied frequently on specific words in responses (Figure 5, Wulff, Mientus, et al., 2022) or similarities between sentences as encoded in the embedding space (Figure 3, Wulff, Mientus, et al., 2022). However, we did not explore a priori if the clusters were easy to detect by humans. The high level of human-human inter-rater reliability indicates, nonetheless, that human analysts can reliably adopt the developed scoring rubric.

Besides, students' written arguments were assigned to one of 20 categories depending on the applied *mode of reasoning* and *level of granularity*. However, automating the analysis of some categories was not possible at all due to the lack of sample arguments. For example, we aggregated the categories of *relational- and linear causal-electronic reasoning* in the sublevel *nucleophiles and electrophiles* or the categories of *descriptive- and relational-electronic reasoning* in the sublevel *acids and bases*. In the future, further student-written arguments need to be collected to expand the scoring rubric to precisely differentiate between all conceivable categories. Here, it may be particularly helpful to gather data at different institutions, from courses with varying instructors and curricula, and from students with diverse demographic backgrounds. Future research can, therefore, increase the validity and reliability of the deep learning architecture with a larger training set.

A limitation related to the absence of arguments in some rubric categories is a low sample heterogeneity in our study since our model is trained on a sample that belongs to a rather

homogeneous demographic group. In other words, the data used to train our deep learning model originated from a single set of students enrolled at the same institution. Hence, we could not determine to what extent the model is generalizable to other institutions or cohorts. Nevertheless, prior research documented that pre-trained large language models applied for analyzing students' reasoning in science subjects could be generalized across contexts, such as institutions and disciplines (Wulff et al., 2023). In detail, a fine-tuned BERT model developed for classifying pre-service physics teachers' written reflections could be reliably utilized across universities as well as for analyzing written reflections of non-physics students (Wulff et al., 2023). This suggests that BERT models as used in this study show strong generalizability across various contexts, enabling researchers to share their models and adapt these models to specific student populations. In the future, we aim to further improve model generalizability by fine-tuning the model to new contexts. Based on more heterogeneous data, the potential bias of our deep learning architecture could also be reduced (cf., Noyes et al., 2020), which may build more confidence in the results produced by our model.

7 | CONCLUSIONS AND IMPLICATIONS

We employed ML at an early stage in the assessment development process to create and refine a holistic scoring rubric for evaluating students' argument complexity when judging the plausibility of competing chemical reactions. Our approach combined data-driven deep learning methods with theory-driven human interpretation, which aligns with emerging recommendations for using ML in conjunction with research-informed theories to get a more comprehensive understanding of complex phenomena (Carlsen & Ralund, 2022; Kubsch et al., 2023; Nelson, 2020; Rosenberg and Krist, 2021; Tschisgale et al., 2023).

In contrast to other projects analyzing students' scientific argumentation with ML, we applied a clustering approach as the first step in our analysis to extend the level of detail of our classification. Here, we used HDBSCAN in an exploratory way to detect common patterns in students' arguments, which gave us an idea of the categories suitable for subsequent supervised automation. Cluster analysis supported us in formulating evidence statements that specify the performance accepted as evidence, which is an integral part of rubric development as postulated by the evidence-centered design approach to educational assessment (Kubsch et al., 2022; Mislevy, 2016; Mislevy et al., 2003; Pellegrino et al., 2016). Although ML is increasingly prevalent in educational assessment, such an unsupervised approach is not frequently applied, yet (Zhai, Yin, et al., 2020).

Cluster analysis informed the process of rubric development, but we did not adopt the findings of the cluster analysis without theory-driven reflection. So, cluster analysis uncovered to some extent how students judged the plausibility of competing chemical reactions but only the theoretical framework enabled the systematic classification of students' argumentation. Combining a data- and theory-driven approach ensured that the rubric categories did not only represent statistical values of similarity evident in the data but also pre-existing, empirical theories and human expertise. A similar methodological approach was used by Anderson et al. (2020), Prevost et al. (2012), Sherin (2013), Wulff, Buschhüter, et al. (2022), Wulff et al. (2023), and Zehner et al. (2016) to justify the validity of their qualitative coding, but not with a focus on rubric development. By contrast, Jescovitch, Doherty, et al. (2019), Jescovitch, Scott, et al. (2019), Jescovitch et al. (2021), Kaldaras et al. (2022), and Mao et al. (2018) focused on rubric development for ML-based scoring, but without the use of unsupervised techniques. In this

article, instead, we provide just like Haudek et al. (2015) and Rosenberg and Krist (2021) an example of how to incorporate unsupervised ML techniques into human qualitative interpretation for rubric development.

After the application of unsupervised techniques, we used large language models, which made significant advancements in addressing a broad spectrum of NLP issues (cf., Koroteev, 2021), as most notably experienced in conversational chatbots like *ChatGPT*. In conjunction with a deep neural network, we could automate the evaluation of students' arguments according to a 20-category holistic scoring rubric. As far as we know, the possibility of applying deep learning to evaluate argumentative patterns in students' responses according to a complex multi-class rubric was not reported before. Other projects in chemistry education research chose either an analytic scoring approach with a single or multiple binary classifiers (e.g., Dood et al., 2018; Watts et al., 2023b; Winograd, Dood, Moon, et al., 2021; Yik et al., 2021) or applied a holistic rubric with three to five levels (e.g., Donnelly et al., 2015; Dood et al., 2020a; Haudek et al., 2009, 2012, 2019; Liu et al., 2014; Maestrales et al., 2021; Noyes et al., 2020; Tansomboon et al., 2017; Vitale et al., 2016). With *computational grounded theory* as a methodological framework, we demonstrated that deep learning techniques can be used to evaluate students' arguments according to an abstract, holistic, multi-class rubric with *almost perfect* machine-human score agreement. Therefore, we assume that by strategically integrating theory-driven considerations into data-driven classifications it is possible to develop a scoring rubric that can be more easily incorporated into supervised analysis.

The diagnosis of students' argumentation skills can, in the future, be used to design instructional settings that realize aspects of adaptive learning (cf., Plass & Pawar, 2020). Here, our analysis enables providing personalized feedback and tailored exercises accounting for the *modes of reasoning* and *levels of granularity* applied in students' argumentation. This ML-assisted approach to adaptive learning seems promising as it increased students' level of explanation sophistication in another chemistry context (Dood et al., 2020a, b).

Along with realizing aspects of adaptive learning, the established deep learning architecture can be used to monitor students' learning progression over time (cf., Swiecki et al., 2022). Specifically, tracking students over a more extended period may help evaluate how argumentation skills change, how students move from one argumentation cluster to another, how and when clusters emerge, and what competencies need to be trained so that students reason in high-level clusters (Zhai et al., 2023). Moreover, this longitudinal data collection may enable measuring the effects of adaptive learning on students' argumentation skills and help determine optimal conditions for adaptive learning including how to fade support. Ultimately, this could allow for new approaches to Just-in-Time Teaching (Novak et al., 1999) realized with adaptively faded scaffolds (McNeill et al., 2006; Noroozi et al., 2018).

Furthermore, students responded in our instructional setting to the same prompt for each of the proposed reactions, that is, they repeatedly judged whether the displayed molecule is a plausible product. In the future, the prompt could be varied to generalize our deep learning model not only across different chemical reactions but also across different prompt types. Here, it may be possible to instantly assess which prompt type is most appropriate for a student and to adapt the prompt afterward. This may allow for providing automatically adapted prompts with varying degrees of scaffolding and guidance.

ACKNOWLEDGMENTS

This publication is part of the first author's doctoral thesis (Dr. rer. nat.) at the Faculty of Biology and Chemistry, Justus-Liebig-University Giessen, Germany. We thank Leonie Lieber, Ira

Caspari-Gnann, and Krenare Ibraj for their groundbreaking research on students' argumentation in chemistry as well as for providing access to their data. We thank Felix Blödtner for scoring students' arguments and all members of the Graulich group for fruitful discussions. Finally, Paul P. Martin thanks Brandon Yik, Amber Dood, and Field Watts for sharing their innovative thoughts on the application of ML in science education. There are no conflicts of interest to declare. Open Access funding enabled and organized by Projekt DEAL.

ORCID

Paul P. Martin https://orcid.org/0000-0001-8648-4250
David Kranz https://orcid.org/0000-0002-2054-6882
Peter Wulff https://orcid.org/0000-0002-5471-7977
Nicole Graulich https://orcid.org/0000-0002-0444-8609

REFERENCES

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

Angelov, D. (2020). Top2Vec: Distributed representations of topics. *arXiv Preprint*, arXiv:2008.09470. https://doi.org/10.48550/arXiv.2008.09470

Bail, C. A. (2014). The cultural environment: Measuring culture with big data. *Theory and Society*, *43*(3/4), 465–482.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv Preprint*, arXiv:1903.10676. https://doi.org/10.48550/arXiv.1903.10676

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26–55.

Biernacki, R. (2012). *Reinventing evidence in social inquiry: Decoding facts and variables*. Palgrave Macmillan.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bodé, N. E., Deng, J. M., & Flynn, A. B. (2019). Getting past the rules and to the WHY: Causal mechanistic arguments when judging the plausibility of organic reaction mechanisms. *Journal of Chemical Education*, *96*(6), 1068–1082.

Brunton, S. L., & Kutz, J. N. (2019). *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in knowledge discovery and data mining: PAKDD 2013* (pp. 160–172). Springer.

Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, *10*(1), 1–51.

Carle, M. S., el Issa, R. J., Pilote, N., & Flynn, A. B. (2021). Ten essential delocalization learning outcomes: How well are they achieved? *ChemRxiv Preprint*, 1–28. https://doi.org/10.26434/chemrxiv.13322771.v1

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

Caruana, R., & Nori, H. (2022). Why data scientists prefer glassbox machine learning: Algorithms, differential privacy, editing and bias mitigation, In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 4776-4777). Association for Computing Machinery.

Caspari, I., & Graulich, N. (2019). Scaffolding the structure of organic chemistry students' multivariate comparative mechanistic reasoning. *International Journal of Physics and Chemistry Education*, *11*(2), 31–43.

Cetin, P. S. (2014). Explicit argumentation instruction to facilitate conceptual understanding and argumentation skills. *Research in Science & Technological Education*, *32*(1), 1–20.

Charmaz, K. (2014). *Constructing grounded theory*. Sage Publications.

Cheuk, T. (2021). Can AI be racist? Color-evasiveness in the application of machine learning to science assessments. *Science Education*, *105*(5), 825–836.

Cheuk, T., Osborne, J., Cunningham, K. R., Haudek, K., Santiago, M. M., Urban-Lurain, M., Merrill, J., Wilson, C. D., Stuhlsatz, M. A. M., Donovan, B., Bracey, Z. B., & Gardner, A. (2019). Towards an equitable design framework of developing argumentation in science items and rubrics for machine learning. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Baltimore, MD.

Chin, C., & Brown, D. E. (2000). Learning in science: A comparison of deep and surface approaches. *Journal of Research in Science Teaching*, *37*(2), 109–138.

Chin, C., & Osborne, J. (2010). Supporting argumentation through students' questions: Case studies in science classrooms. *Journal of the Learning Sciences*, *19*(2), 230–284.

Chinn, C., & Malhotra, B. (2002). Children's responses to anomalous scientific data: How is conceptual change impeded? *Journal of Educational Psychology*, *94*(2), 327–343.

Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., Liu, N., Li, S., Zhu, D., Cai, H., Sun, L., Li, Q., Shen, D., Liu, T., & Li, X. (2023). AugGPT: Leveraging ChatGPT for text data augmentation. *arXiv Preprint*, arXiv:2302.13007. https://doi.org/10.48550/arXiv.2302.13007

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

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & de Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, *162*(104094), 1–43.

Deng, J. M., Carle, M. S., & Flynn, A. B. (2023). Students' reasoning in chemistry arguments and designing resources using constructive alignment. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (pp. 74–89). The Royal Society of Chemistry.

Deng, J. M., & Flynn, A. B. (2021). Reasoning, granularity, and comparisons in students' arguments on two organic chemistry items. *Chemistry Education Research and Practice*, *22*(3), 749–771.

Deng, J. M., & Flynn, A. B. (2023). "I am working 24/7, but I can't translate that to you": The barriers, strategies, and needed supports reported by chemistry trainees from english-as-an-additional language backgrounds. *Journal of Chemical Education*, *100*(4), 1523–1536.

Deng, J. M., Rahmani, M., & Flynn, A. B. (2022). The role of language in students' justifications of chemical phenomena. *International Journal of Science Education*, *44*(13), 2131–2151.

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

Donnelly, D. F., Vitale, J. M., & Linn, M. C. (2015). Automated guidance for thermodynamics essays: Critiquing versus revisiting. *Journal of Science Education and Technology*, *24*(6), 861–874.

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

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

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

Dood, A. J., Winograd, B. A., Finkenstaedt-Quinn, S. A., Gere, A. R., & Shultz, G. V. (2022). PeerBERT: Automated characterization of peer review comments across courses. In *LAK22: 12th International learning analytics and knowledge conference* (pp. 492–499). Association for Computing Machinery.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, *84*(3), 287–312.

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education*, *38*(1), 39–72.

Erduran, S. (2019). Argumentation in chemistry education: An overview. In S. Erduran (Ed.), *Argumentation in chemistry education: Research, policy and practice* (pp. 1–10). The Royal Society of Chemistry.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education*, *88*(6), 915–933.

Faize, F. A., Husain, W., & Nisar, F. (2018). A critical review of scientific argumentation in science education. *EURASIA Journal of Mathematics, Science and Technology Education*, *14*(1), 475–483.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673.

Gerard, L. F., Matuk, C., McElhaney, K., & Linn, M. C. (2015). Automated, adaptive guidance for K-12 education. *Educational Research Review*, *15*, 41–58.

Glaser, B., & Strauss, A. (1999). *Discovery of grounded theory: Strategies for qualitative research* (Vol. 1). Routledge.

Goldberg, Y. (2017). *Neural network methods for natural language processing*. Morgan & Claypool Publishers.

Gombert, S., di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., Grimm, A., Bohm, I., Neumann, K., & Drachsler, H. (2022). Coding energy knowledge in constructed responses with explainable NLP models. *Journal of Computer Assisted Learning*, *39*(3), 767–786.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Graulich, N., & Caspari, I. (2020). Designing a scaffold for mechanistic reasoning in organic chemistry. *Chemistry Teacher International*, *3*(1), 19–30.

Grimm, A., Steegh, A., Çolakoğlu, J., Kubsch, M., & Neumann, K. (2023a). Positioning responsible learning analytics in the context of STEM identities of under-served students. *Frontiers in Education*, *7*(1082748), 1–12.

Grimm, A., Steegh, A., Kubsch, M., & Neumann, K. (2023b). Learning analytics in physics education: Equity-focused decision-making lacks guidance! *Journal of Learning Analytics*, *10*(1), 71–84.

Grootendorst, M. (2020). Topic Modeling with BERT: Leveraging BERT and TF-IDF to create easily interpretable topics. https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6

Harris, C., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice*, *38*(2), 53–67.

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

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

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J. E., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. *CBE: Life Sciences Education*, *11*(3), 283–293.

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

Haudek, K. C., & Zhai, X. (2021). Exploring the effect of assessment construct complexity on machine learning scoring of argumentation. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Virtual.

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A systematic review of deep learning approaches to educational data mining. *Complexity*, *2019*(1306039), 1–22.

Huang, C.-J., Wang, Y.-W., Huang, T.-H., Chen, Y.-C., Chen, H.-M., & Chang, S.-C. (2011). Performance evaluation of an online argumentation learning assistance agent. *Computers & Education*, *57*(1), 1270–1280.

Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal of Philosophy of Science*, *2*(1), 119–135.

Jeong, H., Songer, N. B., & Lee, S.-Y. (2007). Evidentiary competence: Sixth graders' understanding for gathering and interpreting evidence in scientific investigations. *Research in Science Education*, *37*(1), 75–97.

Jescovitch, L. N., Doherty, J. H., Scott, E. E., Cerchiara, J. A., Wenderoth, M. P., Urban-Lurain, M., Merrill, J. E., & Haudek, K. C. (2019a). Challenges in developing computerized scoring models for principle-

based reasoning in a physiology context. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Baltimore, MD.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Doherty, J. H., Wenderoth, M. P., Merrill, J. E., Urban-Lurain, M., & Haudek, K. C. (2019b). Deconstruction of holistic rubrics into analytic bins for large-scale assessments of students' reasoning of complex science concepts. *Practical Assessment, Research & Evaluation*, *24*(7), 1–13.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J. E., Urban-Lurain, M., Doherty, J. H., & Haudek, K. C. (2021). Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression. *Journal of Science Education and Technology*, *30*(2), 150–167.

Jiménez-Aleixandre, M. P., Bugallo Rodríguez, A., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": Argument in high school genetics. *Science Education*, *84*(6), 757–792.

Jiménez-Aleixandre, M. P., & Erduran, S. (2007). Argumentation in science education: An overview. In S. Erduran & M. P. Jiménez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from classroom-based research* (pp. 3–27). Springer.

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing: An Introduction to natural language processing, computational linguistics, and speech recognition* (Vol. 3). Prentice Hall.

Kaldaras, L., & Haudek, K. C. (2022). Validation of automated scoring for learning progression-aligned next generation science standards performance assessments. *Frontiers in Education*, *7*(968289), 1–22.

Kaldaras, L., Yoshida, N. R., & Haudek, K. C. (2022). Rubric development for AI-enabled scoring of three-dimensional constructed-response assessment aligned to NGSS learning progression. *Frontiers in Education*, *7*(983055), 1–15.

Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching*, *41*(7), 748–769.

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: The role of scaffolding in assessment tasks. *Science Education*, *98*(4), 674–704.

Koroteev, M. (2021). BERT: A review of applications in natural language processing and understanding. *arXiv Preprint*, arXiv:2103.11943. https://doi.org/10.48550/arXiv.2103.11943

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

Krist, C., Schwarz, C. V., & Reiser, B. J. (2019). Identifying essential epistemic heuristics for guiding mechanistic reasoning in science learning. *Journal of the Learning Sciences*, *28*(2), 160–205.

Kubsch, M., Czinczel, B., Lossjew, J., Wyrwich, T., Bednorz, D., Bernholt, S., Fiedler, D., Strauß, S., Cress, U., Drachsler, H., Neumann, K., & Rummel, N. (2022). Toward learning progression analytics: Developing learning environments for the automated analysis of learning using evidence centered design. *Frontiers in Education*, *7*(981910), 1–15.

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

Kubsch, M., Rosenberg, J. M., & Krist, C. (2021). Beyond supervision: Human/machine distributed learning in learning sciences research. In E. de Vries, Y. Hod, & J. Ahn (Eds.), *Proceedings of the 15th international conference of the learning sciences: ICLS 2021* (pp. 897–898). International Society of the Learning Sciences.

Küchemann, S., Klein, P., Becker, S., Kumari, N., & Kuhn, J. (2020). Classification of students' conceptual understanding in STEM education using their visual attention distributions: A comparison of three machine-learning approaches. In *Proceedings of the 12th international conference on computer supported education 2020* (pp. 36–46). SciTePress.

Kuhn, D. (2010). Teaching and learning science as argument. *Science Education*, *94*(5), 810–824.

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development*, *74*(5), 1245–1260.

Kuo, C.-Y., & Wu, H.-K. (2013). Toward an integrated model for designing assessment systems: An analysis of the current status of computer-based assessments in science. *Computers & Education*, *68*, 388–403.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lee, H.-S., Gweon, G.-H., Lord, T., Paessel, N., Pallant, A., & Pryputniewicz, S. (2021). Machine learning-enabled automated feedback: Supporting students' revision of scientific arguments based on data drawn from simulation. *Journal of Science Education and Technology*, *30*(2), 168–192.

Lee, H.-S., Liu, O. L., Pallant, A., Roohr, K. C., Pryputniewicz, S., & Buck, Z. E. (2014). Assessment of uncertainty-infused scientific argumentation. *Journal of Research in Science Teaching*, *51*(5), 581–605.

Lee, H.-S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education*, *103*(3), 590–622.

Lieber, L. S., & Graulich, N. (2020). Thinking in alternatives: A task design for challenging students' problem-solving approaches in organic chemistry. *Journal of Chemical Education*, *97*(10), 3731–3738.

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

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

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

Lim, L., Bannert, M., van der Graaf, J., Singh, S., Fan, Y., Surendrannair, S., Rakovic, M., Molenaar, I., Moore, J., & Gašević, D. (2023). Effects of real-time analytics-based personalized scaffolds on students' self-regulated learning. *Computers in Human Behavior*, *139*(107547), 1–18.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, *33*(2), 19–28.

Liu, X. (2020). *Using and developing measurement instruments in science education: A Rasch modeling approach* (Vol. 2). Information Age Publishing.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv Preprint*, arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

Luisi, P. L. (2002). Emergence in chemistry: Chemistry as the embodiment of emergence. *Foundations of Chemistry*, *4*(3), 183–200.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc.

Luo, X., Wei, B., Shi, M., & Xiao, X. (2020). Exploring the impact of the reasoning flow scaffold (RFS) on students' scientific argumentation: Based on the structure of observed learning outcomes (SOLO) taxonomy. *Chemistry Education Research and Practice*, *21*(4), 1083–1094.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25.

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B., & Krajcik, J. (2021). Using machine learning to score multi-dimensional assessments of chemistry and physics. *Journal of Science Education and Technology*, *30*(2), 239–254.

Manning, C. D. (2022). Human language understanding & reasoning. *Daedalus*, *151*(2), 127–138.

Manz, E. (2016). Examining evidence construction as the transformation of the material world into community knowledge. *Journal of Research in Science Teaching*, *53*(7), 1113–1140.

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H.-S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, *23*(2), 121–138.

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

Mathew, A., Amudha, P., & Sivakumari, S. (2021). Deep learning techniques: An overview. In A. E. Hassanien, R. Bhatnagar, & A. Darwish (Eds.), *Advanced machine learning technologies and applications: Proceedings of AMLTA 2020* (pp. 599–608). Springer.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density-based clustering. *Journal of Open Source Software*, *2*(11), 205–206.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv Preprint*, arXiv:1802.03426. https://doi.org/10.48550/arXiv.1802.03426

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

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

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv Preprint*, arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781

Miller, E. C., & Krajcik, J. S. (2019). Promoting deep learning through project-based learning: A design problem. *Disciplinary and Interdisciplinary Science Education Research*, *1*(1), 1–10.

Mislevy, R. J. (2016). How developments in psychology and technology challenge validity argumentation. *Journal of Educational Measurement*, *53*(3), 265–292.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), 1–29.

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundation of machine learning*. MIT Press.

Moon, A., Moeller, R., Gere, A. R., & Shultz, G. V. (2019). Application and testing of a framework for characterizing the quality of scientific reasoning in chemistry students' writing on ocean acidification. *Chemistry Education Research and Practice*, *20*(3), 484–494.

Moreira, P., Marzabal, A., & Talanquer, V. (2019). Using a mechanistic framework to characterise chemistry students' reasoning in written explanations. *Chemistry Education Research and Practice*, *20*(1), 120–131.

Nehm, R. H., & Haertig, H. (2012). Human vs. computer diagnosis of students' natural selection knowledge: Testing the efficacy of text analytic software. *Journal of Science Education and Technology*, *21*(1), 56–73.

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, *49*(1), 3–42.

Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2021). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, *50*(1), 202–237.

Newton, P., Driver, R., & Osborne, J. (1999). The place of argumentation in the pedagogy of school science. *International Journal of Science Education*, *21*(5), 553–576.

Noroozi, O., Kirschner, P. A., Biemanns, H. J. A., & Mulder, M. (2018). Promoting argumentation competence: Extending from first- to second-order scaffolding through adaptive fading. *Educational Psychology Review*, *30*(1), 153–176.

Novak, G. M., Gavrin, A., Patterson, E., & Christian, W. (1999). *Just-In-time teaching: Blending active learning with web technology*. Prentice Hall.

Noyes, K., McKay, R. L., Neumann, M., Haudek, K. C., & Cooper, M. M. (2020). Developing computer resources to automate analysis of students' explanations of London dispersion forces. *Journal of Chemical Education*, *97*(11), 3923–3936.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching*, *41*(10), 994–1020.

Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education*, *95*(4), 627–638.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., & Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. Paper presented at the 33rd Conference on Neural Information Processing Systems, Vancouver, CAN.

Pellegrino, J., DiBello, L., & Goldman, S. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, *51*(1), 1–23.

Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education*, *52*(3), 275–300.

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

Rosenberg, J. M., & Krist, C. (2021). Combining machine learning and qualitative methods to elaborate students' ideas about the generality of their model-based explanations. *Journal of Science Education and Technology*, *30*(2), 255–267.

Ruder, S. (2019). *Neural transfer learning for natural language processing*. National University of Ireland.

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

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction*, *83*(101620), 1–10.

Saldana, J. (2015). *The coding manual for qualitative researchers* (Vol. 3). Sage Publications.

Sampson, V., Grooms, J., & Walker, J. P. (2011). Argument-driven inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments: An exploratory study. *Science Education*, *95*(2), 217–257.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 211–229.

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction*, *23*(1), 23–55.

Sevian, H., & Talanquer, V. (2014). Rethinking chemistry: A learning progression on chemical thinking. *Chemistry Education Research and Practice*, *15*(1), 10–23.

Sherin, B. (2013). A computational study of commonsense science: An exploration in the automated analysis of clinical interview data. *Journal of the Learning Sciences*, *22*(4), 600–638.

Soo, K. W. (2019). *The role of granularity in causal learning*. University of Pittsburgh.

Southard, K., Wince, T., Meddleton, S., & Bolger, M. S. (2016). Features of knowledge building in biology: Understanding undergraduate students' ideas about molecular mechanisms. *CBE: Life Sciences Education*, *15*(1), ar7.1–ar7.16.

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*(100075), 1–10.

Taher Pilehvar, M., & Camacho-Collados, J. (2020). *Embeddings in natural language processing: Theory and advances in vector representations of meaning*. Morgan & Claypool Publishers.

Talanquer, V. (2018). Progressions in reasoning about structure-property relationships. *Chemistry Education Research and Practice*, *19*(4), 998–1009.

Tansomboon, C., Gerard, L. F., Vitale, J. M., & Linn, M. C. (2017). Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education*, *27*(4), 729–757.

Toulmin, S. E. (2003). *The uses of argument (Rev. Ed.)*. Cambridge University Press.

Tschisgale, P., Wulff, P., & Kubsch, M. (2023). Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review Physics Education Research*, *19*(2), 020123-1-020123-24.

Ullmann, T. D. (2019). Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education*, *29*(2), 217–257.

van Mil, M. H. W., Boerwinkel, D. J., & Waarlo, A. J. (2013). Modelling molecular mechanisms: A framework of scientific reasoning to construct molecular-level explanations for cellular behaviour. *Science & Education*, *22*(1), 93–118.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30*. Curran Associates, Inc.

Vitale, J. M., McBride, E., & Linn, M. C. (2016). Distinguishing complex ideas about climate change: Knowledge integration vs. specific guidance. *International Journal of Science Education*, *38*(9), 1548–1569.

Walker, J. P., Van Duzor, A. G., & Lower, M. A. (2019). Facilitating argumentation in the laboratory: The challenges of claim change and justification by theory. *Journal of Chemical Education*, *96*(3), 435–444.

Wang, C., Liu, X., Wang, L., Sun, Y., & Zhang, H. (2021). Automated scoring of Chinese grades 7-9 students' competence in interpreting and arguing from evidence. *Journal of Science Education and Technology*, *30*(2), 269–282.

Watts, F. M., Dood, A. J., & Shultz, G. V. (2023a). Automated, content-focused feedback for a writing-to-learn assignment in an undergraduate organic chemistry course. In *LAK23: 13th international learning analytics and knowledge conference* (pp. 531–537). Association for Computing Machinery.

Watts, F. M., Dood, A. J., & Shultz, G. V. (2023b). Developing machine learning models for automated analysis of organic chemistry students' written descriptions of organic reaction mechanisms. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (pp. 285–303). The Royal Society of Chemistry.

Watts, F. M., Park, G. Y., Petterson, M. N., & Shultz, G. V. (2022). Considering alternative reaction mechanisms: Students' use of multiple representations to reason about mechanisms for a writing-to-learn assignment. *Chemistry Education Research and Practice*, *23*(2), 486–507.

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

Weinrich, M. L., & Talanquer, V. (2016). Mapping students' modes of reasoning when thinking about chemical reactions used to make a desired product. *Chemistry Education Research and Practice*, *17*(2), 394–406.

Wilson, C. D., Haudek, K. C., Osborne, J. F., Buck Bracey, Z. E., Cheuk, T., Donovan, B. M., Stuhlsatz, M. A. M., Santiago, M. M., & Zhai, X. (2023). Using automated analysis to assess middle school students' competence with scientific argumentation. *Journal of Research in Science Teaching*, 1–32.

Winograd, B. A., Dood, A. J., Finkenstaedt-Quinn, S. A., Gere, A. R., & Shultz, G. V. (2021a). Automating characterization of peer review comments in chemistry courses. In C. E. Hmelo-Silver, B. de Wever, & J. Oshima (Eds.), *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning: CSCL 2021* (pp. 11–18). International Society of the Learning Sciences.

Winograd, B. A., Dood, A. J., Moon, A., Moeller, R., Shultz, G. V., & Gere, A. R. (2021b). Detecting high orders of cognitive complexity in students' reasoning in argumentative writing about ocean acidification. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 586–591). Association for Computing Machinery.

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

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2022). Utilizing a pretrained language model (BERT) to classify preservice physics teachers' written reflections. *International Journal of Artificial Intelligence in Education*, *33*(3), 439–466.

Wulff, P., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2023). Enhancing writing analytics in science education research with machine learning and natural language processing: Formative assessment of science and non-science preservice teachers' written reflections. *Frontiers in Education*, *7*(1061461), 1–18.

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

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement*, *76*(2), 280–303.

Zhai, X., Haudek, K. C., & Ma, W. (2023). Assessing argumentation using machine learning and cognitive diagnostic modeling. *Research in Science Education*, *53*(2), 405–424.

Zhai, X., Haudek, K. C., Shi, L., Nehm, R. H., & Urban-Lurain, M. (2020a). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching*, *57*(9), 1430–1459.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020b). Applying machine learning in science assessment: A systematic review. *Studies in Science Education*, *56*(1), 111–151.

Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education*, *39*(12), 1648–1668.

How to cite this article: Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2023). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*, 1–36. https://doi.org/10.1002/tea.21903