Reasoning, granularity, and comparisons: a unit-based method for characterizing students' arguments on chemistry assessments

Jacky M. Deng,[a] Alison B. Flynn*[a]

Abstract

In a world facing complex global challenges, citizens around the world need to be able to engage in argumentation supported by scientific evidence and reasoning. In order for coming generations to have proficiency in this skill, students must be provided opportunities to develop and demonstrate argumentation science classrooms, including on assessments. For example, students can be provided with assessment items that explicitly ask them to reason from evidence. Alongside these assessment items, researchers and educators need methods to evaluate students' written arguments. In this study, we present a unit-based method for characterising students' arguments on chemistry assessments. This unit-based method identifies units (links, concepts, comparisons) within one's argument, and uses these units to evaluate an argument based on three dimensions: reasoning, granularity, and comparisons. To demonstrate this method, we report our findings from using it evaluate two different organic chemistry questions: (1) justifying why one of three bases would drive an equilibrium towards products (*N* = 170), and (2) justifying why one of two reaction mechanisms is more plausible (*N* = 122). Lastly, to translate the method into a rubric for educators, we compare a scoring system based on the unit-based method against a traditional scoring system. As well, we report our findings from interviews with educators (*N* = 4) to invite their feedback on the rubric and its dimensions.

Introduction

Citizens need to be able to argue from scientific evidence

In a world facing complex global issues, such as those highlighted by the United Nations Sustainable Development Goals (United Nations, 2015), citizens around the world need to be able to argue from scientific evidence in order to make informed decisions on various topics. For example, tackling polarizing issues such climate changes will require a citizens who can understand how data related to atmospheric CO_2 levels and ocean acidification are used to justify claims about climate change (Jones and Crow, 2017).

In parallel with these issues, national frameworks for science education in the United States have identified explanations of and arguments about phenomena as a key scientific practice (National Research Council, 2012). The importance of such skills has also been articulated in Europe (European Union, 2006; Jimenez-Aleixandre and Federico-Agraso, 2009) and Canada (Social Sciences and Humanities Research Council, 2018). The Programme for International Student Assessment (PISA) also emphasizes three dimensions of scientific competence, of which "using scientific evidence to draw and communicate conclusion and to identify the assumptions, evidence, and reasoning behind conclusions" is included (Organisation for Economic Cooperation and Development, 2006).

However, despite the need for future citizens to be able to construct scientific arguments and explanations, chemistry education research has found that these scientific practices are largely absent in traditional chemistry assessments. For example, constructing scientific explanations appeared in less than 10% of American Chemical Society (ACS) general chemistry exam items examined in 2016 (Laverty *et al.*, 2016; Reed *et al.*, 2017). Additionally, an ACS Exam for organic chemistry did not assess students' ability to construct scientific explanations or arguments at all (Stowe and Cooper, 2017). From this, curricula have emerged which explicitly include argumentation and explanation (Cooper and Klymkowsky, 2013), as well as research focused on characterizing argumentation and explanation in laboratory settings (Carmel *et al.*, 2019).

Arguments provide insight into students' reasoning

For this work, there is necessary distinction between the practices of argumentation and explanation (Osborne and Patterson, 2011). While a number of definitions have been proposed (Toulmin, 1958; McNeill *et al.*, 2006; Kuhn, 2011; Osborne and Patterson, 2011; Becker *et al.*, 2013), the working definitions which guided this work are as follows. Scientific explanations "explain observed relationships between variables and describe the mechanisms that support cause and effect inferences about them" (National Research Council, 2012); in other words, an explanation is used to explain a consensually agreed-upon fact or phenomenon (Osborne and Patterson, 2011).

In contrast to explanations, arguments seek to persuade by justifying claims with evidence and reasoning (Toulmin, 1958). An argument is an assertion with a justification (McNeill *et al.*, 2006; Kuhn, 2011); the claim is in doubt and must be advanced by constructing an argument about the fit between evidence and claim through reasoning (Osborne and Patterson, 2011). Described in this way, as an educational tool, arguments provide a theoretical foundation to investigate students' reasoning (Emig, 1977; Berland and Reiser, 2009; Grimberg and Hand, 2009).

As such, recent studies in chemistry education research have worked to characterize students' reasoning through analysis of their arguments about chemical phenomena (Sevian and Talanquer, 2014; Weinrich and Talanquer, 2016; Bodé *et al.*, 2019; Moon *et al.*, 2019; Moreira *et al.*, 2019). For example, Sevian and Talanquer (2014) interviewed individuals ranging from high school chemistry students to chemistry experts (*e.g.* academia, industry professionals, *etc.*). The interviewees were asked to construct arguments when deciding on a fuel to power a GoKart, and through their arguments, the researchers characterized students' reasoning as one of descriptive, relational, linear causal, or multi-component causal. These modes of reasoning, known as the Chemical Thinking Learning Progression (CTLP), have since been used in other chemistry education research to characterize students' reasoning through analysis of arguments across a variety of contexts and tasks (Moon *et al.*, 2016; Bodé *et al.*, 2019; Moreira *et al.*, 2019). Building from this literature, we employ these modes of reasoning in our evaluation of students' arguments on two different organic chemistry questions.

As part of this study, we analysed students' written arguments on two organic chemistry assessment items. Specifically, these questions asked students to articulate a claim, and then justify that claim with evidence and reasoning. This structure is similar to other assessment items described in the chemistry education research literature (Cooper *et al.*, 2016; Stowe and Cooper, 2019), and are broadly as constructed-response items, as they ask students to actively construct their responses through a specific medium, such as written text (Emig, 1977; Grimberg and Hand, 2009; National Research Council, 2012; Dood *et al.*, 2020).

Challenges when characterizing students' written arguments

If argumentation from evidence is key scientific practice that students are expected to learn from their science education, educators will not only need to provide students with items prompting them for arguments—educators themselves will also need methods to evaluate and and provide feedback on students' argumentation skills.

Holistic rubrics are one common tool used by educators and researchers to characterize students' arguments (Hogan and Murphy, 2007; Yang *et al.*, 2019). Holistic rubrics often take the form of evaluation grids and allow the user to evaluate an argument by reading the argument and deciding on its sophistication based on the general descriptions for each box in the grid. Holistic rubrics benefit from their ease of use and often allow the user to consider the quality of the response as a whole. However, holistic rubrics also present inherent limitations. First, developing holistic rubrics that characterize the quality of students' arguments with high validity can be challenging, and impractical in time-constrained instructional contexts (Kelly and Bazerman, 2003; Sandoval and Millwood, 2005). Second, applying holistic rubrics reliably is challenging; as holistic rubrics require the user to evaluate students' arguments as a whole, it can be difficult to ensure a level of consistency between users (Ha *et al.*, 2011; Liu *et al.*, 2016). These limitations are aggravated by the variety of constructed-response items that educators can administer to students. In the research alone, students' arguments have been studied in both long-form assessments, such as research reports (Kelly and Takao, 2002; Kelly *et al.*, 2007), and short-assessments (Sandoval and Reiser, 2004; Sandoval and Millwood, 2005; Ha *et al.*, 2011; Liu *et al.*, 2016; Moreira *et al.*, 2019). Lastly, holistic rubrics can also be useful in that they describe the overall quality of a response with a single number, but depending on the context, task, or goals of the educator/researcher, the resulting number may be imprecise or unuseful (Yang *et al.*, 2019).

An alternative to holistic rubrics are unit-based methods (also known as analytical rubrics) (Yang *et al.*, 2019). Unit-based methods determine the quality of a response based on the presence or absence of specific units within the response (Moon *et al.*, 2019); in other words, multiple "sub" numbers are used to generate a final, summative rating for the argument (Yang *et al.*, 2019). For example, to determine the mode of reasoning exhibited in students' arguments about freezing point depression, Moreira *et al.* (2019) identified individual units described in one's argument (*e.g.* entities, properties, activities, organization) and then used these units to decide on the argument's quality (in this case, the argument's mode of reasoning). As unit-based methods make decisions about quality based on the presence or absence of specific units, unit-based methods are generally more reliable than rubric-based methods (Moon *et al.*, 2019). However, because unit-based methods necessitate searching for and identifying individual units within arguments, unit-based methods are generally more time-consuming and tedious, and as a result, may be impractical for use in certain instructional contexts (Moon *et al.*, 2019). As noted by Dood and colleagues (2020): "Though useful for eliciting and developing explanations, constructed-response items are onerous to incorporate in courses, as time is required for an educator to read and score response". As a result, most instances of unit-based methods in evaluating constructed-response items have been focused on shorter constructed-response items in research contexts (Sandoval and Reiser, 2004; Sandoval and Millwood, 2005; Ha *et al.*, 2011; Liu *et al.*, 2016; Moreira *et al.*, 2019).

Additionally, If students are to develop and demonstrate their argumentation through constructed-response items, new methods that allow educators to evaluate *both* the structural (or domain-general) and conceptual (or domain-specific) components of arguments will be essential moving forward (Sandoval and Millwood, 2005; Petritis *et al.*, 2020). To date, most methods, either holistic or unit-based, do not make this distinction explicit. For example, Moreira *et al.* (2019) used a unit-based method to characterize students' reasoning based on entities, properties, activities, and organization. However, this work did not seek to capture the conceptual correctness of within students' arguments. Other work by Kelly and Takao (2002) used a holistic rubric to characterize how evidence was used in students' arguments; again, this work did not seek to capture the conceptual correctness of students' content knowledge.

Lastly, previous work focused on chemistry students' arguments has also almost exclusively focused on evaluating students' reasoning. Though the importance of characterizing students' reasoning is well-reported and argued (Kelly *et al.*, 2007; Talanquer, 2014, 2018a; Caspari, Weinrich, *et al.*, 2018), we propose that an argument can provide insight into additional dimensions of student thinking. For example, being able to compare between claims and constructing an argument at a specific scalar level are both key to the scientific practice of argumentation (Machamer *et al.*, 2000; Darden, 2002), and broader argumentation evaluation frameworks have included counterclaims as a characteristic of high quality argumentation (Erduran *et al.*, 2004).

In this article, we present a unit-based method for characterizing students' arguments on chemistry assessments in terms of three domain-general characteristics—reasoning, granularity, and comparisons—alongside domain-specific characteristics. Herein, we demonstrate our application of this method for evaluating students' arguments on two organic chemistry exam questions. We also discuss our efforts in preparing a rubric based on the dimensions of the unit-based method for use in instructional practice, including comparing scores generated by the unit-based method against scores generated by traditional evaluation methods, as well as collecting educators' perspectives on the rubric in interviews.

Analytical framework

Arguments aim to persuade with evidence and reasoning

Students' responses were interpreted through the lens of Toulmin's argument pattern, which organizes arguments in terms of three components: claim, evidence, and reasoning (Toulmin, 1958).

This approach that has been used in other chemistry contexts, such as physical chemistry (Becker *et al.*, 2013; Moon *et al.*, 2016, 2017) and organic chemistry (Cruz-Ramírez De Arellano and Towns, 2014; Bodé *et al.*, 2019). The questions given to students in this study were explicitly organized in this fashion; in the first part of the question, students were asked to make a claim, and in the second part, they were asked to justify their claim by constructing an argument using evidence and reasoning.

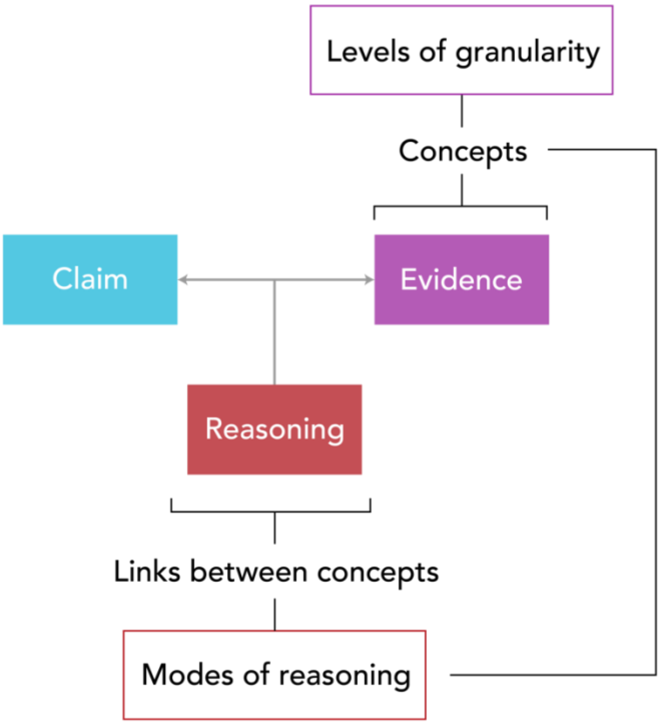


Figure 1: Overview of how Toulmin's structure of argument was used to organize students' responses in terms of concepts and links between concepts. These concepts and links were then used to inform decisions about an argument's mode of reasoning and level of granularity.

In science education (Darden, 2002; Russ *et al.*, 2008; Southard *et al.*, 2017) and chemistry education (Sevian and Talanquer, 2014; Caspari, Kranz, *et al.*, 2018; Caspari, Weinrich, *et al.*, 2018; Talanquer, 2018a), evidence has been considered the entities and activities in a mechanism. Other interpretations of evidence include identified features of entities and phenomena (Kuhn, 2011; Webber and Flynn, 2018), energetic and structural accounts (Kuhn, 2011; Caspari, Kranz, *et al.*, 2018), and dimensions with variations in explanatory power (*i.e.*, chemical mechanism, causality) (Yan and Talanquer, 2015; Weinrich and Talanquer, 2016). In this study, we defined evidence as the key words or concepts the student leveraged in their argument Figure 1. For example, if a student discussed the concept of base strength in their argument, "base strength" was considered a piece of evidence that the student was introducing into their argument to support their claim.

We defined reasoning as how students organized and linked concepts in their arguments (Figure 1). Our discussion of how both concepts and links were used to determine an arguments' mode of reasoning will be discussed in the forthcoming section on modes of reasoning, as well as in our Methods section.

Modes of reasoning, levels of granularity, and levels of comparison

In the following sections we describe the three main dimensions included this study's unit-based method for characterizing students' arguments (Figure 2), as well as the rationale for their inclusion.

Modes of reasoning. Students' reasoning has been analysed through a variety of different lenses and frameworks in chemistry education research. These include Type I and II reasoning (Talanquer, 2007, 2017; McClary and Talanquer, 2011; Maeyer and Talanquer, 2013), teleological reasoning (Talanquer, 2007; Abrams and Southerland, 2010; Caspari, Weinrich, *et al.*, 2018; Trommler *et al.*, 2018; DeCocq and Bhattacharyya, 2019), abstractedness and abstraction (Sevian *et al.*, 2015; Weinrich and Sevian, 2017), rules-, case-, and model-based reasoning (Windschitl *et al.*, 2008; Kraft *et al.*, 2010; DeCocq and Bhattacharyya, 2019), and causal, mechanistic, and causal mechanistic reasoning (Cooper *et al.*, 2016; Crandell *et al.*, 2018). Depending on the goals of the study, the chosen reasoning framework and definitions will vary.

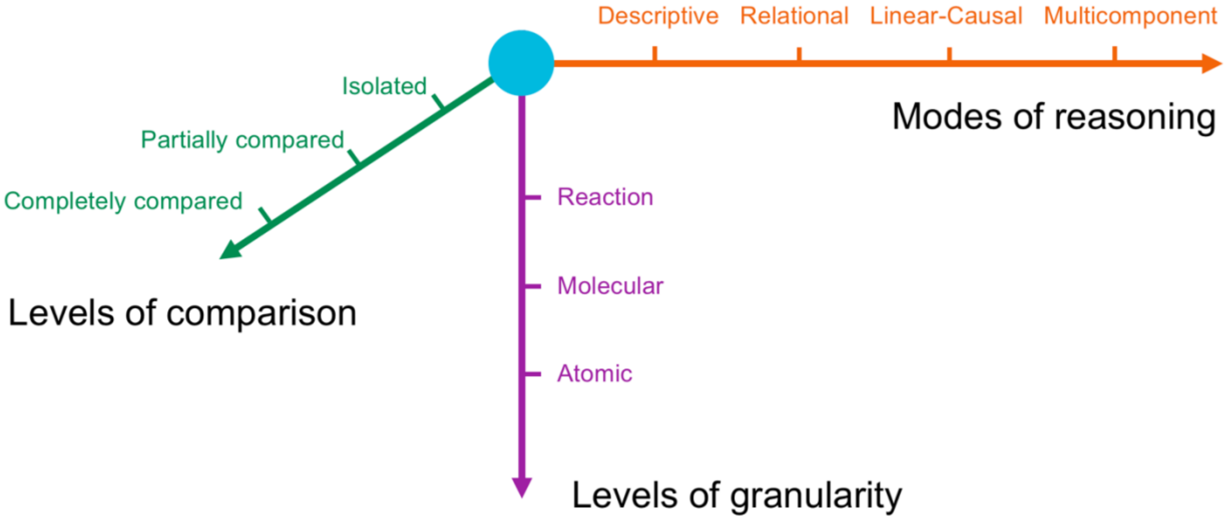


Figure 2: The analytical framework used in this work. This framework allows one to use modes of reasoning, levels of granularity, and levels of comparison to characterize students' evidence and reasoning for their claims.

In this study, we analysed students' arguments using the modes of reasoning framework (Sevian and Talanquer, 2014; Weinrich and Talanquer, 2016; Caspari, Kranz, *et al.*, 2018). We chose this framework due to its alignment with the intended learning outcomes of the course context of this study, including its associated classroom activities that relate to creating scientific arguments (National Research Council, 2012). Within this framework, one categorizes students' arguments into four distinct modes of reasoning: descriptive, relational, linear causal, and multi-component causal (Table 1, with additional examples in Appendix A). Given that we interpreted students' arguments in terms of Toulmin's structure of argument, each mode of reasoning was defined in terms of how claim and evidence were organized and connected within each argument. The following paragraphs will discuss our working definitions for each mode of reasoning, along with examples of how these modes of reasoning might manifest in example arguments in context of global warming.

Descriptive arguments identify evidence without reference to how this evidence relates to the claim. For example, to justify the claim that humans are causing global warming, one might simply state "greenhouse gases" as evidence. However, without an explicit link between the evidence and the claim, it is unclear how the evidence is connected to the claim, if at all.

Relational arguments establish relationships between evidence and claim, but in a correlative fashion absent of causality. In other words, links are stated, but do not get to *why* these links or evidence are appropriate; the statements are a "matter-of-fact". For example, to justify the claim that humans are causing global warming, one might state: "Humans are causing global warming because humans generate greenhouse gases." Compared to the

Table 1: Descriptions and examples of modes of reasoning. Adapted from Sevian and Talanquer (2014).

Mode of Reasoning	Description	Example
Descriptive	Argument contains only descriptions of evidence and/or the claim No relationships established between evidence and the claim.	*NaH is the **strongest base.*** *NaH will **drive the equilibrium towards products.***
Relational	Argument contains descriptions of evidence and the claim Evidence is correlated to other evidence and/or to the claim (*i.e.*, a "matter-of-fact")	*NaH will **drive the equilibrium towards products** because it is the **strongest base.***
Linear Causal	Argument contains descriptions of evidence and the claim Evidence linked causally to other evidence and/or to the claim (*i.e.*, justification for *why*)	*NaH will **drive the equilibrium towards products** because it is the **strongest base**. A **strong base** will drive the equilibrium towards products because its **conjugate acid is weak** and has a **high pK_a value.***
Multi-Component Causal	Multiple causal relationships are described and coalesce to justify a single claim.	*NaH will **drive the equilibrium towards products** because it is the **strongest base**. A **strong base** will drive the equilibrium towards products because its **conjugate acid is weak** and has a **high pK_a value.*** *Additionally, the conjugate acid of NaH is H_2, a **stable gas** that will **drive the equilibrium towards products** due to **Le Chatelier's principle**.*

descriptive example, this argument includes an *explicit* link between the evidence and the claim. However, this argument treats the link as a "matter-of-fact", and the reader is left wondering why or how greenhouse gases contribute to global warming.

Causal arguments describe how claim and evidence are linked through cause-and-effect; links are stated, and additional reasoning explains why or how these links are relevant and/or appropriate by referencing scientific knowledge, principles, additional evidence, *etc*. Linear causal arguments establish a single chain of causal relationships between one or more pieces of evidence to justify a single claim. For example, a linear causal argument to justify the claim that humans are causing global warming may be: "Humans are causing global warming because humans generate greenhouse gases. Greenhouse gases contribute to global warming because they trap heat in the Earth's atmosphere." Here, the second sentence serves as the reasoning that explains the relationship between the claim and evidence in the first sentence.

Lastly, multi-component causal arguments establish multiple chains of causal relationships between more than one piece of evidence to support a single claim. A multi-component causal argument to justify the claim that humans are causing global warming may include the same linear causal example above, but with an added "chain" of causal reasoning to support the original claim, such as: "Human are causing global warming because they produce chlorofluorocarbons. Chlorofluorocarbons contribute to global warming because they damage the ozone layer, making it easier for UV light to penetrate to Earth's surface."

We defined the modes of reasoning in terms of domain-general characteristics of Toulmin's argument pattern due to our desire for the unit-based method to be broadly applicable to various types of questions. As a result, our working definitions for the modes of reasoning vary in some ways from definitions used in previous work (Moon *et al.*, 2016; Weinrich and Talanquer, 2016; Caspari, Kranz, *et al.*, 2018; Bodé *et al.*, 2019; Moreira *et al.*, 2019). For example, Caspari *et al.* (2018) differentiated between the levels of complexity used in their work based on students' relative use of explicit structural differences, implicit structural causes, and electronic effects to justify change. As Toulmin's argument pattern does not necessarily differentiate between explicit and implicit features, our decisions about students' modes of reasoning were focused on how students structured their arguments and the nature of the links used to connect evidence to claims. Where previous chemistry education research may have characterized an argument as linear causal if it established a causal relationship between an *implicit* chemical property and an *explicit* chemical feature, we coded arguments as linear causal if causal reasoning was used to establish a connection between the claim and evidence, regardless of whether implicit or explicit features were discussed. Other analytical frameworks for argumentation have been used to evaluate students' arguments related to socioscientific issues in a similar fashion (Sadler, 2006; Kuhn *et al.*, 2013; Lytzerinou and Iordanou, 2020). This is not to say that capturing implicit and explicit features is not important; whether students leveraged implicit or explicit features in their arguments (*i.e.*, the scalar levels of the evidence provided) is captured separately in the levels of granularity dimension of this work's framework.

We also recognize that the hierarchical descriptions of the four modes of reasoning may imply that some modes of reasoning are "better" than others. Although multi-component causal arguments are the most sophisticated modes of reasoning in this framework, this mode of reasoning is not necessarily "better" than the other three modes. In scientific practice and everyday argument/decision-making, having to construct a multi-component causal argument for every possible argument is unrealistic and impractical; a descriptive argument may be acceptable for accomplishing a particular task (Darden, 2002). Indeed, research suggests that how students and scientists reason depends on the task, learning context, and course expectations (Bernholt and Parchmann, 2011; Weinrich and Talanquer, 2016; Caspari, Weinrich, *et al.*, 2018).

Levels of granularity. By disassociating one's reasoning from one's use of evidence at different scalar levels (*e.g.*, implicit and explicit), the dimension of granularity was developed to capture the different scalar levels that emerged in one's argument (Bodé *et al.*, 2019). Granularity has been described using various terms in previous work, including scales (Talanquer, 2018b), levels (van Mil *et al.*, 2013), nested hierarchies (Southard *et al.*, 2017), emergence (with ideas of downward and upward causality) (Luisi, 2002), and bottom-out reasoning (Darden, 2002).

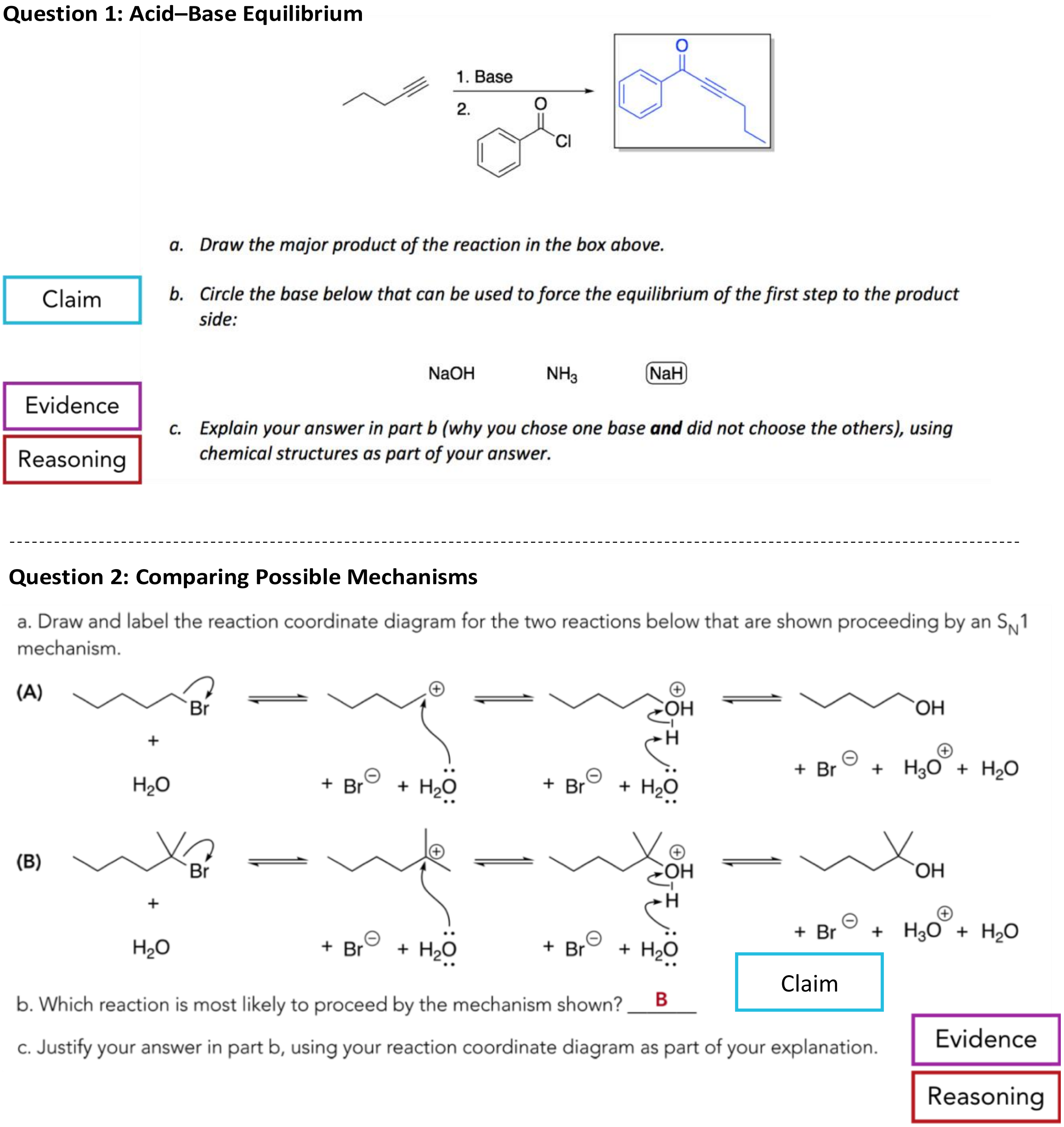
Each discipline has its understood need for particular levels of granularity (Darden, 2002), as different phenomena may be explained from increasingly large macroscopic perspective (*e.g.*, global levels and beyond) or increasingly small submicroscopic perspectives (*e.g.*, atomic levels and beyond). For example, experts or students could be asked to explain how plants can have poisonous and non-poisonous parts; an evolutionary biologist may provide an explanation at the population level (evolutionary explanation for how the differentiation arose), a biologist may provide an explanation tissue/cellular level (cellular differentiation), and a biochemist may provide an explanation at the molecular level (DNA's role) (Southard *et al.*, 2017).

In this study, we categorized students' responses into three distinct levels of granularity: reaction, molecular, and atomic. These three categories were based on the concepts described in students' responses, as well as the intended learning outcomes related to each task. Students' arguments were categorized into a specific level of granularity based on the concepts presented in each argument. We expected the distributions for levels of granularity for the two questions we analysed to be different, because, as is the case for reasoning, people (from experts to students) cannot be expected to provide highly granular arguments for all questions all the time given the variety of contextual factors that impact how they might approach a given task (Darden, 2002). For example, in certain contexts, describing how a reaction proceeds may constitute a sufficient level of granularity for one context, while molecular-level descriptions and reasoning (e.g., resonance effects) may be required in another context.

Levels of comparison. When an argument involves two or more possible claims, a comparison is needed. However, modes of reasoning and levels of granularity do not necessarily capture whether a comparison has been explicitly made between possible claims. Without a comparison, a species cannot be more/less, bigger/smaller, or faster/slower. Comparing between alternatives is key to scientific practice; for example, to justify global warming, one might leverage evidence to refute counterclaims (*e.g.*, claims that global warming does not exist). In the questions used in this study, students had to choose one of multiple claims to argue for, thereby providing an opportunity for students to construct arguments in which they compared their claim to alternatives. In this study, we analysed how students compared between possible claims using definitions for levels of comparison developed in our previous work, shown in Table 2 (Bodé *et al.*, 2019).

Table 2: Descriptions for each level of comparison from Bodé, Deng, & Flynn (2018).

Comparison level	**Description**
Isolated	Concepts in argument for a claim are all discussed in isolation from the other possible claim. Concepts are never used to compare/contrast between the claims.
Partially compared	Some (but not all) concepts in argument for a claim are discussed in relation to the other possible claim. These concepts are used to compare/contrast between the claims.
Fully compared	All concepts in argument for a claim are discussed in relation to the other possible claim. All concepts are used to compare/contrast between the claims.



Question 1: Acid–Base Equilibrium

a. *Draw the major product of the reaction in the box above.*

b. *Circle the base below that can be used to force the equilibrium of the first step to the product side:*

NaOH NH$_3$ [✓ NaH]

c. *Explain your answer in part b (why you chose one base* ***and*** *did not choose the others), using chemical structures as part of your answer.*

Question 2: Comparing Possible Mechanisms

a. Draw and label the reaction coordinate diagram for the two reactions below that are shown proceeding by an S_N1 mechanism.

b. Which reaction is most likely to proceed by the mechanism shown? B

c. Justify your answer in part b, using your reaction coordinate diagram as part of your explanation.

Figure 3: The acid–base equilibrium question (Q1, top), and the comparing mechanisms questions (Q2, bottom). Both questions prompted students for their claim, evidence, and reasoning, and both were taken from the 2017 OCII final exam.

Goals and research questions

Our goals for this work were to (1) analyze students' arguments using the unit-based method and (2) to take initial steps in translating the dimensions of the method into a rubric for educators. To accomplish the first goal, we used the method to evaluate students' arguments on two organic chemistry exam questions. To accomplish the second goal, we developed a rubric based on key words, reasoning, granularity, and comparisons to determine how the evaluating with the unit-based method compared to traditional scoring methods for constructed-response items. We also facilitated interviews with educators to solicit their feedback on the rubric to improve its usability in teaching practice. The research questions that guided this work were:

(1) Using a unit-based method for characterizing arguments, how are students constructing arguments on two different organic chemistry questions? Specifically:

a. How are links made between concepts and what modes of reasoning do students exhibit in their arguments?
b. What concepts are discussed in their arguments and with what levels of granularity?
c. How are concepts compared between possible claims in their arguments?

(2) Using a rubric based on the dimensions of the unit-based method, how do arguments' scores compare against a traditional scoring system focused on keywords?

(3) What are educator's perspectives of a rubric for characterizing students' arguments based on the dimensions of the unit-based method?

Methods

Setting and course

This research was conducted in the Organic Chemistry II course at a large, bilingual, research-intensive university in Canada. At this institution, introductory organic chemistry is provided to students across two semesters as Organic Chemistry I (OCI) and Organic Chemistry II (OCII). OCI is offered in the winter semester of students' first year of studies while OCII is offered in both the summer and fall. Students can take the courses in either English or French. OCII is a 12-week course consisting of two weekly classes (~200 students per section, 1.5 hours each, mandatory, lecture or flipped format) (Flynn, 2015, 2017) and a voluntary tutorial session (1.5 hours). Assessments for the course are comprised of in-class participation via a classroom response system, online homework assignments, two midterms, and a final exam. The course is comprised of ~75% Faculty of Science students, ~17% Faculty of Health Sciences students, and ~8% students from other faculties. General topics addressed in OCII include reactions with σ electrophiles (*e.g.*, $S_N1/S_N2/E1/E2$ and oxidation reactions), introduction to ^1H NMR and IR spectroscopy, reactions of ⍰electrophiles with leaving groups, and reactions with activated ⍰⍰nucleophiles (*e.g.*, aldol reactions) (Flynn and Ogilvie, 2015; Ogilvie *et al.*, 2017).

RQ1: Using a unit-based method for characterizing arguments, how are students constructing arguments on two different organic chemistry questions?

Data source. We analysed students' responses to two final exam questions (shown in Figure 3) from OCII. Question 1 (Q1) asked students to justify the direction of an acid–base equilibrium. Question 2 (Q2) asked students to justify why one of two similar reaction mechanisms (S_N1 vs. S_N2) was more plausible. Both Q1 and Q2 were from a single final exam from 2017 (Ethics approval H03-15-18). pK_a values were not provided to students on Q1, though values for analogues were provided in a data table attached to the exam. Each question followed Toulmin's claim-evidence-reasoning structure, as students were asked to: (a) choose a claim given multiple options, (b) provide an argument for their choice with evidence and reasoning.

Though Q2 had been the subject of analysis in our previous work, this work did not rely on the unit-based method described here. Also, Q1 and Q2 presented a unique opportunity to determine how students' arguments might differ between question types on single summative assessment; therefore, a power analysis revealed that $N = 122$ Q2 arguments would allow for statistically meaningful comparisons between Q1 and Q2.

Importantly, the analysis described here differs in several ways from our previous work (Figure 4), in which we had also used modes of reasoning to characterize Q2 (Bodé *et al.*, 2019) First, the definitions for the modes of reasoning used in our previous work were not aligned with Toulmin's argument pattern, and more aligned with other definitions grounded in discussion of implicit/explicit properties. Our previous work also characterised students' arguments with a holistic rubric in which generic descriptions for the modes of reasoning were used as a rubric to evaluate students' arguments. For example, an argument was evaluated as a whole and said to be linear causal if it linked implicit and explicit properties in a causal fashion. However, in Q1, students were expected to discuss concepts such as conjugate acid strength, pK_a, base strength, the direction of equilibria, and stability; in this case, establishing connections between explicit and implicit features of molecules is less relevant, as students were instead expected to leverage data (*i.e.*, pK_a values), not implicit features of molecules, to justify their claims. Therefore, the analysis is different in this work, which bases decisions about modes of reasoning on the identification of units (concepts and links) within one's argument. Though we did identify units (concepts, links, comparisons) in our previous work, these were not used to make decisions about an argument's mode of reasoning, level of granularity, or level of comparison.

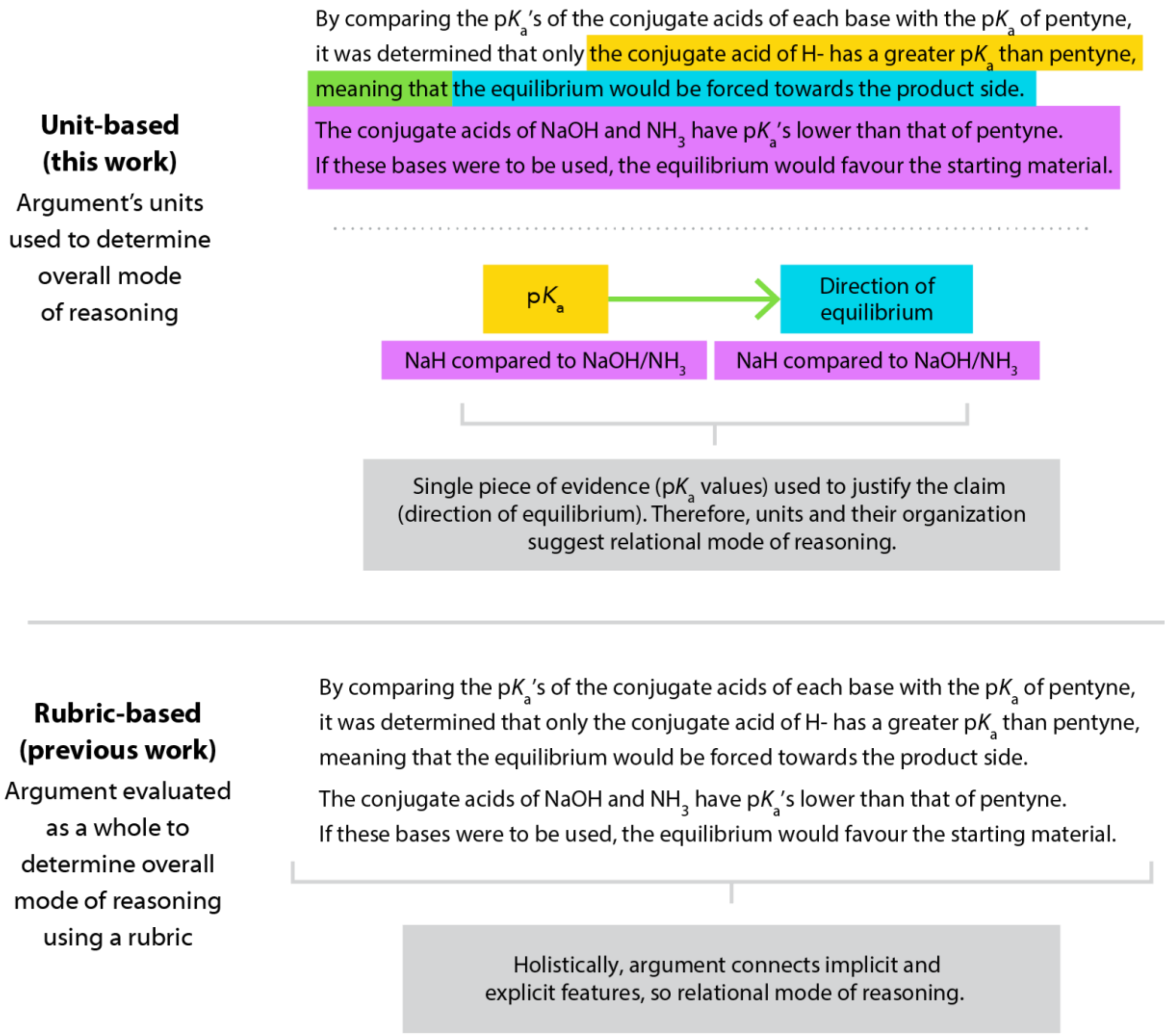

Unit-based (this work)
Argument's units used to determine overall mode of reasoning
By comparing the pK_a's of the conjugate acids of each base with the pK_a of pentyne, it was determined that only the conjugate acid of H- has a greater pK_a than pentyne, meaning that the equilibrium would be forced towards the product side.
The conjugate acids of NaOH and NH_3 have pK_a's lower than that of pentyne. If these bases were to be used, the equilibrium would favour the starting material.
pK_a
Direction of equilibrium
NaH compared to NaOH/NH_3
NaH compared to NaOH/NH_3
Single piece of evidence (pK_a values) used to justify the claim (direction of equilibrium). Therefore, units and their organization suggest relational mode of reasoning.
Rubric-based (previous work)
Argument evaluated as a whole to determine overall mode of reasoning using a rubric
By comparing the pK_a's of the conjugate acids of each base with the pK_a of pentyne, it was determined that only the conjugate acid of H- has a greater pK_a than pentyne, meaning that the equilibrium would be forced towards the product side.
The conjugate acids of NaOH and NH_3 have pK_a's lower than that of pentyne. If these bases were to be used, the equilibrium would favour the starting material.
Holistically, argument connects implicit and explicit features, so relational mode of reasoning.

Figure 4: Comparison of the unit-based method used in this work and the rubric-based method used in Bodé, Deng, & Flynn (2019).

Our decision to switch from a holistic rubric method to a unit-based method was based on challenges associated with consistency and being able to reliably apply the former across a wide range of questions—similar challenges have been described previously in the literature (Ha *et al.*, 2011; Liu *et al.*, 2016; Moon *et al.*, 2019). The unit-based method limits some of the subjectivity by defining each dimension in terms of an argument's units. However, it should be noted that despite a more fine-grained analysis, a unit-based analysis is still subject to variability in interpretation (*e.g.*, different interpretations about whether a specific unit is present or not).

Lastly, though our previous work described granularity as part of its theoretical framework, it did not explicitly analyse the granularity in students' arguments. This was primarily due to the fact that our previous work leveraged students' descriptions of implicit and explicit features to make decisions about their modes of reasoning. In the current study, we disassociate reasoning and granularity and explicitly evaluate students' arguments with these two dimensions separately.

A unit-based method for characterizing arguments. The unit-based method consists of two phases of coding (Figure 5). Phase 1 involves analysing students' arguments for the presence of three types of units: (1) the

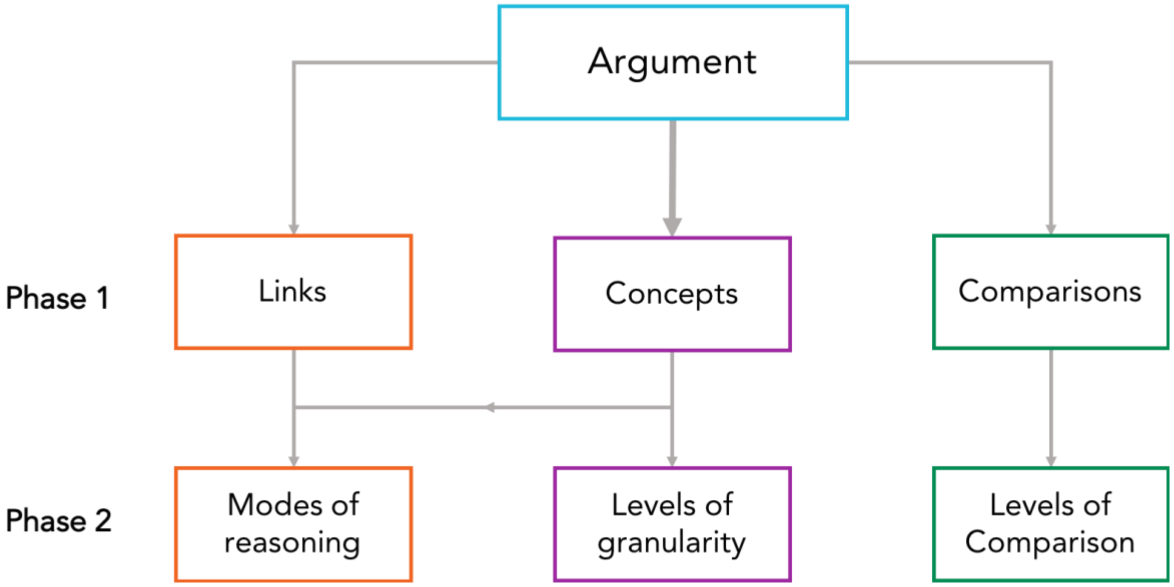


Figure 5: Overview of the unit-based method used to analyse students' arguments. Phase 1 determined the units within arguments, and these units are used to determine the mode of reasoning, level of granularity, and level of comparison (Phase 2).

concepts present within the argument (*i.e.*, the evidence students used in their arguments), (2) links between these concepts, and (3) concepts that are used to compare between claims (Bodé *et al.*, 2019). Phase 2 involved using the units identified in Phase 1 to make decisions about an argument's mode of reasoning, level of granularity, and level of comparison. The following sections will describe each of these steps in detail. For Q1, Phases 1 and 2 were conducted. For Q2, only Phase 2 was conducted; Q2's units had previously been identified in our previous work (Bodé *et al.*, 2019), and because the process for identifying units did not change between the two studies, we drew from our previous data to perform the Phase 2 analysis for Q2.

Phase 1: Identifying units (concepts, links, and comparisons) in arguments. The first phase's codes—concepts, links, and comparisons—were identified based on the expected answers to the questions, which were constructed based on intended learning outcomes from the OCII course (Appendix C). This established content validity for the initial coding scheme, ensuring that we defined our initial scheme based on concepts relevant to course expectations. That is, expected concepts were used to develop the coding protocol, and the coder coded for the presence/absence of these concepts in students' arguments. During the coding process, codes that were not present in the initial coding scheme but were present in students' answers were added to the coding scheme. These additional codes were included even if they were described in error or representative of concepts irrelevant to the question.

Using these codes, Phase 1 analysis involved the following sequence:

(1) Identifying concepts present in the argument and whether these concepts were discussed correctly or with errors.
(2) Identifying links between individual concepts in the argument and whether these links were canonically correct or not.
(3) Identifying which concepts were used to explicitly compare/contrast between possible claims.

Only explicit instances of concepts, links, and comparisons were coded. For example, one would only code for the concept of "base strength" if the argument included phrases like "NaH is a strong base" or "NaH is a stronger base than…".

Links between concepts were said to be present only when the student was explicitly linking between concepts with words like "because", "therefore", "so", *etc.* Lastly, a concept was said to compare between claims if an argument described that concept with reference to one or more of the other possible claims. For example, "NaH is a stronger base than NH_3" or "NaH is a strong base and NH_3 is a weak base" would warrant a comparison code for a "base strength" concept code.

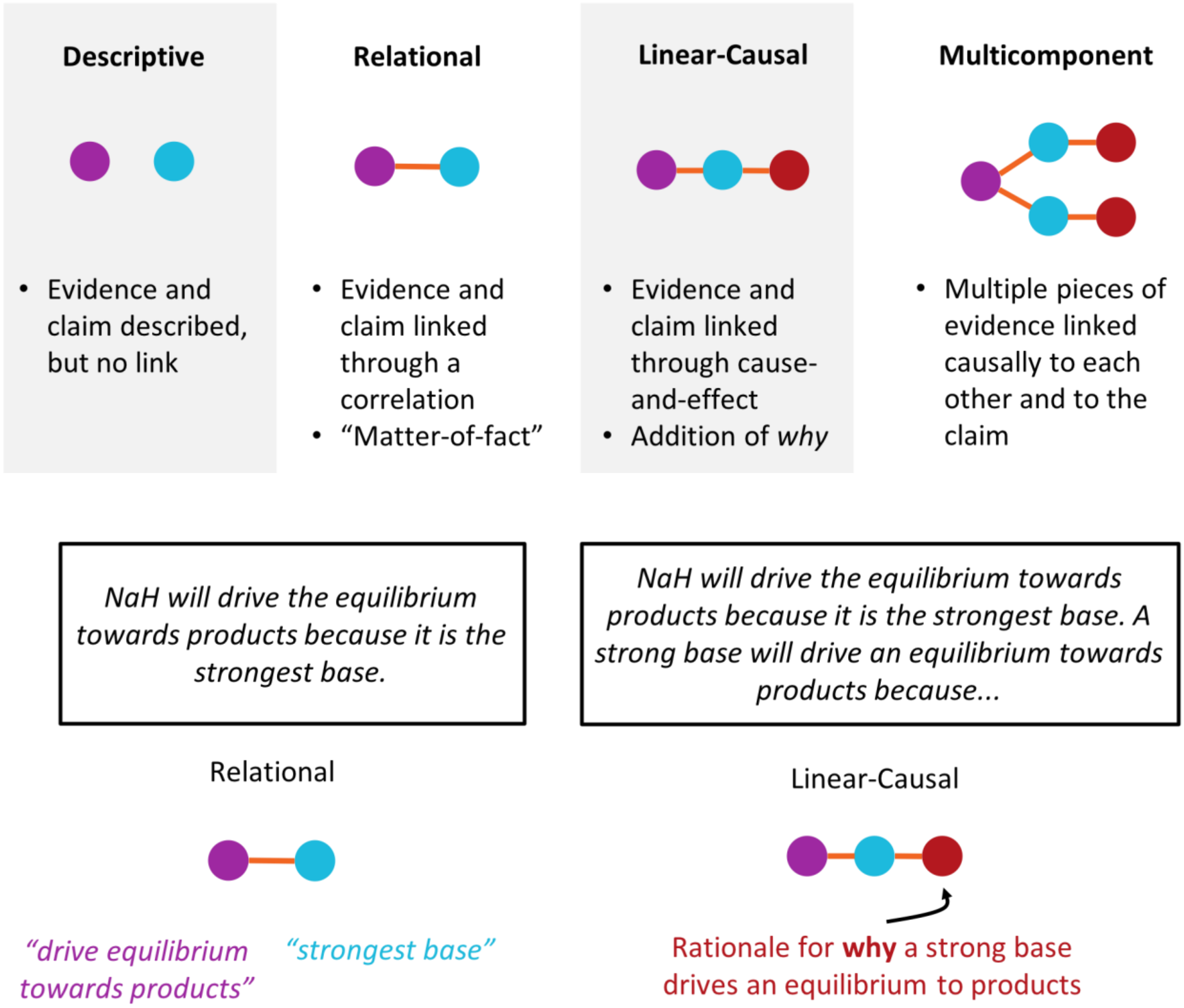
We considered the fact that students may have made implicit references to concepts, links, and comparisons within their arguments. For example, one might be unsure whether to code for the concept of "base strength" if the argument simply stated, "NaH is stronger than NH_3". Though the majority of our analysis focused mainly on the explicit presence of each code, there were instances in which we had to make decisions about the implied presence of a code. However, these instances were in the minority and were resolved by consulting other researchers and/or other aspects of the student's argument to make a decision.

Phase 2a: Modes of reasoning to characterize reasoning in students' arguments

Using the concepts and links identified within students' arguments, we determined the mode of reasoning to be one of descriptive, relational, linear causal, and multi-component causal. For example, in this study, a linear-causal response was said to be present if a student made a claim in their argument (*e.g.*, "The equilibrium will favour products…"), justified that claim with some evidence ("…because NaH is a strong base…"), and further justified that claim by providing reasoning for *why* a strong base drives the equilibrium towards products, with reference to additional evidence ("A strong base drives the equilibrium towards products because it has a conjugate acid with the highest pK_a value"). In contrast, a claim that was justified with only evidence (*e.g.*, "The equilibrium will favour products because NaH is a strong base"), without further justification for why, would be coded as relational. Coded in this way, concept units and link units determined in Phase 1 were used to make decisions about the overall mode of reasoning. Additional examples of how the modes of reasoning were used to characterize arguments are available in Appendix A.

Whether links between concepts were scientifically correct did not inform our coding for the mode of reasoning—with our goal of capturing both domain-specific and domain-general characteristics, an argument could be logically sound but conceptually incorrect (or vice versa). For example, one arguing against the existence climate change may present an argument that is logically sound but relies on evidence and connections that are conceptually false. Indeed, Toulmin's argument pattern describes how an argument can be logical from a structural perspective without being conceptually correct (Toulmin, 1958).

One of the most common and intuitive tools to analyze students' arguments is diagramming, by which the abstract form of an argument can be identified and seen at a glance, and according to which it is then possible to analyze more closely the relationships between the argument's parts. There is a wide range of diagramming techniques; some are very general, while some tailored to particular domains—for instance, the ArguMed and DEFLOG systems are two systems developed analyze the logic of legal arguments (Verheij, 2003). To support or analysis, we drew diagrams to visually represent students' arguments. These diagrams allowed us to visualize both the concept units and link units within students' arguments, which then allowed us to categorize an argument into a specific mode of reasoning by matching the diagram constructed for each argument to a diagram corresponding to a specific mode of reasoning. Examples of how a reasoning diagram was used to characterize an argument can be found in Figure 6, Figure 7,and Appendix A.



- Evidence and claim described, but no link

- Evidence and claim linked through a correlation
- "Matter-of-fact"

- Evidence and claim linked through cause-and-effect
- Addition of *why*

- Multiple pieces of evidence linked causally to each other and to the claim

NaH will drive the equilibrium towards products because it is the strongest base.

NaH will drive the equilibrium towards products because it is the strongest base. A strong base will drive an equilibrium towards products because...

Relational
Linear-Causal
"drive equilibrium towards products"
"strongest base"
Rationale for **why** a strong base drives an equilibrium to products

Figure 6: Reasoning diagrams for each mode of reasoning (top) and example of how reasoning diagrams were used to make decisions about the reasoning within arguments.

Phase 2b: Levels of granularity to characterize the scalar levels of concepts in students' arguments.

The overall granularity of an argument was dependent on the granularity of the concepts provided. For example, for Q1, an argument discussing only the favoured direction of the equilibrium was considered to be at the reaction level of granularity because it was only describing a reaction-level phenomenon. In contrast, an answer discussing the favoured direction of the equilibrium (phenomenon at level of reaction), the strength of the conjugate acids (molecular property), and the electronegativity of relevant atoms in each base (atomic property) was considered to have concepts at all three levels of granularity, with atomic being the "deepest" level appropriate for this context. Coded in this way, concept units characterized in Phase 1 were used to determine an argument's level of granularity (Table 3). Different questions may prompt the need for conceptual knowledge at different levels of granularity. For the comparing mechanisms question, the atomic level of granularity was not expected of students—we have included it in Table 3 to further illustrate how different questions are associated with differing expectations related to granularity.

Table 3: Levels of granularity and examples of concepts at each level for Q1 and Q2.

Acid–base equilibrium question		**Comparing mechanisms question**	
Level of granularity	**Example**	**Level of granularity**	**Example**
Reaction	Favoured direction of an equilibrium	**Reaction**	Reaction likelihood; Activation energy; rate-determining step; transition state
Molecular	Strength of conjugate acid; stability; pK_a values	**Molecular**	Number of carbocation substituents**;** number of α-carbon substituents; Hyperconjugation; steric hindrance
Atomic	Electronegativity; formal charge	**Atomic**	None provided

Phase 2c: Levels of comparison to characterize how students compare claims in their arguments. We coded each argument as one of three levels of comparison—isolated, partially compared, and compared—based on the degree to which concepts in the argument were used to compare between the possible claims (Table 2). For example, if an argument had concept codes “base strength” and “conjugate acid strength", but both these codes were used to discuss only the chosen claim then this argument was coded as “isolated”. If one (but not both) of these concepts was used to compare to another possible claim (e.g., “NaH is a stronger base than NH_3”), then this was coded as “partially compared”. Lastly, if both concepts were used to compare to another base (*e.g.*, “NaH is a stronger base than NH_3, which means H_2 is a weaker conjugate acid than NH_4^{+}”), then this statement was coded as “fully compared”. Coded in this way, concepts used to compare (*i.e.*, comparison units) that were identified in Phase 1 were used to determine an argument’s level of comparison.

In summary, the unit-based method allowed us to characterize students’ arguments first in terms of their units (links, concepts, comparisons), and use these units to make decisions in terms about an argument’s mode of reasoning, level of granularity, and level of comparison. An example of this process is shown in Figure 7.

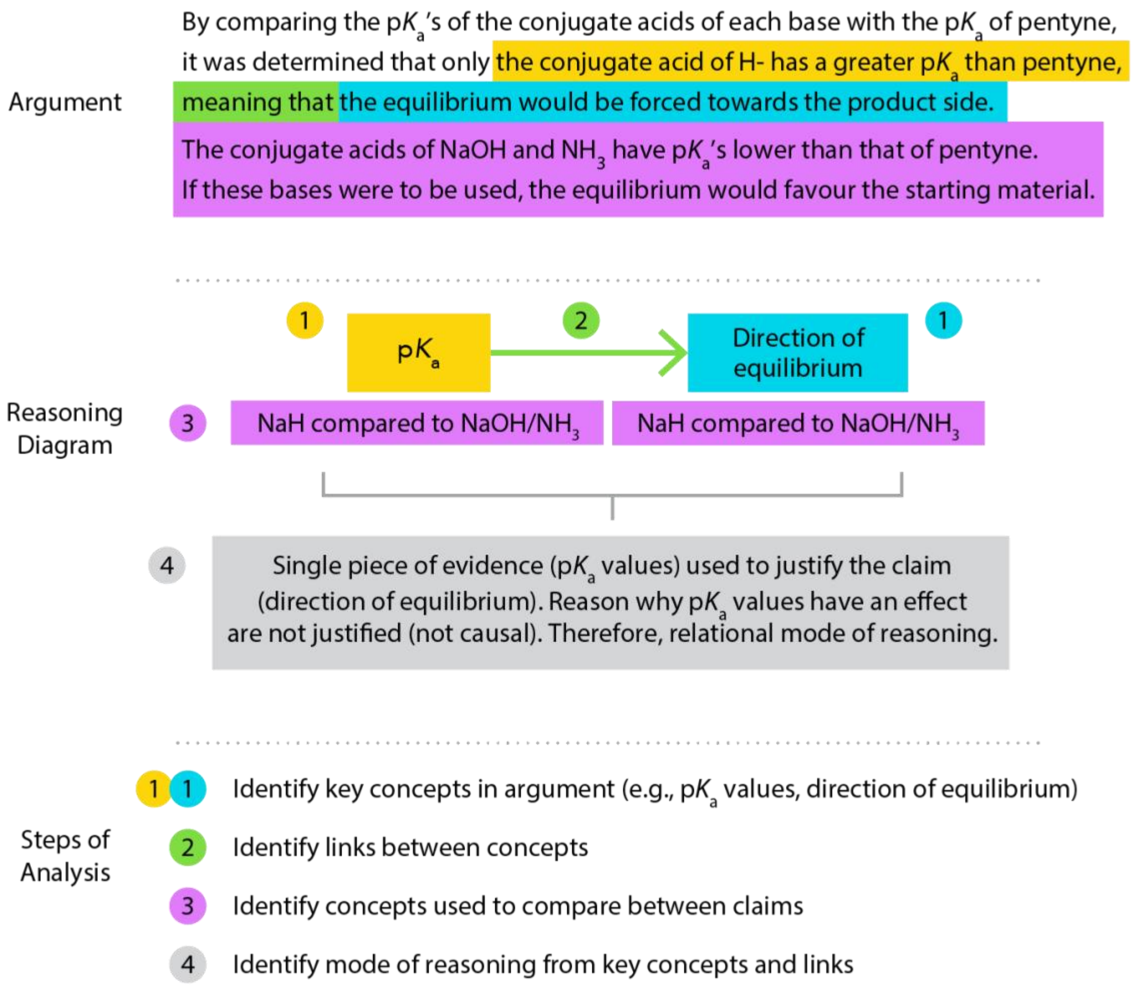

Argument
By comparing the pK_a's of the conjugate acids of each base with the pK_a of pentyne, it was determined that only the conjugate acid of H- has a greater pK_a than pentyne, meaning that the equilibrium would be forced towards the product side.
The conjugate acids of NaOH and NH_3 have pK_a's lower than that of pentyne. If these bases were to be used, the equilibrium would favour the starting material.
Reasoning Diagram
1
pK_a
2
Direction of equilibrium
1
3
NaH compared to NaOH/NH_3
NaH compared to NaOH/NH_3
4
Single piece of evidence (pK_a values) used to justify the claim (direction of equilibrium). Reason why pK_a values have an effect are not justified (not causal). Therefore, relational mode of reasoning.
Steps of Analysis
1 1 Identify key concepts in argument (e.g., pK_a values, direction of equilibrium)
2 Identify links between concepts
3 Identify concepts used to compare between claims
4 Identify mode of reasoning from key concepts and links

Figure 7: Example of analysis of a student argument using the unit–based approach to code students' arguments.

Inter-rater reliability

For each question, a second coder analysed a subset of exams for the units outlined in Table 4 using the method described above to establish inter-rater reliability (Krippendorff, 1970; Hallgren, 2012). Krippendorff's α was used as a statistical measure to evaluate agreement between coders (Krippendorff, 1970); unlike percent agreement, Krippendorff's α accounts for chance agreement between coders. Inter-rater reliability was conducted for Phase 1 codes only (links, concepts, and comparisons), as identification of these codes required analysis of student discourse in their written arguments. Phase 2 codes were not included in the inter-rater process as these codes depended entirely on the Phase 1 codes in the unit-based method. In other words, Phase 1 coding was the only portion of the coding protocol in which interpretation between raters could vary.

For each question, after the primary coder coded the entire set of responses, the second coder used the first iteration of the codebook to code a subset of 20 responses. Both coders then met to discuss differences between their respective analyses, which were mainly a result of overlapping and/or redundant codes and difficulties interpreting the codebook experienced by the second coder. Often, the second coder was unsure what justified a link between concepts being present or not, especially in cases where implicit references to and between units may have been made in a student's argument. Based on these discussions, revisions were made to the codebook, including removal of redundant codes (or combining them with other codes) and refinement of the codebook's criteria for the presence or absence of certain units. This process was repeated until the two coders obtained a Krippendorff's α (Table 4) greater than 0.67, the value described as exceeding the threshold of acceptability for inter-rater reliability (Krippendorff, 1970).

Table 4: Krippendorff α values obtained from inter-rater analysis for units in students' arguments. Acceptable agreement = 0.67.

	Krippendorff's α		
Unit in argument	**Round 1**	**Round 2**	**Round 3**
Concepts	0.58	0.77	0.82
Links between concepts	0.42	0.71	0.86
Concepts used to compare	0.58	0.86	0.95

RQ2: Using a rubric based on the dimensions of the unit-based method, how do arguments' scores compare against a traditional scoring system focused on keywords?

One of our goals was to translate the unit-based method into a rubric for educators. Therefore, we wished to determine how the unit-based method scored students' arguments compared to traditional methods for scoring. Therefore, we compared results from evaluating students' arguments with a rubric based on the unit-based method for Q1 and Q2 against the original scoring system used to evaluate Q1 and Q2 in the OCII course.

To quantify an argument's mode of reasoning, level of granularity, and level of comparison, we scored each dimension out of the number of levels present for each dimension. For example, a descriptive mode of reasoning scored as 1/4, a relational mode of reasoning scored as 2/4, *etc*. Figure 8 shows the scoring breakdown for each dimension. An argument's overall quality was then converted into a percent value adding the scores from each dimension along with the number of expected concepts. We then took these percentage scores and compared them to the percentage scores for the arguments assigned by the original scoring systems. Note that in teaching practice, the values inputted for each dimension of rubric can and likely will vary depending on the goals of the task, course, or educator. For example, certain educators may not place greater value in modes of reasoning than levels of granularity, and choose scoring values to reflect this.

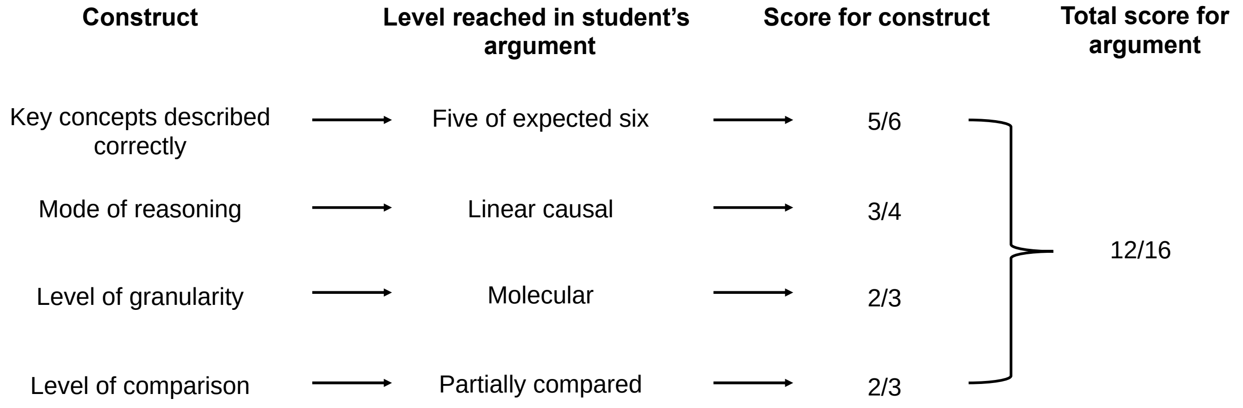


Figure 8: Overview of method to convert each construct analyzed with the unit-based method into a score for the overall argument.

RQ3: What are educators' perspectives of an rubric for characterizing students' arguments based on the unit-based method?

We interviewed educators (*N* = 4) to gather their perspectives on a rubric based on the dimensions of the rubric (key words, reasoning, granularity, and comparisons), with lines of questioning stemming from three broad questions:

(1) How often do educators incorporate argumentation within their courses?
(2) Do educators find the dimensions of the rubric important, if at all?
(3) Would educators use the rubric in their teaching at all?

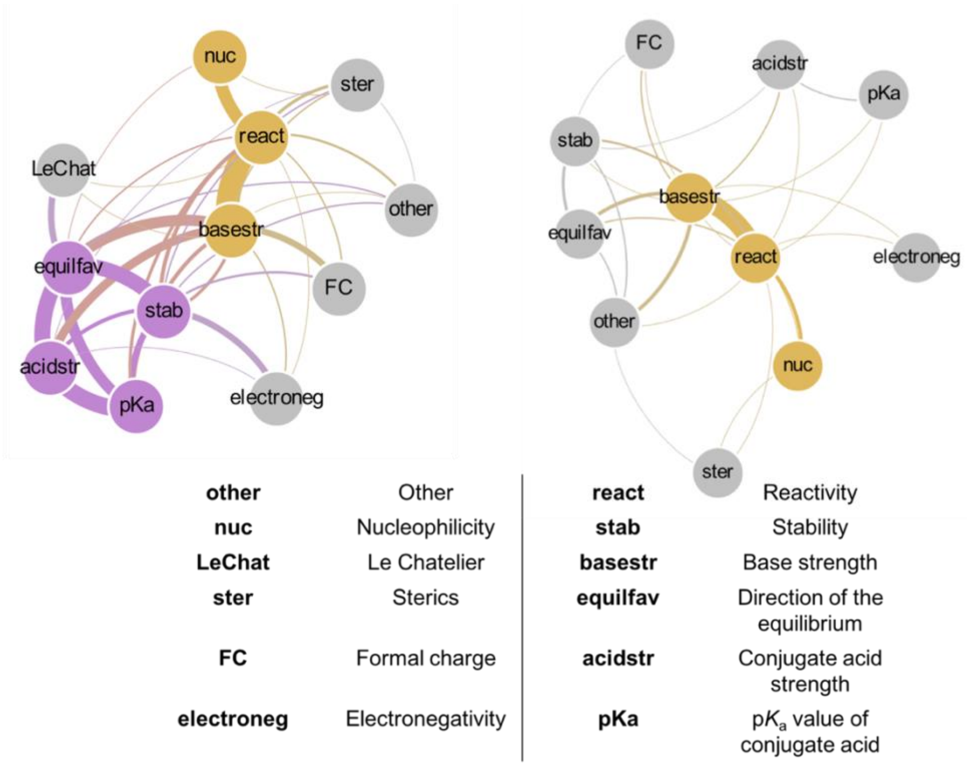
The four educators we interviewed came from the following teaching backgrounds:

- Educator 1: a chemistry graduate student with a degree in education and 6 years of experience teaching high school science and chemistry.
- Educator 2: a professor of education with 30 years of experience teaching a mix of undergraduate and graduate courses in education and environmental science, as well as elementary and high school science.
- Educator 3: professor of chemistry with 25 years of experience teaching a mix of undergraduate and graduate courses in the chemical sciences, with a focus on organic chemistry, biochemistry, and physical organic chemistry.
- Educator 4: a professor of chemistry with 25 years of experience teaching a mix of undergraduate and graduate courses in the chemical sciences, with a focus on introductory chemistry, inorganic chemistry, and biochemistry.

Participants were invited *via* email and ethics approval was granted by the University of Ottawa's Research Ethics Board. Each interview was 25-30 minutes in length and was audio- and video-recorded.

Results and discussion

RQ1: Using a unit-based method for characterizing arguments, how are students constructing arguments on two different organic chemistry questions?



other	Other	**react**	Reactivity
nuc	Nucleophilicity	**stab**	Stability
LeChat	Le Chatelier	**basestr**	Base strength
ster	Sterics	**equilfav**	Direction of the equilibrium
FC	Formal charge	**acidstr**	Conjugate acid strength
electroneg	Electronegativity	**pKa**	pK_a value of conjugate acid

Figure 9: For Q1, connections made between concepts made for correct claims (left) and incorrect claims (right).

How are links made between concepts in students' arguments and what modes of reasoning do students exhibit in their arguments? We used the data visualization software, Gephi, to visualize links made between concepts in students' arguments for Q1. Links in Q2 arguments were previously identified in Bodé *et al*. (2019). In a Gephi visualization, nodes represent concepts; edges (*i.e.*, a line between two nodes) represent links between two concepts. The frequency of links between two concepts was characterized as the thickness of the edge. In other words, two nodes connected by a relatively thick edge represents two concepts that were linked to each other in many arguments, relative to other connections made within the dataset. In contrast, a node with no edges represents a concept that had no links to other concepts in the dataset. To improve clarity, placement of nodes was manually manipulated within Gephi to overlapping nodes and edges.

From our analysis with Gephi (Figure 9), the three concepts most prevalent in correct claims—the favourability of the equilibrium, conjugate acid strength, and pK_a of the conjugate acids—were also the three concepts which exhibited the most frequent connections. Often, arguments for correct claims included a triad of concepts and links that included stating the respective pK_a values of the conjugate acids for the given bases, relating these pK_a values to rank the relative strengths of the conjugate acids, and then using these rankings to justify the extent to which an equilibrium involving each base/conjugate acid would favour a particular direction. For example, Student 116 provided the following argument which included this triad:

Student 116: "I chose NaH as a base because its conjugate acid has a ***pK_a value of around 36****, which makes it* ***a weaker acid*** *than the starting material. The* ***equilibrium will favour the side with the weaker acid****. I did not choose NaOH or NH_3 because their respective conjugate acids would have a* ***pK_a value less than that of the SM [starting material], meaning that the equilibria would favour the starting materials (pK_a ~ 15.7 for H_2O and ~10 for NH_4^+).****"*

In some cases, this type of argument was expanded to include the concept base strength. This included identifying the relationship between the relative strengths of the conjugate acids from the relative strengths of the bases, and then using these ideas in concert to determine the direction of the equilibrium.

The principle connection made in incorrect claims was between base strength and reactivity. The "reactivity" code was present when students' responses described how a base would or might react. In these cases, students often used base strength as the principle concept to justify how their chosen base (or all three bases) would react with the alkyne or the acyl chloride. For example, Student 43 provided the following argument which linked base strength to reactivity but did not discuss further why one base was strong/weak.

Student 43: "[NaOH is] a ***strong base*** *to* ***remove the hydrogen from the alkyl chain****, whereas the other bases are* ***weaker*** *and need more activation energy* ***to remove the hydrogen****.*

Students who invoked a "base strength" and "reactivity" link may have done so in a rote fashion; that is, these students may have memorized a relationship between strength of a base and its reactivity and then used this relationship as the basis for their argument. Though this connection was present in incorrect claims, it was also prevalent in correct claims; however, base strength was more frequently linked to other relevant concepts in correct claims, such as conjugate acid strength. These findings suggest that when using base strength as a concept in their arguments, arguments for correct claims only included the base strength concept when they recognized how to justify *how* and *why* base strength was associated with other concepts relevant to the question.

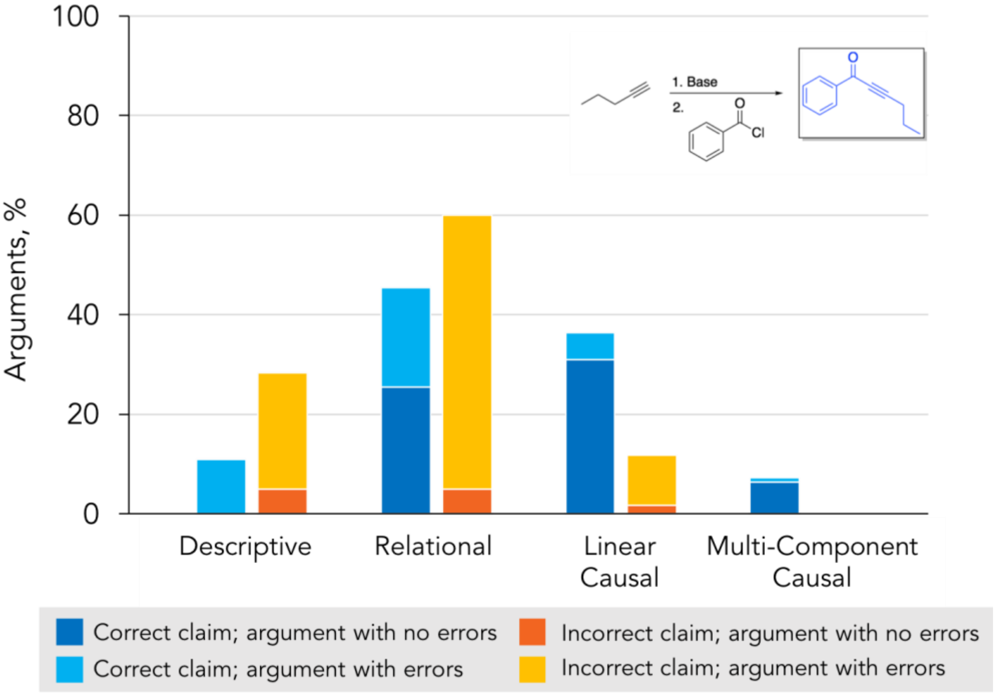


Figure 10: Modes of reasoning for students' arguments to Q1. Students' who were arguing for correct claims were more likely to exhibit causal modes of reasoning.

For Q1, the majority of students (62%) provided the correct claim (*i.e.*, chose the correct base) for which base would drive the equilibrium in question to products (Figure 10). However causal reasoning (either linear causal or multi-component causal) were present in only 31% of all answers. Correct claims more frequently exhibited causal arguments (linear causal and multi-component causal) than incorrect claims, while incorrect claims more frequently exhibited descriptive arguments than correct claims. The frequency of causal arguments was significantly different between arguments with correct and incorrect claims, $\chi^2(1, N = 170) = 18.1$, $p < 0.001$ with a medium effect size, $\phi = 0.33$.

For Q1, relational arguments were the most prevalent across all student answers (48% of all answers). The most common relational argument discussed how the chosen base in the claim was a strong base, allowing it to drive the equilibrium towards products. Other relational arguments included discussions of one of conjugate acid strength or pK_a values in place of base strength. These arguments did not involve discussions of *why* acid strength or pK_a values would affect the formation of product. In comparison, a common linear causal argument discussed how the equilibrium would favour the products due to differences in pK_a values and would then explain why these pK_a values were relevant by linking this idea to the relative strengths of the conjugate acids. For example, the first of Student 19's argument linked the direction of the equilibrium to conjugate acid strength, and justifying this link with pK_a values:

Student 19: "The equilibrium of the first step is dependent on the acid–base reaction and as a result, it is ***dependent on which side does [sic] the stronger acid lie****. Based on the structure of the reactant,* ***the more acidic proton is at the terminal alkyne (pK$_a$ 50 [C-H sp^3] vs 24 [C-H sp]), so the appropriate base must have a weaker conjugate acid."***

Note that although this argument is linear causal, it is a linear causal argument that exhibits a molecular level of granularity. Though Student 19 *does* articulate what they believe to be a reason for why one base is stronger than the other by referencing relevant pK_a values, their argument does not get to the granular levels necessary to describe the true causal reasons (*e.g.*, chemical properties) responsible for the acid strengths. The latter portion of Student 19's argument reaches atomic level of granularity by introducing electronegativity into their argument:

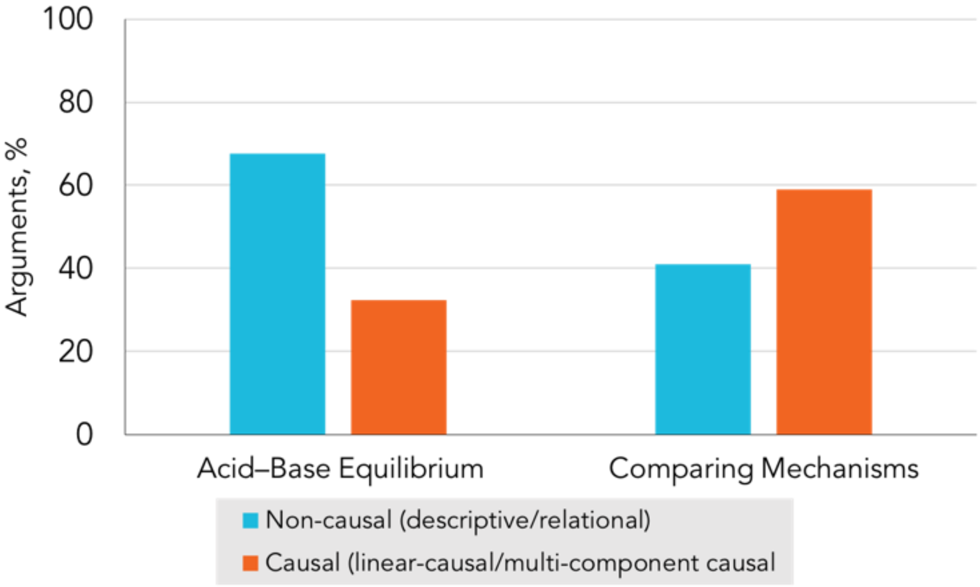


Figure 11: Causal vs non-causal modes of reasoning for the acid–base equilibrium (Q1, *n* = 170) and comparing mechanisms (Q2, *n* = 122) questions.

Student 19: Based on the electronegativity of OH and NH_3, they would serve as better bases than the alkyne as the greater electronegativity of O and N allowing the ionized forms to better stabilize a negative charge (for O, making the $^{-}$OH a more stable base than the ionized alkyne) and less able to stabilize a positive charge (for N, NH_4^+ (CA for NH_3) is more acidic than alkynes and hence, shifts equilibrium to the alkyne). As for NaH, the similar values in electronegativity between H and C would not influence the equilibrium as much as NaOH and NH_3. Also, the reaction results in the production of H_2, which is very stable and hence, the H_2 is less likely to protonate the alkyne, favouring the product side.

Multi-component causal arguments were only present in answers that provided correct claims. The most common multi-component causal arguments discussed how two or more concepts (*e.g.*, pK_a values, relative base stability, conjugate acid strength) influenced the direction of the equilibrium, each with reference to another concept to explain why these concepts were relevant in this discussion.

We compared the respective frequencies of causal and non-causal arguments between the two question types to determine how the question type might impact students' reasoning (Figure 11). Arguments for Q2 had significantly more causal (linear and multi-component) arguments than the acid–base question, which exhibited more non-causal arguments (descriptive and relational), with a medium effect size, $\chi^2(1, N = 292) = 20.456$, $p < 0.001$, $\phi = 0.27$, with a statistical power of $1–\beta = 0.99$. These findings suggest that students' reasoning can depend on the question's content and prompt (Kelly *et al.*, 1998; Sadler, 2004; Sadler and Zeidler, 2005; von Aufschnaiter *et al.*, 2008; Barwell, 2018; Cian, 2020). Therefore, if educators expect their students to exhibit a specific mode of reasoning for a particular question type, they should be explicit about these expectations throughout the course, aligning those expectations with practice, feedback, and other assessment (e.g., exams).

What concepts are discussed in students' arguments and with what levels of granularity? In Phase 1 of the unit-based method, we identified the concept units in students' arguments for Q1. Concept units for Q2 arguments were previously identified in Bodé *et al.* (2019). For Q1, we found differences in the concepts discussed by students depending on whether they provided a correct or incorrect claim (Figure 12). For example, arguments with correct claims more frequently discussed the concepts defined as favourability of the equilibrium, the strength of the conjugate acid, and the pK_a value of the conjugate acid. In contrast, arguments for incorrect claims discussed all of these concepts to a less frequently. In the context of Q1 and the OCII course, all three of these concepts were relevant to the claim and were key concepts employed in the expected answer for this question.

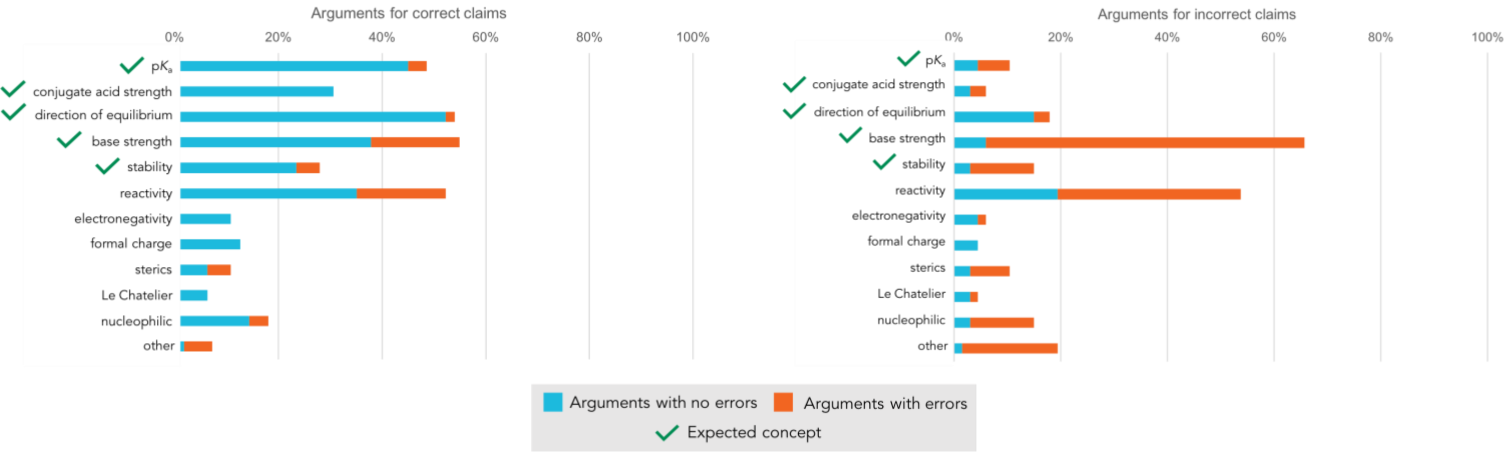


Figure 12: For Q1, concepts discussed in arguments for correct claims (n = 110, left) and incorrect claims (n = 60, right).

For incorrect claims, the two most frequently discussed concepts were base strength and reactivity. For example, Student 10 provided the following argument which used base strength to justify reactivity:

*Student 10: "NaH is **the strong base choice** therefore it is most likely to **deprotonate the carbon**."*

Though base strength was also discussed frequently in correct claims, it was discussed more frequently in connection other concepts (e.g., pK_a value, conjugate acid strength). That is, students who mentioned base strength in their arguments often linked it to other concepts relevant to the question. Furthermore, despite base strength being the most prevalent concept discussed in incorrect claims, the majority of arguments for incorrect claims discussed base strength incorrectly. This was found to be reflective of a broader trend, as correct claims were more frequently justified with concepts that were discussed correctly compared to incorrect claims, which were more frequently justified with concepts that were discussed incorrectly. This finding speaks to the need for students to be able to both identify the relevant concepts for a given task and discuss these concepts appropriately.

Next, we determined the levels of granularity in students' arguments (Figure 13) using the concept units identified for Q1 and Q2. Given that each level of granularity had different numbers of possible concepts (three atomic, four molecular, four reaction for Q1), we normalized the different numbers of concepts that could be described at each level of granularity by dividing the frequency in students' answers for each level of granularity by the number of concepts possible at each level. For example, of the 170 arguments analysed, 52 of the concepts in students' arguments were at the atomic level and 212 were at the molecular level; to correct for the influence of there being 4 possible molecular-level concepts and 3 atomic-level concepts, we divided 52 by 3 and 212 by 4. This adjustment was done to ensure that if, for example, we observed more molecular-level concepts than atomic-level concepts, that this observation was not skewed by their simply being more possible molecular-level concepts than atomic-level concepts.

For Q1, the majority of concepts (N = 500 concepts) proposed by students in their arguments were at the reaction and molecular levels of granularity (Figure 13). Of all concepts presented in students' arguments, 13% were atomic-level concepts. We also conducted similar analyses on correct claims and incorrect claims individually and found similar relative distributions: 14% atomic-level concepts out of all concepts in correct claims (n = 361) and 13% atomic-level concepts out of all concepts in incorrect claims (n = 139).

We then compared the levels of granularity Q1 and Q2. Because Q1 and Q2 assessed different conceptual knowledge and required different levels of granularity, our comparisons of granularity expressed in students' arguments for the two questions are qualitative.

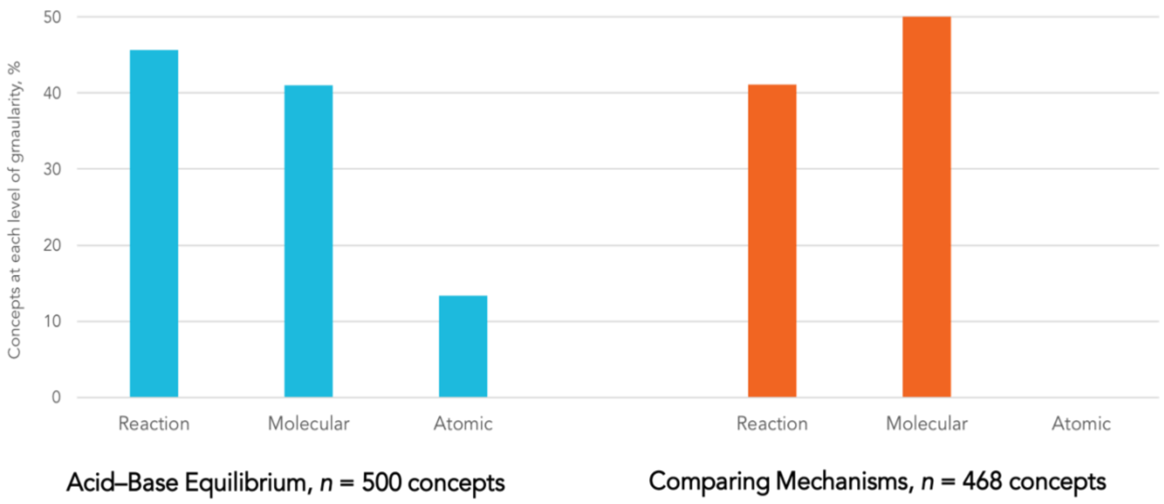


Figure 13: The proportion of concepts exhibited at each level of granularity for both the acid–base question (n = 500) and the comparing mechanisms question (n = 468). Descriptions for each level of granularity are described in Table 5.

For Q1, students produced arguments primarily at a molecular-level of granularity (e.g., arguments focused on pK_a values, conjugate acid strength), though some students included concepts in their arguments at a more granular, atomic-level (e.g., electronegativity, formal charge) (Figure 13). For Q2, the majority of arguments were at the "molecular" level, which included concepts such as the number of α-carbon substituents and number of carbocation substituents. As was the case for the modes of reasoning, this finding reinforces the idea that how one reasons in their argument is dependent on the task at hand; educators expect their students to be able argue with concepts at a specific level of granularity, these expectations need to be made clear both throughout the course and on the assessment task.

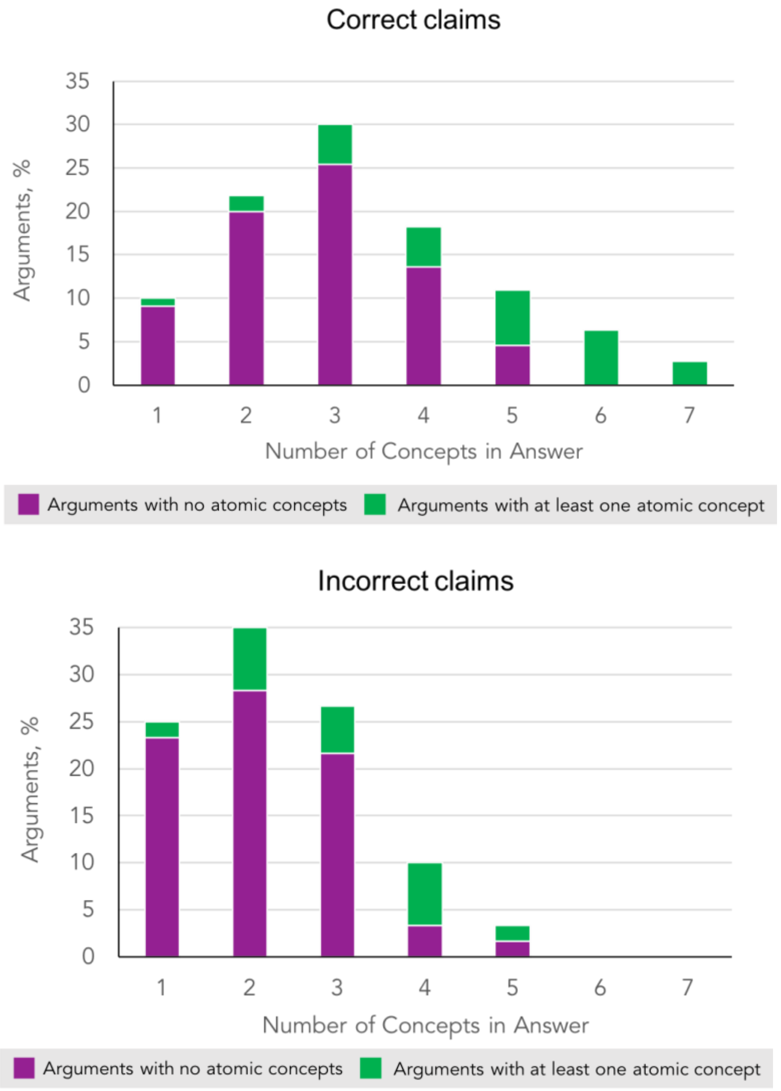


Figure 14: For Q1, the proportion of arguments with/without atomic concepts against the number of concepts in those arguments. Students rarely exclusively discussed atomic concepts; rather, atomic concepts were discussed alongside molecular and reaction concepts in 95% of arguments with atomic concepts.

Lastly, for Q1, we sought to investigate how atomic-level concepts arose in students' arguments. In other words, were students describing atomic-level concepts immediately in their arguments, or did they only describe atomic-level concepts when molecular- or reaction-level concepts were present? We found that 95% of answers with atomic-level concepts also included discussion of one or more reaction- and/or molecular-level concepts (Figure 14). That is, the majority of answers that provided atomic-level concepts did so alongside concepts of molecular or reaction-level granularity —few answers provided atomic-level concepts outright or in the absence of the other levels of granularity. This finding may suggest that students who provide concepts at atomic-level granularity do so only when they recognize the relevance of these concepts to reaction- and/or molecular-level concepts.

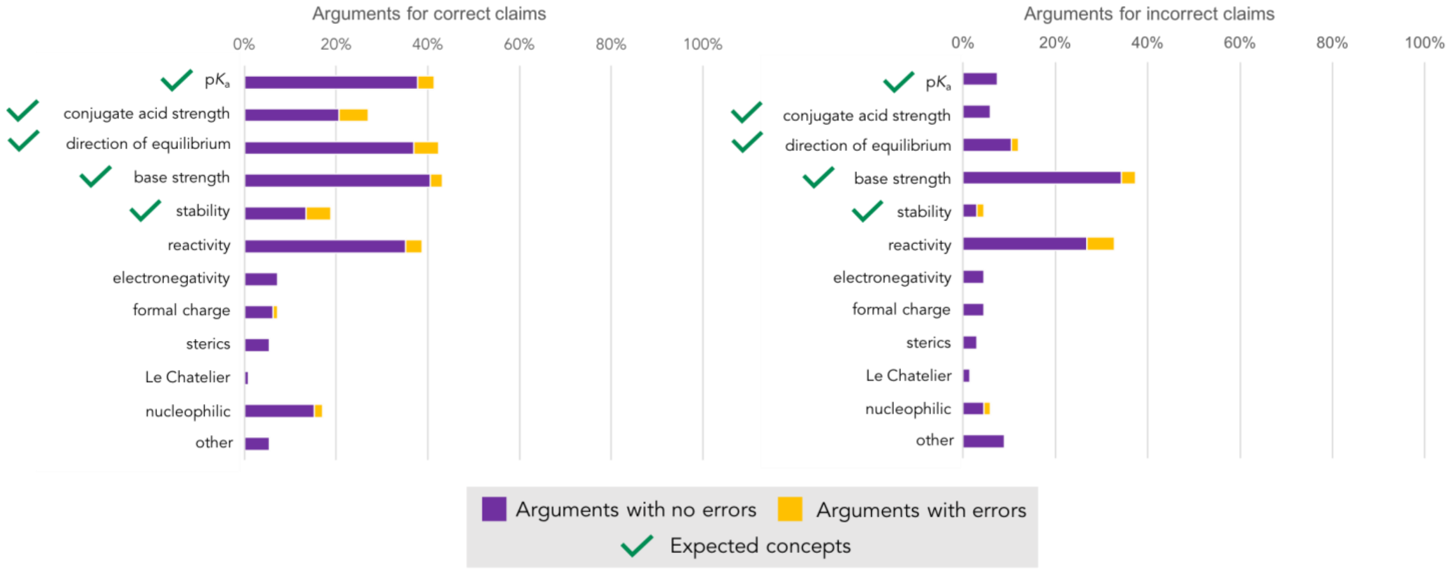


Figure 15: For Q1, how often each concept was used to compare between claims in arguments for both correct (left, *n* = 110) and incorrect (bottom, *n* = 60) claims.

How are concepts compared between possible claims in students' arguments? Figure 15 shows how often a given concept was used in a comparison between claims. Correct claims primarily compared between claims during discussions of pK_a values, conjugate acid strength, and the favourability of the equilibrium. For example, Student 14 listed the pK_a values for all three conjugate acids, compared the relative strength of the acids based on these values, and then described which direction the equilibrium would favour in each case:

Student 14: "I chose NaH as the base because its conjugate acid has a *higher pK_a value than the alkyne. That means that the conjugate acid is a weak acid, weaker than the alkyne, so the reaction will favour the products. I did not choose NaOH or NH_3 because their conjugate acids has [sic] smaller values* ***than the alkyne,*** *driving the equilibrium towards the starting materials.*

In contrast, incorrect claims primarily compared between claims using base strength and reactivity. A common example, as shown in the following excerpt from Student 55's argument, was a student stating that one base was stronger than the other two bases, leading them to conclude that the stronger base would be able to react as a base with the alkyne. Comparisons of relative stability were also present, with some students stating that the other two bases would either not react or produce undesired reactions due to their being weaker bases, such was the case with Student 55:

Student 55: NaH will take the H of the bonding end of the triple bond to make H_2(g). ***NaH is a much stronger base than NaOH and NH_3.*** *NaOH and NH_3 are* ***too weak to deprotonate the alkyne***. ***NH_3 would break the triple bond and add NH_2 to the end of the triple bond. NaOH wouldn't react at all.*** *NaH when a solution has H^+ floating around, which are extremely reactive.*

For Q1, arguments for correct claims were found to more frequently compare concepts in one claim to other possible claims than arguments for incorrect claims, $\chi 2(1, N = 170) = 11.2$, $p = 0.001$, $\phi = 0.257$ (Figure 16). Students who provided correct claims were more likely to compare and contrast between claims, while students who provided incorrect claims were more likely to discuss their claim in isolation of the other possible claims.

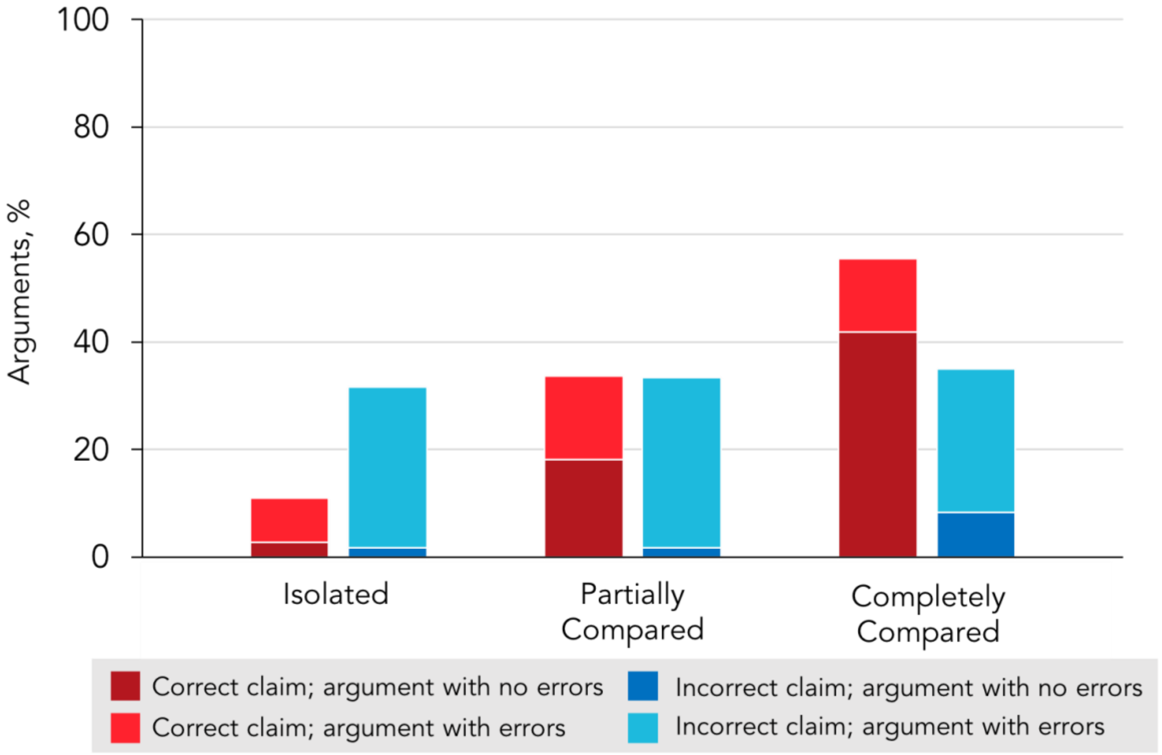


Figure 16: Levels of comparison for Q1. Students who provided correct claims (*n* = 110) were more likely to compare and contrast between claims, while students who provided incorrect claims (*n* = 60) were more likely to discuss their claim in isolation of the other possible claims.

We also investigated how students compared between claims in Q1 versus Q2 (Figure 17). Students more frequently compared (either partially or fully) on Q2 than Q1, $\chi^2(1, N = 329) = 10.748$, $p = 0.001$, $\phi = 0.18$. Additionally, when investigating the relative frequencies of partial versus full comparisons, we found that students more frequently made full comparisons on the mechanisms question than the acid–base question, $\chi^2(1, N = 329) = 36.170$, $p < 0.001$, $\phi = 0.354$. It may be that there being two possible claims for Q2 more effectively cued students to compare between claims than if there were three possible claims, such as in Q1.

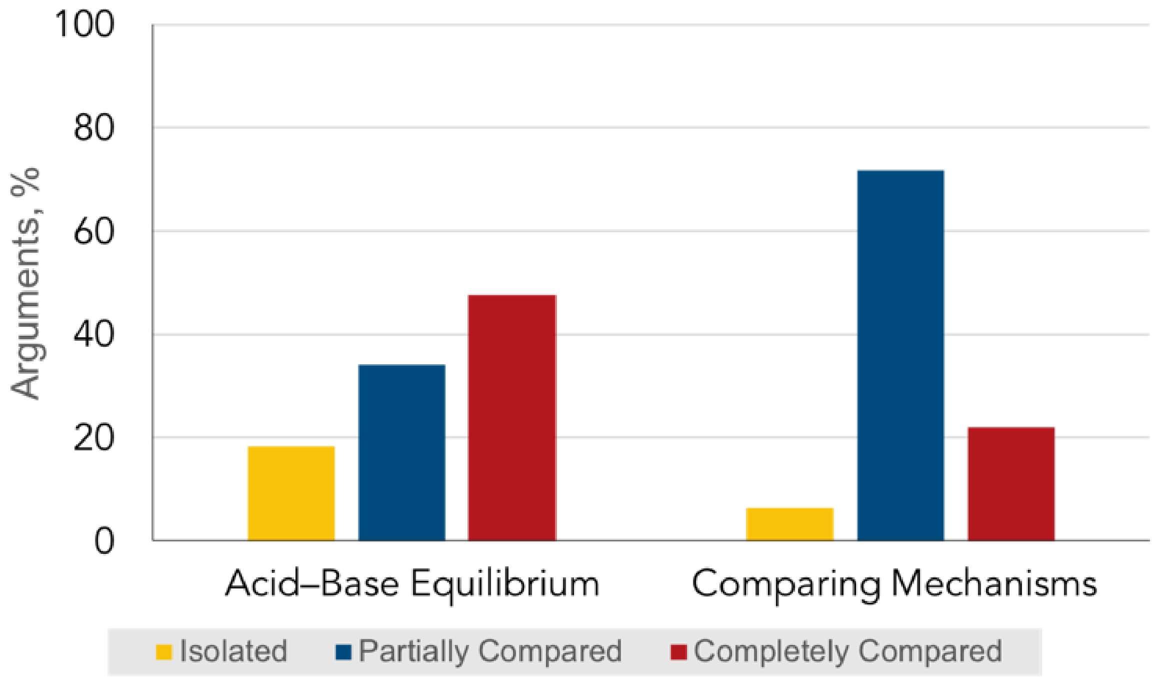


Figure 17: Comparison of the levels of comparison for Q1 (acid–base equilibrium, *n* = 170) and Q2 (comparing mechanisms, *n* = 122).

RQ2: Using a rubric based on the dimensions of the unit-based method, how do arguments' scores compare against a traditional scoring system focused on keywords?

We compared a rubric based on the unit-based method described in this work against a traditional scoring system to determine the impact that using the former could have on evaluation in instructional contexts. In their original exam contexts, Q1 and Q2 were evaluated primarily on the presence of key words/concepts and did not explicitly evaluate the domain-general characteristics of students' arguments (reasoning, granularity, and comparisons).

We evaluated students' arguments for Q1 and Q2 using the unit-based method, and then used the number of expected concepts used, the mode of reasoning, level of granularity, and level of comparison to determine a numerical score. From Figure 18, we found that using the unit-based method to score students' responses for both questions was associated with higher scores for students' responses compared to scoring with the traditional methods. Our findings suggest that the unit-based method captures structural aspects of students' arguments that are not explicitly captured by the traditional methods.

To more explicitly compare students' abilities to construct arguments *structurally* (in terms of the domain-general characteristics of reasoning, granularity, and comparisons) versus their abilities to include the necessary *conceptual* units (the domain-specific characteristics, such as key words/concepts), we independently scored students' arguments in terms of structure and content and compared these two broad dimensions. For Q1 and

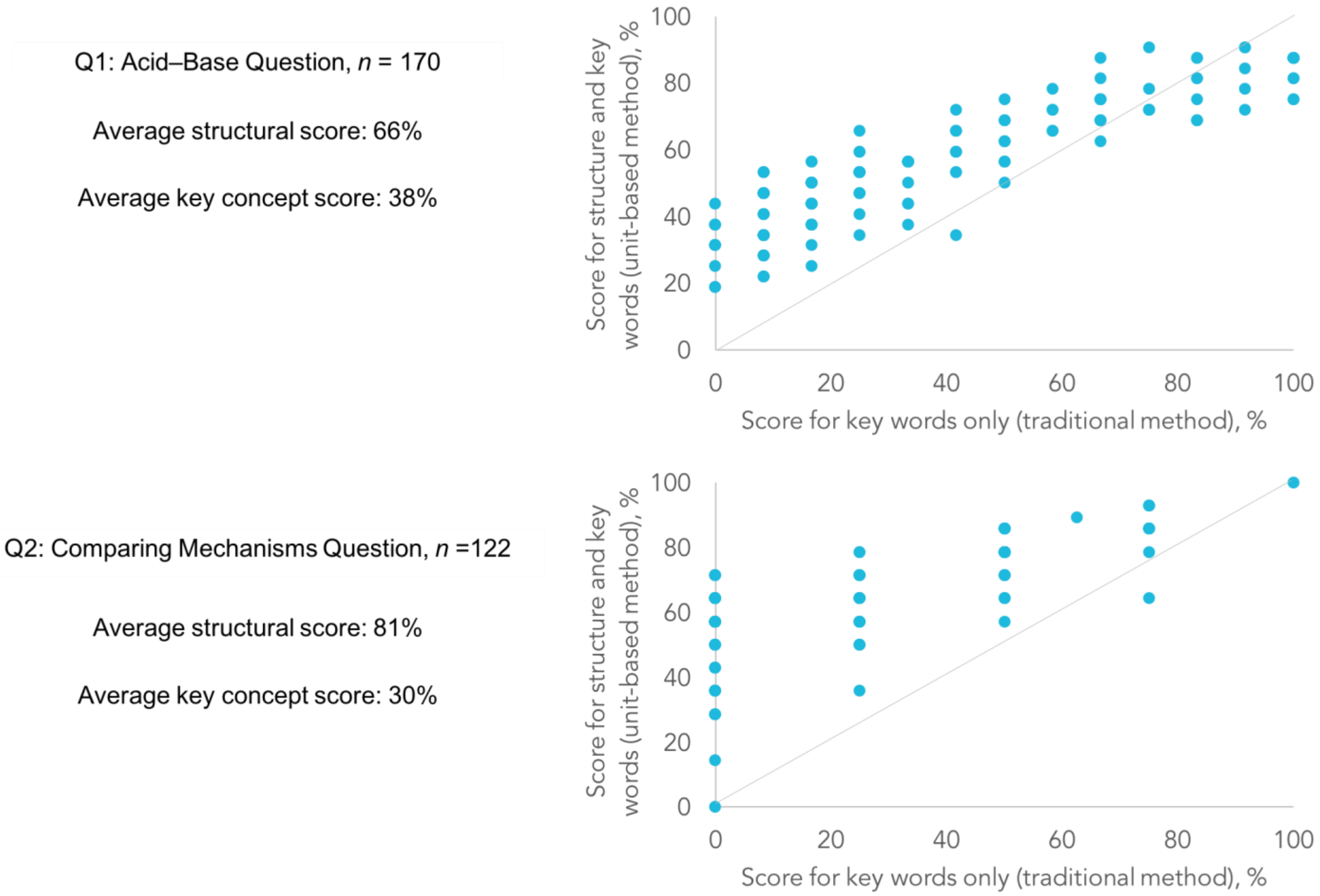


Figure 18: Comparing scores for students' arguments evaluated using the unit-based method vs. traditional method focused on key words only.

Q2, students' arguments had average structural scores of 66% and 81%, respectively, and average key concept scores of 38% and 30%, respectively. These findings suggest that, in these contexts, although students may not have provided conceptually correct content in their arguments, the arguments provided were structurally sound and logical with repsect to our domain-general characteristics of reasoning, granularity, and comparisons.

Importantly, readers should note that we weighed all three domain-general dimensions (reasoning, granularity, comparison) equally when calculating the quality of one's argument using the unit-based method. Educators may choose to weigh each dimension differently, depending on the goal of the assessment task and/or instructional context. Furthermore, our criteria for scoring using the unit-based method set the highest modes of reasoning (multi-component for Q1; linear causal for Q2), the deepest levels of granularity (atomic for Q1; molecular for Q2), and the highest level of comparison (completely compared) as the upper limits (*i.e.*, the maximum possible score) for each dimension. Again, in instructional practice, the upper limit will depend on the context; it may be that on a given question in a specific instructional context, students will be expected, for example, to provide only a relational mode of reasoning, molecular level of granularity, *etc*. Figure 19 presents an example of the impacts of varying the expectations for each dimension depending on the instructional context. In

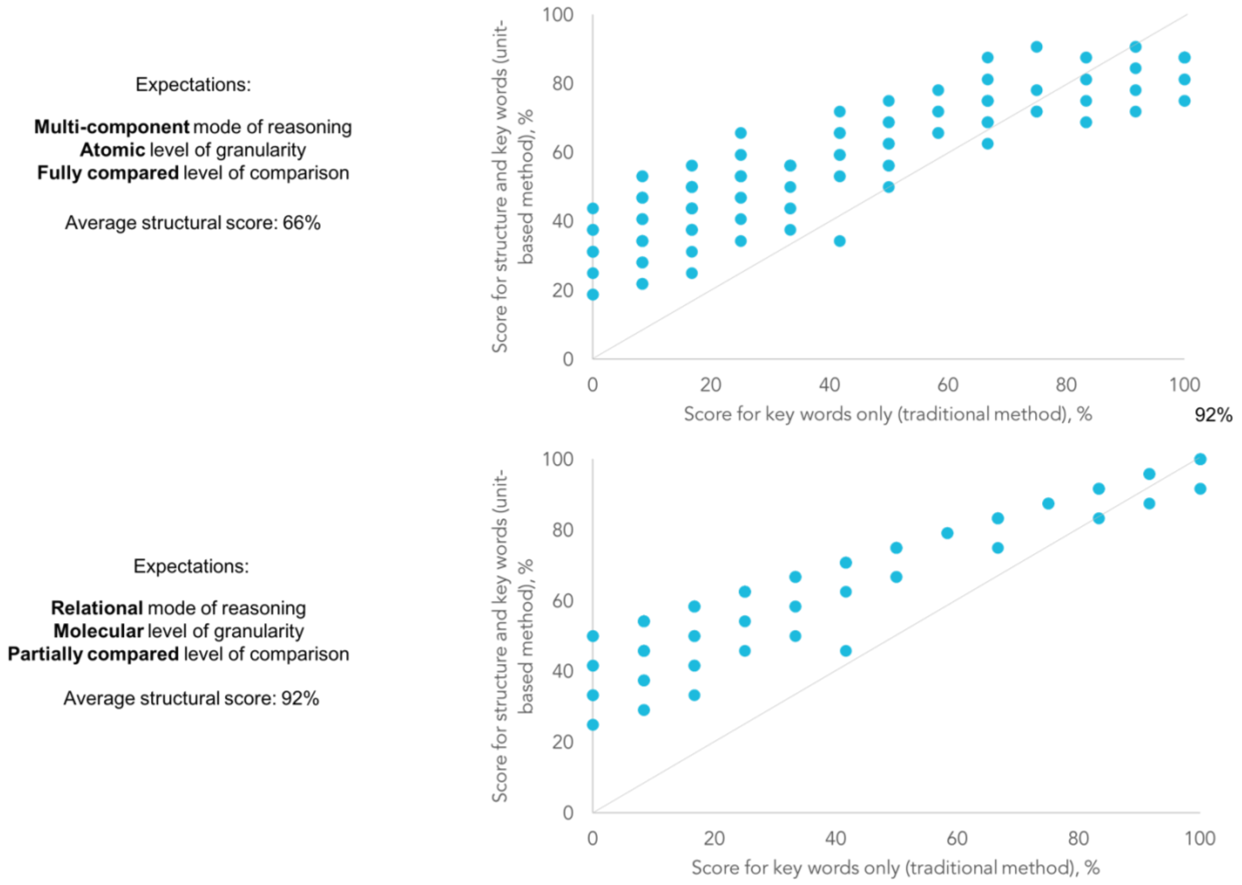


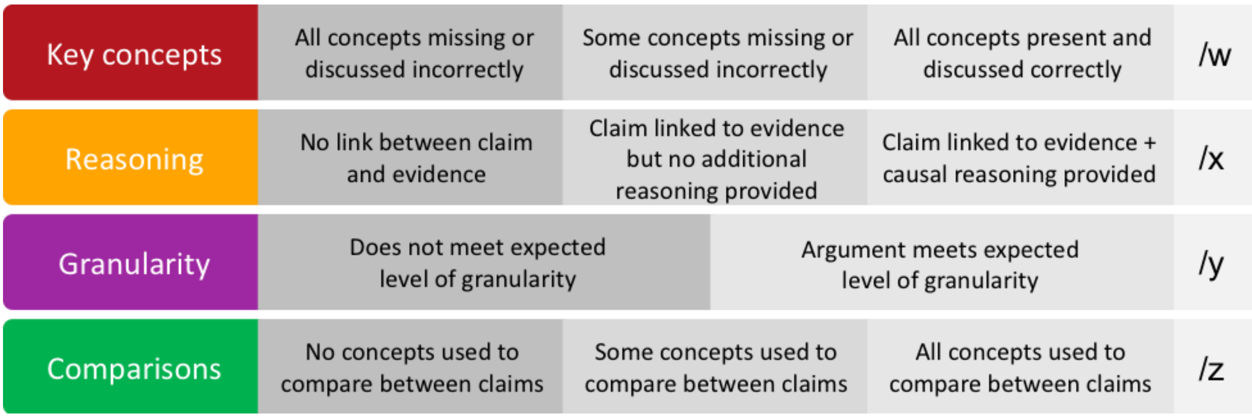
Figure 19: Impact of different structural expectations on scoring for students' arguments.

this example, when expectations for Q1 shift from multi-component to relational, atomic to molecular, and fully to partially compared, scores for students' arguments shift upwards as a reflection of these different expectations.

Lastly, we do not imply that traditional scoring systems such as those used to evaluate Q1 and Q2 are not "lesser" than the rubric we have described in this work. It is likely that in certain contexts, key words/concepts are all that is expected of students, which makes evaluation methods focused on key words/concepts completely appropriate. We simply present the rubric described here as a means to support educators interested in incorporating its associated dimensions into their teaching.

RQ3: What are educator's perspectives of a rubric for characterizing students' arguments based on the dimensions of the unit-based method?

We interviewed educators (*N* = 4) to gather their perspectives on a rubric based on the dimensions of the rubric (key words, reasoning, granularity, and comparisons). The goal of these interviews was to examine how educators might respond to the rubric, and to solicit their feedback on the rubric so we could refine it for future use. We have framed the results from these interviews in terms of the overarching questions that guided our interview protocol.



Key concepts	All concepts missing or discussed incorrectly	Some concepts missing or discussed incorrectly	All concepts present and discussed correctly	/w
Reasoning	No link between claim and evidence	Claim linked to evidence but no additional reasoning provided	Claim linked to evidence + causal reasoning provided	/x
Granularity	Does not meet expected level of granularity		Argument meets expected level of granularity	/y
Comparisons	No concepts used to compare between claims	Some concepts used to compare between claims	All concepts used to compare between claims	/z

Figure 20: Rubric to evaluate arguments in terms of key concepts, reasoning, granularity, and comparisons.

How often do educators incorporate argumentation within their courses?

Educator 1 said that they never explicitly ask students to construct arguments, but that they often ask students to provide justifications for claims for short- and long-answer written questions. When asked how they would evaluate these responses, Educator 1 described how they often follow a pre-determined rubric or marking scheme, such as the Achievement Chart, which focuses on four dimensions: knowledge and understanding, thinking, communication, and application (Ontario Ministry of Education, 2004).

When asked if they ever asked students to construct arguments in their courses, Educator 2 said they never explicitly ask, but certainly incorporate questions that ask students to justify their claims with evidence and reasoning and/or through examples. Educator 2 said that when evaluating arguments, they did often do not use a rubric. Instead, they said they simply determine if they think the argument provided by the student was thoughtful, thorough, and supports the student's claim.

Educator 3 stated that although they never explicitly describe what an argument was to their students nor did they explicitly ask students to "construct arguments", they frequently ask students to justify their claims on assessments. When asked how they evaluate students' responses to these questions, Educator 3 said they focus mainly on "key words" and that they "read the argument and decide whether [the student] has cited the appropriate evidence or concepts required to fully to justify their response."

Educator 4 said that they heavily emphasize argumentation in their courses by building in opportunities to provide justify their claims using observed data provided to or collected by students. Similar to the other interviewees, Educator 4 stated that their use of argumentation was "informal" in that they never explicitly referred to these activities as "argumentation activities" in the learning outcomes for courses nor when communicating these tasks to students. Educator 4 said that they generate an expected response for the question based on relevant course learning outcomes, and then evaluate students' responses based on the claim and evidence provided, as well as how components pieces are connected.

In summary, all four of the participants stated that their assessments often included tasks asking students to construct arguments to justify their claims. However, all four interviewees stated that they never thought of these tasks as "argumentation activities". In terms of evaluation, interviewees seemed to rely on a mix of approaches, with some taking more unit-based approaches (Educators 3 and 4) and others taking more holistic and rubric-based approaches (Educators 1 and 2).

Do educators find the dimensions of the rubric important, if at all?

All four educators believed that all dimensions of the rubric were important. Educator 1 stated that providing students with a framework that considered and was explicit about different levels of granularity would help students "see the bigger picture" when learning about chemistry and science. Educator 1 also stated that the comparison dimension would be valuable for students and educators in that it would allow both groups to consider and argue for or against alternative claims and opinion—a skill Educator 1 believed to be key for practicing scientists.

Educator 2 noted that although they believed the dimensions of the rubric were important, they also noted the importance of understanding the context in which the rubric might be used. They stated: "In certain instances, it might be sufficient to simply provide evidence, and an educator might consider reasoning to be implicit." When told that the rubric would be flexible in its relative weightings for each dimension (*i.e.*, educators are welcome to assign as little or as many points as they like to each dimension depending on the context, task, values, *etc*.), Educator 2 stated that they really enjoyed this aspect of the rubric and believed that this freedom was an advantage that would for the rubric to be used in various contexts as (1) a way to guide student and educator thinking about arguments and (2) as a way to guide evaluation of students' arguments. All four educators expressed similar sentiments regarding the flexibility of the rubric, and also added that the broad definitions provided were beneficial in that an educator could adapt them to be more specific for their course context (for example, the levels of granularity relevant to a course or topic). Educators 2, 3, and 4 all stated that the dimensions of the framework served as a great "starting point" for both educators and students to learn about the value of argumentation and what goes into a "good argument".

Would educators use the rubric in their teaching at all?

All four educators said that they would use the rubric their courses both as a way to communicate their expectations about arguments to their students. For example, Educator 2 said would give the rubric to their students early on in a course or before a major assessment to ensure that their expectations about arguments were being clearly communicated.

There was a mix of beliefs about whether educators would use the rubric as a stand-alone form to evaluate students' responses. When asked to describe what they believed to be the worst aspects of the rubric, Educators 1 and 2 expressed concerns about their own and their colleagues' abilities to consistently use the rubric in evaluating students' arguments. Moreover, Educator 1 also noted that as a graduate teaching assistant, it would be difficult for them to engage in a unit-based approach when grading multiple student reports; therefore, an easy-to-use rubric would be advantageous for their context. Educators 3 and 4, the chemistry professors in the sample, both stated that it would be more practical for educators to have more specific rubrics for given questions and course topics. Educator 3 was primarily concerned with the time required to apply the rubric in evaluation; though they liked the rigour of the rubric, they believed it would be difficult to convince a team of teaching assistants to be as thorough as possible when using the rubric to evaluate a large number of student responses. Therefore, Educator 3 suggested that the rubric should be used as a general framework for educators to create their own context-specific rubrics. These specified rubrics could then be given to teaching assistants for grading. These sentiments were shared by Educator 4, who said that the rubric was a "great starting point for instructors to build off of to develop more context-specific assessments, and a great way to communicate expectations to students." Educator 4 also stated that it would be helpful to create a resource with examples of (1) how to use the general rubric to generate more context-specific rubrics and (2) how to apply a context-specific rubric to evaluate student responses.

In response to these comments, in Appendix D we provide examples of how the rubric might be specified for certain disciplinary contexts, as well as how one might apply these specified rubrics when evaluating student responses. Currently, the examples provided only describe how the rubric might be adapted for organic chemistry tasks. We invite educators and/or researchers from a variety of disciplines to adapt the rubric for tasks relevant to their own teaching such that we might provide the teaching community with a rich and diverse database of items and associated rubrics.

Conclusions

In this work, we demonstrate a unit-based method for characterizing students' arguments on chemistry assessments and created a rubric that educators can use and adapt to their own contexts. The method determines an argument's mode of reasoning, level of granularity, and level of comparison based on the presence and organization of units (links, concepts, and comparisons) within the argument. We also applied this method to evaluate students' arguments on two items from an organic chemistry final exam: (Q1) justifying which base would drive an equilibrium towards products and (Q2) justifying which of two similar reaction mechanisms is more plausible. We also compared scores generated by evaluating with rubric based on the unit-based methods versus scores from evaluating with traditional scoring systems focused on key words. Lastly, we conducted interviews with educators to gather their perspectives on potential directions, uses, and refinements for this rubric.

We found that correct arguments in Q1 (acid–base) exhibited causal reasoning and comparisons between claims more frequently than incorrect claims (RQ1a). Furthermore, correct and incorrect claims focused on different sets of concepts; correct claims more frequently discussed pK_a, conjugate acid strength, and the direction of equilibria, while incorrect claims more frequently discussed base strength and reactivity (RQ1b). Both claim types exhibited primarily molecular levels of granularity in their arguments (RQ1b). For the acid–base question, students who argued for correct claims provided links between pK_a, conjugate acid strength, and the direction of equilibria, while incorrect claims generated links between base strength and reactivity (RQ1c). Lastly, arguments for correct claims more frequently compared between claims than arguments for incorrect claims. How students compared between claims in their arguments was also found to differ between Q1 and Q2 (RQ1d).

The unit-based method can be extended beyond a single question type, as we demonstrated through our analysis of Q2 (comparing mechanisms). Students' arguments differed between the two question types. Arguments in answers to Q2 were more often causal and more frequently compared between possible claims than arguments in Q1. We found that arguments for both questions were similar in that the majority did not discuss concepts past the molecular level of granularity. Taken together, these findings reinforce the notion that how one constructs an argument heavily depends on the task and expectations for that task. In professional settings, the context frames or dictates the structure of an argument that is required. In educational contexts (e.g., courses), educators need to explicitly communicate expectations or context through learning outcomes, lectures, examples, assessments, and feedback (Bernholt and Parchmann, 2011; Stoyanovich *et al.*, 2015; Weinrich and Talanquer, 2016; Caspari, Weinrich, *et al.*, 2018; Carle and Flynn, 2020).

We found that traditional methods of scoring scientific arguments effectively captured the quality of the scientific **content** but not the quality or sophistication of arguments' **structures**, leading to different scores if using one method over the other (RQ2). Specifically, traditional methods of scoring often focus on key words, and do not explicit evaluate for structural aspects of students' arguments, such as reasoning, granularity, and/or comparisons. In an instructional context, educator's goals for a particular assessment and other contextual factors will affect how they choose to use the unit-based method. Our findings provide evidence that using the unit-based method allows the user to capture aspects of students' arguments perhaps overlooked by scoring systems focused on key words.

Lastly, to support educators, we share a rubric based on the unit-based method than can be used to guide evaluation of students' arguments in terms of reasoning, granularity, and/or comparisons. Findings from interviews with educators suggest that the rubric serves as a helpful framework for educators to consider different dimensions of students' arguments, a strong starting point for educators communicate expectations about arguments with students, and a flexible foundation that can be adapted into more task-specific evaluation methods and rubrics (RQ3). We invite educators and/or researchers from all disciplines to adapt the rubric for tasks relevant to their own teaching such that we might provide the teaching community with a rich database of items and rubrics to support the development of students' argumentation.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank Myriam Carle for her assistance with inter-rater reliability. This project was funded in part by the Natural Sciences and Engineering Research Council's Canadian Graduate Scholarship (Master's) and the Ontario Graduate Scholarship.

References

(1) Abrams E. and Southerland S., (2010), The how's and why's of biological change : How learners neglect physical mechanisms in their search for meaning. *Int. J. Sci. Educ.*, **23**(12), 1271–1281.

(2) von Aufschnaiter C., Erduran S., Osborne J., Simon S., Education P., and Giessen J., (2008), Arguing to Learn and Learning to Argue: Case Studies of How Students' Argumentation Relates to Their Scientific Knowledge. *J. Res. Sci. Teach.*, **45**(1), 101–131.

(3) Barwell R., (2018), Word problems as social texts. *Numer. as Soc. Pract. Glob. Local Perspect.*, 101–120.

(4) Becker N., Rasmussen C., Sweeney G., Wawro M., Towns M., and Cole R., (2013), Reasoning using particulate nature of matter: an example of a sociochemical norm in a university-level physical chemistry class. *Chem. Educ. Res. Pract.*, **14**(1), 81–94.

(5) Berland L. K. and Reiser B. J., (2009), Making Sense of Argumentation and Explanation. *Sci. Educ.*, **93**, 26–55.

(6) Bernholt S. and Parchmann I., (2011), Assessing the complexity of students ' knowledge in chemistry. *Chem. Educ. Res. Pract.*, **12**(2), 167–173.

(7) Bodé N. E., Deng J. M., and Flynn A. B., (2019), Getting Past the Rules and to the WHY: Causal Mechanistic Arguments When Judging the Plausibility of Organic Reaction Mechanisms. *J. Chem. Educ.*, **96**(6), 1068–1082.

(8) Carle M. S. and Flynn A. B., (2020), Essential learning outcomes for delocalization (resonance) concepts: How are they taught, practiced, and assessed in organic chemistry? *Chem. Educ. Res. Pract.*, **21**(2), 622–637.

(9) Carmel J. H., Herrington D. G., Posey L. A., Ward J. S., Pollock A. M., and Cooper M. M., (2019), Helping Students to "do Science": Characterizing Scientific Practices in General Chemistry Laboratory Curricula. *J. Chem. Educ.*, **96**(3), 423–434.

(10) Caspari I., Kranz D., and Graulich N., (2018), Resolving the complexity of organic chemistry students' reasoning through the lens of a mechanistic framework. *Chem. Educ. Res. Pract.*, **19**(4), 1117–1141.

(11) Caspari I., Weinrich M. L., Sevian H., and Graulich N., (2018), This mechanistic step is "productive": organic chemistry students' backward-oriented reasoning. *Chem. Educ. Res. Pract.*, **19**(1), 42–59.

(12) Cian H., (2020), The influence of context: comparing high school students' socioscientific reasoning by socioscientific topic. *Int. J. Sci. Educ.*, **42**(9), 1–19.

(13) Cooper M. and Klymkowsky M., (2013), Chemistry, life, the universe, and everything: A new approach to general chemistry, and a model for curriculum reform. *J. Chem. Educ.*, **90**(9), 1116–1122.

(14) Cooper M. M., Kouyoumdjian H., and Underwood S. M., (2016), Investigating Students' Reasoning about Acid-Base Reactions. *J. Chem. Educ.*, **93**(10), 1703–1712.

(15) Crandell O. M., Kouyoumdjian H., Underwood S. M., and Cooper M. M., (2018), Reasoning about Reactions in Organic Chemistry: Starting It in General Chemistry.

(16) Cruz-Ramírez De Arellano D. and Towns M. H., (2014), Students' understanding of alkyl halide reactions in undergraduate organic chemistry. *Chem. Educ. Res. Pract.*, **15**(4), 501–515.

(17) Darden L., (2002), Strategies for Discovering Mechanisms: Schema Instantiation, Modular Subassembly, Forward/Backward Chaining. *Philos. Sci.*, **69**(S3), 354–365.

(18) DeCocq V. and Bhattacharyya G., (2019), TMI (Too much information)! Effects of given information on organic chemistry students' approaches to solving mechanism tasks. *Chem. Educ. Res. Pract.*, **20**(1), 213–228.

(19) Dood A. J., Dood J. C., Cruz-Ramírez De Arellano D., Fields K. B., and Raker J. R., (2020), Analyzing explanations of substitution reactions using lexical analysis and logistic regression techniques. *Chem. Educ. Res. Pract.*, **21**(1), 267–286.

(20) Emig J., (1977), Writing as a Mode of Learning. *Coll. Compos. Commun.*, **28**(2), 122–128.

(21) Erduran S., Simon S., and Osborne J., (2004), TAPping into argumentation: Developments in the application of Toulmin's Argument Pattern for studying science discourse. *Sci. Educ.*, **88**(6), 915–933.

(22)European Union, (2006), Recommendation of the European Parliament and of the Council of 18 December 2006 on key competences for lifelong learning. *Off. J. Eur. Union*, L 394/19-L 394/18.

(23)Flynn A. B., (2017), Flipped Chemistry Courses: Structure, Aligning Learning Outcomes, and Evaluation, in *Online Approaches to Chemical Education*, American Chemical Society, pp. 151–164.

(24)Flynn A. B., (2015), Structure and evaluation of flipped chemistry courses: Organic & spectroscopy, large and small, first to third year, English and French. *Chem. Educ. Res. Pract.*, **16**(2), 198–211.

(25)Flynn A. B. and Ogilvie W. W., (2015), Mechanisms before Reactions: A Mechanistic Approach to the Organic Chemistry Curriculum Based on Patterns of Electron Flow. *J. Chem. Educ.*, **92**(5), 803–810.

(26)Grimberg B. I. and Hand B., (2009), Cognitive pathways: Analysis of students' written texts for science understanding. *Int. J. Sci. Educ.*, **31**(4), 503–521.

(27)Ha M., Nehm R. H., Urban-Lurain M., and Merrill J. E., (2011), Applying computerized-scoring models ofwritten biological explanations across courses and colleges: prospects and limitations. *CBE Life Sci. Educ.*, **10**(4), 379–393.

(28)Hallgren K. A., (2012), Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor. Quant. Methods. Psychol.*, **8**(1), 23–34.

(29)Hogan T. P. and Murphy G., (2007), Recommendations for preparing and scoring constructed-response items: What the experts say. *Appl. Meas. Educ.*, **20**(4), 427–441.

(30)Jimenez-Aleixandre M. P. and Federico-Agraso M., (2009), Justification and persuasion about cloning: arguments in Hwang's paper and journalistic reported versions. *Res. Sci. Educ.*, **39**(3), 331–347.

(31)Jones M. D. and Crow D. A., (2017), How can we use the "science of stories" to produce persuasive scientific stories. *Palgrave Commun.*, **3**(1), 1–9.

(32)Kelly G. J. and Bazerman C., (2003), How students argue scientific claims: A rhetorical-semantic analysis. *Appl. Linguist.*, **24**(1), 28–55.

(33)Kelly G. J., Druker S., and Chen C., (1998), Students' reasoning about electricity: Combining performance assessments with argumentation analysis. *Int. J. Sci. Educ.*, **20**(7), 849–871.

(34)Kelly G. J., Regev J., and Prothero W., (2007), Analysis of Lines of Reasoning in Written Argumentation, in *Argumentation in Science Education*, pp. 137–157.

(35)Kelly G. J. and Takao A., (2002), Epistemic levels in argument: an analysis of university oceanography students' use of evidence in writing. *Sci. Educ.*, **86**(3), 314–342.

(36)Kraft A., Strickland A. M., and Bhattacharyya G., (2010), Reasonable reasoning: multi-variate problem-solving in organic chemistry. *Chem. Educ. Res. Pract.*, **11**(4), 281–292.

(37)Krippendorff K., (1970), Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educ. Psychol. Meas.*, **30**(1), 61–70.

(38)Kuhn D., (2011), *The skills of argument*, Cambridge University Press.

(39)Kuhn D., Zillmer N., Crowell A., and Zavala J., (2013), Developing norms of argumentation: Metacognitive, epistemological, and social dimensions of developing argumentive competence. *Cogn. Instr.*, **31**(4), 456–496.

(40)Laverty J. T., Underwood S. M., Matz R. L., Posey L. A., Carmel J. H., Caballero M. D., et al., (2016), Characterizing College Science Assessments: The Three-Dimensional Learning Assessment Protocol. *PLoS One*, **11**(9), 1–21.

(41)Liu O. L., Rios J. A., Heilman M., Gerard L., and Linn M. C., (2016), Validation of automated scoring of science assessments. *J. Res. Sci. Teach.*, **53**(2), 215–233.

(42)Luisi P. L., (2002), Emergence in Chemistry: Chemistry as the Embodiment of Emergence. *Found. Chem.*, **4**(3), 183–200.

(43)Lytzerinou E. and Iordanou K., (2020), Teachers' ability to construct arguments, but not their perceived self-efficacy of teaching, predicts their ability to evaluate arguments. *Int. J. Sci. Educ.*, **0**(0), 1–18.

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

(45) Maeyer J. and Talanquer V., (2013), Making Predictions About Chemical Reactivity: Assumptions and Heuristics. *J. Res. Sci. Teach.*, **50**(6), 748–767.

(46) McClary L. and Talanquer V., (2011), Heuristic reasoning in chemistry: making decisions about acid strength. *Int. J. Sci. Educ.*, **33**(10), 1433–1454.

(47) McNeill K. L., Lizotte D. J., Krajcik J., and Marx R. W., (2006), Supporting Students' Construction of Scientific Explanations by Fading Scaffolds in Instructional Materials. *J. Learn. Sci.*, **15**(2), 153–191.

(48) van Mil M. H. W., Jan D., Arend B., and Waarlo J., (2013), Modelling Molecular Mechanisms : A Framework of Scientific Reasoning to Construct Molecular-Level Explanations for Cellular Behaviour. *Sci. Educ.*, **22**(1), 93–118.

(49) Moon A., Moeller R., Gere A. R., and Shultz G. V., (2019), Application and testing of a framework for characterizing the quality of scientific reasoning in chemistry students' writing on ocean acidification. *Chem. Educ. Res. Pract.*, **20**(3), 484–494.

(50) Moon A., Stanford C., and Cole R., (2017), Analysis of Inquiry Materials to Explain Complexity of Chemical Reasoning in Physical Chemistry Students ' Argumentation. *J. Res. Sci. Teach.*, **54**(10), 1322–1346.

(51) Moon A., Stanford C., Cole R., and Towns M., (2016), The nature of students' chemical reasoning employed in scientific argumentation in physical chemistry. *Chem. Educ. Res. Pract.*, **17**(2), 353–364.

(52) Moreira P., Marzabal A., and Talanquer V., (2019), Using a mechanistic framework to characterise chemistry students' reasoning in written explanations. *Chem. Educ. Res. Pract.*, **20**(1), 120–131.

(53) National Research Council, (2012), *A Framework for K-12 Science Education*, National Academies Press.

(54) Ogilvie W. W., Ackroyd N., Browning S., Deslongchamps G., Lee F., and Sauer E., (2017), *Organic Chemistry: Mechanistic Patterns*, 1st ed. Nelson Education Ltd.

(55) Ontario Ministry of Education, (2004), The Ontario Curriculum: Achievement Charts,.

(56) Organisation for Economic Cooperation and Development, (2006), Assessing scientific, reading and mathematical literacy: a framework for PISA 2006,.

(57) Osborne J. F. and Patterson A., (2011), Scientific Argument and Explanation: A Necessary Distinction? *Sci. Educ.*, **95**(4), 627–638.

(58) Petritis S. J., Kelley C., and Talanquer V., (2020), Exploring the impact of the framing of a laboratory experiment on the nature of student argumentation. *Chem. Educ. Res. Pract.*

(59) Reed J. J., Brandriet A. R., and Holme T. A., (2017), Analyzing the Role of Science Practices in ACS Exam Items. *J. Chem. Educ.*, **94**(1), 3–10.

(60) Russ R. S., Scherr R. E., Hammer D., and Mikeska J., (2008), Recognizing Reasoning in Student Scientific Inquiry: A Framework for Discourse Analysis Developed From Philosophy of Science. *Sci. Educ.*, **92**(3), 499–525.

(61) Sadler T. D., (2004), Informal reasoning regarding socioscientific issues: A critical review of research. *J. Res. Sci. Teach.*, **41**(5), 513–536.

(62) Sadler T. D., (2006), Promoting discourse and argumentation in science teacher education. *J. Sci. Teacher Educ.*, **17**(4), 323–346.

(63) Sadler T. D. and Zeidler D. L., (2005), The significance of content knowledge for informal reasoning regarding socioscientific issues: Applying genetics knowledge to genetic engineering issues. *Sci. Educ.*, **89**(1), 71–93.

(64) Sandoval W. A. and Millwood K. A., (2005), The Quality of Students' Use of Evidence in Written Scientific Explanations. *Cogn. Instr.*, **23**(1), 23–55.

(65) Sandoval W. A. and Reiser B. J., (2004), Explanation-driven inquiry: Integrating conceptual and epistemic

scaffolds for scientific inquiry. *Sci. Educ.*, **88**(3), 345–372.

(66)Sevian H., Bernholt S., Szteinberg G. A., and Auguste S., (2015), Use of representation mapping to capture abstraction in problem solving in different courses in chemistry. *Chem. Educ. Res. Pract.*, **16**(3), 429–446.

(67)Sevian H. and Talanquer V., (2014), Rethinking chemistry: a learning progression on chemical thinking. *Chem. Educ. Res. Pr.*, **15**(1), 10–23.

(68)Social Sciences and Humanities Research Council, (2018), Truth Under Fire in a Post-Fact World.

(69)Southard K. M., Espindola M. R., Zaepfel S. D., and Molly S., (2017), Generative mechanistic explanation building in undergraduate molecular and cellular biology. *Int. J. Sci. Educ.*, **39**(13), 1795–1829.

(70)Stowe R. L. and Cooper M. M., (2019), Arguing from Spectroscopic Evidence. *J. Chem. Educ.*, **96**(10), 2072–2085.

(71)Stowe R. L. and Cooper M. M., (2017), Practicing What We Preach: Assessing "Critical Thinking" in Organic Chemistry. *J. Chem. Educ.*, **94**(12), 1852–1859.

(72)Stoyanovich C., Gandhi A., and Flynn A. B., (2015), Acid-base learning outcomes for students in an introductory organic chemistry course. *J. Chem. Educ.*, **92**(2), 220–229.

(73)Talanquer V., (2017), Concept Inventories: Predicting the Wrong Answer May Boost Performance. *J. Chem. Educ.*, **94**(12), 1805–1810.

(74)Talanquer V., (2014), DBER and STEM education reform: Are we up to the challenge? *J. Res. Sci. Teach.*, **51**(6), 809–819.

(75)Talanquer V., (2007), Explanations and Teleology in Chemistry Education. *Int. J. Sci. Educ.*, **29**(7), 853–870.

(76)Talanquer V., (2018a), Importance of Understanding Fundamental Chemical Mechanisms. *J. Chem. Educ.*, **95**(11), 1905–1911.

(77)Talanquer V., (2018b), Progressions in reasoning about structure – property relationships. *Chem. Educ. Res. Pract.*, **19**(4), 998–1009.

(78)Toulmin S., (1958), *The Uses of Argument*, Cambridge University Press.

(79)Trommler F., Gresch H., Hammann M., Trommler F., Gresch H., and Hammann M., (2018), Students' reasons for preferring teleological explanations. *Int. J. Sci. Educ.*, **40**(2), 159–187.

(80)United Nations, (2015), Transforming our World: the 2030 Agenda for Sustainable Development.

(81)Verheij B., (2003), Dialectical argumentation with argumentation schemes: An approach to legal logic. *Artif. Intell. Law*, **11**(2–3), 167–195.

(82)Webber D. M. and Flynn A. B., (2018), How Are Students Solving Familiar and Unfamiliar Organic Chemistry Mechanism Questions in a New Curriculum? *J. Chem. Educ.*, **95**(9), 1451–1467.

(83)Weinrich M. L. and Sevian H., (2017), Capturing students' abstraction while solving organic reaction mechanism problems across a semester. *Chem. Educ. Res. Pract.*, **18**(1), 169–190.

(84)Weinrich M. L. and Talanquer V., (2016), Mapping students' modes of reasoning when thinking about chemical reactions used to make a desired product. *Chem. Educ. Res. Pract.*, **17**(2), 394–406.

(85)Windschitl M., Thompson J., and Braaten M., (2008), Beyond the Scientific Method: Model-Based Inquiry as a New Paradigm of Preference for School Science Investigations. *Sci. Educ.*, **92**(5), 941–967.

(86)Yan F. and Talanquer V., (2015), Students' Ideas about How and Why Chemical Reactions Happen: Mapping the conceptual landscape. *Int. J. Sci. Educ.*, **37**(18), 3066–3092.

(87)Yang J. S., Morell M., and Liu Y., (2019), Constructed-Response Items, in *SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*, pp. 381–383.

Appendix A: Codebook and Examples

Concepts

Concepts are the components of students' arguments that are described in the text of a student's argument. They represent the evidence that students find relevant in their arguments and can be linked to each other within in argument to generate the student's chain of reasoning. Additionally, they can be used to compare and contrast between the possible claims to create students' chains of reasoning.

Table 1. Codes for *concepts* proposed in arguments, with notes to guide appropriate application.

#	Code	Definition	Notes
1	pKa	pK_a of conjugate acids	Specific reference to the pK_a of any (or all) of the conjugate acids described. • May include explicit discussion of relative pK_a values without referring to actual values (*e.g.*, pK_a of conjugate acid X is lower than that of conjugate acid Y). • Correct if the pK_a value for the chosen base (Claim) discussed correctly relative to other bases. • Incorrect if the pK_a for the chosen base (Claim) discussed incorrectly relative to other bases.
2	acidstr	Conjugate acid strength	Specific reference to the strength of the conjugate acid of any (or all) of the bases provided. • Must explicitly discuss acid strength of the conjugate acid(s) (*i.e.* using terms like "strong" and/or "weak"). • Do not code if argument only mentions "conjugate acid" without reference to the strength of that conjugate acid.
3	equilfav	Direction of equilibrium	Specific reference to the direction favoured by the equilibrium involving the base(s) of interest • Do not code for asymmetrical equilibrium arrows; must be explicitly stated. Correctness of the code dependent on the linked concepts used to justify favourability towards a specific direction. No need to get the direction correct, as long as justification is correct. • pKa, basestr, Acidstr, and stab would all be correct justifications
4	basestr	Base strength	Reference to the strengths of the given bases. • Correct if C > A > B discussed explicitly or implicitly (*i.e.*, C is the strongest base, *etc.*) • C > A,B acceptable. For example, "NaH is the strongest base of the three" would be coded as correct. • Incorrect if discussed without any reference to other bases or if comparisons between bases are incorrect.
5	stab	Stability of molecule	Reference to the stability of a molecule. Stability can be thought of as an internal characteristic of the molecule(s) (the molecules themselves are being described as stable). • Does not need to include discussion of energetics; simply saying "stable" warrants inclusion of code. • Discussions related to the favourability of lower-energy species should be coded. Correctness of code dependent on the linked concepts used to rationalize stability of the molecule.
6	react	Reactivity	Reference to the general reactive behaviour of a molecule. Reactive behaviour can be seen as an *activity* the molecule engages in (how the molecule will/might interact with other molecules). • *e.g.* "Because NaH is a strong base, **it will be able to deprotonate the alkyne."** Not mechanistic because it

			does not describe how electrons move from the base to the alkyne. • Different from “mech” because mechanism focuses on explicit discussion around electron movement and arrow-pushing. “react” is more general and used to capture arguments that *do not involve* more detailed discussions of reactions. Correct if linked concepts used to rationalize reactivity of the molecule (*e.g.*, basestr, ster, FC) are correct Correct if base(s) described with correct reactive behaviour (*e.g.*, “NaOH will deprotonate the alkyne.”) Incorrect if described with incorrect reactive behaviour (*e.g.*, “NaOH will react nucleophilically with the acyl chloride. “)
7	electroneg	Electronegativity	Specific reference to electronegativity of specific atoms on the base(s), or to the difference in electronegativity across a bond. • Code if differences in electron density are discussed or if an atom is said to have a greater capacity to attract electrons; argument does not need to explicitly say “electronegativity”
8	FC	Formal charge	Specific reference to the charges (or lack thereof) on molecules • Do not code if formal charges simply drawn onto structures. Must be explicitly discussed as a concept in-text (*e.g.* “The oxygen has a negative charge, which means…”).
9	ster	Steric hindrance	Specific reference to the size/bulk of the bases. • Correct if B > A > C discussed explicitly or implicitly. • B > A, C acceptable. For example, “NH_3 is the largest base of the three” would be coded as correct. • Incorrect if discussed without any reference to other bases.
10	LeChat	Le Châtelier’s principle	Discussion involving the equilibrium “shifting” to compensate for external effects on the equilibrium • Correct if terminology which describes Le Châtelier’s principle is used. No need to say “Le Châtelier” explicitly in the argument. • Commonly used to capture arguments which include discussion of “H_2 leaving the reaction mixture”.
11	other	Other	Concept discussed that is unique and found not to be recurring (possible outlier). Provide supplementary comments regarding what “other” is referring to.
12	nuc	Nucleophilicity	Discussion of “nucleophilicity” of bases. • Correct if A, B > C discussed explicitly or implicitly (*i.e.*, C is less nucleophilic and A, B are more nucleophilic) • Incorrect if discussed without any reference to other bases or if comparisons between bases are incorrect.

1. *Concepts*

   For each concept, the following considerations should be made:

- Is it discussed correctly [y(g)] (y = yes, concept is present; g = correct) or with errors [y(e)] (y = yes, concept is present; e = error)?
- If the discussion of the concept contains no errors, it is considered correct
- There may be cases where the discussion of the concept itself is correct while the link to another concept is incorrect

    - *e.g.*, Relative base strengths of the bases are discussed correctly but are incorrectly linked to concept of steric hindrance.
- In some cases, the correctness of the concept is dependent on the links to the concept
    - *e.g.*, The proposed direction of an equilibrium (“equilfav”) is correct if it is correctly rationalized with Le Chatêlier’s principle (“lechat”). It is incorrect if it is at all rationalized with incorrectly linked concepts, such as electronegativity (“electroneg”) or steric hindrance (“ster”).

2. *Links Between Concepts*

- Is the basis for the link correct (g) or incorrect (e)?
- In most cases, a link will be established in an argument as a student describes how Concept 1 impacts or explains Concept 2.
- If two concepts are linked, but the argument moved through *an additional concept* to link them, be sure to code all three concepts as being linked.
    - *e.g.*, if someone gives an explanation where pK_a is linked to conjugate acid strength which is then linked to base strength, don’t code pK_a as being linked to base strength (unless the two ideas are explicitly linked elsewhere). Links in this case would be pKa - Acidstr – basestr.

3. *Comparisons Between Bases*

- A = NaOH; B = NH_3; C = NaH
- Implicit: “NaH is a very strong base” or “NaH is the strongest base”
- Explicit: “NaH is a stronger base than NaOH which is a stronger base than NH_3”
- If *general* comparisons are being made with a single concept across all three bases, code as “X to Y, Z” where X is the base chosen in the claim. These general comparisons are coded as implicit.
- If no comparisons are made, still include the base for which the concept is being discussed in reference to.
    - If the argument only discusses the pK_a of A, then the comparison portion of the code would simply have “A”

In a given cell in an Excel spreadsheet (in this example, coding for “Concept X”):

Indicate if Concept X is present, and if it is discussed correctly (g) / incorrectly (e)

Indicate what concepts are linked to Concept X, and whether the basis for the link is correct (g) / incorrect (e)

Concept X
y(g/e) – [linked Concept Y](g/e),…,[linked concept *n*](g/e) – A to B

Be sure to provide codes for all concepts linked to Concept X

Indicate whether a comparison is made between the three bases with respect to Concept X

Features

Features differ from Concepts as the former are usually drawn out and not linked to other codes, while the latter are written and presented in-text. For example, codes “equildraw”, “struct”, and “mech” provide additional context to question as they capture portions of the question that are not in-text. Because they are not in-text, they are not explicitly linked to other concepts.

Table 2. Codes for *Features* present in arguments, with notes to guide appropriate application.

13	equildraw	Equilibrium drawn	Drawing the equilibrium for one of or all of the bases provided. Can be as simple as the equilibrium for the three bases and their respective conjugate acids (no need to include other reactants/products).
14	mech	Mechanism	Explicit description of how the reaction will proceed mechanistically. Either explicit discussion of electron movement or drawn with electron-pushing formalism. • Mechanisms drawn on top of the original question can be used • Correct as long as the mechanism described/drawn with the *chosen base* is correct. • Incorrect if mechanism described or drawn for the *chosen base* is incorrect or if the described mechanism is too vague to convey meaningful mechanistic information.
15	struct	Structure	Drawing a structure to represent the base(s).

			• Does not need to be a full Lewis structure • Any attempt to re-draw the bases with additional features (*i.e.*, lone pairs, formal charges, *etc.*) is coded as correct. • Incorrect if structures presented are explicitly incorrect (*e.g.*, incorrect connectivity, atoms, number of lone pairs, formal charges, *etc.*)
16	alkpKa	Alkyne pK_a	Reference to the pK_a value of the alkyne in relation to the pK_a values of the other bases. • Correct if pK_a value of alkyne is correctly identified as lower than the conjugate acid of the chosen base (does not need to be the exact value of 24).
17	alkCAstr	Alkyne acid strength	Reference to the acid strength of the alkyne in relation to the strength of the conjugate acids of the base(s). • Correct if relative acid strength of alkyne is described correctly in comparison to the conjugate acids of the base(s)
18	pKa	pK_a values listed (outside of text)	Code if pK_a values of the conjugate acids of the bases are provided in the argument in isolation from the main text and/or other concepts. The “pK_a” code from *Concepts* captures references to pK_a within the main text and/or other concepts. • Correct if the pK_a value listed for the chosen base (Claim) is correct. • Incorrect if the pK_a value listed for the chosen base (Claim) is incorrect.

These codes can be coded in a given Excel spreadsheet cell simply as (for an example with Feature X):

Indicate if the feature is portrayed correctly (g) / incorrectly (e)

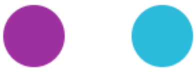
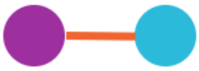
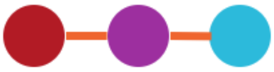
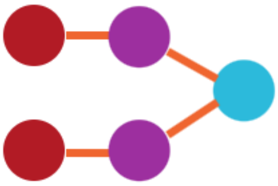
Indicate for which bases the feature is provided for

Feature X
y(g/e) – A, B, C

Modes of Reasoning

- Does the argument include a descriptive, relational, linear causal, or multi-component mode of reasoning?

Table 3. Descriptions of the four levels of reasoning used in this study present in arguments, with illustrations to guide appropriate application.

Level of Reasoning	Description	Reasoning Diagram
Descriptive (D)	Argument contains descriptions of evidence and the claim No relationships established between evidence and the claim.	
Relational (R)	Argument contains descriptions of evidence and the claim Evidence is correlated to other evidence and/or to the claim (*i.e.*, a "matter-of-fact")	
Linear causal (L)	Argument contains descriptions of evidence and the claim Evidence linked causally to other evidence and/or to the claim (*i.e.*, justification for *why*)	
Multi-component causal (M)	Multiple causal relationships are described and coalesce to justify a single claim.	

Note that there may be portions of the explanation that exhibit higher/lower levels of reasoning than other portions. For this study, the argument was ascribed the highest level of reasoning identified in the argument.

Reasoning Diagrams

To visualize the concepts and concept links present in students' explanations, reasoning diagrams were constructed. In addition to providing a visual aid for describing how students construct their explanations, the diagrams also serve as a means to identify the level of reasoning for each student explanation based on the illustrations shown in the aforementioned "Levels of Reasoning" section.

In a given reasoning diagram, concepts identified in students' arguments are represented by nodes in the diagrams. Links identified between these concepts are then linked by lines. Generally, links describe how Concept 1 explains Concept 2 (regardless of whether this link is correct). Once links have been established, one can then identify the level of reasoning using the illustrations for each level.

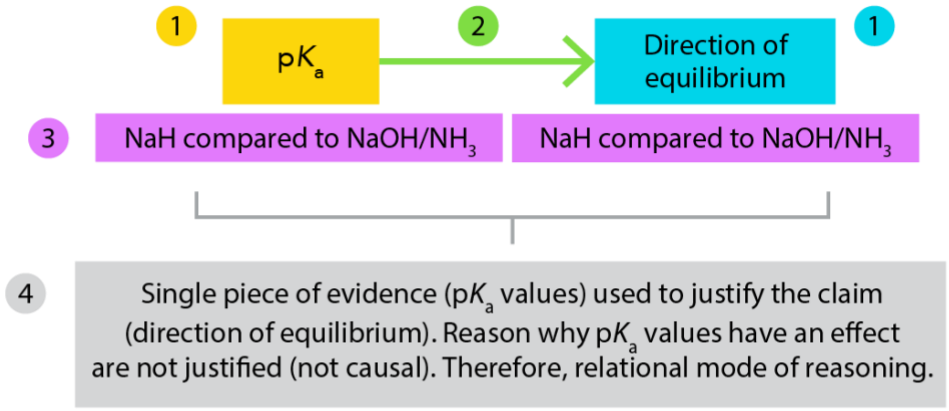
In the following example, the concepts identified in the student's argument are pK_a (pKa) and the favourability of the equilibrium (equilfav). These two concepts are linked, so a line is drawn between the two concept nodes. Because only evidence (pKa) is being used to explain an activity (equilfav), this explanation is coded as overall relational.

Argument

By comparing the pK_a's of the conjugate acids of each base with the pK_a of pentyne, it was determined that only the conjugate acid of H- has a greater pK_a than pentyne, meaning that the equilibrium would be forced towards the product side.

The conjugate acids of NaOH and NH_3 have pK_a's lower than that of pentyne.
If these bases were to be used, the equilibrium would favour the starting material.

Reasoning Diagram



1 pK_a 2 Direction of equilibrium 1

3 NaH compared to NaOH/NH_3 NaH compared to NaOH/NH_3

4 Single piece of evidence (pK_a values) used to justify the claim (direction of equilibrium). Reason why pK_a values have an effect are not justified (not causal). Therefore, relational mode of reasoning.

Steps of Analysis

1 1 Identify key concepts in argument (e.g., pK_a values, direction of equilibrium)

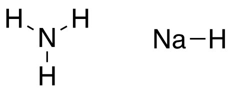
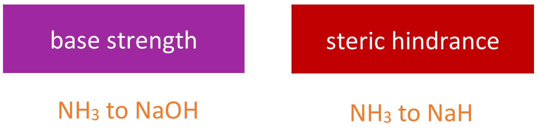
2 Identify links between concepts

3 Identify concepts used to compare between claims

4 Identify mode of reasoning from key concepts and links

Coding Examples:

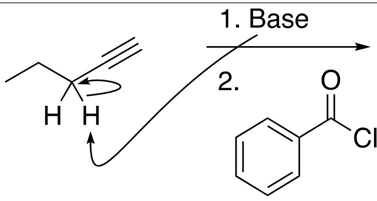
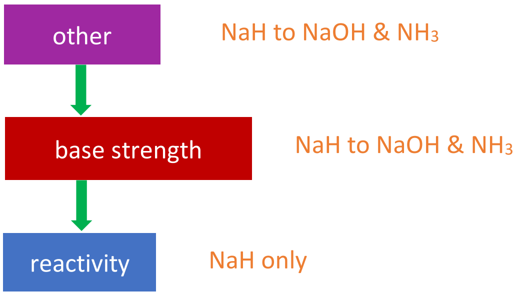
Student 9 (Claim: B)

Student Argument	Reasoning Diagram
I chose NH_3 because it is a strong base compared to NaOH. As well it is bulkier than NaH.  ***Pentyne is not very crowded and therefore NH_3 can access it very easily.***	 base strength steric hindrance NH_3 to NaOH NH_3 to NaH

	base strength	steric hindrance	structure drawn
	y(e) - B to A	y(g) - B to C	y(g) – B,C
Concepts	Incorrect because NH_3 is not a stronger base than NaOH.	Correct because it is true that NH_3 is bulkier than NaH	Both structures are drawn without error.
Concept links	No clear, explicit links are made between base strength and sterics.	No clear, explicit links are made between base strength and sterics.	-
Comparisons	NH_3 (B) is explicitly said to be a stronger base than NaOH (A).	NH_3 (B) is explicitly said to be bulkier than NaH (C).	-

Mode of reasoning	Level of granularity	Level of comparison
Descriptive (with errors). No explicit links are made between base strength and sterics.	Both concepts are discussed at the molecular level of granularity.	**Fully compared**. Both concepts are used to compare between bases.

Student 10 (Claim: C)

Student Argument	Reasoning Diagram
 ***NaH is the strong base choice therefore it is most likely to deprotonate the carbon. NH$_3$*** **(structure)** ***is an extremely weak base because the lone pairs on nitrogen are very weak. NaOH*** **(structure)** ***is much stronger because Na increases the strength of the lone pair. However, NaH*** **(structure)** ***is an even better option because the negative H has a greater affinity to bond to protons because of the strength of its lone pairs.***	 other NaH to NaOH & NH3 base strength NaH to NaOH & NH3 reactivity NaH only

	base strength	reactivity	other	mechanism drawn	structure drawn
	y(g) - [o](e),[react](g) - C to A,B	y(e) - [basestr](g) - C	y(e) - [basestr](e) - C to A,B	y(e) - A,B,C (gen)	y(g) - A,B,C
Concepts	Correct because it is true that NaH is the strongest of the three bases.	Incorrect because it is unclear which carbon is being deprotonated.	Incorrect because it is unclear what “strength of lone pairs” means or represents.	Incorrect because the base does not deprotonate the terminal alkyne.	All three structures are drawn without error.
Concept links	other: Relative base strengths are justified by relative “strengths of lone pairs” for each base. incorrect because strength of lone pairs is not an appropriate justification for base strength. react: Reactive behaviour justified by base strength. Correct because base strength can be used to determine how molecules will react with each other.	basestr: Reactivity justified by base strength. Correct because base strength can be used to determine how molecules will react with each other.	basestr: Relative base strengths are justified by relative “strengths of lone pairs” for each base. incorrect because strength of lone pairs is not an appropriate justification for base strength.	-	-
Comparisons	The relative strengths of all three bases are explicitly stated.	Reactivity discussed only with respect to NaH (C)	The relative “strength of lone pairs” of all three bases explicitly stated.	-	-

Mode of reasoning	Level of granularity	Level of comparison

Linear causal (with errors). Base strength is used to justify the reactive behaviour, and the “strength of lone pairs” is used to justify why base strength is relevant.	Reactive behaviour is a reaction-level concept. Base strength is a molecular-level concept. “Other” concepts were not coded for their granularity due to their variety.	**Partially compared**. All concepts except “react” are used to compare between bases.

Student 19 (Claim: C)

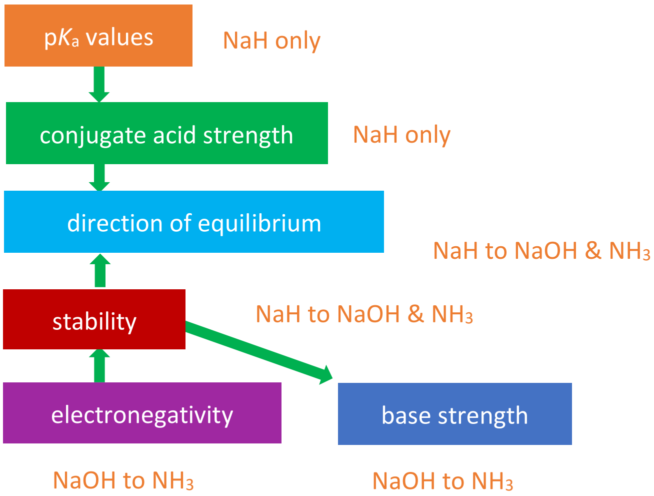
Student Argument

"The equilibrium of the first step is dependent on the acid-base reaction and as a result, it is dependent on which side does the stronger acid lie. Based on the structure of the reactant, the more acidic proton is at the terminal alkyne (pKa 50 [C-H sp^3] vs 24 [C-H sp]), so the appropriate base must have a weaker conjugate acid.

Based on the electronegativity of OH and NH$_3$, they would serve as better bases than the alkyne as the greater electronegativity of O and N allowing the ionized forms to better stabilize a negative charge (for O, making the $^-$OH a more stable base than the ionized alkyne) and less able to stabilize a positive charge (for N, NH$_4^+$ (CA for NH$_3$) is more acidic than alkynes and hence, shifts equilibrium to the alkyne). As for NaH, the similar values in electronegativity between H and C would not influence the equilibrium as much as NaOH and NH$_3$.

Also, the reaction results in the production of H$_2$, which is very stable and hence, the H$_2$ is less likely to protonate the alkyne, favouring the product side.

Reasoning Diagram



	pK_a values	conjugate acid strength	direction of equilibrium	base strength
	y(g) - [Acidstr](g) - C	y(g) - [pKa](g),[equilfav](g) - C	y(g) - [Acidstr](g),[stab](g) - C to A,B	y(g) - [stab](g) - A to B
Concepts	Correct because the pK_a value of the bases is correctly discussed *relative* to the pK_a value of the alkyne.	Correct because it is true that NaH will have the weakest conjugate acid.	Correct because both conjugate acid strength and stability are both correctly used to justify for the direction of the equilibrium.	Correct because as the relative base strengths are compared without error.
Concept links	Acidstr: pK_a value of a conjugate acid is correctly linked as an indicator of conjugate acid strength	pKa: pK_a value of a conjugate acid is correctly linked as an indicator of conjugate acid strength equilfav: Relative conjugate acid strength is correctly linked as an indicator of the direction of the equilibrium	Acidstr: Relative conjugate acid strength is correctly linked as an indicator of the direction of the equilibrium stab: Stability of species on either side of the equilibrium is correctly linked as an indicator of the direction of the equilibrium	stab: Stability of a base is correctly linked to the strength of that base (more stable = weaker base)
Comparisons	pK_a values discussed generally, but implied through links that H$_2$ (C)	Conjugate acid strength discussed generally but implied through links that H$_2$ (C) is a weak conjugate acid.	The favoured direction of the equilibrium in the presence of each possible base is discussed explicitly.	Both NaOH (A) and NH$_3$ (B) are correctly stated to be weaker bases than the alkyne anion.

	has a sufficiently high pK_a value

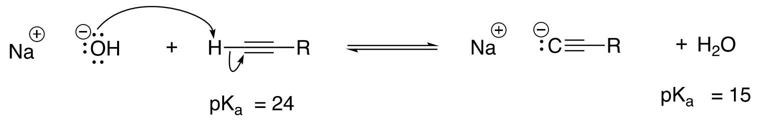
	stability	electronegativity	pK_a value of alkyne	acid strength of alkyne
	y(g) - [basestr](g),[equilfav](g),[electroneg](g) - C to A,B	y(g) - [stab](g) - A to B	y(g)	y(g)
Concepts	Correct because each stability is correctly justified through each of the linked concepts.	Correct because it is true that oxygen is more electronegative than nitrogen.	Alkyne pK_a value reported without error.	Relative conjugate acid strength of the alkyne was correctly stated to be stronger than NaH.
Concept links	basestr: Stability of a base is correctly linked to the strength of that base (more stable = weaker base) equilfav: Stability of species on either side of the equilibrium is correctly linked as an indicator of the direction of the equilibrium electroneg: Stability of a molecule correctly justified using electronegativity (oxygen more electronegative, so more stable anion)	stab: Stability of a molecule correctly justified using electronegativity (oxygen more electronegative, so more stable anion)	-	-
Comparisons	The relative stabilities of NaOH (A) and NH_3 (B) are compared correctly (with reference to electronegativity). H_2 (C) is correctly referenced as “very stable”, but in a sequence that is in isolation from the other claims.	The electronegativities of the O and N atoms on NaOH (A) and NH_3 (B) are explicitly discussed and compared.	-	-

Mode of reasoning	Level of granularity	Level of comparison
Multi-component causal (with errors). First Linear causal sequence explains the direction of the equilibrium using relative conjugate acid strengths, which are justified by different pK_a values. Second Linear causal sequence explains the direction of the equilibrium using stability, which is justified by differences in electronegativity.	Favoured direction of the equilibrium and pK_a value are reaction-level concepts. Base strength, conjugate acid strength, and stability are molecular-level concepts. Electronegativity is an atomic-level concept.	**Partially compared**. pK_a values and conjugate acids strength are discussed only in relation to NaH (C). All other concepts are used to compare between two or all bases.

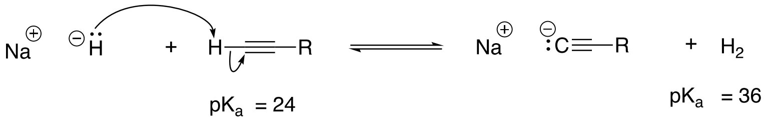
Student 29 (Claim: C)

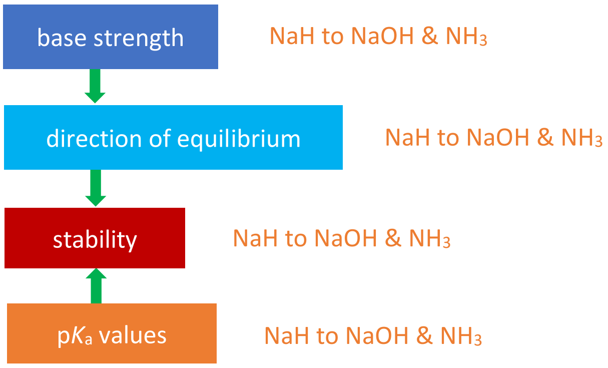
Student Argument	Reasoning Diagram

The rate determining step is the formation of the nucleophile [alkyne anion]. In order to really favour the product side, the base must be very strong. This is because alkyne, due to its triple bond, is very stable.



Since the proton donor of the SM is higher in pK_a (means more stable) than the product acid (H_2O) this will favour the SM. This is also the case for NH_3.





	p*K*ₐ values	**direction of equilibrium**	**base strength**	**stability**
	y(g) - [stab](g) - C to A,B	y(g) - [basestr](g),[stab](g) - C to A,B	y(g) - [equilfav](g) - C to A,B	y(g) - [pKa](g),[equilfav](g) - C to A,B
Concepts	Correct because the p*K*ₐ values of the bases is correctly discussed *relative* to the p*K*ₐ of the alkyne.	Correct because both base strength and stability are both correctly used to justify for the direction of the equilibrium.	Correct because as the relative base strengths are compared without error.	Correct because each stability is correctly justified through each of the linked concepts.
Concept links	stab: Though p*K*ₐ value is not directly indicative of a molecule's stability, it is true that acids with larger p*K*ₐ values are weaker, more stable acids.	basestr: Base strength is explicitly stated to be a reason for an equilibrium to favour products. stab: Similar to base strength, stability is explicitly stated to be a reason fro an equilibrium to favour products.	equilfav: Base strength is explicitly stated to be a reason for an equilibrium to favour products.	pKa: Though p*K*ₐ values are not directly indicative of a molecule's stability, it is true that acids with larger p*K*ₐ values are weaker, more stable acids. equilfav: Similar to base strength, stability is explicitly stated to be a reason fro an equilibrium to favour products.
Comparisons	p*K*ₐ values discussed generally between all three claims.	Rationale for equilibrium favouring products discussed generally between all three claims.	Base strength discussed generally between all three claims.	Stability discussed generally between all three claims.

equilibrium drawn	**mechanism**	**structure drawn**	**p*K*ₐ value of alkyne**	**p*K*ₐ values listed**
y(g) - A,C	y(g) - A,C	y(g) - A,C	y(g)	y(g) - A,C

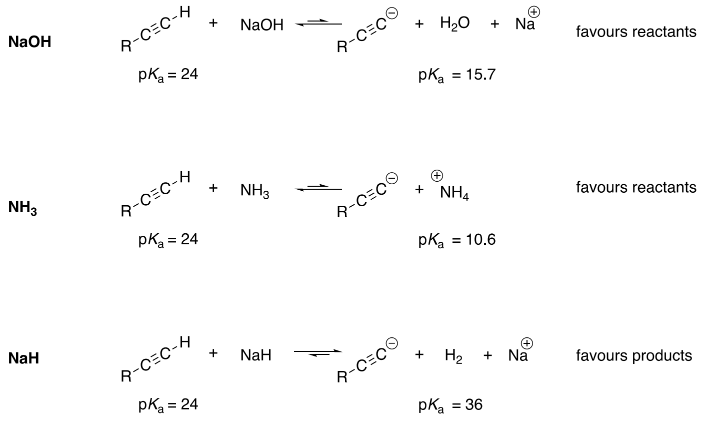
Concepts	Equilibria are correctly drawn (including the alkyne) for both NaOH and NaH	Mechanistic arrows correctly depict the acid—base reaction within the equilibria for both NaOH and NaH	Correct Lewis structures are shown for both NaOH and NaH.	The pK_a value of the alkyne is listed and correctly larger or smaller than the corresponding value for H_2O and H_2, respectively.	The pK_a values of both H_2O and H_2 are correctly listed (relative to each other).
Concept links	-	-	-	-	-
Comparisons	-	-	-	-	-

Mode of reasoning	**Level of granularity**	**Level of comparison**
Linear causal. pK_a value is used to justify stability, which is used to justify the direction of the equilibrium. Base strength is also used to justify the direction of the equilibrium, but this chain of reasoning is relational, so there is only one causal chain.	Favoured direction of the equilibrium and pK_a values are reaction-level concepts. Base strength and stability are molecular-level concepts.	**Fully compared**. All concepts (not including features) presented in the argument are used to compare (implicitly) between all three claims.

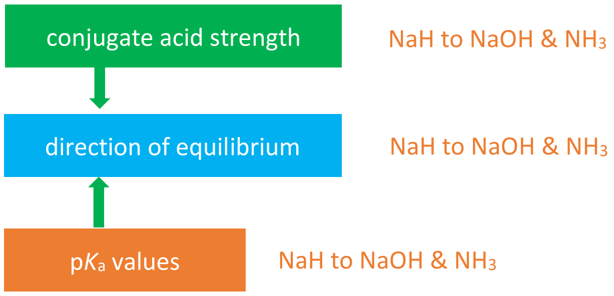
Student 39 (Claim: C)

Student Argument

NaH was chosen because the pK_a of the hydrogen connected to carbon [alkyne with arrow pointing to terminal C-H alkyne bond] is 24. In order for one of the above bases to push the first step to equilibrium, their conjugate acid in the reaction needs to have a lower pK_a than the 24 of the hydrogen (equil favours side with weakest species).



Reasoning Diagram


conjugate acid strength — NaH to NaOH & NH_3

direction of equilibrium — NaH to NaOH & NH_3

pK_a values — NaH to NaOH & NH_3

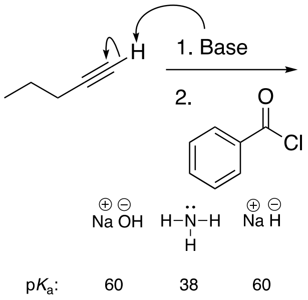
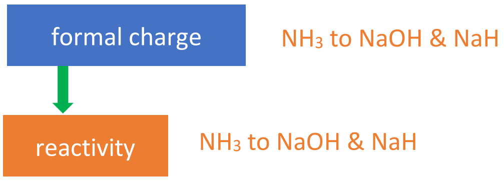
	pK_a values	**conjugate acid strength**	**direction of equilibrium**
	y(g) - [equilfav](g) - C to A,B	y(g) - [equilfav](g) - C to A,B	y(g) - [Acidstr](g),[pKa](g) - C to A,B
Concepts	Comparisons between the relative pK_a values of the conjugate acids are correct.	Correct because it is true that NaH will have the weakest conjugate acid.	Correct because both conjugate acid strength and pK_a values are both correctly used to justify for the direction of the equilibrium.
Concept links	equilfav: Though not true chemical rationale, pK_a value is used heuristically to correctly justify the direction of the equilibrium.	equilfav: Relative conjugate acid strength is correctly linked as an indicator of the direction of the equilibrium	pK_a values: Though not true chemical rationale, pK_a values are used heuristically to correctly justify the direction of the equilibrium. Acidstr: Relative conjugate acid strength is correctly linked as an indicator of the direction of the equilibrium
Comparisons	pK_a values discussed generally between all three claims.	Conjugate acid strength discussed generally. Implied through illustrations and additional concepts that H_2 (C) is weakest conjugate acid.	The favoured direction of the equilibrium in the presence of each possible base is explicitly stated.

	equilibrium drawn	**pK_a of alkyne**
	y(g) – A,B,C	y(g)
Concepts	Equilibria are correctly drawn (including the alkyne) for all three bases.	The pK_a of the alkyne is listed and correctly larger or smaller than the corresponding value for each of the three bases.

Concept links	-	-
Comparisons	-	-

Mode of reasoning	Level of granularity	Level of comparison
Relational. Two concepts – pK_a values and conjugate acid strength – are used to justify the direction of the equilibrium. However, a relationship between these two concepts is never ex**plicitly established (if this relationship was established**, the argument would be Linear causal).	Favoured direction of the equilibrium and pK_a are reaction-level concepts. Conjugate acid strength is a molecular-level concept.	**Fully compared**. All concepts (not including features) presented in the argument are used to compare (implicitly) between all three claims.

Student 42 (Claim: B)

Student Argument	Reasoning Diagram
 H 1. Base 2. O Cl ⊕ ⊖ Na OH H–N–H H ⊕ ⊖ Na H pK_a: 60 38 60 ***The NH_3 would be the best base in this case out of all the choices because it is the least reactive.*** ***The N on the NH_3 has a slightly negative part which makes it less of a base compared [than] NaOH which has a strong negative charge and NaH will also have a strong negative charge. Therefore, these will react very rapidly compared to NH_3.***	 formal charge NH_3 to NaOH & NaH reactivity NH_3 to NaOH & NaH

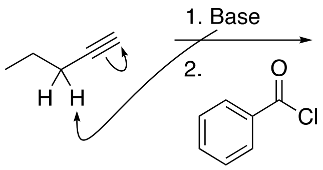
	reactivity	formal charge	structure drawn	pK_a values listed
	y(g) - [FC](g) - B to A,C)	y(g) - [react](g) - B to A,C	y(g) - A,B,C	y(e) - A,B,C

Concepts	Though NH_3 is not necessarily more reactive than either NaOH or NaH, the argument correctly uses formal charge as justification for reactivity.	Though NH_3 is not necessarily more reactive than either NaOH or NaH, the argument correctly uses formal charge as justification for reactivity.	An expanded structure for NH_3 is drawn. NaOH and NaH are both re-drawn with appropriate formal charges.	pK_a values are listed for all three bases. It is unclear mislabelling these values onto the bases themselves. If they are representative of the values for the conjugate acids, they are also incorrect relative to each other (NaOH and NaH should not have identical pK_a values).
Concept links	FC: See “Concepts” above.	react: See “Concepts” above.	-	-
Comparisons	The relative reactivity of each base is explicitly compared in the final statements of both the first and second sections of the argument. Because it is true that NH_3 is the least reactive of the three bases, this comparison is correct.	The charges on each base are explicitly and correctly discussed in the second section of the answer.	-	-

Mode of reasoning	Level of granularity	Level of comparison
Relational. Formal charge is used as the sole justification for reactivity. However, there is no discussion for why formal charge is even relevant to reactivity (nor why even reactivity is relevant in this context).	Reactivity is a molecular-level concept. Formal charge is an atomic-level concept.	**Fully compared**. All concepts presented in the argument are used to compare between all three claims.

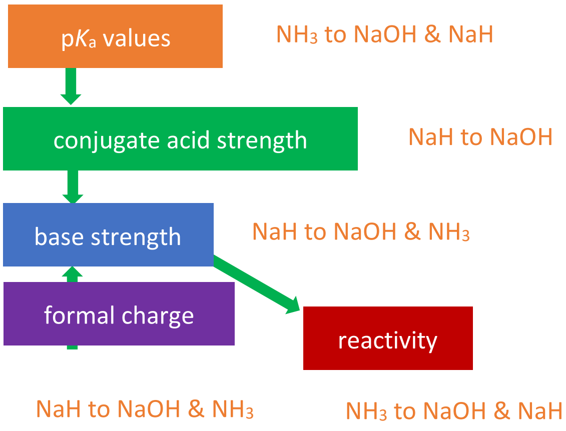
Student 118 (Claim: B)

Student Argument	Reasoning Diagram



NH$_3$ -> pK$_a$ ~ 38, strong base
H$^-$ -> H$_2$ -> pK$_a$ ~ 36 (weak CA = strong base)
$^-$OH -> OH$_2$ -> pK$_a$ ~ 17 (weak CA = strong base)

You need a strong base but weak Nu because otherwise Nu attack would occur and you need to avoid. NH$_3$ is a strong base and weak Nu. NaH/NaOH [with charges] are both strong bases but strong Nu (- charge)



NH$_3$ to NaOH & NaH

NaH to NaOH

NaH to NaOH & NH$_3$

NaH to NaOH & NH$_3$

NH$_3$ to NaOH & NaH

	pK_a values	conjugate acid strength	base strength	reactivity
	y(e) - [Acidstr](g) - B to A,C	y(g) - [pKa](g),[basestr](g) - C to A	y(e) - [react](g),[FC](e) - B to A,C	y(e) - [basestr](g) - B to A,C
Concepts	The pK_a values for the conjugate acid are compared, but are incorrect relative to each other (*i.e.*, H$_2$ should have the largest pK_a).	Though it is unclear which of H$_2$ or H$_2$O are the weaker acid, the argument correctly relates the pK_a and base strength of the conjugate acids and bases, respectively, to their relative conjugate acid strengths.	The argument's discussion of base strength is vague and erratic, jumping from point to point. It is unclear which base the student believes to be strongest.	The argument describes how "Nu attack" will occur *unless* you have a strong base and "weak Nu". It is unclear why this reaction will occur nor what the alternative is.
Concept links	Acidstr: It is true that pK_a can be used to infer the relative strength of the conjugate acid.	pKa: It is true that pK_a can be used to infer the relative strength of the conjugate acid. basestr: it is true that conjugate acid strength can be used to infer the relative strength of the original base.	FC: formal charge is mentioned briefly at the end as justification for relative base strengths, but it is unclear why this idea (formal charge) is relevant. react: it is true that the strength of a base can be related to its overall reactivity (stronger base = more reactive base).	basestr: it is true that the strength of a base can be related to its overall reactivity (stronger base = more reactive base).
Comparisons	The pK_a values for each conjugate acid are explicitly provided in-text.	The argument explicitly compares the relative conjugate acid strength of bases A and C (though ultimately no meaningful comparison is made).	Similar to pK_a, the relative base strength for each claim are explicitly compared. Like conjugate acid strength, however, no meaningful comparison is made.	It is implied that the statement related to reactivity applies to all claims in this context.

formal charge	mechanism drawn	structure drawn

	y(g) - [basestr](e) - C to A,B	y(e) - A,B,C (gen)	y(g) - A,C
Concepts	It is true that NaH and NaOH both have formal negative charges on the non-Na portions of these molecules.	A general mechanism using “Base” is drawn, but this mechanism is canonically incorrect.	In the last sentence of the argument, NaOH and NaH are drawn with their correct formal charges.
Concept links	basestr: formal charge is mentioned briefly at the end as justification for relative base strengths, but it is unclear why this concept (formal charge) is relevant.	-	-
Comparisons	NaOH and NaH have negative charges, while NH_3 is neutral.	-	-

Mode of reasoning	Level of granularity	Level of comparison
Linear causal (with errors). Reactivity is justified by base strength. Evidence for base strength includes both conjugate acid strength and formal charge (briefly), with the former being inferred with pK_a values. However, the argument overall is vague and several concepts are discussed with error and seemingly heuristically.	pK_a values are a reaction-level concept. Reactivity, conjugate acid strength, and base strength are molecular-level concepts. Formal charge is an atomic-level concept.	**Fully compared**. All concepts presented in the argument are used to compare between all two or more possible claims.

Appendix B: Intended Acid–Base Learning Outcomes

Be able to:

Required skills

- Draw the mechanism of an acid–base reaction, given the starting materials
- Identify the acid, base, conjugate acid, and conjugate base, given the starting materials
- Deprotonate a given molecule
- Protonate a given molecule
- Draw the conjugate acid and conjugate base, given a molecule
- Estimate the pK_a value of a given molecule

Key concepts

- Apply the following ideas
    - The stronger the acid, the weaker its conjugate base (and vice versa)
    - An equilibrium favours the direction with the weaker (most stable) species (acid or base)
    - The lower the pK_a value, the stronger the acid
    - The following terms are synonymous: stronger = less stable = higher energy and weaker = more stable = lower energy

Applying required skills and key concepts to:

- Compare pK_a values of acids
- Compare relative stabilities of two species (e.g., bases or acids), analyzing the effects of the following chemical principles: electronegativity, atom size, resonance, hybridization, inductive effects, charge, solvent

In the following contexts:

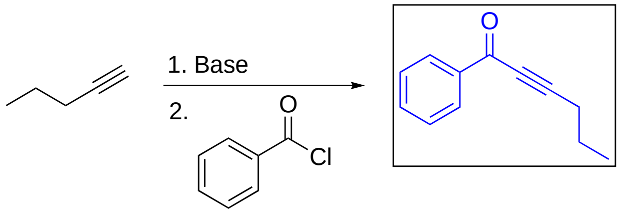
- Within a single molecule
- Between multiple molecules
- In an acid–base equilibrium

Integrate acid–base concepts in situations including:

- Identify the strongest acid and base that can exist in a given solvent
- Identify the predominant form of a compound at a given pH (Henderson-Hasselbalch equation)
- Draw (or select from a list) a base/acid that could quantitatively deprotonate/protonate a given acid/base

Note: For associated explanations and activities, see http://www.flynnresearchgroup.com/acid-base and/or the Acid–base module in https://orgchem101.com/

Appendix C: Expected Answer from Intended Acid–Base Learning Outcomes



1. Draw the major product of the reaction in the box above.
2. Circle the base below that can be used to force the equilibrium of the first step to the product side:

   NaOH NH_3 [✓ NaH]

3. Explain your answer in part b (why you chose one base **and** did not choose the others), using chemical structures as part of your answer.

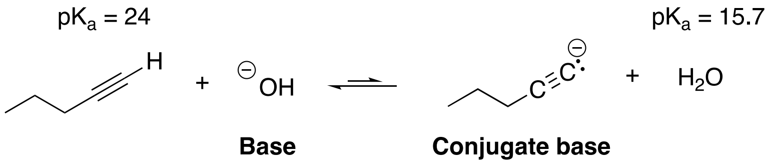
Relevant p*K*$_a$ values in H_2O:

NH_3 (NH_4^+): 9.2
NaOH (H_2O): 15.7
sp C–H: 24 (acetylene; estimated)
NaH (H_2): 35 (estimated*,* see *J. Chem. Soc.., Chem. Comm*., 1976, 648.)

Arguments

Expected (based on learning outcomes and expectations in that specific question): For one of the indicated bases to drive the acid–base equilibrium in the reaction illustrated above towards the products, it must be a less stable and stronger base than the carbanion. A base with a pK_a value of its conjugate acid greater than that of the alkyne (24) will be sufficiently strong. The pK_a values of the conjugate acids of NaH, NaOH, and NH_3 are 35, 15.7, and 9.2, respectively. Therefore, NaH is the only base capable of driving the equilibrium to products as it is the only base with a pK_a of its conjugate acid (35) greater than that of the alkyne (24). Using either NaOH or NH_3 would instead result in the equilibrium favouring reactants.

More granular causal argument (with chemical reasons as justification, molecular and atomic): Chemical properties and pK_a values together can be used to justify why NaOH and NH_3 will not be suitable bases to drive the acid–base equilibrium to products, and why H_2 will be suitable. These properties are used to explain the relative stability of the base (sodium hydroxide, ammonia, or sodium hydride) and conjugate base (acetylide).
The base and conjugate base involved in the equilibrium with sodium hydroxide have two competing chemical factors affecting their stability: the oxygen atom in the hydroxide (base) is more electronegative than the carbon atom in the conjugate base, which stabilizes the negative charge of the base more than the conjugate base. The electrons in the conjugate base are in an sp-hybridized orbital, which stabilizes the electrons more than the oxygen atom's electrons in the base that are in an sp^3-hybridized orbital (by virtue of being closer to the protons in the nucleus and therefore having a lesser effective negative charge). pK_a evidence indicates that the electronegativity factor dominates over the hybridization factor, with H_2O being a stronger acid (pK_a = 15.7) than the acetylene (pK_a = 24). The equilibrium favours the side with the weaker (more stable) species.

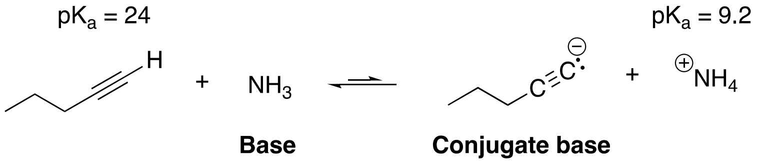


Electronegativity:
O is more electronegative than C
Oxygen atom is better stabilized than the negatively charged carbon atom (closer to the protons in the nucleus)

Hybridization:
C's electrons are in an sp-hybridized orbital, while O's electrons (base) are in an sp^3-hybridized orbital
Electrons are better stabilized in an sp-hybridized orbital as they are closer to the protons in the nucleus)

pK$_a$ values (experimental evidence) indicate that electronegativity dominates over hybridization in this case.

Ammonia will similarly not be a suitable base to drive the equilibrium to products. Hybridization factors stabilize the electrons on the C in the conjugate base more than on the N in the base (sp versus sp^3-hybrized, as above). However, both electronegativity and charge factors contribute to greater stability of the base than the conjugate base. The greater electronegativity of N (base) compared to C (conjugate base) stabilizes the base more than the conjugate base. Since charge decreases the stability of a species, the neutral nitrogen atom (base) is more stabilized than the negatively charged carbon atom (conjugate base). pK$_a$ evidence indicates that the electronegativity and charge factors dominate over the hybridization factor, with ammonium being a stronger acid (pK$_a$ = 9.2) than the acetylene (pK$_a$ = 24). The equilibrium favours the side with the weaker (more stable) species.



Electronegativity:
N is more electronegative than C.
Nitrogen atom is better stabilized than the negatively charged carbon atom (closer to the protons in the nucleus)

Charge:
Charge decreases the stability of a species, making the neutral N in the base more stable than the negatively charged C in the conjugate base.

Hybridization:
C's electrons are in an sp-hybridized orbital, while N's electr (base) are in an sp^3-hybridized orbital
Electrons are better stabilized in an sp-hybridized orbital as are closer to the protons in the nucleus)

pK$_a$ values (experimental evidence) indicate that electronegativity and charge dominate over hybridization in this case. The equilibrium favours the side with the weaker (more stable) species.

Sodium hydride will be a suitable base to drive the equilibrium to products. Hybridization/Orbital factors stabilize the electrons on the H in the base more than on the C in the conjugate base (s versus sp-hybrized). However, both electronegativity and atom size contribute to greater stability of the conjugate base. The greater electronegativity of C (conjugate base) compared to H (base) stabilizes the conjugate base more than the base. Because carbon is a larger atom than hydrogen, the larger atom can better disperse the electron density (and negative charge), stabilizing the conjugate base more than the base. pK$_a$ evidence indicates that the electronegativity and atom size factors dominate over the hybridization/orbital factor, with the acetylene being a stronger acid (pK$_a$ = 25) than the acetylene (pK$_a$ = 35). The equilibrium favours the side with the weaker (more stable) species, in this case, the products.

Appendix D: Rubric to evaluate arguments in terms of key concepts, reasoning, granularity, and/or comparisons

Key concepts	All concepts missing or discussed incorrectly	Some concepts missing or discussed incorrectly	All concepts present and discussed correctly	/w
Reasoning	No link between claim and evidence	Claim linked to evidence but no additional reasoning provided	Claim linked to evidence + causal reasoning provided	/x
Granularity	Does not meet expected level of granularity		Argument meets expected level of granularity	/y
Comparisons	No concepts used to compare between claims	Some concepts used to compare between claims	All concepts used to compare between claims	/z

Description of rubric and terms

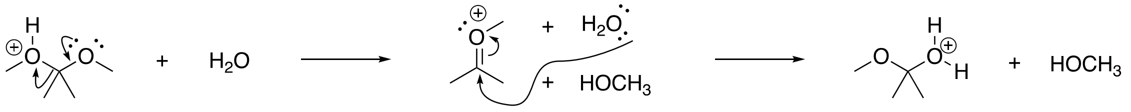
The rubric frames arguments in terms of one dimension focused on **content** (key words) and three dimensions focused on **structure** (reasoning, granularity, and comparisons.

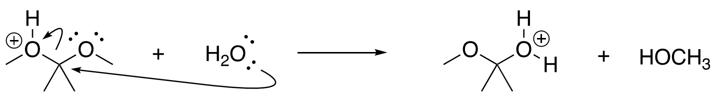
- Key concepts: the evidence, concepts, words, phrase units that students are expected to include in their argument
- Reasoning: how a student makes connections between their claim and evidence
- Granularity: the scalar level students reach in their arguments
- Comparisons: how a student compares between different possible claims using the concepts they have provided

Adapting the rubric to evaluate students' context-specific arguments

Users are welcome to use the rubric and its dimensions as a general framework to guide the creation of more task-specific rubrics. The following provides users with a list of examples of how the rubric can be adapted for different types of questions in chemistry. Users are welcome to include as many or as few dimensions of the framework to best suit their teaching goals.

Example 1: Comparing mechanisms

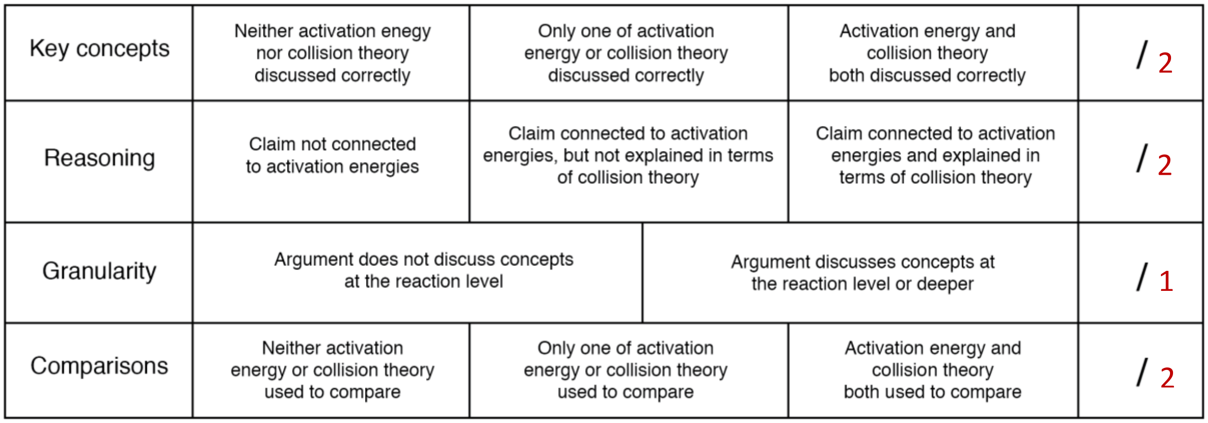




a. Which mechanism is most plausible for the reaction shown? ___________ **(1 point)**

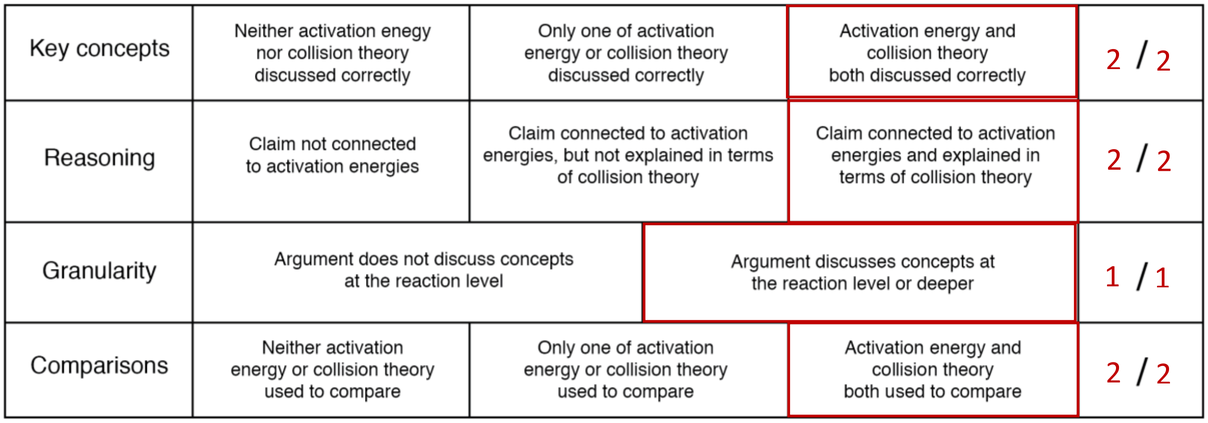
b. Justify your answer in part b with reference to activation energies and collision theory **(5 points)**

Adapted rubric for “Comparing mechanisms”



Key concepts	Neither activation enegy nor collision theory discussed correctly	Only one of activation energy or collision theory discussed correctly	Activation energy and collision theory both discussed correctly	/ 2
Reasoning	Claim not connected to activation energies	Claim connected to activation energies, but not explained in terms of collision theory	Claim connected to activation energies and explained in terms of collision theory	/ 2
Granularity	Argument does not discuss concepts at the reaction level		Argument discusses concepts at the reaction level or deeper	/ 1
Comparisons	Neither activation energy or collision theory used to compare	Only one of activation energy or collision theory used to compare	Activation energy and collision theory both used to compare	/ 2

Examples of how to use rubric to evaluate responses to “Comparing mechanisms”
Example Response 1: “A is more likely to proceed than B because it has a **lower activation energy than B**. Based on collision theory, A requires a lower activation energy because B proceeds through an **intermolecular reaction**, while A proceeds through an **intramolecular reaction**. It takes more energy for two molecules to collide and react than for one molecule to react, so A is more likely to proceed.”

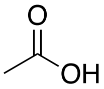


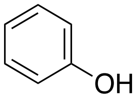
Key concepts	Neither activation enegy nor collision theory discussed correctly	Only one of activation energy or collision theory discussed correctly	Activation energy and collision theory both discussed correctly	2 / 2
Reasoning	Claim not connected to activation energies	Claim connected to activation energies, but not explained in terms of collision theory	Claim connected to activation energies and explained in terms of collision theory	2 / 2
Granularity	Argument does not discuss concepts at the reaction level		Argument discusses concepts at the reaction level or deeper	1 / 1
Comparisons	Neither activation energy or collision theory used to compare	Only one of activation energy or collision theory used to compare	Activation energy and collision theory both used to compare	2 / 2

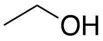
Student Response 2: “A is more likely to proceed than B because it has a **lower activation energy than B**. A requires a lower activation energy because A is less sterically hindered than B.”

Key concepts	Neither activation enegy nor collision theory discussed correctly	[✓ Only one of activation energy or collision theory discussed correctly]	Activation energy and collision theory both discussed correctly	1 / 2
Reasoning	Claim not connected to activation energies	[✓ Claim connected to activation energies, but not explained in terms of collision theory]	Claim connected to activation energies and explained in terms of collision theory	1 / 2
Granularity	Argument does not discuss concepts at the reaction level	[✓ Argument discusses concepts at the reaction level or deeper]		1 / 1
Comparisons	Neither activation energy or collision theory used to compare	[✓ Only one of activation energy or collision theory used to compare]	Activation energy and collision theory both used to compare	1 / 2

Example 2: Ranking organic acids







a. Rank the acids above from most acidic (1) to least acidic (3)

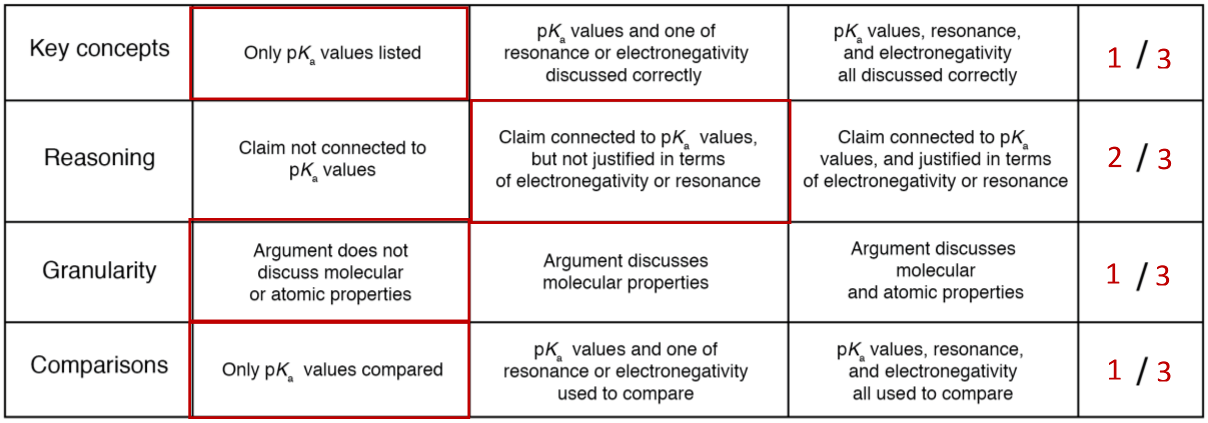
b. Justify your ranking with reference to pK_a values and properties of the conjugate bases.

Adapted rubric for "Ranking organic acids"

Key concepts	Only pK_a values listed	pK_a values and one of resonance or electronegativity discussed correctly	pK_a values, resonance, and electronegativity all discussed correctly	/ 3
Reasoning	Claim not connected to pK_a values	Claim connected to pK_a values, but not justified in terms of electronegativity or resonance	Claim connected to pK_a values, and justified in terms of electronegativity or resonance	/ 3
Granularity	Argument does not discuss molecular or atomic properties	Argument discusses molecular properties	Argument discusses molecular and atomic properties	/ 3
Comparisons	Only pK_a values compared	pK_a values and one of resonance or electronegativity used to compare	pK_a values, resonance, and electronegativity all used to compare	/ 3

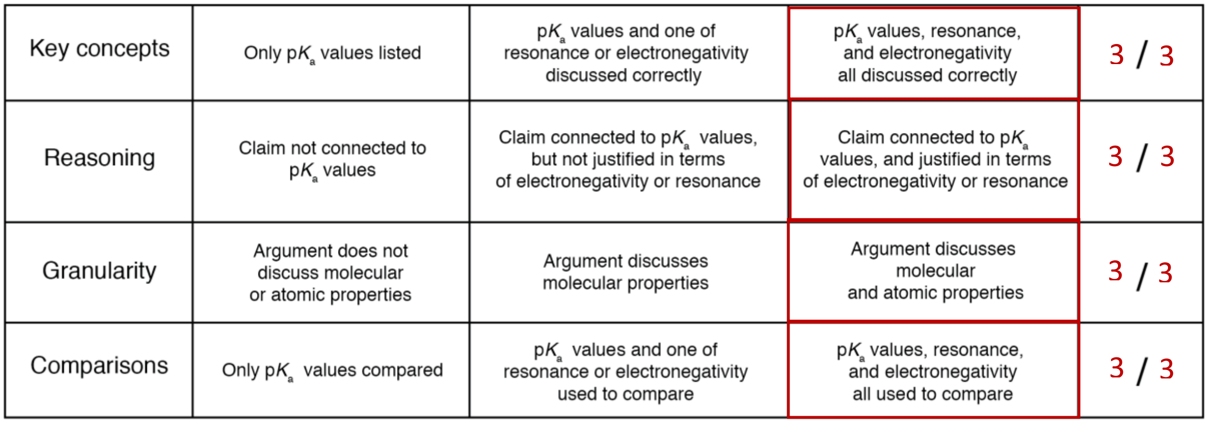
*Note that different course expectations may result in different rubrics. For example, other rubrics/takss expect students to discuss hybridization and inductive effects as part of their answer.

Examples of how to apply rubric to responses to "Ranking organic acids" question
Example Response 1: "Ethanoic acid is the strongest acid because it has the **lower pK_a value** (4.76), while phenol and ethanol have progressively **higher pK_a values** (10 and 16, respectively). The higher the pK_a value, the weaker the acid."



Key concepts	[✓ Only pK_a values listed]	pK_a values and one of resonance or electronegativity discussed correctly	pK_a values, resonance, and electronegativity all discussed correctly	1 / 3
Reasoning	Claim not connected to pK_a values	[✓ Claim connected to pK_a values, but not justified in terms of electronegativity or resonance]	Claim connected to pK_a values, and justified in terms of electronegativity or resonance	2 / 3
Granularity	[✓ Argument does not discuss molecular or atomic properties]	Argument discusses molecular properties	Argument discusses molecular and atomic properties	1 / 3
Comparisons	[✓ Only pK_a values compared]	pK_a values and one of resonance or electronegativity used to compare	pK_a values, resonance, and electronegativity all used to compare	1 / 3

Example Response 2: "Ethanoic acid and phenol are stronger acids than ethanol because both of their conjugate bases exhibit **resonance**. Bases in which negative charge is delocalized are more stable, weaker bases, making their conjugate acids stronger. Ethanoic acid is a stronger acid than phenol because its negative charge is distributed primarily across two oxygen atoms, while phenol's negative charge is distributed across oxygen and carbon atoms. Oxygen atoms are better able to stabilize negative charge because they are more **electronegative** than carbon atoms. The pK_a also supports these conclusions, as ethanoic acid has the **lowest pK_a** (strongest acid), while ethanol has the **highest pK_a** (weakest acid)."



Key concepts	Only pK_a values listed	pK_a values and one of resonance or electronegativity discussed correctly	[✓ pK_a values, resonance, and electronegativity all discussed correctly]	3 / 3
Reasoning	Claim not connected to pK_a values	Claim connected to pK_a values, but not justified in terms of electronegativity or resonance	[✓ Claim connected to pK_a values, and justified in terms of electronegativity or resonance]	3 / 3
Granularity	Argument does not discuss molecular or atomic properties	Argument discusses molecular properties	[✓ Argument discusses molecular and atomic properties]	3 / 3
Comparisons	Only pK_a values compared	pK_a values and one of resonance or electronegativity used to compare	[✓ pK_a values, resonance, and electronegativity all used to compare]	3 / 3

Example Response 3: "Ethanoic acid and phenol are stronger acids than ethanol because both of their conjugate bases exhibit **resonance**. Bases in which charge is delocalized are more stable, weaker bases, making their conjugate acids stronger. Ethanoic acid is a stronger acid than phenol because its negative charge is distributed primarily across two oxygen atoms. The pK_a also supports these conclusions, as ethanoic acid has the **lowest pK_a** (strongest acid), while ethanol has the **highest pK_a** (weakest acid)."

Key concepts	Only pK_a values listed	[✓ pK_a values and one of resonance or electronegativity discussed correctly]	pK_a values, resonance, and electronegativity all discussed correctly	2 / 3
Reasoning	Claim not connected to pK_a values	Claim connected to pK_a values, but not justified in terms of electronegativity or resonance	[✓ Claim connected to pK_a values, and justified in terms of electronegativity or resonance]	3 / 3
Granularity	Argument does not discuss molecular or atomic properties	[✓ Argument discusses molecular properties]	Argument discusses molecular and atomic properties	2 / 3
Comparisons	Only pK_a values compared	[✓ pK_a values and one of resonance or electronegativity used to compare]	pK_a values, resonance, and electronegativity all used to compare	2 / 3