Title

Problem solving in organic chemistry and undergraduate research: Characterizing and catalyzing the transition from novice to expert

Author

Helix, Max Ryan

Publication Date

2021

Peer reviewed|Thesis/dissertation

Problem solving in organic chemistry and undergraduate research:
Characterizing and catalyzing the transition from novice to expert

By

Max R Helix

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Science and Mathematics Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Anne M. Baranger, Chair
Professor Angelica M. Stacy
Professor Richmond Sarpong

Spring 2021

Abstract

Problem solving in organic chemistry and undergraduate research:
Characterizing and catalyzing the transition from novice to expert

by

Max R Helix

Doctor of Philosophy in Science and Mathematics Education

University of California, Berkeley

Professor Anne M. Baranger, Chair

I have created a general workflow that describes how students reason through non-trivial organic predict-the-product problems. This was accomplished through an iterative process that combined the experiences of instructors with holistic impressions of think-aloud interviews, incorporating feedback from both undergraduate focus groups and graduate students. The workflow serves as both a potential instructional tool and a model for student thinking. An analysis of think-aloud interview data showed that the workflow describes both undergraduate and graduate student thought processes. Successful and unsuccessful problem solvers did not differ in which problem-solving actions they took. However, the successful problem solvers were more likely to name relevant functional groups, which is a simple and concrete action that can be recommended to students. Graduate student approaches were not qualitatively different than those of undergraduate students, but an increase in focus on deciding between multiple pathways suggests that more expert-like practitioners may frame the problem in a different way. If we want to help students to utilize more high-level reasoning when predicting organic reactivity, they need to be exposed to more situations in which multiple reasonable solutions are possible.

Additionally, I have developed and implemented two organic chemistry lessons, one on the broad topic of acid-base chemistry, and another on the narrower topic of directing group effects in electrophilic aromatic substitution. These lessons were designed based on the Preparation for Future Learning (PFL) framework, which involves students collectively exploring data to find contrasting cases and "invent" chemical principles. Student performance was measured by pre-tests, immediate post-tests, and scores on relevant exam questions. While the acid-base lesson did not result in an immediate benefit relative to lecture alone, there appeared to be a delayed effect. Students who attended the PFL lesson scored significantly higher on acid-base questions on the final exam, even though this was not the case on the midterms. Assessment on directing group effects suffered from a ceiling effect, making it difficult to draw any conclusions about that lesson. Student feedback on both lessons was overwhelmingly positive. In general, students who attended the lessons felt much more prepared for when the material covered in these lessons was subsequently introduced in lecture. Students also enjoyed the overall format of working with the data to discover trends. Overall, PFL lessons show promise as a useful and active way to familiarize students with various chemical principles prior to their "official" introduction in class.

Understanding the impact of undergraduate research experiences (UREs) and course-based undergraduate research experiences (CUREs) is crucial as universities debate the value of allocating scarce resources to these activities. I have designed and tested the BURET instruments, a new set of tools designed to assess the learning outcomes of UREs and CUREs in the sciences. To study the BURET instruments, they were administered to 89 undergraduate students, and the performance of students who had less than one year of undergraduate research was compared to those with more than one year of research experience. Students were assessed on four primary dimensions based on written reflections and poster presentations for their own research project: communicating the significance of their project, analyzing their experimental design, interpreting their data, and proposing future research. The instruments were found to yield reliable scores and helped clarify the impacts of undergraduate research, providing insight into the strengths and weaknesses of undergraduate researchers at this institution. Students with at least a year of research experience were able to use disciplinary evidence more effectively than those with less than one year of experience. Novice students excelled at explaining the societal relevance of their work, but they incorporated only minimal discussion of prior research into their reflections and presentations. Students at all levels struggled to critique their own experimental design. These results have important implications for undergraduate learning, suggesting ways for faculty members, graduate student research mentors, and CURE or URE programs to optimize undergraduate research experiences.

TABLE OF CONTENTS

For my Mom and Dad

ACKNOWLEDGEMENTS

Graduate school is not a solo pursuit. It took the combined efforts of dozens of people, maybe hundreds, to get me across the finish line. There will inevitably be some unfortunate omissions in the following list of supporters, but I would like to thank everyone who has contributed to my education, research, and general survival over the last 7 years.

I'll start by thanking my research group, which is made up *exclusively* of helpful, kind, intelligent, and amazing people. One of the strengths of our research group is the wealth of experience and insight offered by the various lecturers who participate in our group meetings, particularly Michelle Douskey, Alexis Shusterman, MaryAnn Robak, and Terry Johnson. I appreciate all the helpful advice and interesting ideas over the years. Thanks to the grad students ahead of me (Geri Kerstiens, Erin Palmer, Sara Tischhauser, and Beatriz Brando), who gave me endless advice on how to successfully traverse the SESAME program and its often nebulous requirements. Thank you also to Laura Armstrong for figuring out and getting through the dissertation writing process with me. And thank you to Angy Stacy, for providing guidance on my research and agreeing to sit on my committees, even though some of that work was post-retirement.

Many members of our research group provided not just friendship and general support, but actively worked to make my research projects into a reality. Thank you to Katie Blackford for providing substantial intellectual contributions to my organic projects and making sure I didn't make any silly chemistry mistakes. Thanks also to the undergraduates who worked with me (Zach Firestein, Katia Gibson, and Julia Greenbaum), who helped me transcribe and code excessive amounts of data. On the BURET project, thanks especially to Laleh Coté, who worked tirelessly on organization, coding, lit searches, and more, making sure that the full weight of this enormous project wasn't fully resting on my shoulders. Thanks also to Chrissy Stachl and Elisa Stone for their central contributions to the project. Thank you to Marcia Linn for helpful feedback in the planning and writing stages, and thanks to the undergraduates (Jiho Kim, Seth van Doren, and Yongbo Wang) who worked with various aspects of our data.

Beyond my research group, I was helped by a wide variety of other Berkeley people. Thank you to Richmond Sarpong and Michael Ranney for being on my committees and giving me valuable feedback from two very different areas of expertise. Thanks to Kate Capps and Rosa Garcia for helping me navigate all the administrative work required to get my degree. Thank you to Lloyd Goldwasser for facilitating my introduction to the field of education. I also want to thank my doctors (Christina, Kat, and Alison), who seem to have finally found a solution to the issues that have plagued me throughout the past 15+ years. Thank you to my SESAME cohort (Elena Durán, Anna Weltman, and Anna Zarkh), who helped me get through my first year of the program and beyond. And thank you to the hundreds of undergraduate students who volunteered to participate in my research, even though most of you were never compensated in any way.

While on the subject of students, I was blessed to have the opportunity to work with some amazing ones through my involvement with the ChemScholars program. Research was hard work, but teaching kept me motivated to continue. There are too many to name, but thank you to everyone for reminding me why I was doing this in the first place. And thank you to the ChemScholars leaders (Michael Curtis, Ryan Khalaf, Tahoe Fiala, Franco Faucher, Sarah Buzsaki, and Tanya Lazarus) and administrator (Maura Daly), who allowed me to barge my way in and help out.

Although most of my close friends during this period of my life were the aforementioned graduate students I worked with, I also had the opportunity to live with some wonderful people that helped remind me that my Ph.D. program was not the entirety of life on Earth. Thank you to Hannah Pfahl, Amara Pfahl, Ariella Powers, Hugo Martinez-Bernardino, and others for putting up with my messes and providing extra care when I had major deadlines to meet. Thanks also to honorary housemates like Monica Cereseto, Cara Spangler, Clint Brody, and Jill Tokutomi.

I especially want to thank Heather Thams, who has been my oldest and closest friend over these many years. Even though we weren't close in terms of physical distance, she has never been far from my thoughts, and I would not have made it through without her emotional support. I also want to thank my cat, Jojo, for being my constant companion for 18 years. He didn't quite make it to see the completion of my Ph.D., but nearly all the work that went into it was done with him curled up by my side. I miss him every day.

Thank you to my advisor, Anne Baranger, who gave me a bewildering amount of freedom to pursue my interests, allowing me to create new research projects even though neither of us really knew where they might end up. Thanks for patient guidance on all of my research and degree requirements, and for never once complaining that I wasn't progressing fast enough, even when it was warranted. I could not have made it through a doctoral program with just any advisor, so I am exceptionally grateful to have been able to work with you.

Finally, thank you to my parents. They have provided me with constant support throughout my entire life, from my best moments to my worst. There's no period of my life in which they didn't play a major role. They have always been there for me, and I am extremely lucky. I love you both, and I dedicate this dissertation to you.

PROBLEM SOLVING IN ORGANIC CHEMISTRY AND UNDERGRADUATE RESEARCH: CHARACTERIZING AND CATALYZING THE TRANSITION FROM NOVICE TO EXPERT

As educators, we are tasked with helping relative novices develop into more expert-like practitioners. Although this process was once envisioned as the simple transmission of information to a blank slate, decades of educational research have shown that effective teaching requires a more collaborative effort between instructor and student. A successful educator will help their students build new knowledge on top of existing mental structures and provide guidance on how to apply that knowledge to solve domain-specific problems. In the classroom, problem-solving skills often take a back seat to content knowledge, in part because of the sheer volume of material that students are expected to learn. However, another contributing factor is that the problem-solving process is not as well understood, and there are no “one size fits all” approaches to answering some of the more important types of questions. Complicating matters further is the fact that experts may solve problems in ways that are inaccessible to novices, due to their lack of familiarity with the problem type (Bodner, 2003). Understanding how both novices and experts solve problems and developing ways to guide students toward more successful and more expert-like strategies is a common thread that links together the research projects described in this dissertation.

Expertise and Problem Solving

A central feature of expertise is the ability to recognize meaningful patterns of information (Bransford et al., 2000a). For example, expert chess players will perceive certain configurations of pieces by how they are functionally or strategically related, whereas such “chunks” are not generally noticed by inexperienced players. Such expert pattern recognition has been documented across a variety of domains, including chess (deGroot, 1965), electronic circuitry (Egan & Schwartz, 1979), math (Hinsley et al., 1977), and even teaching (Sabers et al., 1991). In fact, it has been proposed that the nebulous concept of expert intuition is nothing more than the recognition of familiar combinations of cues, often below the level of conscious awareness (Kahneman & Klein, 2009).

This sort of recognition plays a key role in problem solving. The way in which a problem is initially perceived substantially influences how a solution begins to develop in the mind of the solver. This has been most extensively researched in the area of physics problems, often using card sort tasks, in which both novices and experts are asked to sort problems into categories (Chi et al., 1981; Mason & Singh, 2011; Snyder, 2000). It has consistently been found that experts base their categories on the underlying physics principles that are most relevant to reaching a solution, whereas novices tend to focus on surface features, such as “block on an inclined plane” (Chi et al., 1981). Similar findings have been reported with general chemistry problems (Krieter et al., 2016). Card sort studies in organic chemistry have also been conducted, and students vary in whether they sort reactions based on structural features or underlying mechanisms (Galloway et al., 2019).

Problem Solving in Organic Chemistry

Investigating how students initially perceive a problem and identify meaningful chunks of information is one important focus of the work presented in Chapter 1. For example, whether a student identifies the first problem they are asked to solve as an acetal hydrolysis followed by a Horner-Wadsworth-Emmons reaction has a large effect on their eventual success. Such students are likely to be much more successful than those who do not explicitly identify the structural patterns corresponding to those mechanistic pathways. This work investigates key differences in what more novice and more expert-like students notice when problem solving.

More generally, Chapter 1 attempts to fully characterize student problem solving in the context of complex predict-the-product problems. I have chosen to focus on these problems because of the ubiquity of this type of question in organic chemistry and its potential to allow students to participate in open-ended mechanistic reasoning. A detailed understanding of how different students approach these problems would be valuable to us as instructors. Breaking down student thought processes can help identify precisely where problems arise and may suggest what types of scaffolding would be most appropriate for students who are first learning to answer these questions. An investigation of graduate student reasoning provides an expert-like “target” that novices can be encouraged to move towards.

Preparing to Effectively Learn New Concepts

When students are first starting their year of introductory coursework on organic chemistry, they are faced with many different problem types that are often not explicitly asked later. These questions test their mastery of the basic skills that will later help them solve more fundamental problems like predicting reactivity, drawing mechanisms, and proposing syntheses. For example, students are asked to draw resonance forms, not because it is an inherently useful activity, but because it is a skill that could eventually help them answer a mechanism question or rationalize an experimental observation. One issue that arises is that novice students compartmentalize this knowledge to the portion of the class in which it is learned, reducing their ability to transfer it to new contexts later (Bransford et al., 2000c). They can recall a fact or procedure when explicitly asked to do so, but when that same fact or procedure would be implicitly useful to solve a different type of problem, the knowledge is not activated. Conditionalized knowledge, or knowledge that specifies the contexts in which it is useful, is another commonly observed trait of expert thinking (Bransford et al., 2000a; Simon, 1980).

How can we teach concepts in such a way that they will be recalled and applied when they are most relevant, generally in the context of more complex problems? The Preparation for Future Learning (PFL) framework suggests that certain types of exploratory activities may prime students to see problems through the lens of the material being taught, thus activating knowledge when it is most applicable (Bransford & Schwartz, 1999). These activities take place before the student has formally “learned” the material through lecture or text, hence the name. To determine whether such an approach would work in organic chemistry, two PFL lessons were designed around the key topics of acid-base chemistry and directing group effects in electrophilic aromatic substitution. The development and implementation of these lessons are discussed in Chapter 2.

Expertise and Learning in Scientific Research

Developing expertise is not something that only happens in the classroom. Students who wish to become scientific researchers often apprentice with a faculty member or graduate student to gain hands-on knowledge of scientific practices and a greater understanding of the overall research process. Research mentors are also attempting to help develop a more expert-like student, but unlike classroom instructors, they often don’t assess student progress in a consistent way. One learning outcome that some researchers have tried to measure is experimental design ability (A. P. Dasgupta et al., 2014; J. A. Harsh, 2016a; Sirum & Humburg, 2011a). In general, these instruments require the student to transfer what they have learned in their individualized research experiences to a standard hypothetical research scenario.

My collaborators and I were interested in creating instruments that would not require as much transfer, in which a student is given the opportunity to exhibit their knowledge around experimental design and the research process more generally, but in the context of their own research project. Characterizing the differences between novice and advanced students and having a way to identify areas of strength and weakness is the first step toward making evidence-based improvements in mentoring practices. The Knowledge Integration (KI) framework was used to help develop instruments to assess this sort of expertise. KI postulates that as students become more experienced, they start to make deeper and more numerous links between domain-specific concepts, eventually reaching a fully integrated understanding of the field. Chapter 3 of this dissertation focuses on the development and validation of these instruments.

Taken together, the research described in the following chapters sheds light on how novices differ from experts in both organic chemistry and undergraduate research. A particular focus is placed on problem solving, which ranges from small questions like “which molecule is more acidic?” to large questions like “how do I develop a novel research project that contributes to my scientific field?” Based on the results obtained, a variety of suggestions are made about how to expedite the process of developing expertise, and methods for assessing this progress are presented. Evidence-based practices already inform the way that we conduct scientific research, and the work presented here will help provide well-supported methods for improving our teaching and mentoring as well.

CHAPTER 1

Characterizing student problem solving and development of a general workflow for predicting organic reactivity

I have created a general workflow that describes how students reason through non-trivial organic predict-the-product problems. This was accomplished through an iterative process that combined the experiences of instructors with holistic impressions of think-aloud interviews, incorporating feedback from both undergraduate focus groups and graduate students. The workflow serves as both a potential instructional tool and a model for student thinking. An analysis of think-aloud interview data showed that the workflow describes both undergraduate and graduate student thought processes. Successful and unsuccessful problem solvers did not differ in which problem-solving actions they took. However, the successful problem solvers were more likely to name relevant functional groups, which is a simple and concrete action that can be recommended to students. Graduate student approaches were not qualitatively different than those of undergraduate students, but an increase in focus on deciding between multiple pathways suggests that more expert-like practitioners may frame the problem in a different way. If we want to help students to utilize more high-level reasoning when predicting organic reactivity, they need to be exposed to more situations in which multiple reasonable solutions are possible.

INTRODUCTION

As instructors of organic chemistry, we aspire to help each of our students develop a foundational understanding of the subject. Learning organic chemistry is more than just memorizing a long list of facts, much to the surprise of many undergraduates. On a deeper level, the facts of organic chemistry are organized around a highly interconnected set of concepts that students need to internalize. However, the primary purpose of this internalization is to be able to *use* these concepts in the context of problem solving. In organic chemistry, assessment of student knowledge relies heavily on just a few types of questions, all of which correspond to authentic questions that practicing chemists must routinely answer:

- How do I make that? (Synthesis)
- What will happen when I mix these? (Predict-the-product)
- What did happen when I mixed those? (Spectroscopy)
- How did it happen? (Mechanism)
- Why did it happen? (Rationalize experimental observation)

Of these questions, 3 are given special priority on organic exams: predict-the-product, mechanism, and synthesis. An analysis of two years of second-semester organic chemistry exams at this institution showed that these three question types (and their combination into “roadmap” problems) accounted for 74% of the possible points, with an additional 16% accounted for by “rationalize this observation” questions. Note that this does not include the laboratory portion of the course, which has its own set of fundamental questions. Because there are so few of these central questions, an important learning outcome for organic chemistry students is to develop general strategies for approaching each one. Focusing on how the same problems manifest with different types of reactivity throughout the year provides a common thread that students often fail to see.

A detailed understanding of how different students approach these problems would be valuable to instructors. More specific descriptions of student thought processes help identify precisely where problems arise and suggest types of scaffolding that would be most appropriate for students who are first learning to answer these types of questions. In-depth studies of how more successful students are approaching problems reveal strategies that can be conveyed to other students as well. Additionally, investigations of graduate student reasoning provide an expert-like “target” that novices can be encouraged to move towards.

Problem Solving in Organic Chemistry

Researchers have taken two main approaches to investigating how students reason about the central questions in organic chemistry. One method is to take advantage of the large pool of data produced when students turn in exams and problem sets as part of their coursework. Some researchers have used this to identify strategies that more successful students are more likely to use, such as atom mapping and clearly identifying bonds to be formed during synthesis questions (Bode & Flynn, 2016; Flynn & Featherstone, 2017). However, only a small fraction of student thinking is captured by these written artifacts. Another approach that provides a more detailed description of student thought processes is to use think-aloud interviews (Bowen, 1994), in which students are instructed to vocalize their thoughts as they have them while attempting to solve problems. Think-aloud interviews and protocol analysis are well-established research methods for studying problem solving in a variety of disciplines (Charters, 2003; Ericsson & Simon, 1980, 1993; Fonteyn et al., 1993).

Several types of organic chemistry questions have been investigated with think-aloud protocols, such as synthesis (Flynn, 2014), spectroscopy (Cartrette & Bodner, 2010), and acid-base (Petterson et al., 2020), but the most heavily researched are problems in which the student is instructed to provide a reasonable mechanism for a given organic transformation (Bhattacharyya, 2014; Bhattacharyya & Bodner, 2005; Caspari et al., 2018; Ferguson & Bodner, 2008; Kraft et al., 2010; Weinrich & Sevian, 2017). Complete answers to these questions involve using the electron-pushing formalism (EPF), or "arrow-pushing," to show the flow of electrons throughout the reaction. Students at both the undergraduate and graduate levels commonly take a means-end analysis or difference reduction approach to solving mechanism problems in which the final product is given. In a study involving graduate students, solvers focused almost exclusively on steps that "get me [closer] to the product" (Bhattacharyya & Bodner, 2005). Students at various levels tend to start problems of this type by mapping atoms between the starting material and the product, even when this was not an explicit strategy introduced during their coursework (Bhattacharyya, 2014; Ferguson & Bodner, 2008).

This focus on "getting to the product" appears to be very robust, leading to some surprising results. DeCocq and Bhattacharyya (2019) conducted interviews in which students were asked to propose the product of a single mechanistic step, and later in the interview were given the same exact step but also shown the overall transformation it is a part of. Even after correctly predicting the product of that step earlier, most students changed their answers to something less correct when shown the full transformation for certain problems. An important conclusion from these investigations is that students' reasoning depends dramatically on what initial information they are given, suggesting that traditional mechanism questions may not fully capture student abilities to engage in more open-ended mechanistic reasoning.

Predict-the-product (PtP) is another major problem type that has been studied via think-aloud interviews. Although students are not always asked to provide a mechanism along with their predictions, mechanistic reasoning is a central tool for generating and justifying predictions. However, as suggested by DeCocq and Bhattacharyya (2019), the reasoning that occurs on these questions may look quite different from mechanism questions in which the final product is given. Of the existing publications investigating student reasoning on open-ended PtP questions, many use either particularly simple transformations (Grove, Cooper, & Cox, 2012; Grove, Cooper, & Rush, 2012) or a limited subset of reaction types (Cruz-Ramírez de Arellano & Towns, 2014; Finkenstaedt-Quinn et al., 2020), although this is not exclusively the case (Webber & Flynn, 2018). The work reported here examines student thinking on complex, potentially ambiguous PtP questions that cover a range of different reaction types.

Because students will often not draw mechanisms for PtP questions, or will draw one only after making a prediction, some researchers have argued that the EPF is an exercise in symbol manipulation and that these arrows hold no physical meaning for the students (Bhattacharyya & Bodner, 2005; Grove, Cooper, & Rush, 2012). For example, in the Grove et al. studies, student use of the EPF when solving PtP problems was sparse, and for the simplest problems, it did not appear to be significantly helpful unless the problem involved an intramolecular step (Grove, Cooper, & Cox, 2012). However, little is known about student thinking when approaching PtP problems that involve more complex molecules and potentially longer reactive pathways. It seems probable that students may engage in more explicit mechanistic reasoning when approaching scenarios that cannot be matched exactly to the canonical reactions they have memorized.

While the major problem types in organic chemistry have been studied extensively through a variety of theoretical lenses, few studies attempt to model the overall flow of student reasoning. One exception is Bhattacharyya's meta-analysis (2014) that proposes a model for how students work through mechanism problems. One potential way to solve a mechanism problem would be to begin with the starting material and reason mechanistically in the forward direction until reaching the desired product. However, this is not how mechanism problems are solved according to this model. Instead, students map the reactant onto the product and look for key differences to identify what type of reaction is occurring. Bhattacharyya proposes a dual path model depending on whether the reaction is a single-step canonical reaction or a multi-step functional group transformation. On both of these paths, EPF arrows are not filled in until after the intermediates or other key reaction elements are already drawn.

Evidence already exists that mechanistic reasoning looks different depending on whether the ultimate product is known (DeCocq & Bhattacharyya, 2019). Therefore, a model of how students work through open-ended predict-the-product problems may look quite different than a model for mechanism questions. Developing such a model is the primary focus of this work. In particular, I focus on relatively complex problems that are difficult to answer through a purely memorization-based approach.

Theoretical Frameworks

Researchers have been attempting to create general models of human problem solving for decades. One of the earliest was Polya's four-stage model, which involves defining or understanding the problem, making a plan, implementing the plan, and reflecting on the implementation (Polya, 1945). Many other more specific models followed, but most of them included the presence of these same basic stages. An early example in chemistry is Bunce et al.'s explicit method of problem solving (EMPS) model, which focuses on mathematical problems in general chemistry (Bunce et al., 1991). The basic steps are identifying what information is given and what is asked for (i.e., defining the problem), recalling relevant rules and equations, making a schematic plan, implementing the plan mathematically, and reviewing the overall solving process.

One example of a qualitatively different model for how people engage in problem solving is the anarchistic model proposed by Bodner (2003), based on earlier work by G.H. Wheatley (Wheatley, 1984):

- Read the problem
- Now read the problem again
- Write down what you hope is the relevant information
- Draw a picture, make a list, or write an equation or formula to help you begin to understand the problem
- Try something
- Try something else
- See where this gets you
- Read the problem again
- Try something else
- See where this gets you
- Test intermediate results to see whether you are making any progress toward an answer
- Read the problem again
- When appropriate, strike your forehead and say, "son of a …"
- Write down *an* answer (not necessarily *the* answer)
- Test the answer to see if it makes sense
- Start over if you have to, celebrate if you don't

Under some circumstances, a model with very distinct stages might seem most appropriate, but most of us also recognize in the anarchistic model an accurate description of some of our own problem solving.

The type of model that most accurately describes a given instance of problem solving is closely related to the concept of routine *exercises* and novel *problems* (Bodner, 2003). Any given chemistry question can potentially be either an exercise or a problem, depending on who is solving it. The difference between an exercise and a problem lies not in how difficult the question is, but rather in how *familiar* the solver is with the material and question type. For example, a practicing analytical chemist would be able to solve a titration question by recalling algorithms for getting to a solution, whereas a freshman taking general chemistry for the first time would likely take a longer, more circuitous approach to solving the same problem. The question is an exercise for the experienced chemist but a problem for the student. Using this framework, it is proposed that linear models with distinct stages are accurate descriptions of how chemists solve exercises, including worked examples in textbooks and lectures. However, for a true problem, a more anarchistic model would be more appropriate (Bodner, 2003). In this work, it is assumed that some students will recognize "what is happening" on a given predict-the-product question, whereas others will need to try a variety of approaches before moving closer to a solution. My analysis attempts to characterize both types of solving.

Research Questions

The goal of this work is to characterize the problem-solving approaches students use when presented with relatively complex predict-the-product questions. This type of problem was chosen due to its centrality in organic chemistry courses and the opportunity for students to engage in open-ended mechanistic reasoning. Characterizing the different actions that students take when problem solving is useful for both instructors and students. As instructors, knowing the specific components of problem solving used by students would allow us to identify areas of difficulty and tailor interventions to assist them. From the student's point of view, understanding the components of problem solving would aid in the metacognitive regulation of the student's own thought processes, which is a key attribute of effective problem solving (Schoenfeld, 1987).

To achieve this goal, I interviewed undergraduate students who had completed their first year of organic chemistry using a think-aloud protocol in which they were asked to solve non-trivial predict-the-product problems. Major problem-solving actions were identified, and a model was developed to capture student reasoning processes. This model was then developed into a workflow for how to predict organic reactivity.

Ultimately, we want our students to move over time to a more successful, expert-like thought process when problem solving. Student problem-solving actions were quantified and compared to identify differences between the approaches of students who were able to reach a reasonable solution and those who were not. Additionally, "key features" were identified that appeared more often in successful solution pathways. Graduate students specializing in organic chemistry were then interviewed to gain insight into more expert-like approaches.

This chapter seeks to answer the following research questions:

1. What are the common characteristics of problem solving in the context of non-trivial predict-the-product problems?
2. Does the workflow model I developed accurately reflect student problem solving?
3. What differentiates successful from unsuccessful problem solvers?
4. What differences are there between the approaches of sophomore undergraduates and more experienced organic graduate students?

Problem Design

The problems used in this study are shown in Figure 1. The first question involves the hydrolysis of a cyclic acetal, followed by reaction of the aldehyde intermediate with a Horner-Wadsworth-Emmons (HWE) reagent. The resulting product can then potentially undergo an intramolecular oxa-Michael addition, reforming a six-membered ring. The second question is an acid-catalyzed cycloaddition, which could plausibly proceed through either a concerted or stepwise mechanism. The substrates in the third problem are set up to undergo a Mannich reaction, though an amine-catalyzed intramolecular aldol reaction would also be a reasonable response. In the final problem, the conditions mimic those used in electrophilic aromatic substitution reactions, but bromination is likely to occur first on the more reactive alkene substituent. To ensure that all students would have time to work on each problem during the hour-long interview, the number of problems was capped at four.

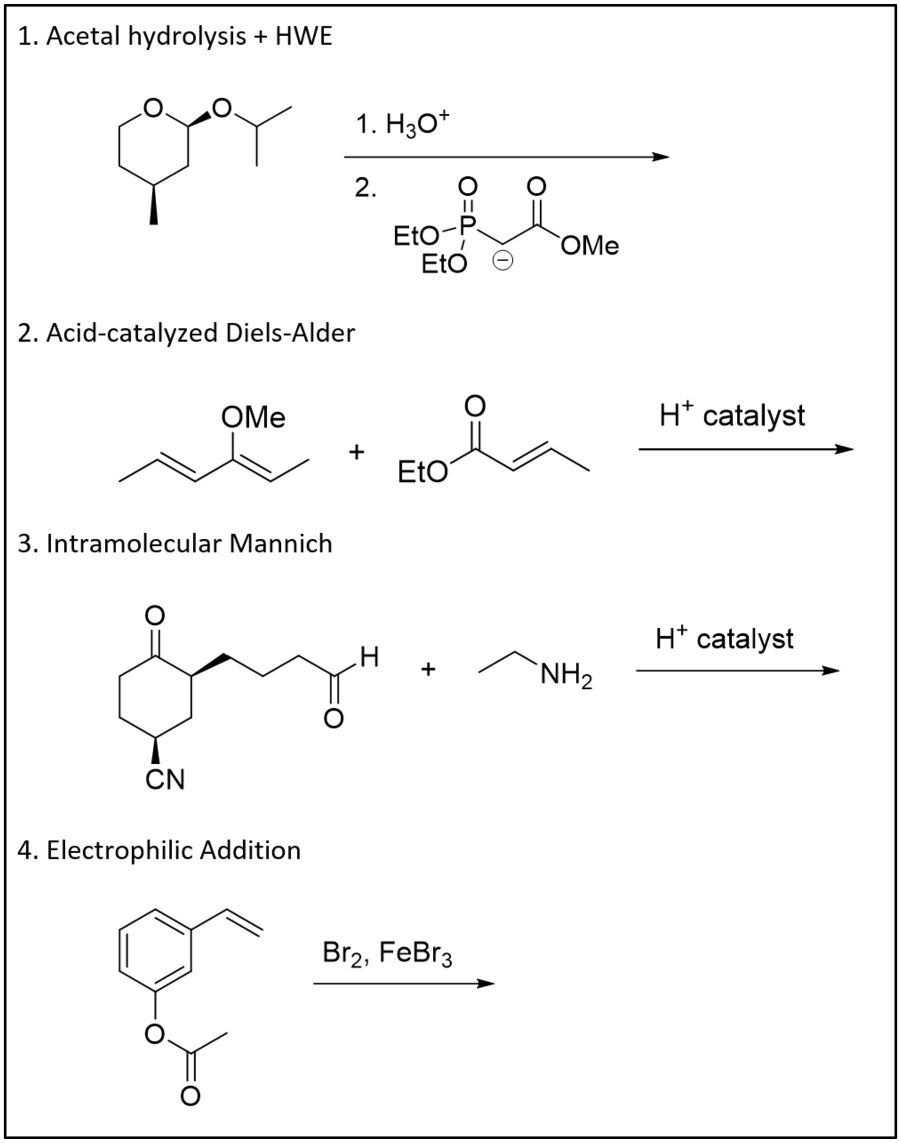


Figure 1. Problems used in think-aloud interviews. Answers that were considered to be correct for the purposes of this study can be found in the Results section (Figures 3-6).

A primary goal when designing these problems was to make the questions function as problems for most students, rather than exercises. I was interested in probing how students think when pushed beyond questions that could be solved by a purely memorization-based approach. To achieve this goal, various aspects of the problems were designed to be potentially ambiguous, and elements were included that might make the reactions seem unfamiliar to the participants. However, specialized reagents were mostly avoided so that students would be able to reason through some reactivity even if they forgot what a given reagent typically does. All of these problems have the potential to be difficult, but it is unlikely that students will have no ideas on how to start working through them.

Multiple types of ambiguity were included in the design of these questions. One source of ambiguity is that there is more than one reasonable answer or solution pathway for each problem. It was hypothesized that as students gain experience, they are more likely to discuss multiple competing pathways, weighing their relative likelihoods against each other. If there is a clear, unambiguous answer to all problems, this type of reasoning would not be observed. An additional source of ambiguity is that no conditions (solvents, temperature, equivalents) are listed, which is typical of how these types of problems are presented to students in their coursework. Whether students at different levels of experience make assumptions about conditions, explicitly ask for clarification, or ignore them entirely is an open question. In particular, the number of equivalents of the HWE reagent in problem 1, the amine in problem 3, and the bromine in problem 4 may have an effect on the outcome of the reaction. Finally, the problems use relatively complex, generally polyfunctional molecules, in which not every functional group plays a major role in the reaction. Ambiguity about which portions of a molecule are reactive in a given context is often not something that students gain much experience with until undertaking advanced coursework or research projects.

Unfamiliar elements are also a key part of the question design. For example, a cyclic acetal in which one of the alcohols does not leave the molecule upon hydrolysis can lead to confusion when compared to a more prototypical memorized substrate, like the ethylene glycol acetal of acetone. The Diels-Alder question included two features that make the question seem less familiar to students. The first is that the diene is drawn in the *s*-trans conformation, which can be quite disruptive to the recognition of a possible cycloaddition. Another key feature is the catalyst; Bronsted acids are not the most common catalysts for Diels-Alder reactions, though it was conjectured that some students would propose a stepwise ring formation even if they did not notice the potential concerted reaction. The alkene bromination question makes use of unexpected conditions, more generally seen when attempting an electrophilic aromatic substitution.

Previous work on predict-the-product problems has often focused on much simpler transformations in which the most efficient method for answering may be a simple recall of how the given reagent changed the given functional group on the monofunctional starting material. Responses to these problems led these researchers to conclude that students do not value mechanisms as a way to reason about chemical reactivity, but it seemed plausible that students would make more use of mechanisms when there are more possibilities to take into account. The problems used in this study provide an opportunity to investigate this hypothesis.

METHODS

<u>Participants and Context</u>

All work was conducted at a large, research-intensive institution in the Western United States. Undergraduate student participants were recruited from Chem 12B, the second-semester organic chemistry course for students majoring in chemistry, chemical biology, or chemical engineering. A recruiting announcement was made to the entire Chem 12B course near the end of each semester in which interviews were conducted. Additionally, regular attendees of the ChemScholars discussion section were sent recruitment emails. ChemScholars is an optional discussion section run by advanced undergraduates that tracks along with the Chem 12B course. I attended most ChemScholars discussion sections to provide support for the undergraduate leaders, so many of these students were familiar with me prior to the interview. Procedures used in this research were approved by the University of California, Berkeley Committee for Protection of Human Subjects, Protocol #2015-08-7858.

Participants for Research Questions 1-3

The first three research questions were addressed using data collected from 35 Chem 12B students recruited in Spring 2018 and Spring 2019. A majority of the students interviewed (77%) were regular attendees of the ChemScholars discussion section. A range of student "abilities" are represented, as measured by their grade on the Chem 12B final exam (see Figure 2). It should be noted that this population is not a random sample of all students in Chem 12B. The average interviewee scored 0.6 standard deviations above the mean on the final exam, and only 23% of the interviewees scored below the mean on their final.

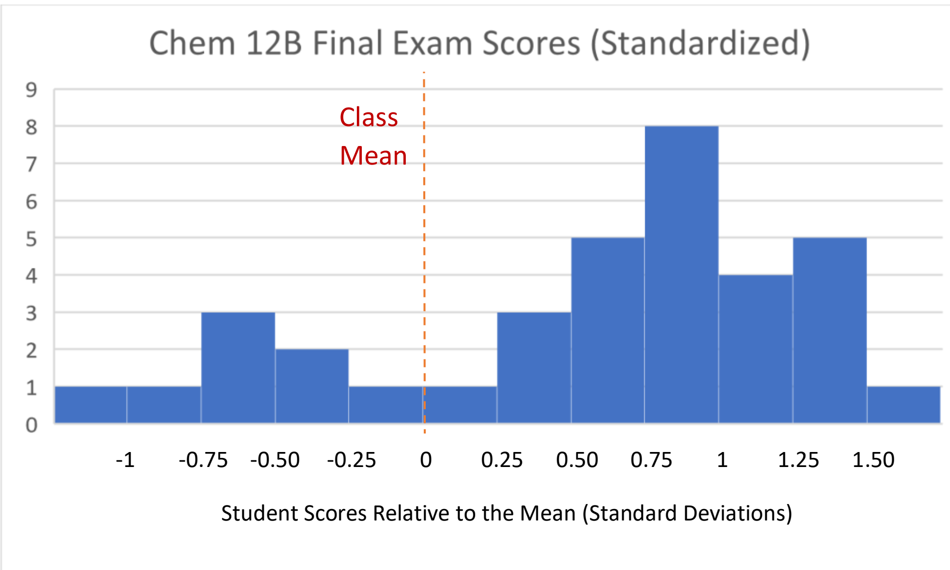


Figure 2. Grade distribution of undergraduate interview participants on the Chem 12B final. All exam scores are converted to z-scores (standard deviations above the mean) to combine data across semesters.

Participants for Research Question 4

To compare the undergraduate interviews with more advanced solvers, 9 graduate students were recruited by email from a variety of synthetic organic, organometallic, and chemical biology research groups on campus, using a convenience sampling method. Students ranged from the 2nd through 5th year of their program, and all but one had been a teaching assistant for an organic course in the academic year prior to the interviews.

<u>Think-Aloud Interviews</u>

All students participated in a think-aloud interview, during which they were asked to solve the four predict-the-product problems in Figure 1 while attempting to vocalize their thought processes. All questions were presented on separate sheets, one at a time. Each one had the instructions "Predict the major organic product(s) of the following reaction(s). Please indicate stereochemistry where appropriate." A complete interview protocol can be found in Appendix 1.1. Students were allowed to work uninterrupted until indicating that they had reached a final answer, after which they were asked a few follow-up questions, some of which prompted them to consider other possible outcomes.

Interviews were audio recorded, and video recordings were taken of student writing. Interviews were transcribed and annotated with what the student was writing while they were talking.

Coding and Model Development

Characterizing Student Thinking

The interview transcripts from the Spring 2018 interviews were initially coded to identify the primary problem-solving actions that students were taking. Through a constant comparative method, a coding scheme was developed to classify the most common student actions. Saturation was achieved with a set of 15 codes, which were then collapsed into 12 total codes.

Transcripts for all think-aloud interviews were then fully coded with the primary set of 12 codes using MaxQDA software. Only the students' spontaneous thought processes (i.e., prior to any significant prompting by the interviewer) were coded. Beginning with the Spring 2018 and 2019 interviews, transcripts were coded independently by 3 researchers. Coding was periodically compared, and discrepancies were resolved by discussion among researchers. After 40% of the data was coded in this way, intercoder agreement between pairs of researchers reached levels of 50-60% between pairs of coders, which was considered sufficient based on the complexity of the data and coding system. Subsequent coding for the remaining interviews was completed by a single researcher.

Proposed Workflow

The primary codes were arranged into a flowchart, subsequently referred to as a "workflow," which served both as a model for student thought processes during complex PtP problems and as a potential instructional guide for how to approach such problems. The exact form of the workflow was developed over time through discussions among the research team. The pathways outlined on the first complete draft of the workflow (Appendix 1.2) were informed partially by holistic impressions of the Spring 2018 interviews and partially by my previous experience as an instructor and tutor. After many years of working with organic chemistry students, I had already developed advice for how to think about predicting reactivity prior to conducting any think-aloud interviews. These ideas are summarized in Appendix 1.3 in the form of an outline.

Feedback and Workflow Revisions

The first draft of the workflow was introduced to advanced undergraduate students who agreed to participate in focus groups. Two focus groups, each containing 6 students, met to discuss the format of the workflow and its potential usefulness as an educational resource for sophomore undergraduates in the middle of their first year of organic chemistry. To accompany the workflow, a one-page document that briefly explained each step was generated (see Appendix 1.4). Additionally, a set of 7 example PtP problems typical of first-semester organic chemistry were provided so that students could better assess the usefulness of the workflow (see Appendix 1.5). This document also contained information on the types of feedback that would be most useful, to help guide the resulting discussion.

A set of revisions was made to the workflow based on feedback from the focus groups. The resulting draft was shown to 5 graduate students in organic research groups after they participated in think-aloud interviews. Feedback from these students was incorporated into the final draft of the workflow. Additionally, the explanatory document was replaced by sets of questions for students to ask themselves at each step of the process.

Quantitative Analysis

After all interviews were coded, subsets of the data were averaged to create profiles enumerating what percentage of the transcripts corresponded to each code for various groups of students (e.g., all Spring 2018 interviews, all successfully solved problems, etc.). To determine similarities and differences between different sets of students, the average transcript percentages corresponding to each code were compared using t-tests. The average interview length, as measured by character count, was also a point of comparison between groups. Because multiple tests are run for each group of students (one per code), the conservative Bonferroni correction is used to reduce the rate of false positives. All statistical tests were conducted using Stata software. Further trends in the data were identified using holistic coding. Evidence for the resulting claims was then gathered by quantifying the presence or absence of various features in each interview.

RESULTS AND DISCUSSION

The think-aloud interviews are rich sources of data that can be analyzed in a variety of ways, and numerous interesting observations can be made about student discussions of the problem-solving process. Before addressing more general characteristics of student problem solving and attempting to answer my primary research questions, summaries of "what students did" on each problem are presented to provide context for the subsequent analysis. This is followed by a description of how the coding system and workflow model were developed. After presenting evidence for the validity of my model, comparisons are made between different groups of students to identify how more successful and more expert-like participants approached the problems utilized in this study.

General Overview of Student Problem Solving

Problem 1

Student thought processes on the first problem focused heavily on the reactivity of the HWE reagent. Students often discussed it before addressing the first step at all, and 23 students (66%) specifically stated that they needed to form a carbonyl in the first part of the problem in order for the HWE reagent to react properly. However, they struggled to hydrolyze the acetal. Most (69%) figured out that they could protonate and then eliminate one of the two oxygens, but many struggled to continue the reaction. Students often did not treat the two steps separately, and as a result, they tried to directly react the unstable oxocarbenium intermediate with the nucleophilic HWE reagent (31%). This approach generally caused confusion when it did not produce the expected betaine.

Another common occurrence was for students to recognize the similarity of the starting material to a THP-protected alcohol (37%). Because of this, students often broke the acetal down into a methyl-DHP group and isopropanol, neither of which provided an appropriate substrate for the HWE reaction. This cognitive dissonance was resolved by 7 students (20%) through assuming that isopropanol must somehow become oxidized to acetone. Because students knew something about where they needed to get to (a carbonyl), they proposed chemically unreasonable transformations to get there as directly as possible, similar to what DeCocq & Bhattacharyya (2019) found in their study.

Student answers were considered "correct" if they proposed either of the product molecules indicated in Figure 3. Only 4 students (11%) gave perfectly correct answers, but an additional 5 students (15%) gave nearly correct answers with only minor errors, such as inverting the methyl stereocenter when drawing it in a different orientation. Of the 9 students who were correct or nearly so, 8 drew the unsaturated ester, while only 1 recognized the formation of the oxacycle. Recognizing that water was present and might participate in the reaction seemed to be key for reaching a reasonable solution; 63% of students who drew water as a nucleophile proposed a correct answer, compared to 15% of students who did not. Most (74%) students drew a mechanism, but drawing a mechanism was not associated with greater success on this problem.

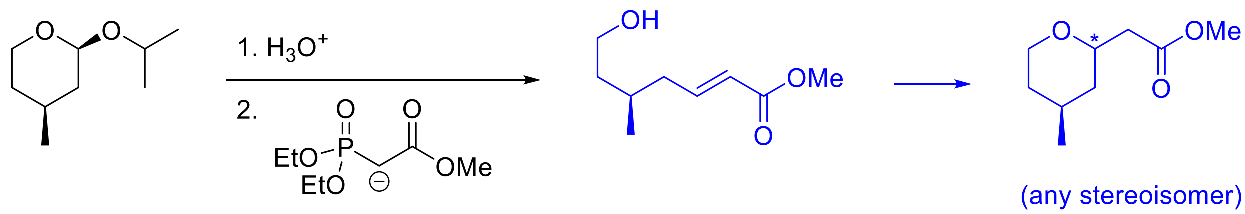


Figure 3. Accepted solutions to Problem 1

Problem 2

Student success on this problem depended largely on whether they recognized a potential Diels-Alder reaction, but noticing this possibility did not guarantee a correct answer. About half (54%) of the students did not identify the substrates as a well-matched diene and dienophile pair, largely due to the *s*-trans conformation of the diene. Overall, 16 students (46%) suggested a Diels-Alder reaction, and of the 19 who did not, 13 recognized it immediately upon being asked “would your thinking have differed if the starting material was drawn this way [showing them the diene in the *s*-cis conformation]?” Interestingly, of the 16 who recognized the Diels-Alder possibility, 6 said that it could not be a cycloaddition, remarking that, “I was going to say Diels-Alder but there’s no heat,” or, “I'm thinking of a Diels-Alder but that needs heat, so maybe it's a no.” This was an unexpected outcome; I did not anticipate how important the presence of a written Δ was for students to consider the possibility of a thermally allowed pericyclic reaction. In their coursework, students had briefly seen a Lewis acid-catalyzed Diels-Alder reaction, though never a Bronsted acid-catalyzed one.

The 19 students who did not recognize the cycloaddition generally proceeded by finding nucleophiles and electrophiles to pair up. Of these students, 8 (23% of total) saw only the methoxy lone pairs and the protonated carbonyl and attempted to do a transesterification, encountering problems once they reached an acylated oxonium. The Michael addition was recognized by 7 others (20% of total), and while some balked at the fact that they still had a positively charged oxygen, 3 students (9% of total) identified the second Michael addition and completed the stepwise cycloaddition to form a six-membered ring. The success rate on this problem was 26%, with an additional 14% recognizing the possible Diels-Alder reaction but drawing an incorrect regiochemical outcome. Most (77%) of the students drew a mechanism, though generally not after they recognized the possibility of a Diels-Alder reaction.

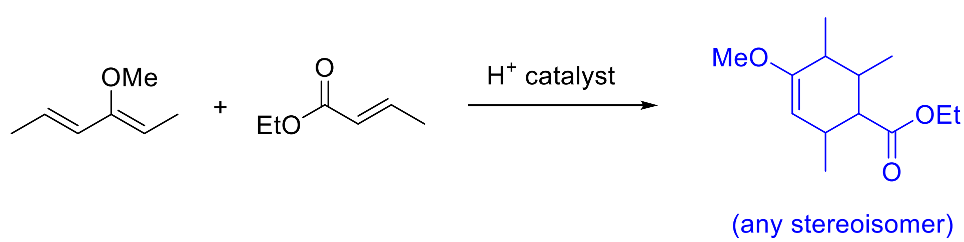


Figure 4. Accepted solutions to Problem 2

Problem 3

Students explored a variety of pathways for this problem, but most successfully started by condensing the amine and the aldehyde. Essentially all (97%) students recognized the aldehyde as the most reactive electrophile and at some point had the amine add to the protonated aldehyde. From there, 29 students (85%) continued on to form an iminium or imine, while the other 5 students (15%) had the nitrogen of the resulting hemiaminal react with the ketone to form a seven-membered ring. A few (18%) students left the imine as their final product, but most assumed that there must be additional possible reactivity.

At this point, students proposed multiple different mechanistic paths. Some (18%) attempted cyclization by having the imine nitrogen add to the ketone. Others (47%) recognized the possibility for one or both possible enol tautomers of the ketone, although only 9 students (26%) mentioned the Mannich reaction by name. In total, 14 students (41%) completed the Mannich transformation to generate a 6-5 spirocycle (Figure 5). However, half of these students were unhappy with this structure, because "it just looks kinda funky." In fact, 5 students rejected it in favor of much less reasonable answers, assuming the reaction was not yet complete. Unexpectedly, 4 students (12%) remembered that one place they had seen spirocycles was as an intermediate in the Bischler-Napieralski reaction, which undergoes a 1-2 shift to form a 6-6 fused ring system. As a result, they did an analogous shift, ignoring the fact that they were shifting to a ketone or saturated carbon center and not a carbocation. Overall, 9 students (26%) successfully drew the Mannich product and stopped at that point. An amine-catalyzed aldol between the two carbonyls was also considered acceptable, but no students gave this as their final answer.

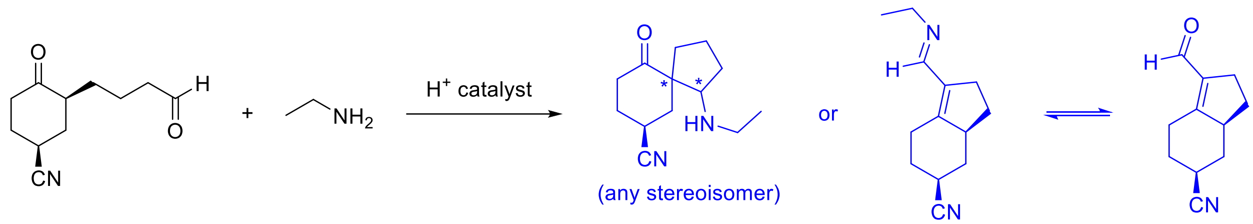


Figure 5. Accepted solutions to Problem 3

Problem 4

Students mostly focused on brominating the aromatic ring in this problem, and very few attempted to react the bromine with the alkene substituent. As expected, all students recognized the conditions for an electrophilic aromatic substitution reaction. As a result, much of the discussion revolved around directing group effects to determine which position on the ring would be brominated. Most students (>80%) explicitly referred to both the vinyl and ester substituents as electron-donating, activating, or ortho/para directing. In the case of the vinyl group, this was often (37%) done by categorizing it as a generic alkyl group. All but 6 students (82%) brominated the ring at one or more of the ortho/para positions, and 3 of the remaining students (9%) brominated at the meta position. Although 7 students (21%) recognized the possibility of the alkene reacting with the bromine, only 3 drew the product of this addition, and only 2 (6%) settled on this as their final product.

Students were reluctant to consider possibilities other than an electrophilic aromatic substitution. All students were eventually led with prompting to the idea that bromine could react with the alkene. However, even after recognizing this transformation, students were still more likely to say that the aromatic ring would react first. Some gave erroneous but chemically based reasons, such as identifying the ester as an "activating group", which the alkene did not have. However, a more common reason for discounting the alkene addition was "that's 12A [first semester] material!" Another reason given for this decision was that the alkene addition reaction is "too simple" to be correct. This type of non-chemical reasoning has also been found to be prevalent among students in other organic courses at this institution (Brando, 2019).

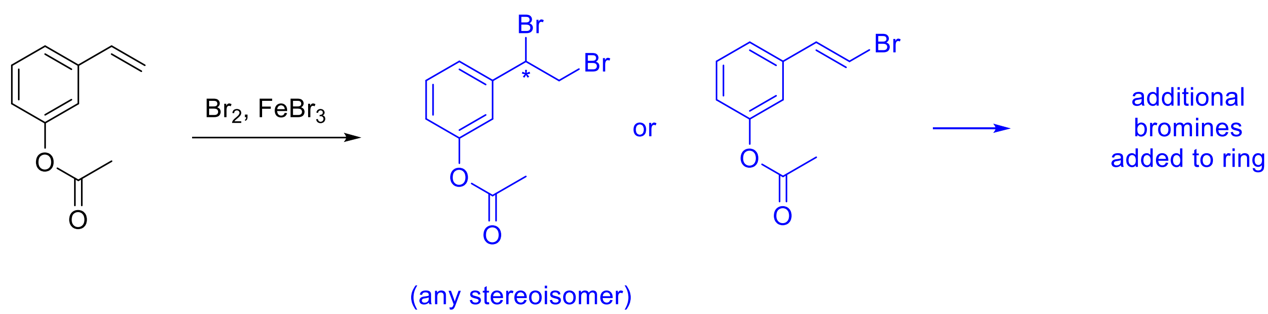


Figure 6. Accepted solutions to Problem 4

Use of Mechanisms

Grove et al. (2012) and others have suggested that students do not value mechanisms as a method for reasoning through PtP problems, but mechanisms were widely used during problem solving in this study. Students were considered to have used a mechanism only if they drew out at least one complete step (starting material, arrows, and product). On problems 1, 2, and 3, students drew a mechanism 74%, 77%, and 94% of the time, respectively. Only on problem 4, in which the mechanism is generally assumed to be a straightforward electrophilic aromatic substitution, did students largely forego the use of mechanisms (only 2 students (6%) drew them). These results suggest that students *do* value mechanisms as a problem-solving tool, but they may only use them when the problem is sufficiently long or complex.

Research Question 1

What are the common characteristics of problem solving in the context of non-trivial predict-the-product problems?

Primary Codes for Characterizing Student Problem Solving

Exploratory coding of a subset of the think-aloud transcripts was conducted to identify general problem-solving actions exhibited by multiple students. Unlike the actions described in the previous section that are specific to a single problem, this analysis focused on more abstract themes, like deciding between competing pathways or checking work, that might be applicable to a broad range of questions. An effort was made to code all student discussion pertaining to chemistry, to avoid only identifying expected actions.

Students frequently started a problem with an initial planning stage. For example, students often started by naming functional groups and reagents, identifying nucleophiles and electrophiles, and noting any unusual structural features. In essence, they were *collecting information* about the problem, specifically outlining the starting conditions for their solving process. “Collecting Information” was one of the first codes to be identified.

After the initial planning stage, students often took one of two pathways, the first of which occurs when the student recognizes the conditions for a specific reaction they have learned. This *recognition of a known reaction* is generally followed by the application of relevant knowledge. *Mapping* of general knowledge about a reaction onto the current problem was a major component of many interviews. This mapping process allowed students to *identify an endpoint* that they considered reasonable. Reaching a possible solution in this way can occur either by directly jumping to a product by analogy with the known reaction, or by working through the mechanism for the identified reaction using the given substrate. Occasionally, students would draw chemical structures that very strongly suggested that they had a specific reaction in mind, but they did not identify it out loud. In these cases, the *implied mapping* code was used.

When a student could not identify any known reaction, they had to take a more step-by-step approach to solving the problem. After considering which potential *proton transfers* might take place, students taking this path would use the EPF to identify the *first elementary steps* that might occur. This first step generally resulted in a reactive intermediate, from which students could *follow the reactive pathway* typical of that sort of structure. Determining the resulting *stereochemistry* sometimes occurred in the middle of the solving process and sometimes at the end.

For some students, this was the end of their problem solving, but others took this opportunity to *check their work*, by looking for *errors*, *further reactivity* of the proposed product, or *alternate reactivity* of the given substrates. Students who *proposed alternate reactivity* would then need to *decide between the major pathways or products* to reach a final solution. Finally, throughout the solving process, students would stop to *assess their progress*.

In total, initial coding resulted in a set of 15 primary codes that were used to characterize student problem solving throughout the study. Due to the similarity between "Mapping onto current problem" and "Implied mapping", these are combined into a single code for further analysis (abbreviated "MapTotal"). Similarly, the three types of checking work are collapsed into a single "CheckTotal" code. These 12 primary codes and their descriptions are summarized in Table 1, and examples of student responses corresponding to each code can be found in Appendix 1.8.

Table 1. Codes for student problem-solving actions during PtP problems

Code	Abbrev.	Description
Collect Information	CollInfo	Student gathers information that might help them solve the problem, both what is on the page and relevant prior knowledge.
Acid-Base Equilibria	AcidBase	Student talks about proton transfer steps and/or equilibria between different protonation states.
Identify First Steps (non-H+)	IdentFirst	This code is used to identify the first elementary step proposed after any initial proton transfers.
Follow Reactive Pathway	FRP	Student starts with a reactive intermediate and goes through one or more steps in an attempt to reach a stable intermediate/product.
Recognize Similarity to Known Reaction	RecogRxn	Student mentions the name or a description of a reaction or transformation.
Map onto Current Problem	MapTotal	Students generally then map what they know about that reaction onto the current problem they are solving.
Identify Reasonable Endpoints	IdentEnd	Student does a last step to reach a final answer, indicating that they are finished.
Propose Alternate Reactivity	PropAlt	After discussing one possible path/reaction, student backs up and considers a different path/reaction.
Decide Between Pathways/ Products	Decide	Student has two (or more) identifiable paths or molecules to choose between and makes a decision, sometimes with rationale.
Stereochemical Analysis	Stereo	Student talks about determining stereochemistry outcomes.
Assess Progress	AssessProg	Student stops and comments on how correct or certain they are.

Check Work	CheckTotal	- *After* reaching something they seem to consider to be a final answer, student looks back to see whether alternate chemical pathways or further reactivity are possible. - Student checks for mistakes at any point during the interview

Quantifying Student Problem-Solving Actions

Once the coding system was developed, all interview transcripts from Spring 2018 and 2019 were fully coded to get a quantitative overview of what students were most commonly thinking about as they worked through the problems. With 82% of the transcripts identifiable as one of these 12 codes, the large majority of student thinking seems to be captured with this coding system. The percentage of the transcripts coded with each action is shown in Figure 7. The most common way students spent their time was mapping knowledge about a known reaction onto the given substrates. They also spent a significant amount of time assessing their own progress through the problem. Conversely, this group of students often neglected to consider stereochemistry in their think-aloud discussion, with only about 1% of the total discussion time devoted to it.

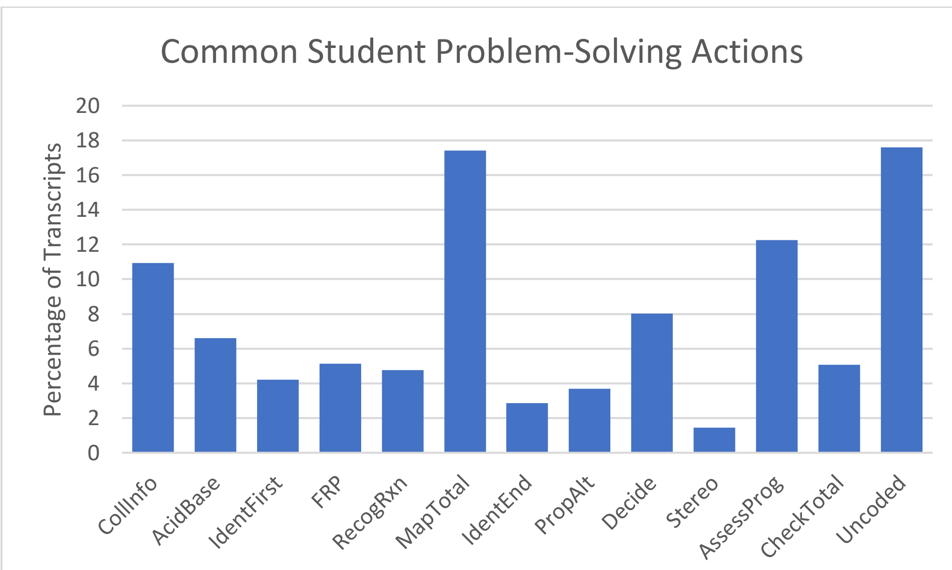


Figure 7. Frequencies of problem-solving actions by percentage of interview transcripts

Overall, about 18% of the transcripts were left "uncoded." What was going on during those periods of time? The largest part is made up of quotations like "[inaudible muttering], and this [???] looks like this, so...", in which the student is mumbling to themselves and/or making statements with unclear referents, so they cannot be unambiguously coded. Other common uncoded segments include comments or questions to the interviewer (e.g., "am I allowed to use notes?", or "Is this supposed to be a really hard problem or should I be worried?"). Most of the uncoded segments do not include clear discussions around chemistry. However, one relatively common occurrence that was not part of the current coding system was for a student to state what they would do if the situation were different (e.g., if water was present, if heat was specifically indicated, if an alcohol were a carbonyl instead).

This analysis pools student data from 2018 and 2019 and assumes that there are no systematic differences between these years. Both years of Chem 12B students had very similar instructional experiences, and from a holistic perspective, no systematic differences were noted between the interviews in 2018 and 2019. As a result, very similar problem-solving profiles were expected between the two years. This is indeed the case; the percentages in Figure 8 are nearly indistinguishable from year to year, and only one difference (RecogRxn) meets the $p < 0.05$ cutoff. Because multiple comparisons are being made, it is appropriate to apply the Bonferroni correction and adjust the threshold for significance to 0.05/12 = 0.004. By this criterion, none of the differences are significant. However, there is one significant difference between years: the 2019 cohort spoke about 40% more than students in 2018 ($p = 0.002$). The reason for this difference is not clear, but one possibility is that the interviewer was more experienced the second year and may have allowed students more time to speak before interjecting with follow-up questions.

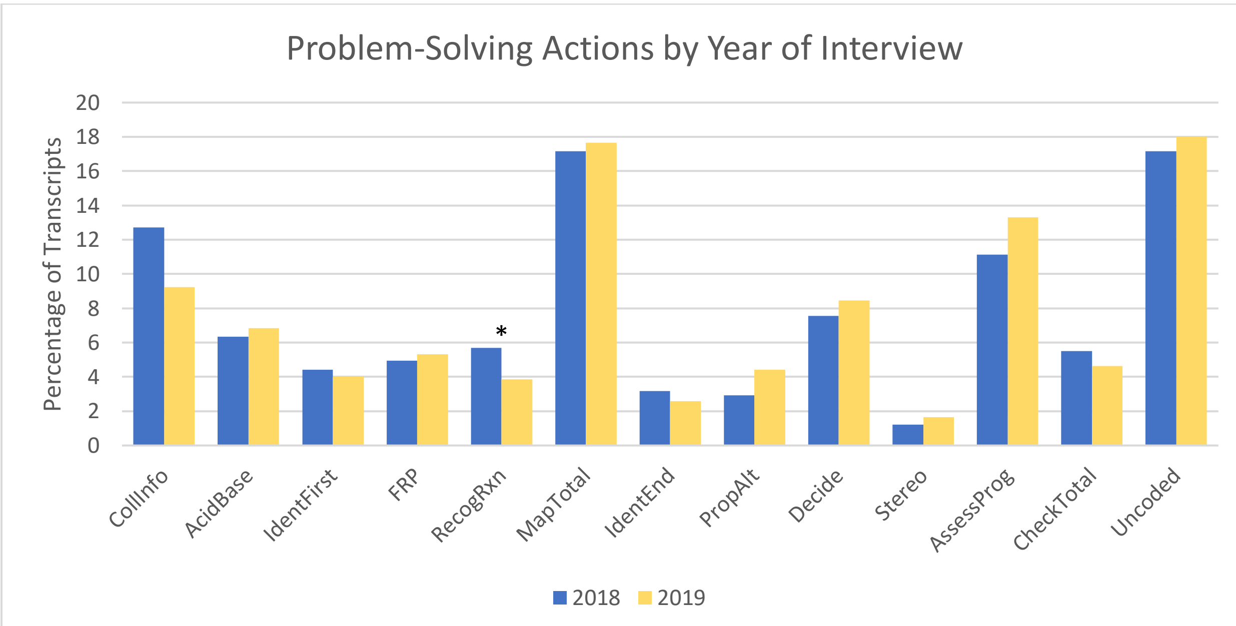


Figure 8. Comparison of problem-solving actions by two cohorts of interviewees (*$p < 0.05$)

Workflow Development

The set of 12 codes in Table 1 was arranged into a first draft of the workflow (Appendix 1.2), which is intended to both model student thinking and serve as a guide for approaching predict-the-product questions. As described in the Methods section, the form of the workflow was developed through holistic impressions of the think-aloud interviews and discussions among the research team, most of whom have extensive experience as organic instructors. The workflow model is not fully self-explanatory, so an additional one-page document was written to clarify the intended meaning of each component.

The first draft of the workflow, the one-page explanatory sheet, and example problems were presented to two focus groups of advanced undergraduates. Feedback on this first draft was largely positive, and several students expressed a desire to have a copy for themselves, despite having already completed their organic coursework. The inclusion of different pathways depending on whether a known reaction is recognized was highlighted as particularly important. Students felt that the workflow matched their own thought processes, especially after trying example problems. Student comments included, “It's pretty close to my thinking,” “It's a good match for how we think about it,” and “this is literally what I do.” The focus groups also had some critiques on what to modify to make it less confusing and more streamlined. For example, some were confused by certain phrases, and many students thought the use of color should be more explicitly defined. The workflow was modified based on these comments, and a second draft was generated.

The additional one-page document that briefly explains each workflow bubble was considered to be an essential and useful accompaniment to the workflow itself. In particular, focus group participants liked the descriptions that included questions for students to ask themselves while on that step of the process. Comments included "to introduce this, you could make these into questions, so they know what they're asking themselves," "focusing on being more question-based might be good," and "I like that you put the thought questions." In response to this feedback, the explanatory sheet was expanded to a two-page document composed of prompting questions that students could ask themselves at each step (Appendix 1.6).

Most focus group participants felt that the workflow and accompanying materials would be a useful resource for organic students if introduced at the appropriate time. One student commented that, "I think this will be helpful. For the people that use it, they'll get a lot out of it." However, some expressed concerns that the full workflow would be overwhelming, especially if given out too early. One student remarked that, "It's kind of scary when you first see it though. It might be a progression situation." By "progression situation," they were referring to the idea of first introducing a simplified version before giving out the complete workflow. Other participants agreed, and this was the approach taken when the workflow was eventually introduced to students.

A second draft of the workflow was shown to 5 graduate students with expertise in organic chemistry after they participated in think-aloud interviews. Again, feedback was largely positive. Students generally felt it captured their own solving process, with one student remarking, "This is, with a little bit more thoroughness, more or less what I do when I solve a problem." One student thought the workflow seemed "useful", and another said, "I don't think I have anything I would change or improve." One student liked it but also pointed out that the scope of the workflow was really limited to 2-electron chemistry: "I think this is pretty comprehensive. I mean, I'm reading a lot of these elements, and a lot of them are things that I did for these problems. I guess my only comment on this that could be taken as a critique or whatever is that these are all really based on 2-electron pathways…. But I think for 2-electron chemistry this is... I think this kinda nails it." An additional "Determine most likely pathway(s)" bubble was included based on these interviews, but few other changes were made at this point.

Final Workflow

The final workflow included two different forms. The first was a complete version intended to capture the full range of student solving (Figure 9). The other was a simplified form (Appendix 1.7) intended for use as a student resource, along with the prompting questions (Appendix 1.6). This version emphasized step-by-step mechanistic reasoning and de-emphasized the "shortcut" of predicting products based on the recognition of a known reaction. At the point when this workflow would be introduced, mechanisms are generally short and best considered one step at a time; it is not until later that multi-step mechanisms need to be "chunked" to efficiently determine the outcome(s) of a given reaction. The use of the simplified version of the workflow was demonstrated to students in a large lecture class, and preliminary indications are that it can be a helpful resource. However, instructors must be thoughtful and explicit about when, why, and how students should make use of it.

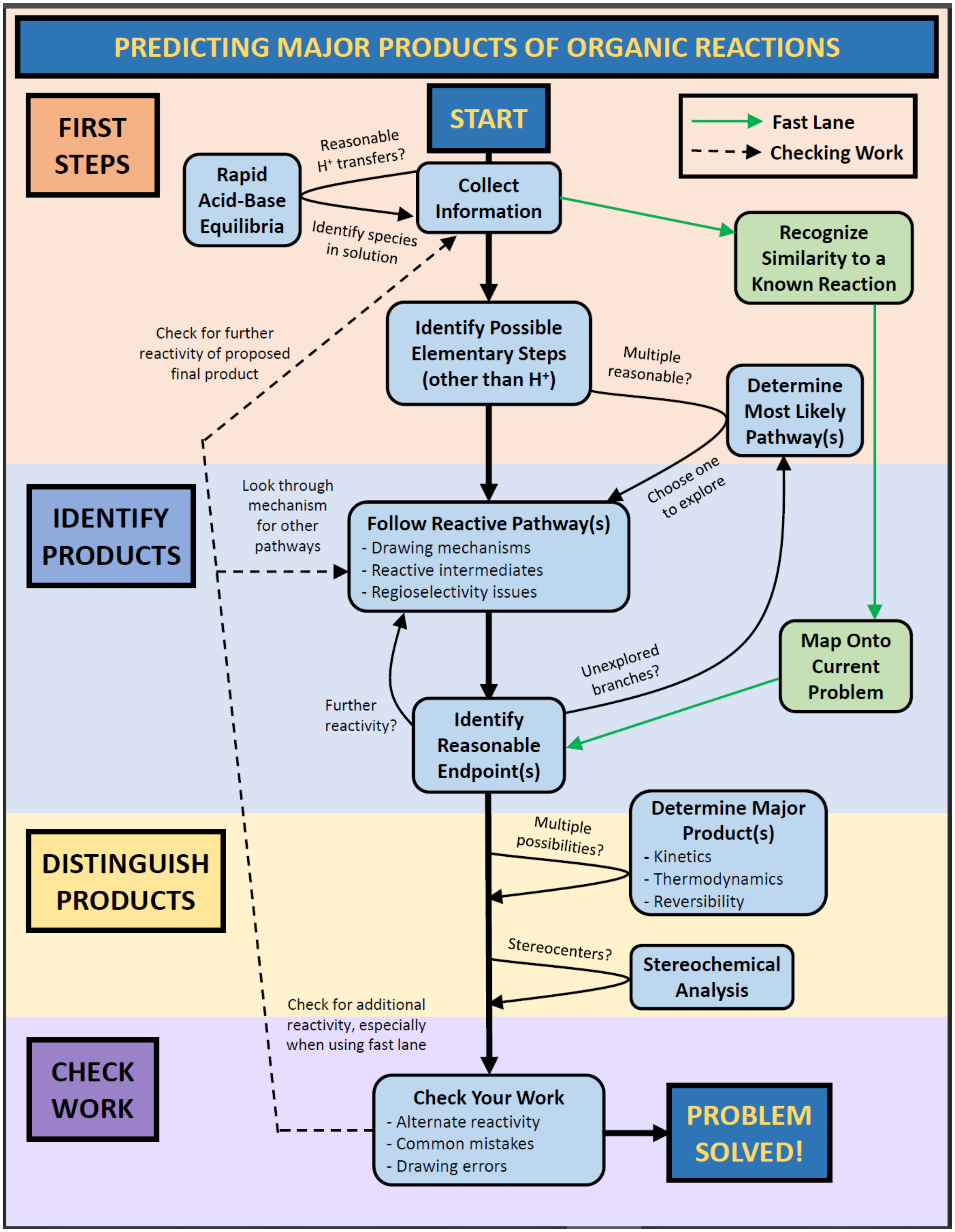


Figure 9. Finalized workflow

Model Comparisons

It is interesting to compare this workflow to Bhattacharyya's (2014) model of how students approach mechanism problems. Although both models are attempting to characterize mechanistic reasoning, they are ultimately quite different, due to the information given as part of the problem. Students do not need to identify an endpoint, distinguish between products, or do stereochemical analysis as part of a mechanism problem. Conversely, one cannot "map the reactant onto the product" in a predict-the-product question. Both models do feature "mapping knowledge onto the current problem" steps, but other than that, similarities are surprisingly low.

One surface feature that appears similar between the two models is that they both have two pathways through the diagram. However, although they both feature diverging paths, the reason for divergence is different. In Bhattacharyya's model, the path chosen (single- or multi-step transformation) is based on features of the reaction that are recognized when comparing the starting material and product. In my model, the path chosen depends on whether a reaction is recognized in the first place. Both of Bhattacharyya's pathways would fit best within the workflow "fast lane," in which a reaction is recognized and knowledge about that reaction is mapped onto the given substrate. The Bhattacharyya model does not attempt to characterize student reasoning when cues in the problem are not sufficient for them to recall the appropriate reaction. Because this work was based on graduate students, it is possible that nearly everyone was able to recall a reaction, which would not provide enough data to propose a model for those who did not recall one.

Research Question 2

Does the workflow model I developed accurately reflect student problem solving?

Assessing the Workflow Model – Two Paths

Roughly speaking, there are two primary paths through the workflow: investigating a mechanism step-by-step (the central column) and recognizing a known reaction and applying relevant knowledge (the "fast lane"). The interview transcripts do not cleanly divide into one or the other, as many students (46%) incorporate elements of both paths in their approach. However, the interview transcripts can be divided into two groups based on students' initial approaches to solving the problem. In one group, students identified a first elementary step (not counting simple proton transfers) before identifying any known reaction. In the other group, students identified a known reaction and did not identify a first elementary step as a separate action. Problem-solving action profiles for these two types of interviews are contrasted in Figure 10.

There are a clear set of differences between these two types of solution processes. Students with an Identify First Steps code in their transcript were significantly more likely to spend their time discussing acid-base chemistry ($p < 0.005$) or following reactive pathways ($p < 0.05$). In contrast, students without an Identify First Steps code spent significantly more time discussing known reactions ($p < 0.01$) and mapping their knowledge onto the given problem ($p < 0.0001$). Additionally, no items that appear on the workflow after the two paths merge (Identify Reasonable Endpoint and beyond) showed differences. The fact that specific items cluster together with other items on the same pathway, but there is no difference after the two paths merge, is evidence in favor of considering the workflow to be an accurate model of student work.

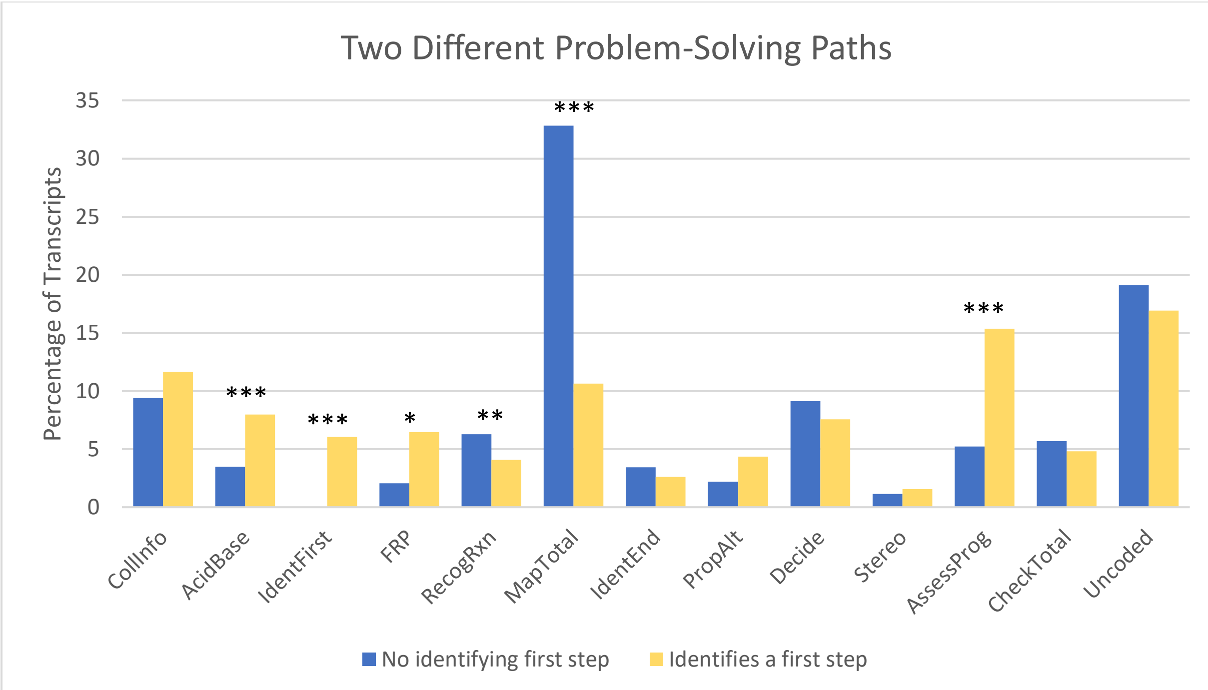


Figure 10. Problem-solving action profiles for two different workflow pathways (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.004$)

There is a second large difference between these two groups; students who took the Identify First Steps path spent nearly three times as long on assessing their progress ($p < 0.0005$). A common observation made during the interviews was that once students recognized a known reaction, regardless of whether it was the most appropriate reaction, they tended to be overconfident in moving quickly towards an answer. The dramatic difference on assessing progress between these two pathways is consistent with this observation.

It is notable that there was no difference in success rate between students who identified a first step (19%) and those who did not (22%). This is somewhat surprising, because students who recognize a reaction might be expected to perform better than students who do not. However, in these interviews, there was a high rate of students recognizing a reaction that was not the most relevant one. In particular, most students recognized the potential for electrophilic aromatic substitution on Problem 4, but that was not the most likely mechanism. Similarly, many students became fixated on treating the substrate in Problem 1 as a THP-like alcohol protecting group, leading them astray when identifying the important product of the acetal hydrolysis. It is probable that with more straightforward questions, the success rate would be higher for students who recognize a known reaction.

Assessing the Workflow Model – Individual Pathways

The analysis above showed that the workflow model and associated coding system captures student thinking on average, but further investigation was needed to determine whether the pathways shown on the workflow represent how individual students approach each problem. To visualize student problem-solving pathways, the progression of codes was drawn directly onto the workflow. For example, one student's approach to Problem 4 is represented in Figure 11a. After collecting information, this student quickly recognized a reaction that they knew and mapped what they knew about that reaction onto the given substrate, reaching what at first glance were multiple reasonable answers. The student then decided which of those answers was most likely. However, upon checking their work, they came up with another, even more reasonable solution and decided upon that as their final answer. This student went through the phases of the workflow in a relatively orderly fashion, with just one loop back up after checking their work. Compare this to the reasoning path of the student in Figure 11b, who was attempting the same problem but went through many different loops before reaching a final solution.

a) b)

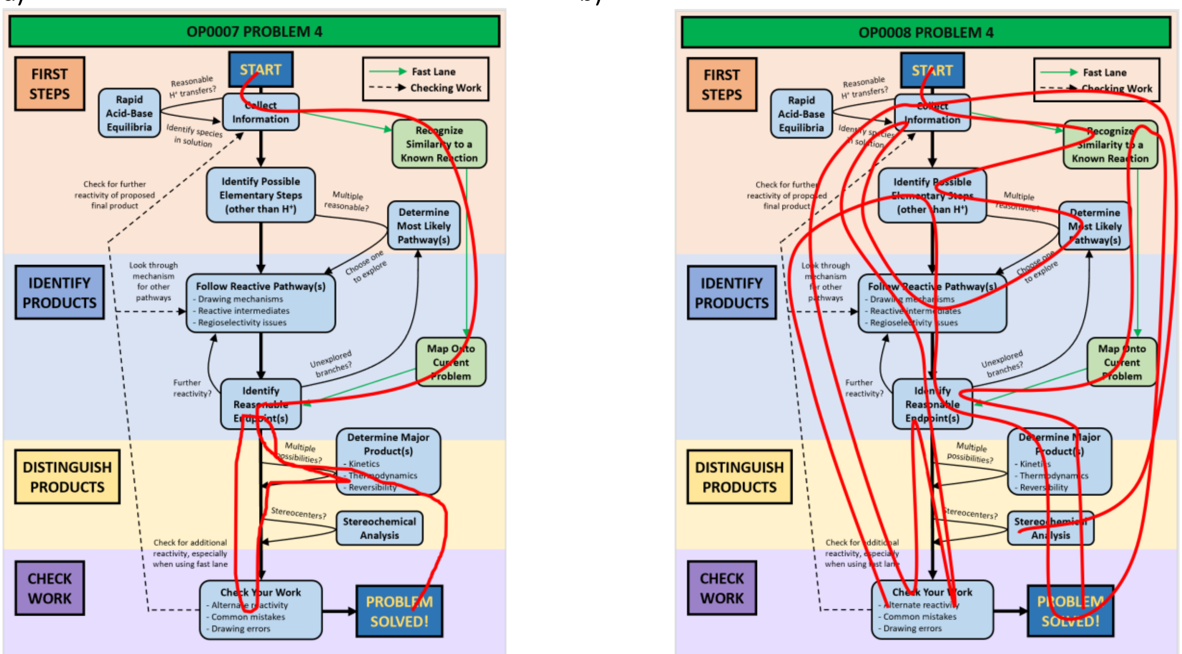


Figure 11. Student reasoning pathways for problem 4 mapped onto the workflow. Red lines indicate the order in which students engaged in each step of the process.

The two different pathways depicted in Figure 11 are reminiscent of the idea that different types of models may need to be used depending on whether the solver is treating the question as a problem or an exercise (Bodner, 2003). The student in Figure 11a seems to mostly have an idea about how to work through the problem, and they work through the various phases of problem solving in a relatively linear way. It seems likely they are treating the question as an exercise, and as a result, a model with distinct phases matches their solving process. In contrast, the student in Figure 11b seems to be treating the question as a genuine problem, and modeling such a pathway would require a more anarchistic model. Student solution pathways varied quite a bit in complexity, and there was not a clear division between exercises and problems. These examples were chosen to represent common paths at different ends of the spectrum.

Overall, the data suggests that student reasoning is well captured by the workflow model. Because pathways varied so widely, it is not feasible to capture all of them exactly with a single diagram. However, the components of the workflow describe the majority of student actions while problem solving, accounting for 82% of the interview transcripts. Additionally, the clustering of certain problem-solving actions provides support for the dual path model proposed.

Research Question 3
What differentiates successful from unsuccessful problem solvers?

Successful Solutions

Due to the overall difficulty of the problems, the undergraduate students struggled to propose reasonable answers. Out of the 138 total problems solved by the 35 undergraduate participants, only 19 resulted in fully correct solutions, with another 10 nearly correct answers (i.e., inverted stereochemistry, extra or missing carbons, etc.). Including all 29, the success rate was 21%.

To identify differences in how successful and unsuccessful solutions were generated, the problem-solving action profiles for these two groups were compared (Figure 12). Interestingly, no statistically significant differences were found. At most, there may be a small trend that successful students were more likely to take the recognized reaction branch than the step-by-step one (as measured by the presence of an "Identify First Step" code), but the effect was not significant with this sample size. Of the data I quantified, there was only one significant difference between the more and less successful students. The successful students talked about 35% more than the unsuccessful ones, as measured by character count. While students who eventually proposed correct answers were distributing their time in the same ways, they were simply doing more talking.

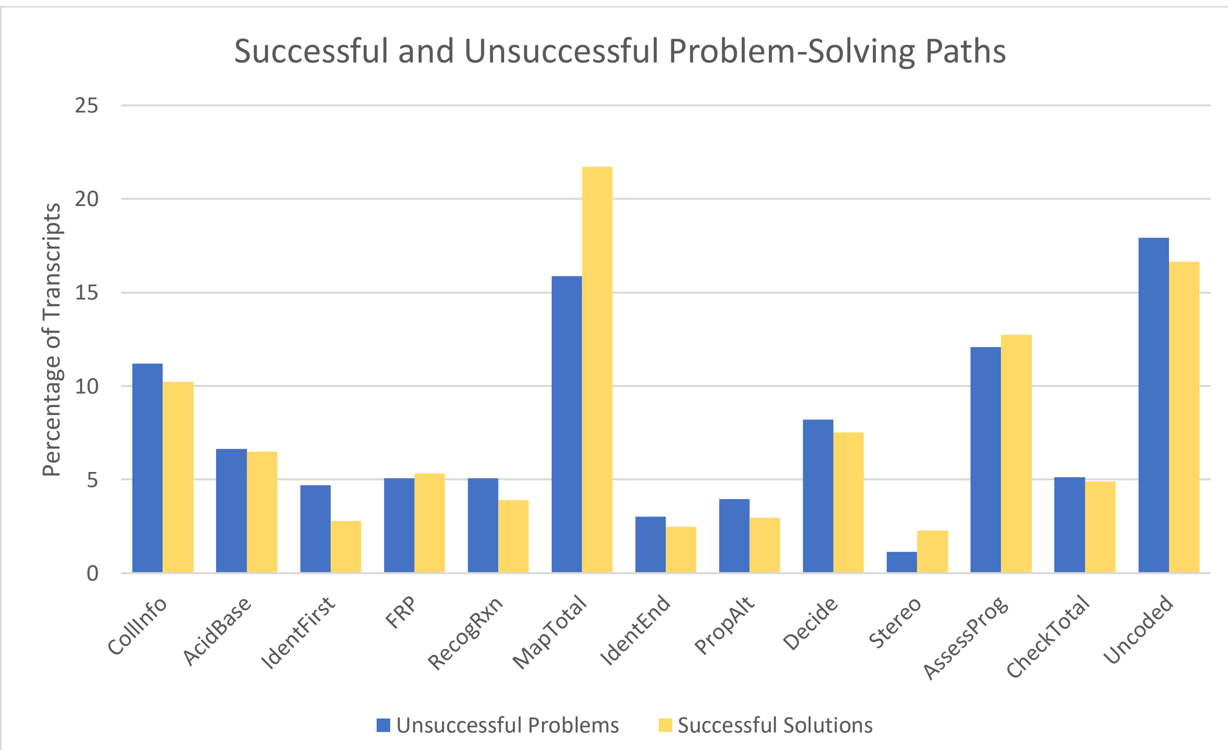


Figure 12. Problem-solving action profiles for more and less successful solvers

High Exam Scores

The interview questions are somewhat specialized and not necessarily what students are most frequently assessed on in classes. In fact, there is only a weak, non-significant correlation (r = 0.11) between the number of problems correct during this interview and students' grades on the Chem 12B final exam, which was held within a week of the interview. In addition to comparing students who were more or less successful on interview questions, I also wanted to look at differences between students who performed better or worse on exams.

Dividing students into "high" and "low" exam scores resulted in the comparison shown in Figure 13. Again, no significant differences were observed. The analysis is similar when comparing the highest 10 and lowest 10 exam scorers. The lack of association between problem-solving actions and success on either interview questions or exams was unexpected. I had hypothesized that certain behaviors like assessing progress or checking work might be more common among successful students, but this was ultimately not the case.

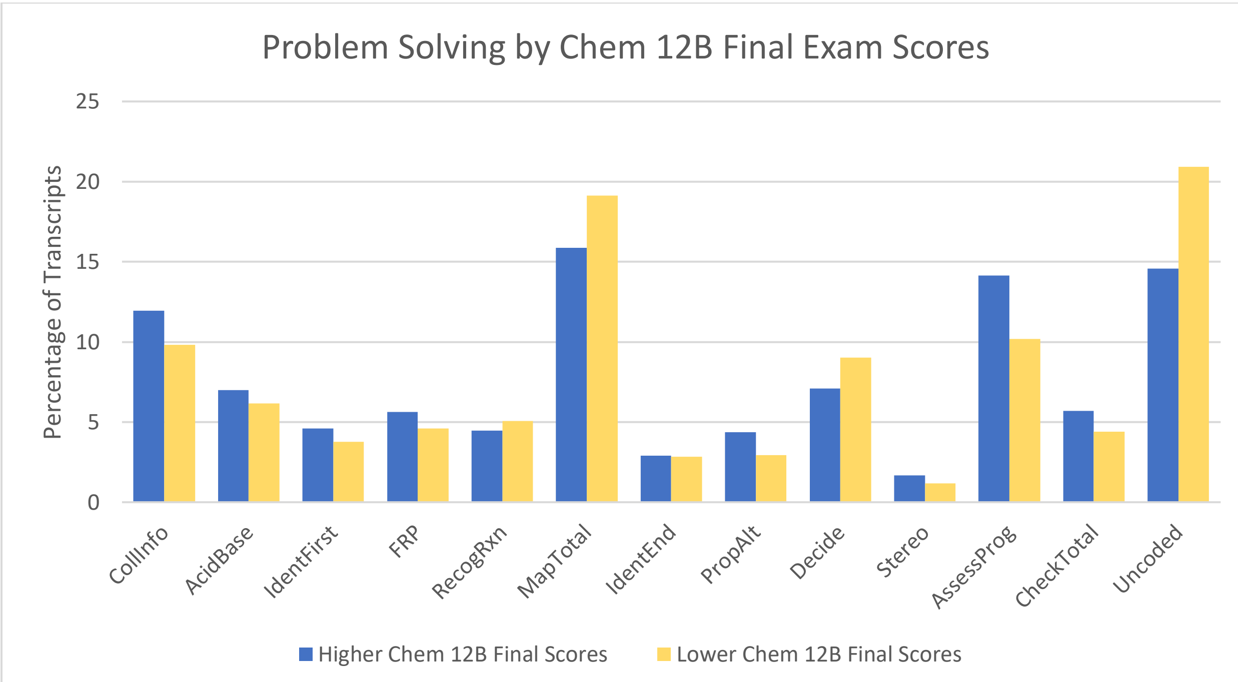


Figure 13. Problem-solving action profiles for students with higher or lower exam scores in Chem 12B

Investigating Other Differences

What is it that successful students are doing differently? The problem-solving action profiles above do not differentiate between successful and unsuccessful problem solvers. However, as mentioned earlier, students who spoke more tended to be more successful. One possible conclusion from this data is that simple persistence, working on a problem a little longer, is a key to being successful. Of course, this may not always be feasible, as is often the case on timed exams. There are certainly other possible interpretations. It may be that putting one's thoughts into words is beneficial to the solving process, so students who spoke more were more likely to succeed.

Another notable trend involves naming functional groups. Students who explicitly identified the starting material in Problem 1 as an acetal were more likely to be successful (55% vs. 13%). Similarly, all 6 students (100%) who specifically identified a "diene" in Problem 2 noticed the Diels-Alder, compared to 34% of students who didn't. Additionally, naming the alkene in Problem 4 was associated with recognizing the possibility for reactivity on that side chain. All students referred to the vinyl group, but only 6 students (18%) actually named it an alkene, while most just pointed and called it "that." Of these 6 students, 5 noticed the potential reactivity at the alkene (83%), whereas only 2 of the remaining 28 (7%) said anything about reaction at the alkene before being prompted to consider it. A possible hypothesis is that the act of naming these functional groups activates relevant knowledge more effectively than just looking at the structure. Overall, these results suggest that it may be good practice for students to explicitly name the given functional groups as they are starting a problem.

Other keys to success can be identified, but they vary by problem. Seeing the water "hidden" in the H_3O^+ was helpful for Problem 1; students who drew H_2O on their page were much more likely to be successful (63% vs. 15%). Recognizing the possibility of enolization was necessary for getting to the most likely answers for Problem 3. Being open to further reactivity (Problem 3) or alternate reactivity (Problem 4) also made a difference in some cases. Individual take-away lessons can be drawn from these trends (e.g., always consider whether your solvent might be involved in the reaction, ketones can be "secret nucleophiles," check whether your proposed product can keep reacting). In addition to learning general strategies, student success may partially depend on having a broad repertoire of these more specific themes and concepts at their disposal. Based on these results, it is suggested that having students create a list of the take-away lessons from problems they have encountered may be a useful exercise.

Research Question 4

What differences are there between the approaches of sophomore undergraduates and more experienced organic graduate students?

Graduate and undergraduate students took largely similar approaches to the problems used in this study. For many actions, especially those that occur early in the problem-solving process (collecting information, identifying first steps, recognizing known reaction, etc.), there is no real difference between these two groups (Figure 14). This suggests that, like the undergraduate students, graduate students use both the step-by-step and known reaction pathways, and in about the same ratios. Note that this might not be the case for simpler problems, for which graduate students would be more likely to have the exact transformation shown committed to memory. Also notable is that undergraduate and graduate interviews were nearly the same length on average (by character count).

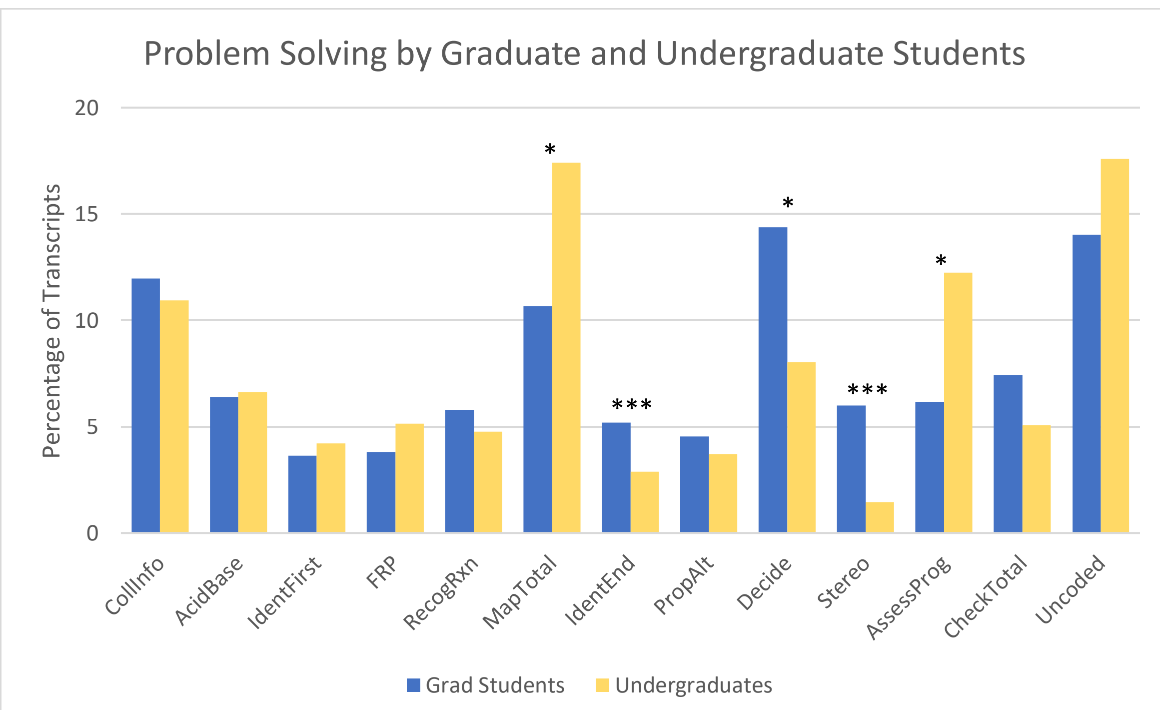


Figure 14. Problem-solving action profiles for undergraduates and graduate students (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.004$)

However, there are some differences in the relative amount of time spent on various problem-solving actions (Figure 14). All of these differences fit the $p < 0.05$ criteria, though using the stricter Bonferroni corrected $p < 0.004$ criteria, only identifying reasonable endpoints (IdentEnd) and stereochemistry (Stereo) can be pointed to confidently as real differences. However, it is informative to examine all of these differences, because the small sample size of 9 graduate students can make it difficult for smaller effects to reach the level of statistical significance.

Graduate students spent nearly twice as much time as undergraduates identifying reasonable solutions (IdentEnd) and deciding between them (Decide). This is consistent with the fact that graduate students were more likely to identify more than one possible solution to the problem. This could be because graduate students recognize more possible paths, but it could also point to the different ways that graduate students think about predicting reactivity. In a course, there is generally one and only one correct answer to a predict-the-product problem. As a result, undergraduates are used to finding a reasonable answer and then stopping, moving quickly to the next question, because they are often under a time crunch. However, in the research lab, such questions are generally approached by seeking all reasonable answers and narrowing down which is/are most likely.

Graduate students also spent substantially more time in these interviews discussing stereochemistry (Stereo). Even though the directions for every problem asked students to indicate stereochemistry, undergraduates often left newly formed stereocenters undefined unless prompted by the interviewer. (As a reminder, only the unprompted responses are included in the coding.) Anecdotally, stereochemistry seems to be given a lower priority in the Chem 12B curriculum, allowing students to focus more on getting the transformations correct and less on exactly which stereoisomers are formed. This relative emphasis might be an alternative explanation to a more sweeping generalization like "graduate students consider stereochemistry more important."

In contrast, undergraduates spent more of their time assessing their progress (AssessProg) and mapping their knowledge about a known reaction onto the current problem (MapTotal). Undergraduates are not more likely to recognize a known reaction, but they do spend more time figuring out how to apply their knowledge. This is particularly evident on the Diels-Alder question, in which undergraduates slowly work through a long algorithm to determine the correct regio- and stereochemistry. The difference in assessing progress is unexpected. One hypothesis is that the graduate students are more confident and feel less need to stop and assess along the way. Additionally, general statements like "this looks so wrong" were extremely common among undergraduates, which may account for a substantial portion of their assessing progress.

If the goal is to transition students to a more expert-like approach when solving these problems, a key area for intervention might be increasing the time students spend developing alternate solutions and then deciding between those solutions. To accomplish this, problem sets might include a more complex question with multiple reasonable answers, and students could be instructed to generate two possible answers and rationalize why one is more likely. Alternately, to focus even more closely on the decision-making process, students could be shown a reaction and asked to explain which of several different given solutions is most likely.

Limitations

Self-selection bias among research participants is a common limitation for this type of study. Students who struggled with their organic coursework are understandably less likely to volunteer their time to be observed while solving extra chemistry problems. As a result, over 75% of the sample was above average, in terms of their final exam score, and few students who scored very poorly participated in this study.

Another limitation is that many of the research participants were at some point directly taught by a member of the research team in the ChemScholars discussion section. This may have influenced their approach to problem solving in a particular direction, possibly making it more similar to the resulting workflow. However, solutions by non-ChemScholars undergraduates and graduate students did not seem to differ substantially in their ability to be modeled by the workflow.

Some outcomes from this study should be generalized with caution. Raw numbers on how much time is spent on each action vary widely across problems, so it should not be assumed that similar values would be reproduced using a different set of questions. Additionally, the workflow is proposed to be a valid model for solving relatively complex predict-the-product problems, but student approaches may differ on a more straightforward set of questions. Indeed, preliminary work with students at the end of their first semester of organic chemistry suggests that a more collapsed model may be applicable in some cases.

CONCLUSIONS AND IMPLICATIONS

Through an iterative process, I have developed a workflow for predicting organic reactivity. It is shown to be an appropriate model for how both more novice undergraduates and more expert-like graduate students solve complex predict-the-product problems, to the extent that one can accurately model a sometimes anarchistic process. Evidence is provided at both the aggregate and individual level to support this claim. The workflow deconstructs a complex problem into individual skills, each of which can be improved through teaching. Various potentially useful interventions can be easily imagined when problems are viewed from this perspective. Encouraging students to name all the functional groups before starting a problem would be a useful way of training the "collecting information" skill that is consistent with the results of this study. If students are struggling to "follow reactive pathways," a lesson on reactive intermediates might be helpful. Teaching these skills also includes simple reframings of what we are already doing. For example, teaching students to reason with reaction coordinate diagrams is one way of developing the "determine major product(s)" skill.

I observed that students who spoke more were more likely to be successful, and it may be that putting thoughts into words can be a helpful strategy, as illustrated by the observed benefit of naming the functional groups for the problems used in this study. Studies explicitly comparing student problem solving with and without thinking aloud have not been conducted in organic chemistry. Work in other disciplines suggests that the act of vocalizing one's thoughts may itself affect the problem-solving process (Schooler, 2002; Schooler et al., 1993). If thinking aloud is inherently beneficial for organic chemistry problems, this would be a concrete suggestion that we can give to our students that they can easily implement and are unlikely to be doing already.

While there were some commonalities, the key moments that led to successful solutions were largely idiosyncratic to the individual problem. However, concepts can be abstracted from each of these problems, and these concepts are likely to appear again eventually in future problems. Having students identify and record these take-away lessons for each problem as a study technique is a strategy that I have used in my own teaching. Requiring that students engage in this type of analysis helps to reinforce the idea that "course content" is introduced not only through lectures or textbook readings, but also through the assigned problems, which are carefully selected to assess specific learning goals valued by the instructor.

The approaches taken by graduate students were not qualitatively different from those taken by the undergraduates. However, they did spend more time identifying solutions and deciding between them. If we consider this to be expert-like behavior that we want to guide our students towards, it would benefit us to give students more opportunities to work on complex, open-ended problems that may have more than one reasonable solution.

Predicting reactivity is a fundamental skill in organic chemistry, but it is often taught gradually over the course of a year, focusing on whatever functional group or reaction type is the subject of the current unit. I believe that due to the central nature of this problem type, it should at some point be addressed as its own topic. The workflow model developed here provides a deeper perspective on how both novices and experts approach complex predict-the-product questions, and the conclusions made about student problem-solving actions provide guidance for how to best support our students to be successful solvers.

CHAPTER 2

Design and Implementation of "Preparation for Future Learning" Lessons in Organic Chemistry

I have developed and implemented two organic chemistry lessons, one on the broad topic of acid-base chemistry, and another on the narrower topic of directing group effects in electrophilic aromatic substitution. These lessons were designed based on the Preparation for Future Learning (PFL) framework, which involves students collectively exploring data to find contrasting cases and "invent" chemical principles. Student performance was measured by pre-tests, immediate post-tests, and scores on relevant exam questions. While the acid-base lesson did not result in an immediate benefit relative to lecture alone, there appeared to be a delayed effect. Students who attended the PFL lesson scored significantly higher on acid-base questions on the final exam, even though this was not the case on the midterms. Assessment on directing group effects suffered from a ceiling effect, making it difficult to draw any conclusions about that lesson. Student feedback on both lessons was overwhelmingly positive. In general, students who attended the lessons felt much more prepared for when the material covered in these lessons was subsequently introduced in lecture. Students also enjoyed the overall format of working with the data to discover trends. Overall, PFL lessons show promise as a useful and active way to familiarize students with various chemical principles prior to their "official" introduction in class.

INTRODUCTION

At many institutions, simple transmission of information through lectures and a corresponding textbook is still the standard method for teaching organic chemistry. These are passive approaches to learning, which generally perform poorly when directly compared to methods of teaching that involve more active engagement with the material (Freeman et al., 2014). However, lectures and textbooks are not completely useless. In fact, there are times when students can benefit from these formats, but they generally must be prepared to do so. A common complaint from students is that they cannot follow organic chemistry lectures, because the rapid introduction of new ideas requires frantic transcription with no time to think about the material as it is being presented. Students are encouraged to read the text prior to lecture, but even students who take this advice often struggle, suggesting that students may benefit from more guided exposure to the course content before it is more formally introduced.

Interacting with the material prior to "officially" learning it can potentially provide the cognitive architecture needed to readily assimilate the new information when presented as a lecture or text. This is the central idea of the Preparation for Future Learning (PFL) framework (Bransford & Schwartz, 1999; Schwartz et al., 2005; Schwartz & Bransford, 1998; Schwartz & Martin, 2004). Lessons based on this framework typically involve students exploring sets of data, paying particular attention to contrasting cases. After students have had a chance to explore and "invent" principles to explain the trends they see, the canonical explanations for these phenomena are discussed (Schwartz & Martin, 2004). Schwartz and Bransford (1998) theorize that actively noticing distinctions between contrasting cases helps students focus on the most important features and generate differentiated knowledge structures. Having this introduction to the material creates a "time for telling," in which "learners are prepared to be told the significance of the distinctions they have discovered" (Schwartz & Bransford, 1998). It is at this point when lectures and expository texts are at their most useful.

Meta-analyses of research on contrasting cases and the related Problem Solving followed by Instruction (PS-I) strategy provide some guidance as to best practices when developing this type of lesson. In particular, building on student solutions during the lesson and the use of contrasting cases that differ by only one "deep" feature (rather than rich cases that have a variety of contrasts) seem to be helpful for subsequent problem solving (Loibl et al., 2017). Focusing more on similarities than differences and explicitly presenting the principle to be learned after the lesson are also associated with greater effects relative to various types of controls (Alfieri et al., 2013). These meta-analyses cover a variety of academic disciplines, with much of the focus on problems from physics, mathematics, and psychology. However, none of the studies reviewed in these publications focus on chemistry, indicating the need for further research in this domain.

More recently, it has been suggested that learning through contrasting cases might be fruitfully applied to organic chemistry (Graulich & Schween, 2018). In subsequent studies, case comparisons have been used to elicit student reasoning about which of two reactions proceeds faster (Caspari et al., 2018; Rodemer et al., 2020), and case studies have been reported on how this type of reasoning changes over time (Watts et al., 2021). However, to my knowledge, no PFL lessons based on contrasting cases have been developed and assessed within the field of organic chemistry. During the 2019-2020 academic year, I developed and conducted two PFL lessons that utilize contrasting cases. (A third lesson was also designed and implemented but will not be discussed in this chapter, see Appendix 2.1.) Quiz scores, exam scores, and student feedback were examined to address the following research questions:

1. How do students who have participated in PFL interventions perform on quiz and exam questions relevant to the topics being covered?
2. To what extent and in what ways do students perceive PFL lessons as being beneficial?

LESSON DESIGN

Lesson 1 – Acid-Base Chemistry

The first PFL lesson focused on having students identify structural features that help predict relative acidity or basicity. The six target trends were charge, atomic size, electronegativity, hybridization, resonance, and induction. Students were divided into groups and provided a large dataset of 40 molecules with their pK_a values. The molecules were chosen so that many comparisons could be made between molecules that differ by only one variable. Students were not explicitly told which molecules to compare; they had to create their own contrasting cases. Rather than a simple table of values, students were provided pK_a data in the form of playing cards like those shown in Figure 1 (see Appendix 2.3 for the complete set). The design of these cards is explained more fully in the following section. Working with cards rather than a single sheet of data allows for more interactive group work, and I have generally observed that students enjoy working with data in this way.

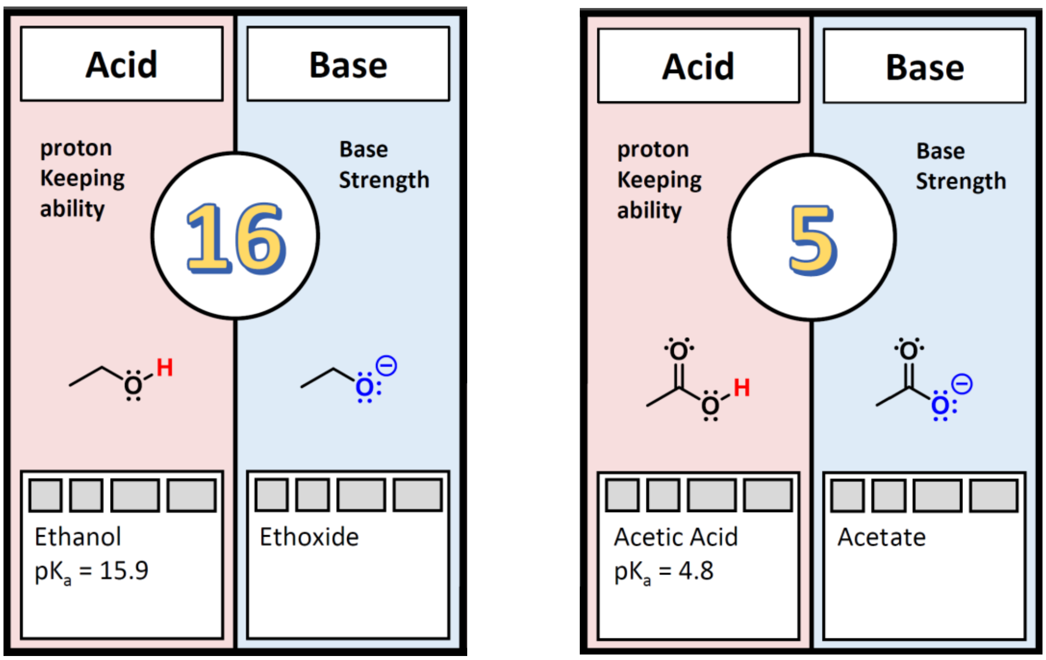


Figure 1. Example pK_a cards

In addition to the cards, students were given one sheet explaining the different pieces of information on the pK_a cards (see Appendix 2.4), another with a "pK_a ruler" (Appendix 2.5), and a third for them to take notes. The notetaking sheet included the following instructions and guiding categories:

> <u>Instructions</u>*:* Examine the acid-base cards given and look for patterns that might help you determine what factors affect acidity and basicity, and how changing those factors modifies the acid-base properties of a molecule. As you discuss, record any ideas that are proposed by members of your group. (They do not have to be correct!)
> a. Primary Factors affecting acidity and basicity
> b. "Rules" or trends for how changes in these factors affect acidity and basicity
> c. Chemical rationale for observed rules and factors
> d. Miscellaneous

Students were allowed to explore and discuss the data together for 15 minutes before being brought back together for a full-section discussion. Preliminary ideas were then solicited from the groups and jotted down on the board. During this time, students were prompted to think a bit more broadly. For example, a given rule is that nitrogen is more basic than oxygen, all else being equal, but is there some property that could be invoked to generate a more general rule than this? After another 15-minute exploratory period, more ideas were recorded on the board in this way. This was followed by a short ~7-minute mini-lecture (the "learning" that students had been preparing for). The focus was on identifying the six target trends by building on student answers, showing how all of the comparisons they noticed are more specific examples of one of these broader trends, or simply another way of stating the same thing (e.g., row in the periodic table, rather than atomic size). All of these trends were then further connected by showing how they are manifestations of two fundamental chemical principles: allowing electrons/charge to spread out is stabilizing (charge, atomic size, resonance, induction), and electrons are more stable in lower energy orbitals (electronegativity, hybridization).

pK_a Cards

The design of the cards themselves is based on ideas about how to familiarize students with common acids and bases in a way that would lead them to think about pK_a values in a useful manner. The overall structure of the cards corresponds to ideas that I have tried to instill in my students over many years of teaching organic chemistry. Early drafts were discussed with undergraduates, graduate students, and organic faculty, and the specific information included in the design was refined in subsequent versions.

A central goal when designing these cards was for students to think of pK_a values as corresponding to a conjugate pair of molecules, not a single acid. This is especially necessary for amphoteric compounds like water, which students need to associate with different pK_a numbers depending on whether it is acting as an acid (16) or a base (-2). Additionally, I postulated that referring to -2 as the "base strength" of water would be more helpful than calling it the mouthful "the pK_a of the conjugate acid." I also wondered whether there was a more direct interpretation of pK_a than "the reverse of the acidity." Acidity can be thought of as proton-donating ability, but pK_a is a more direct measure of proton-*keeping* ability. Delighted by the acronym, I used this phrase on the cards.

In addition to thinking of acids and bases in pairs, I wanted students to group acids and bases into tiers, rather than attempting to memorize a continuous spectrum of pK_a values. Each tier has a simple prototype molecule (acetic acid, water, acetone, etc.) that represents the approximate pK_a for the whole family. Because this organizes molecules into clusters instead of lists in which each acid has a unique value, this method for learning approximate pK_a values is consistent with research on human memory, which has found clustering to be a key strategy for efficient recall (Bousfield, 1953). Reflecting the emphasis placed on this organization, the pK_a values associated with each tier are prominently displayed in the center of the card.

The boxes at the bottom of the cards contain additional information. The common names for both the acid and base are provided to familiarize students with common names, which can be a slow and haphazard process when not done intentionally. They also contain exact pK_a values, in order to provide more accurate information and to show that inductive effects do modify the exact pK_a value but generally do not change the tier of a given functional group. In their original form, these cards had structural features in place of the gray boxes (Figure 2). The intention was to highlight the fact that atom identity, charge, hybridization, and presence/absence of resonance were the main features contributing to pK_a values. However, because these were the very features I wanted students to "discover" in the PFL lesson, they needed to be covered up, so empty boxes were used instead.

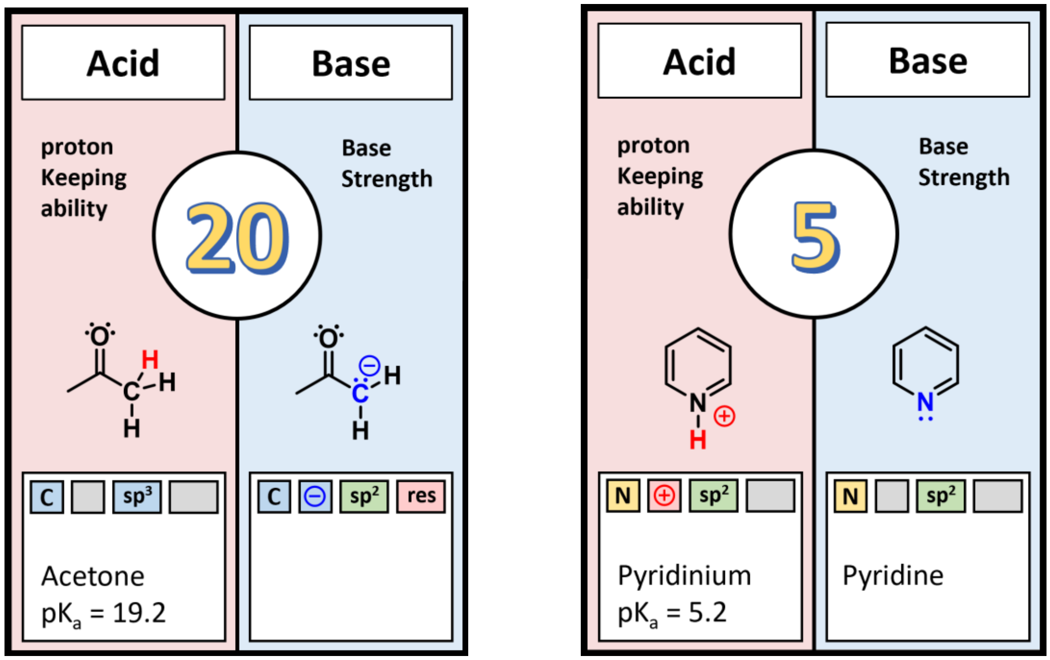


Figure 2. Example pK_a cards with additional data

<u>Lesson 2 – Electrophilic Aromatic Substitution</u>

The goal of the second PFL lesson was for students to gain an understanding of directing group effects in electrophilic aromatic substitution (S_EAr) reactions. Students often approach this topic with brute force memorization, simply remembering for each functional group whether it is an *ortho*/*para* or *meta* director. However, this strategy becomes problematic when encountering a novel functional group, and students who take this approach can be confused by groups like esters that have different effects depending on how they are connected to the ring. Directing group effects are generally taught by simply showing students the resonance structures that help rationalize the observed reactivity. However, students already have the tools they need to figure out this rationalization for themselves. Ideally, at this point students are comfortable drawing resonance structures, and they understand concepts like the instability of carbocations or the unfavorability of placing like charges close together. Combining those concepts with data on the regioselectivity of S_EAr reactions, it is possible for students to rationalize the data in ways that are similar to expert interpretations.

For this exercise, students were given a set of 20 electrophilic aromatic substitution reactions on monofunctionalized benzenes and a straightforward question: How can the regioselectivity of these reactions be predicted? A set of example reactions is shown in Figure 3 (see Appendix 2.6 for the full set). Reactions were chosen to illustrate a variety of *ortho*/*para* and *meta* directors, and two sets of conditions were given for each substrate to illustrate that the regioselectivity is primarily determined by the aromatic substituent and not the functional group being added. Because students were not yet familiar with this type of reaction, a mechanism for the bromination of nitrobenzene was written on the board, but no additional information was supplied. As with the acid-base lesson, data was provided on small strips of paper that could be easily sorted and moved around by the small groups. Once most groups had developed hypotheses that were on the right track (~20-30 minutes), the instructor discussed the basic principles of directing group effects.

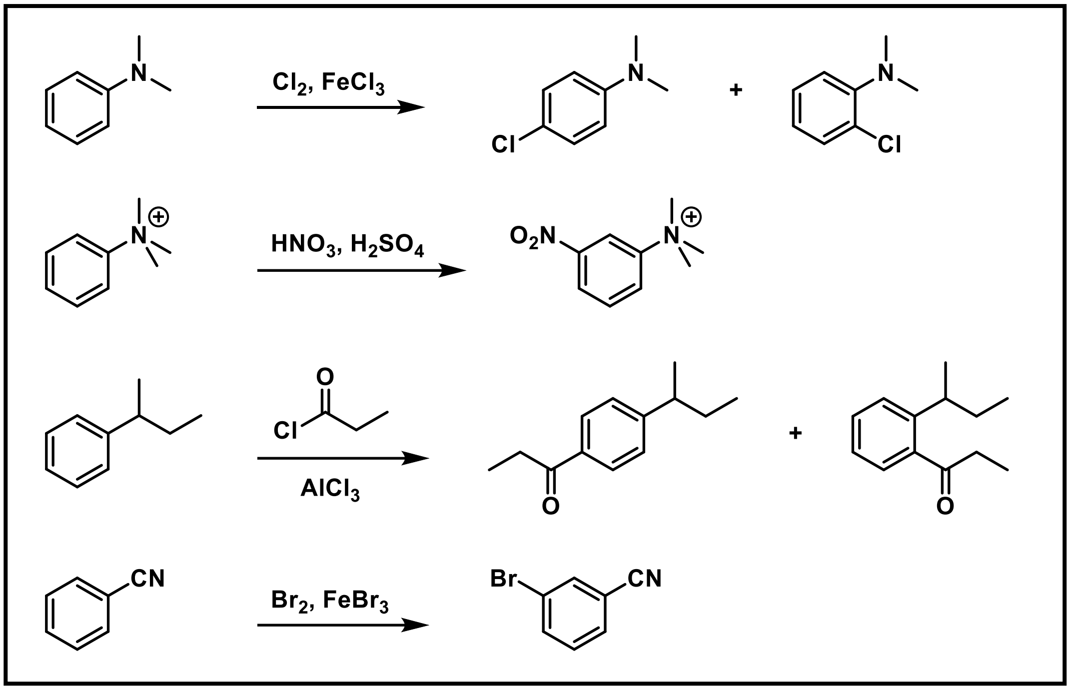


Figure 3. Example data for S_EAr lesson

METHODS

Lesson Implementation

Participants and Context

All work was conducted at a large, research-intensive institution in the Western United States. Undergraduate student participants were recruited from Chem 12A and 12B, the two-semester organic chemistry sequence for students majoring in chemistry, chemical biology, or chemical engineering. A recruiting announcement was made to the entire Chem 12A/B course in the week prior to each lesson. Additionally, an announcement was made to the ChemScholars discussion section. ChemScholars is an optional discussion section run by advanced undergraduates that tracks along with the Chem 12A and 12B courses. Both lessons were conducted by the research team during the regular ChemScholars time slot, and as a result, most of the participants were frequent ChemScholars attendees. This includes 58% of Lesson 1 participants and 79% of Lesson 2 participants. On average, students who attended these lessons scored slightly above the mean on their final exams. Participants in Lesson 1 averaged 74% on their final exam, which was significantly higher than the 66% mean for the full class ($p < 0.05$). Lesson 2 participants scored an average of 62% on their final, which was not significantly higher than the 57% class mean.

Lesson 1 – Acid-Base Chemistry

The first lesson took place in the first few weeks of Chem 12A, before students had covered any acid-base chemistry in lecture. A set of 30 students, 26 of whom agreed to participate in the study, was randomly assigned to either a PFL or a control lesson. Students in both sections were first given a 10-minute pre-test containing 5 questions relevant to acid-base chemistry (Appendix 2.7). The PFL lesson was then conducted as described above. The control lesson involved a typical lecture format (~10 minutes) followed by small group problem solving for the remainder of the session (see Appendix 2.8 for the worksheet used). Both sections used similar molecules in their examples to avoid introducing extra variables. After the lesson was complete, students were given another 10-minute quiz with 5 questions analogous to the pre-test (Appendix 2.9). A total of 22 students completed the post-test, including 12 in the experimental section and 10 in the control.

Lesson 2 – Electrophilic Aromatic Substitution

The second PFL lesson took place the following semester, just before the students started learning about electrophilic aromatic substitution. Students were not given a pre-test in this case, because they had not learned about this overall class of reaction prior to the lesson, so they would only have been able to guess randomly about the correct answers. The lesson was implemented as described above, and then students were given a 10-minute, 5-question quiz on the regioselectivity of aromatic bromination (see Appendix 2.10). None of the molecules used as substrates for the quiz questions had been used as examples during the lesson. Because the second semester course is smaller, only 14 students were recruited to participate, and no control lesson was offered.

Data Collection

Observational Data. During the group work portions of both lessons, graduate student researchers and/or undergraduate student assistants walked around and took notes on what they overheard students talking about. (See Appendix 2.11 for a discussion.)

Pre/Post Tests. Students were given short quizzes at the beginning of Lesson 1 and the end of both lessons, as described in the lesson implementation. The questions on the pre- and post-tests for Lesson 1 were aligned so that the same basic topics (e.g., circle the strongest acid, predict the position of this acid-base equilibrium) were covered on each quiz. The substrates on the quiz for Lesson 2 were chosen to be analogous to but different from the molecules that students worked with during the lesson.

Exams. Midterms and final exams from Chem 12A and 12B were examined to identify questions relevant to the learning goals of each lesson. The full list of relevant questions for both lessons can be found in Appendices 2.12 and 2.13.

Student Feedback. To determine whether these lessons had any long-term impact, a Qualtrics survey was sent out the following academic year to all students who attended one of the lessons. The timing of this survey was 15 months after Lesson 1 and 10 months after Lesson 2. The survey consisted of the following 4 questions:

1. What do you remember about the lesson(s) you attended?
2. Can you recall something specific that you learned? If so, please describe.
3. Did you find this type of preview lesson useful when later learning the material in lecture?
4. Would you recommend optional lessons of this type to other Chem 12 students?

Students were asked in the recruitment email whether they remembered either lesson. For those who did not, it was requested that they reply to the email saying so. Only students who remembered the lesson(s) were asked to take the survey. Of the 30 students contacted, 10 emailed back to say they did not recall the lesson, while 11 took the survey, for a response rate of 70%.

Data Analysis

Acid-Base Lesson. Student performance on the pre- and post-tests in the acid-base lesson were compared by a paired t-test. Pre-test and post-test scores between the PFL section and the control section were also compared by a t-test. For the exam data, the total score of all questions requiring acid-base knowledge were calculated for each student on each exam. Comparisons on total acid-base scores were made using pairwise t-tests between the PFL section, the control section, and those who had attended neither lesson.

Electrophilic Aromatic Substitution Lesson. Because there were no pre-tests or control lesson, the primary comparison to make was between exam data for students who did and did not attend the lesson. Independent t-tests were run on each of the four subsequent exam questions that tested students on S_EAr regioselectivity.

Student Feedback. Student survey responses were descriptively coded to identify common themes. After five major categories of comments were identified, the dataset was fully recoded to identify all instances of each category and representative quotations.

RESULTS

<u>Lesson 1 – Acid-Base Chemistry</u>

Pre/Post-Tests

Overall, students in both sections scored significantly higher on the post-test than on the pre-test ($p < 0.001$) (Table 1). Most of these gains are due to improvement on problem 1a (circle the stronger acid) and problem 2 (circle the stronger base). Students in the control section scored slightly higher on average for both the pre- and post-test, but neither of these differences were significant. The experimental section did improve more than the control group from pre- to post-test, but this was not a significant difference.

Table 1. Student quiz performance by section

Test	PFL Lesson (n = 12)	Control Lesson (n = 10)
Pre-lesson	30%	46%
Post-lesson	65%	70%

Table 2. Student quiz performance by question

Question	Pre-Test	Post-Test
1a. Circle the stronger acid	41%	100%
1b. Circle the stronger acid (given pK_a)	81%	86%
2. Circle the stronger base	14%	86%
3. Favored side of acid-base equilibrium	50%	64%
4. Rank protons by acidity	41%	41%
Total	37%	67%

Exam Questions

Examples of the Chem 12A exam questions most relevant to acid-base chemistry are shown in Figure 4 (see Appendix 2.12 for the full set). Some of the problems *explicitly* ask students to consider acidity or basicity, whereas others require such knowledge *implicitly* but do not mention acids or bases in the problem. Most of the explicit questions ask students to identify the stronger acid or base, while the implicit ones include mostly predict-the-product questions that require noticing favorable proton transfers.

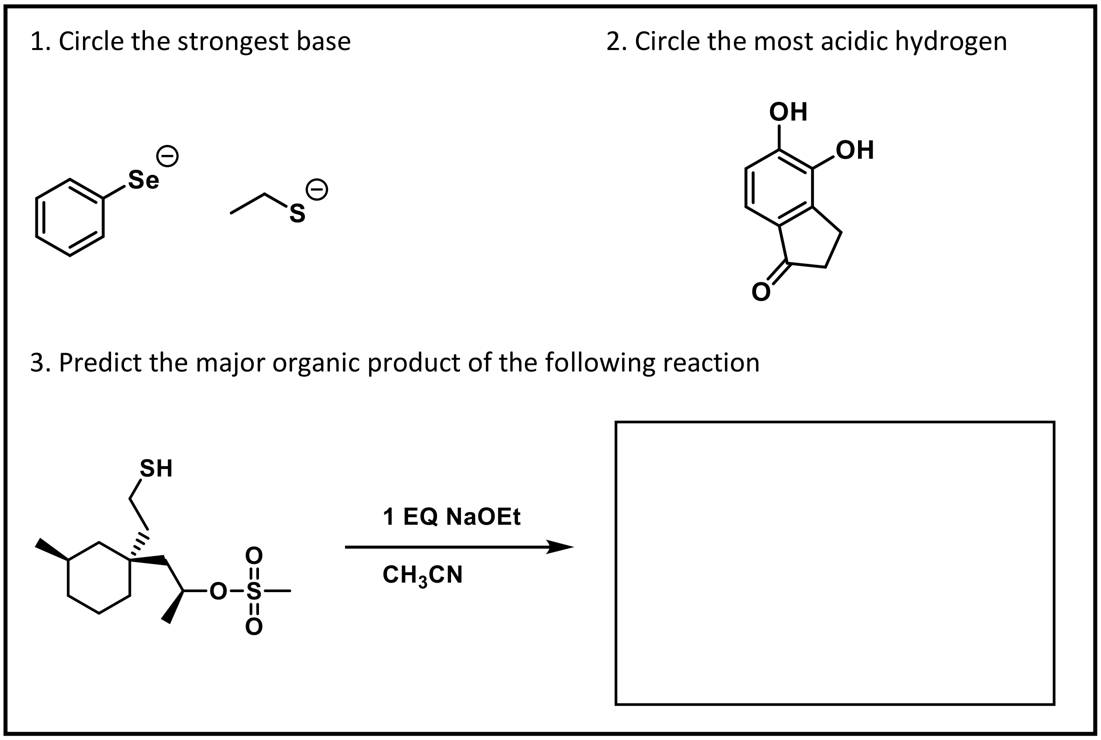


Figure 4. Example exam questions requiring acid-base knowledge

An interesting trend in the data is that the students who attended the PFL lesson gradually started to do better on acid-base questions relative to their peers. Looking at the first midterm, it was disappointing to see that neither the PFL nor the control lesson seemed to have a measurable benefit on acid-base questions, relative to lecture alone. However, on the final, the difference between PFL students and those who attended neither lesson was large enough to be statistically significant.

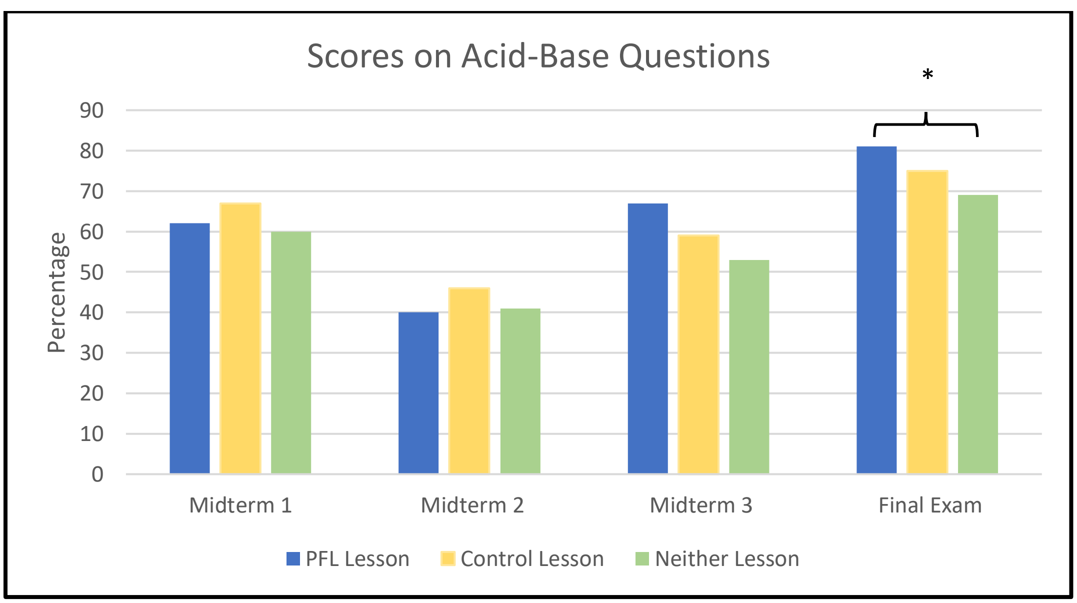


Figure 5. Student scores on exam questions requiring acid-base knowledge (*$p < 0.05$)

Lesson 2 – Electrophilic Aromatic Substitution

Post-Test

Overall, students did quite well on the quiz taken directly after the lesson, with half of the students achieving a perfect score. All but one of the 14 students correctly identified a C-linked ester as a *meta* director, and the rationale given for this choice included either resonance structures, the phrase "electron-withdrawing group," or both. Benzenesulfonic acid turned out to be the most difficult for these students, with only 57% of them correctly classifying the $-SO_3H$ substituent as electron-withdrawing, essentially flipping a coin. Getting a correct answer did not seem to depend on whether the student drew resonance structures, nor did correct answers seem to cluster within any particular discussion group(s). Students did well on predicting amino and phenyl group directing effects, with only one student on each problem misclassifying them as *meta* directors. No one misclassified the methyl groups on the final problem (*m*-xylene), although two students did not write an answer.

Exam Questions

The midterm following this lesson was the first exam after the COVID shutdown. It was written as an in-class exam, but because of the circumstances, students took the exam at home and were allowed to use notes and the textbook. As a result, scores were very high; the average score on this exam was 84%. Consequently, any comparisons made between students are limited by a ceiling effect, so no significant differences were observed between students who did and did not attend the S_EAr lesson.

The problems from Midterm 2 that required students to determine the regioselectivity of an electrophilic aromatic substitution reaction are shown in Figure 6. Gratifyingly, students who attended the lesson all received perfect scores on question 1. However, so did 90% of the class, and no one scored lower than 5/6. Additionally, the entire class received full credit on question 2. For the synthesis problem, all tests were examined for errors in S_EAr regioselectivity, and only 2 instances out of 110 students were found, neither of whom participated in the lesson.

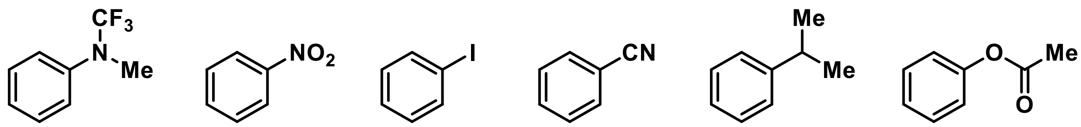
The final exam was designed with the extra student resources in mind, and so student performance was more typical for an organic exam; the average score was 57%. Only one question involving S_EAr regioselectivity issues was included on the test (see Figure 7). The average scores for this question were marginally higher for those who attended Lesson 2 (82% vs. 71%), but this difference was not significant.

Student Feedback

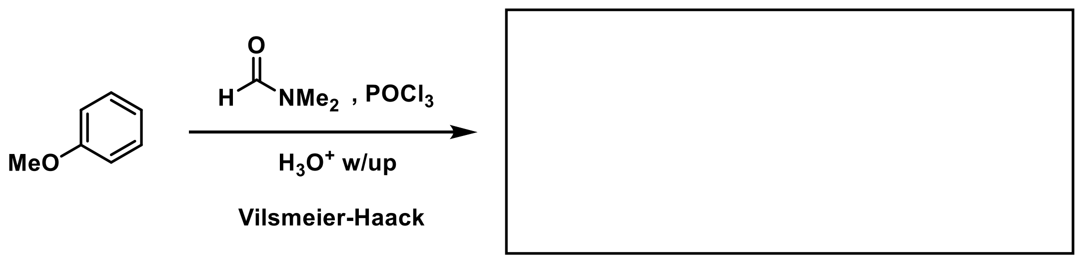
Of the 30 students who attended one of these lessons, 21 students responded to the survey request, 11 of which (52%) could recall the general format of the lesson(s) they attended. Many of them also remembered specific details, such as, "I remember sorting the compounds according to their acidity/basicity on the blackboard," and "I remember working in pairs/groups to take a set of reactions (printed on slips of paper) and come up with a set of rules (or an explanation) to determine the product (e.g., would the substitution occur *ortho*, *meta*, or *para*). At the end of the lesson we reconvened as a class to talk about the results we came up with." Some students could even state the learning goals of the lessons: "A more stable conjugate base/conjugate acid = a weaker acid/base. There are some factors that influence the stability of the conjugate acid or base such as charge, size of atom, resonance, and hybridization," and "I remember figuring out from the SeAr reaction images that electron donating substituents are *ortho*/*para* directors and that electron withdrawing substituents are *meta* directors."

All 11 students would recommend these lessons to their peers, primarily because of how much easier it was to follow along with subsequent lectures. In a recommendation to their peers, one student wrote, "these lectures will give you the background knowledge necessary to not fall behind the next day

1. Indicate which of the following substituents would direct *ortho*/*para* (o/p) or *meta* (m).



2. Predict the major organic products of the following transformation.



3. Propose a reasonable synthesis of 3-ethylbenzonitrile starting from benzene and any inorganic or organic reagent with two or less carbon atoms.

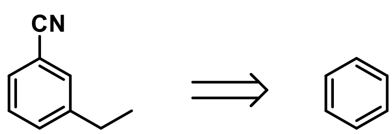


Figure 6. Midterm 2 questions involving directing group effects

1. S_EAr bromination of the following substrate with NBS yields one product. Predict the structure and justify your answer with a clear drawing of resonance structures.

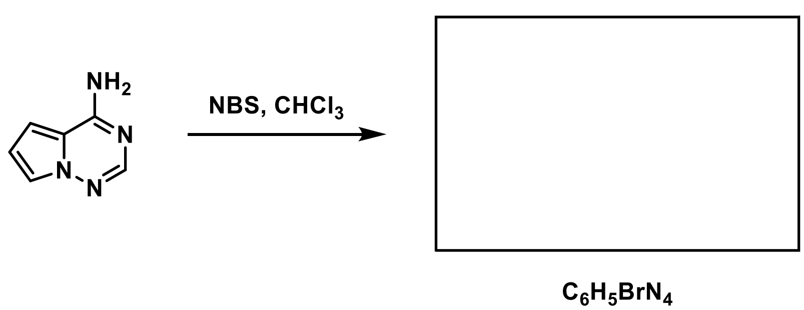


Figure 7. Final exam question focused on S_EAr regioselectivity

in class. More of these lectures would help students feel useful in their lecture, and not like they wasted their time trying to keep up with the pace of the professor." This sentiment was echoed in many other survey responses.

Some students specifically identified the exploratory nature of the lessons as both fun and helpful for deeper learning: "It was fun to find the patterns myself as though I was analyzing results of the original experiment, and it helped the resulting concepts stick in my head."; "I actually felt like I understood these topics better because I wasn't immediately taught some rule but instead had to come up with a reasoning myself. I also felt like I had greater 'trust' in the rules/concepts governing these reactions because I had taken a minute to reason these concepts on my own and ask questions in a smaller learning environment." These are additional factors that should be taken into account when assessing the effectiveness of these interventions.

DISCUSSION AND CONCLUSIONS

The first research question for this study concerned how students who have participated in PFL interventions perform on quiz and exam problems relevant to the topics being covered. It was found that student performance is improved when assessed immediately after the lesson, but medium-term benefits (i.e., better scores on the next midterm relative to students who did not attend a PFL lesson) were not observed. However, those who attended the PFL lesson on acid-base chemistry did score significantly higher on relevant questions on the final than those who attended neither lesson. A plausible interpretation of this outcome would suggest that while PFL lessons may not show immediate benefits relative to a control lesson or the lecture alone, they appear to result in a more persistent understanding of the topic. It is possible that the deeper learning proposed to happen during PFL lessons is most detectable after the knowledge obtained via other methods of learning the material starts to fade. A similar effect was seen in a study comparing PFL and "tell-and-practice" lessons in the context of calculus (Stratton, 2020). Stratton (2020) found that "although participants' performance tended to decrease on PFL problems from the post-test to the delayed post-test across groups, their performance decreased less for the [PFL] condition." Even more evidence of persistent learning is indicated by the fact that survey respondents could recall the learning goals for these lessons a year after the fact. One wonders how many experimental lessons have been deemed ineffective by not waiting long enough to see their true value.

Students who attended the electrophilic aromatic substitution lesson did quite well on relevant questions both immediately after the lesson and on subsequent exams. However, students who only attended lecture also did well on these exam questions, resulting in a ceiling effect. This is consistent with the PFL literature in which the benefit of PFL lessons relative to more passive approaches to learning appear only on more complex tasks, not on simple recall of factual information (Schwartz & Bransford, 1998). Assessment questions specifically designed to detect PFL learning outcomes may be required to observe a clear positive effect from this type of lesson. Additionally, because it was not possible to include a pre-test or a control group, isolating the effects of the lesson itself proved difficult. To increase the impact of this lesson, it may be worth expanding on this relatively narrow topic to also include data on the reactivity of polyfunctionalized aromatic molecules, in which directing group effects must be weighed against each other to determine regioselectivity.

The second research question for this study addressed the extent to which students perceive PFL lessons as being beneficial. A search of the PFL literature did not reveal any studies in which students were asked to directly comment on their experience participating in this type of lesson. Student feedback on the lessons was universally positive. During the lessons, students were continuously engaged in discussion over the data, and as mentioned in the surveys, some found this to be a fun activity. Multiple students commented that the lessons made them more confident in the material, which allowed them to follow along in lecture rather than just trying to rapidly transcribe without thinking about the chemistry. This common reaction suggests that, at least for some students, the goal of preparing students for future learning was achieved.

Two primary limitations of this study involve the populations of students who attended each lesson. Self-selection bias is an issue, with more motivated and potentially higher performing students being more likely to participate. However, the very highest performing students are less likely to participate, and overall exam scores were not drastically different between students who did and did not attend the PFL lessons. Another limitation is that most lesson participants were also regular attendees of the ChemScholars discussion section, so it is difficult to determine whether the positive effects seen in the PFL students from Lesson 1 were due to a single lesson or a more long-term effect. However, because this was the primary lesson on the topic of acid-base questions and a corresponding increase was not seen in the control lesson, it is likely that the observed effect is largely due to the PFL lesson itself.

Building engagement and confidence in our students is an elusive goal, and lessons that accomplish this feat while also encouraging deep learning should be prioritized. I agree with the sentiment expressed by one student, “I think the fact that I can speak to these lessons a year after they were given is a real testament to their strength.” Overall, PFL lessons on these topics were well-received by the students, and PFL attendees performed as well or better on exam questions than students who only attended traditional lessons. Instructors who primarily teach via lecture might want to consider augmenting their approach with PFL lessons, particularly in areas like acid-base chemistry that are fundamental to the entire course.

CHAPTER 3

Evidence of knowledge integration in the context of undergraduate research: Assessing oral and written research artifacts

Understanding the impact of undergraduate research experiences (UREs) and course-based undergraduate research experiences (CUREs) is crucial as universities debate the value of allocating scarce resources to these activities. I have designed and tested the BURET instruments, a new set of tools designed to assess the learning outcomes of UREs and CUREs in the sciences. To study the BURET instruments, they were administered to 89 undergraduate students, and the performance of students who had less than one year of undergraduate research was compared to those with more than one year of research experience. Students were assessed on four primary dimensions based on written reflections and poster presentations for their own research project: communicating the significance of their project, analyzing their experimental design, interpreting their data, and proposing future research. The instruments were found to yield reliable scores and helped clarify the impacts of undergraduate research, providing insight into the strengths and weaknesses of undergraduate researchers at this institution. Students with at least a year of research experience were able to use disciplinary evidence more effectively than those with less than one year of experience. Novice students excelled at explaining the societal relevance of their work, but they incorporated only minimal discussion of prior research into their reflections and presentations. Students at all levels struggled to critique their own experimental design. These results have important implications for undergraduate learning, suggesting ways for faculty members, graduate student research mentors, and CURE or URE programs to optimize undergraduate research experiences.

INTRODUCTION

Research experiences are a critical component of undergraduate education for many students majoring in science, technology, engineering, and mathematics (STEM) disciplines, allowing them to engage with the larger scientific enterprise while still completing relevant coursework. Although research experiences vary widely in nature, they generally share common goals across settings, such as developing research skills, improving and applying understanding of scientific content knowledge, expanding scientific reasoning skills, increasing confidence for doing science, and integrating students into scientific culture (Linn et al., 2015; National Academies of Sciences & Medicine, 2017; Robnett et al., 2015; Rodenbusch et al., 2016). Research experiences can serve as a positive influence for career aspirations involving science, despite other challenges students may have faced since entering their college or university. As a result, many institutions and funding organizations across the U.S. have dedicated considerable resources to support these programs each year (Auchincloss et al., 2014; Krim et al., 2019; Laursen et al., 2010).

To determine how students progress over time and to identify how research experiences can be improved to better serve participants, it is important to assess the impact of science research experiences on student learning, including both course-based undergraduate research experiences (CUREs) and undergraduate research experiences that take place in research laboratories (UREs) (Auchincloss et al., 2014). Most previous studies that assess learning outcomes of science research experiences are limited to a description of the research experience or self-report data; fewer studies validate self-reports with analysis of research products, direct measures of mastery of scientific content or practice, or observations of student activities (Krim et al., 2019; Lin et al., 2019; Linn et al., 2015; National Academies of Sciences & Medicine, 2017). There is a need for additional assessment tools that can be applied to undergraduate research experiences for a variety of different STEM research projects both inside and outside of the classroom.

Research literature on undergraduate research and educational policy documents have identified the following scientific practices as foundational to the URE experience (Laursen et al., 2010; Sadler et al., 2010): formulating research questions or hypotheses, designing experiments, analyzing and interpreting data, making conclusions, iteratively planning next steps, and explaining the significance of the research project. Educators have moved away from the idea that such skills involve a single cognitive activity, viewing them instead as a “set of different but coordinated skills” (Opitz et al., 2017). Thus, the goal of this study was to develop assessment tools to be used with students across STEM disciplines that measure the extent to which they understand research as a set of connected practices. A specific aim for the study tool design was to focus on assessing students in the context of their own research project, rather than investigating their ability to answer questions about a hypothetical scenario. In this study, I address the following research questions:

1. Do the tools I developed distinguish between students with different levels of prior research experience?
2. What do these tools reveal about what students understand about research and what they are still learning at different stages of their undergraduate careers?

<u>Theoretical Framework</u>

The theoretical basis for this work comes from Knowledge Integration (KI), a framework that has been used extensively in the design of learning environments and instruments to assess K-12 student knowledge of scientific content and practices (Linn et al., 2018; Linn, 1995; Linn & Eylon, 2011; Ryoo & Linn, 2012; Stone, 2014). This learning science framework emphasizes that coherent understanding occurs when students make deep connections between their prior and new ideas. KI specifies four key components that support student learning (Linn & Eylon, 2011). The first is to elicit student ideas and prior understandings about a given topic. Students already have a repertoire of knowledge to draw on, and new knowledge will ultimately be built on these existing structures. Second, as students engage more with a particular concept, they add new, scientifically normative ideas, some of which may challenge or contradict existing ideas. Third, students begin to distinguish between competing ideas and the contexts in which they are applicable, in order to create a more nuanced understanding of the topic. Finally, students reflect upon their new knowledge to consolidate it into a coherent narrative.

The process of conducting research generates knowledge in a way that parallels the KI framework (Linn et al., 2015). Activities such as predicting and hypothesizing allow for *eliciting* undergraduate students' initial ideas. Undergraduate researchers then begin *adding* new ideas over time as they gather data (Linn & Eylon, 2011; White & Gunstone, 2014), and they gradually learn to *distinguish* between possible interpretations for their data. Later, *reflecting* on their research will consolidate knowledge and generate new ideas for future work (Brown et al., 1989; Linn & Eylon, 2011). KI guides an expectation that as undergraduates progress in research, they will become more proficient in understanding and discussing their research project, integrating relevant discipline-specific content knowledge when appropriate.

Literature Review

Impacts of Student Participation in Science Research Experiences

A number of studies suggest that gains related to retention in STEM (e.g., graduation rates, entry into the STEM workforce, graduate school attendance) are supported through participation in research experiences, especially for students from groups historically underrepresented in STEM fields (Carpi et al., 2017; Estrada et al., 2016; Schultz et al., 2011). There is increasing evidence to link participation in authentic scientific research with the development of science identity through immersive learning of discipline-specific practices, referred to as "legitimate peripheral participation" in situated learning theory (Lave & Wenger, 1991; Robnett et al., 2015). Factors such as a positive science identity, self-efficacy development, access to mentoring, and engagement in research at the undergraduate level have been found to be important for persistence in STEM and are critical to supporting students from historically underrepresented minority groups (Carlone & Johnson, 2007; Chang et al., 2011; Mondisa & McComb, 2018; Ortiz et al., 2020).

Measures of Student Learning in Science Research Experiences

Various performance assessments have been developed to directly measure multiple dimensions of student knowledge and skills gained from participation in science research experiences (Butz & Branchaw, 2020). For example, the Biological Experimental Design Concept Inventory (BEDCI) measures knowledge and diagnoses non-expert-like thinking in experimental design by analyzing open-ended responses to different scenarios (Deane et al., 2014). The Assessment of Critical Thinking Ability (ACTA) is an open-ended survey that assesses critical thinking skills in biology and chemistry students (White et al., 2011). The Rubric for Experimental Design (RED) identifies areas of experimental design in which biology students struggle (Annwesa P. Dasgupta et al., 2014, 2016). The Performance assessment of Undergraduate Research Experiences (PURE) instrument measures experimental problem solving and quantitative literacy skills in chemistry students participating in UREs through a series of multipart questions about real-world scientific problems (Harsh et al., 2017; Harsh, 2016). Designed for use in biology classrooms, the Test of Scientific Literacy Skills (TOSLS) consists of multiple-choice questions about real-world problems and measures student skills related to scientific literacy (Gormally et al., 2012). Recently, Crawford and Kloepper (2019) developed an exit interview involving a series of written and oral exercises that assess the ways in which chemistry students connect course content to laboratory activities.

Many of the instruments developed are “authentic assessments,” which are meaningful opportunities for students to integrate and apply their knowledge to novel, complex, and/or realistic situations that simulate typical activities of scientists (Doğan & Kaya, 2009; Laungani et al., 2018; Wiggins, 1998). For example, the Experimental Design Ability Test (EDAT) gives introductory biology students a real-world scenario and research question and tasks them with designing an appropriate experiment (Sirum & Humburg, 2011b). To assess scientific reasoning skills in science classrooms, Timmerman and colleagues (2011) created the Rubric for Science Writing to assess written laboratory reports across various undergraduate-level biology courses, which simulate how scientists record and communicate their findings. The tool to assess interrelated experimental design (TIED) was developed for use in an introductory biology CURE, in which students are tasked with designing an experiment to answer a specific research question (Killpack & Fulmer, 2018).

Studies that measure student learning gains typically consider these gains only over the course of a semester-long research experience, though there is evidence to suggest that undergraduates need to participate in high-impact research experiences spanning more than one semester to develop their understanding of the research process (Corwin et al., 2015; Deane et al., 2014; Griffeth et al., 2015; Harsh, 2016; Hernandez et al., 2018; Remich et al., 2016). There is compelling evidence to suggest that participation in a CURE leads to significant gains in research skills and academic outcomes and can support the subsequent advancement to (and success in) a URE (Krim et al., 2019; Rodenbusch et al., 2016).

Prior Gaps Identified in Learning for Undergraduate Researchers

Previous studies point to a lack of mastery among undergraduate researchers in fully understanding their research projects in several areas (Airey & Linder, 2009; Coil et al., 2010; Gormally et al., 2012). Prior findings suggest it is possible for a student to participate in research without understanding the scientific or societal significance of their work, though this skill supports more expert-level reasoning in the discipline of the research project (Bransford et al., 2000b; Coil et al., 2010). Many students, and in particular those from groups underrepresented in STEM fields, choose a STEM pathway in order to make a positive contribution to their communities and/or society, and this interest is likely to influence their commitment to a career in STEM (Bonous-Hammarth, 2000; Chang et al., 2014; Harackiewicz & Hulleman, 2010). However, undergraduates do not always develop the ability to articulate answers to questions about the context of their research project, such as: “Why is this question important to others in this discipline?” or, “What is the ‘big picture?’” (Timmerman et al., 2011).

In order to become independent researchers, undergraduates are also expected to develop an understanding of experimental design (Killpack & Fulmer, 2018; Sirum & Humburg, 2011b). Undergraduates are typically presented with narratives about experiments as part of their STEM coursework, but training in experimental design is less common (Gormally et al., 2012). When reading scientific papers, undergraduates commonly struggle with evaluating and critiquing the experimental design used (Coil et al., 2010; Varela et al., 2005). Guided-inquiry laboratories, CUREs, and UREs, in which students design their own experiments, can be used to support experimental design skills; this is especially true for lower-performing students, who are documented to make the greatest gains in this area (Blumer & Beck, 2019; Peteroy-Kelly et al., 2017; Thiry et al., 2011). Multiple studies make the case that instruments are needed to measure this and other skills critical for the development of students as scientists (e.g., Dasgupta et al., 2014, 2016; Sirum & Humburg, 2011).

Science research experiences often require that students contribute to data interpretation, but many undergraduates enter introductory-level STEM courses with insufficient skill in understanding how to work with data (e.g., reading graphs, interpreting data, creating data visualizations), and STEM coursework does not necessarily cover this content (Coil et al., 2010; Maltese et al., 2015). Comparing the data analysis skills of various researchers, it has been found that novices are more heavily reliant on personal beliefs, while those with more expertise focus on empirical consistency to draw conclusions from their observations (Hogan & Maglienti, 2001). Students need to be taught explicitly how to read and generate the kinds of graphs they will need for a particular project, and although there is consensus that data interpretation is critical for developing students into scientists, relatively few studies have focused on assessing student skill level in this area (Maltese et al., 2015; Peteroy-Kelly et al., 2017).

Undergraduate students should also be able to develop hypotheses and conduct appropriate experiments to test these hypotheses by the time they graduate with a STEM bachelor's degree (H. B. White et al., 2013). When students are provided with the space to encounter challenges, revise their research goals, and repeat their work, this iterative process can have a powerful impact on their sense of ownership as they learn to navigate obstacles in their scientific discipline (Corwin et al., 2018; Gin et al., 2018).

The BURET study

My collaborators and I have drawn from this literature to develop four *BURET Indicators* that describe how undergraduates are expected to integrate foundational scientific practices:

1. Communicate the significance of their specific project to the overarching research questions of the laboratory and the broader scientific field
2. Justify their experimental design as appropriate for their research question
3. Analyze and interpret data in order to construct explanations and models that are relevant to their research question
4. Generate hypotheses and plan future experiments relevant to their research question in response to their analysis and interpretation of data

These Indicators provide the underlying construct for a set of instruments I have created, called the Berkeley Undergraduate Research Evaluation Tools (BURET), to assess how undergraduates develop an integrated knowledge of scientific content and practices as they engage in research. The first is the BURET Reflection instrument (BURET-R), which is a pair of prompts and assessment rubrics for student written reflections about the progress of their research project. The second is the BURET Poster Presentation instrument (BURET-P), which is a short interview protocol and assessment rubric designed to be used at capstone poster sessions. Both tools were administered to students in a variety of research settings, and the resulting findings are presented below.

METHODS

Participants and Context

Undergraduate participants were recruited from CUREs and UREs at a public, research-intensive university in the western United States. Undergraduate researcher volunteers came from five different populations (Table 1). It should be noted that nearly all students in UREs had previously taken a CURE, as is typical for URE students in many science departments at this institution.

The students participating in the study ranged from having zero to four or more semesters of research experience prior to study participation. The study population of 89 undergraduates contained a mixture of identities, including gender, race, ethnicity, and first language (Table 2). Students who self-identified as American Indian/Alaska Native, Black/African-American, Hispanic/Latinx, or other Pacific Islander, collectively referred to as underrepresented minorities (URM), were intentionally oversampled.

Table 1. Study populations

Group	n	Response Rate	Type	Duration	Prior Research Experience	Student Description
1	35	58%	CURE	Semester	Mostly none	Freshman chemistry students
2	6	90%	CURE	Summer	None	New chemistry transfer students
3	28	59%	URE	Ongoing	Variable	Department of chemistry students
4	5	65%	CURE	Semester	None	Pre-service STEM teachers
5	15	88%	URE	Summer+	Variable	Pre-service STEM teachers

Table 2. Descriptive statistics for undergraduates included in this study

Descriptive Statistics					
			CURE	**URE**	
Type of Research Experience			52	48	
			Female	**Male**	
Gender			58	42	
			Yes	**No**	
English is not their first language			24	76	
Identifies as an underrepresented minority			19	81	
	0	**1**	**2**	**3**	**4+**
Previous Research Experience (Semesters)	29	12	11	13	24
	Chemistry	**Chemical Engineering**	**Chemical Biology**	**Biology**	**Other**
Major	39	22	17	17	4
***All numbers are percentages; n = 89**					

Graduate student participants were mostly recruited from a workshop series for improving mentoring skills in the context of undergraduate STEM research. A total of 22 participants were recruited from this workshop, and an additional 6 advanced graduate students were recruited via convenience sampling.

Instrument Development

Expert Review of the BURET Indicators

To confirm that the BURET Indicators were aligned with the goals of faculty advisors working with undergraduate researchers, 33 faculty were interviewed. All chemistry faculty with undergraduate researchers (response rate 41%) and 30 faculty in other STEM departments (response rate 40%) were invited to participate in interviews. Over half of the faculty interviewed were chemists, but professors in various subfields of biology and engineering were also sampled. The faculty in this study ranged from assistant professors to full professors, and a wide variety of research group sizes is represented. During a 1-hour interview, faculty were asked to describe their goals for their undergraduate researchers, discuss mentoring practices, and review the BURET Indicators to comment on whether these were appropriate goals for their undergraduates. Nearly all of the responses were positive, with some faculty members expressing that the Indicators “exactly” described their overall goals for undergraduate researchers.

Assessment Protocols

Data collection and assessment approaches were chosen that would both support student learning and allow for direct measures of the Indicators across both CUREs and UREs. Many undergraduate researchers create a poster and present their research project as a capstone requirement, providing an opportunity to assess student integration of scientific content and practices in the context of their own work. A set of interview questions targeting the BURET Indicators were developed to ask at the end of each student's prepared presentation. This interview protocol coupled with a rubric to assess several aspects of these verbal presentations make up the BURET Poster Presentation instrument (BURET-P, Figure 1).

1. Can you please summarize why your research project and what you've learned is important?

2. Can you explain more about why you (and your lab) chose this general strategy for your research project?

3. Can you choose one experimental technique that is central to this work and say why you used it, rather than other options?

    3b. What are the limitations of this technique?

4. Could you expand on how you interpret these results?

    4b. How confident are you in your data and your conclusions?

5. What would you do if you had another year to work on this project, and why?

Figure 1. BURET-P interview protocol

Additionally, a pair of prompts (BURET-R, Figure 2) to complement the poster presentation assessment were developed to elicit written reflections that could be used at different points in the research experience to provide information on students' developing competencies with regard to scientific content and practices. In this study, BURET-R was administered a few weeks prior to their poster session. These prompts targeted Indicators 3 and 4, respectively, but many students also incorporate discussions of Indicators 1 and 2 in their responses.

1. *Data Analysis Prompt*: Think about the ways you have analyzed data recently.

    (a) Describe one example of data analysis you have done.

    (b) Reflect on how you used this data analysis to create or change an explanation or a model.

    Frame your response for an experienced scientist who is unfamiliar with your project.

2. *Next Steps Prompt*:

    (a) If you had another month or two to work, what would be your next steps and why?

    (b) What about if you had another year?

Figure 2. BURET-R reflective prompts

Scoring Rubrics

Preliminary rubric development was conducted with a small group of 7 undergraduates and 5 graduate students. All participants responded to BURET-R prompts, and a few also presented posters to the research team. Four undergraduates were also interviewed, during which they were asked to expand on BURET-R and BURET-P responses. Written responses and audio recordings were reviewed to determine what both novice and expert researchers discussed, focusing on concepts most relevant to the BURET Indicators. The emergent themes from initial rounds of coding and a review of the relevant literature were used to develop an overlapping set of specific items aligned with the BURET Indicators, resulting in a set of 6 items for the BURET-R and 11 items for the BURET-P scoring instruments (Table 3).

Table 3. Scoring rubric items for the BURET instruments

<table>
<tr><th>Item</th><th>BURET-R</th><th>BURET-P</th></tr>
<tr><td>Placing their work in a broader scientific context</td><td rowspan="2">X</td><td>X</td></tr>
<tr><td>Placing their work in a broader societal context</td><td>X</td></tr>
<tr><td>Providing rationale for an experimental design choice</td><td>X</td><td>X</td></tr>
<tr><td>Addressing limitations of an experimental design choice</td><td></td><td>X</td></tr>
<tr><td>Comparing alternatives to an experimental design choice</td><td></td><td>X</td></tr>
<tr><td>Number of experimental design choices with some rationale</td><td></td><td>X</td></tr>
<tr><td>Identifying and discussing the key variables</td><td>X</td><td></td></tr>
<tr><td>Describing their data analysis procedures</td><td rowspan="2">X</td><td></td></tr>
<tr><td>Interpreting their data</td><td>X</td></tr>
<tr><td>Analyzing sources of error and uncertainty</td><td></td><td>X</td></tr>
<tr><td>Proposing next steps for the project</td><td>X</td><td>X</td></tr>
<tr><td>Incorporating references to previous work</td><td></td><td>X</td></tr>
<tr><td>Integrating additional content knowledge</td><td>X</td><td>X</td></tr>
</table>

The KI framework typically scores items on a 0-5 scale according to progressively more integrated and connected ideas expressed by student respondents. Applying the KI framework to BURET scoring, descriptions were written for each possible score on each item (see Table 4 for an example rubric, see Appendices 3.2 and 3.3 for complete rubrics). These were anchored by the idea that a 2 should be a correct statement about an isolated part of the research process, and a 4 should be a clear, basic link from the relevant part of the research process to other scientific content and research practices. For example, a score of 2 on "Addressing the limitations of an experimental design choice" could be obtained by simply noting a drawback for a particular technique, whereas a score of 4 would require that students explain that limitation in terms of underlying scientific principles or with clear reference to the research question. The remaining levels were defined as follows: 0 indicates the item is absent; 1 indicates a vague statement; 3 indicates a partial link, generally a vague version of a 4; and 5 indicates a complex link of 3 or more isolated concepts. The highest level descriptions were informed in part by the graduate students who responded to the BURET-R and BURET-P assessments as scoring categories were being refined.

Table 4. Coding rubric for "Addressing limitations of an experimental design choice"

<table>
<tr><th colspan="3">DESIGN CHOICES (LIMITATIONS)
What are the limitations or drawbacks of the approach or technique they used?</th></tr>
<tr><td>Score</td><td>Description</td><td>Examples</td></tr>
<tr><td>0</td><td>- Does not discuss any limitations of design choice</td><td></td></tr>
<tr><td>1</td><td>- Vague reference to limitations</td><td>- "Again, part of the main problem is that graphite furnace is really temperamental."</td></tr>
<tr><td>2</td><td>- Clear statement of a generic limitation, OR</td><td>- "In terms of that technique, I think it depends on the accuracy in which the solutions are prepared. So if the standards aren't</td></tr>
</table>

	- Vague description of thoughtful limitation	prepared correctly or if they're too high on concentration, it may negatively, it definitely will negatively affect our data. So I think that's a big limitation. And also you have to produce a lot of different samples, which can be time consuming."
3	- One or more thoughtful limitations mentioned, but content knowledge only implied	- "The limitations of Congo Red is that it is visual. So it is qualitative even though we can't measure the radius. The radius isn't really going to tell us anything numerical about how much cellulose the bacteria digests."
4	- Gives at least one explicit limitation that demonstrates domain-specific content knowledge	- "One experimental technique that we use is hydrating the sample and then putting them under nanoindentation. … So the limitations of that technique are that you're running it under PBS, which is phosphate buffered saline, and that only mimics the ionic concentrations, it doesn't mimic the chemical functionality you'd encounter in in vivo synovial fluid."
5	- Basic Link (4) AND (- Discusses how limitations affected conclusions OR - Discusses how limitation was addressed, minimized, avoided, etc.)	- "The main limitation is that the scaled particle theory ignores the entropic consideration in the energy of interaction here, so it's hard to say what would happen at different temperatures. In order to predict the temperature dependence, you need an approximate value of the entropy of dissolution, which isn't known for a lot of these molecules. However, we found that that's actually very easy to predict. For each group of molecules it's approximately constant for a certain chlorination number so you know that if you have a PCB and it has three chlorines that you will know the entropy very well."

Instrument Testing

Data Collection

To determine whether the BURET instruments could detect a difference between novice and advanced researchers, students enrolled in the three target CUREs (Groups 1, 2, and 4 in Table 1) were invited to be part of this study. A few weeks before their corresponding poster session, student responses to BURET-R were collected in class from all who agreed to participate (see Table 1 for response rates). At the final poster session for those courses, a sample of the consenting students were interviewed using the BURET-P protocol (see Appendix 3.1 for sampling procedures).

Additionally, students presenting at one of the two target URE poster sessions (Groups 3 and 5 in Table 1) were invited by email to participate in this study. Responses to BURET-R were collected via Qualtrics a few weeks prior to the poster sessions, and all consenting students who provided responses to BURET-R were interviewed at their poster session. From this complete dataset, 80 BURET-R responses and 55 BURET-P interviews were selected for further analysis.

Graduate students from the mentoring workshop were asked to respond to BURET-R as part of a workshop assignment. BURET-P was not administered to these students. Additional advanced graduate students completed BURET-R and BURET-P during individual meetings with one of the researchers.

Coding and Rubric Reliability

Data from BURET-R and BURET-P were scored for each study participant according to the corresponding rubrics. 60% of the written responses were coded by two different researchers, and discrepancies were resolved through subsequent discussion. A weighted Cohen's kappa of 0.73 was calculated, and subsequent coding was completed individually. Each poster transcript was coded independently by at least two people, and any discrepancies between coders were discussed and resolved. Two different pairs of coders assessed these transcripts, and a weighted Cohen's kappa of 0.65 was calculated between coding pairs, using posters that were coded by all researchers.

Quantitative Analysis

The primary evidence for the validity of the BURET instruments is based on their ability to successfully distinguish between responses from undergraduates with more or less prior research experience. Participants were divided into novice (0-1 semesters) and advanced (2+ semesters) groups based on how many semesters of research they had completed prior to the one in which they were presenting a poster. For each instrument, all items were averaged to produce a single test statistic, and comparisons between novice and advanced participants were made using a t-test. Additionally, students were collapsed into either low scorers (KI score of 0-3) or high scorers (KI score of 4-5), and chi-squared tests were performed to determine whether high scores were significantly associated with increased research experience for each item.

Item-response theory (IRT) analysis was conducted to gather validity evidence based on internal structure at the instrument level. Because the sample size was not sufficient to run the analysis using all thresholds from the rubrics, data were collapsed into scores of low (0-2), moderate (3), or high (4-5), and Wright maps for each instrument were generated from the collapsed data. As a measure of internal consistency, Cronbach's alpha was calculated for each instrument. Additionally, exploratory factor analysis was performed and item-test correlations were calculated to determine whether the construct being measured is uni- or multi-dimensional. All statistical analysis was conducted on Stata except for the IRT analysis, which was performed on Conquest.

RESULTS

<u>Research Question 1</u>

Do the BURET instruments distinguish between students with different levels of prior research experience?

An analysis of student responses showed that both the BURET-R ($n = 80$) and BURET-P ($n = 55$) instruments are able to distinguish between more and less experienced undergraduate researchers. Total scores for each instrument revealed statistically significant differences between students with 2 or more semesters of prior research experience and students with less experience ($p < 0.001$; Tables 5 and 6). Average scores on each item also increased with more research experience, with 9 of the 17 items showing statistically significant gains.

The original BURET-R instrument counted describing data analysis and interpreting data as two separate items, but they were combined into a single item for the final analysis. A notable feature of the responses to BURET-R is that students often chose to either talk about the details of their data analysis procedures *or* give an interpretation of the data, but generally not both. This is consistent with the negative correlation of the scores on those two sub-items ($r = -0.17$, $p < 0.10$), whereas all other pairs of items showed positive correlations ($r > 0.25$, $p < 0.05$ for all but one pair). Because of this unexpected tendency in the observed responses, I decided to instead use the higher of the two data analysis/interpretation scores rather than one score for each, to account for the different ways in which the BURET-R prompt was interpreted by the students.

Factor analysis provided evidence that a unidimensional construct is being measured. For each instrument, only one factor had an eigenvalue greater than 1, and the ratio of the first two eigenvalues was well above 4. Additionally, all items except one had a correlation of at least $r = 0.55$ with the overall score on the corresponding instrument. As a measure of reliability, Cronbach's alpha is calculated to be 0.78 for both instruments, which is in the range considered acceptable for science education research instruments (Taber, 2018).

Table 5. Mean scores on BURET-R rubric items.

Semesters of Previous Research Experience	0-1	2+	Sig
Sample size (n)	42	38	
Placing work in a broader context	2.9	3.6	*
Providing rationale for expt. design choice	1.9	2.8	**
Identifying and discussing the key variables	2.0	2.4	
Describing OR interpreting data analysis	2.6	3.5	**
Proposing next steps for the project	2.6	3.2	
Integrating additional content knowledge	0.8	2.4	**
Average Score	**2.1**	**3.0**	***
* p < 0.05; ** p < 0.01; *** p < 0.001;			

Table 6. Mean scores on BURET-P rubric items.

Semesters of Previous Research Experience	0-1	2+	Sig
Sample size (n)	24	31	
Placing work in a broader scientific context	2.3	3.5	*
Placing work in a broader societal context	3.6	3.7	
Providing rationale for an expt. design choice	3.5	3.9	
Addressing limitations of an expt. design choice	2.8	3.3	
Comparing alternatives to an expt. design choice	2.7	3.4	
Expt. design choices with some rationale (max. 5)	2.5	3.3	*
Interpreting data	3.1	3.5	*
Analyzing sources of error and uncertainty	2.3	2.5	
Proposing next steps for the project	3.1	3.2	
Incorporating references to previous work	1.9	2.9	*
Integrating additional content knowledge	2.3	3.5	*
Average Score	**2.7**	**3.3**	***
* p < 0.05; ** p < 0.01; *** p < 0.001			

Item-response theory (IRT) analysis was conducted to gather validity evidence based on internal structure at the instrument level. The resulting Wright maps (Figure 3) show that the range of instrument item logit values span nearly the entire distribution of respondent logit values, with only a few students falling below all item thresholds on the BURET-R instrument, and a few Thurstonian thresholds located below the lowest respondent logit value for the BURET-P instrument. The reliability of partial credit model analysis carried out on the data is 0.774 for BURET-R and 0.758 for BURET-P. These values indicate an acceptable consistency of the items to measure respondent performance (Bond & Fox, 2007; Wright & Masters, 1982).

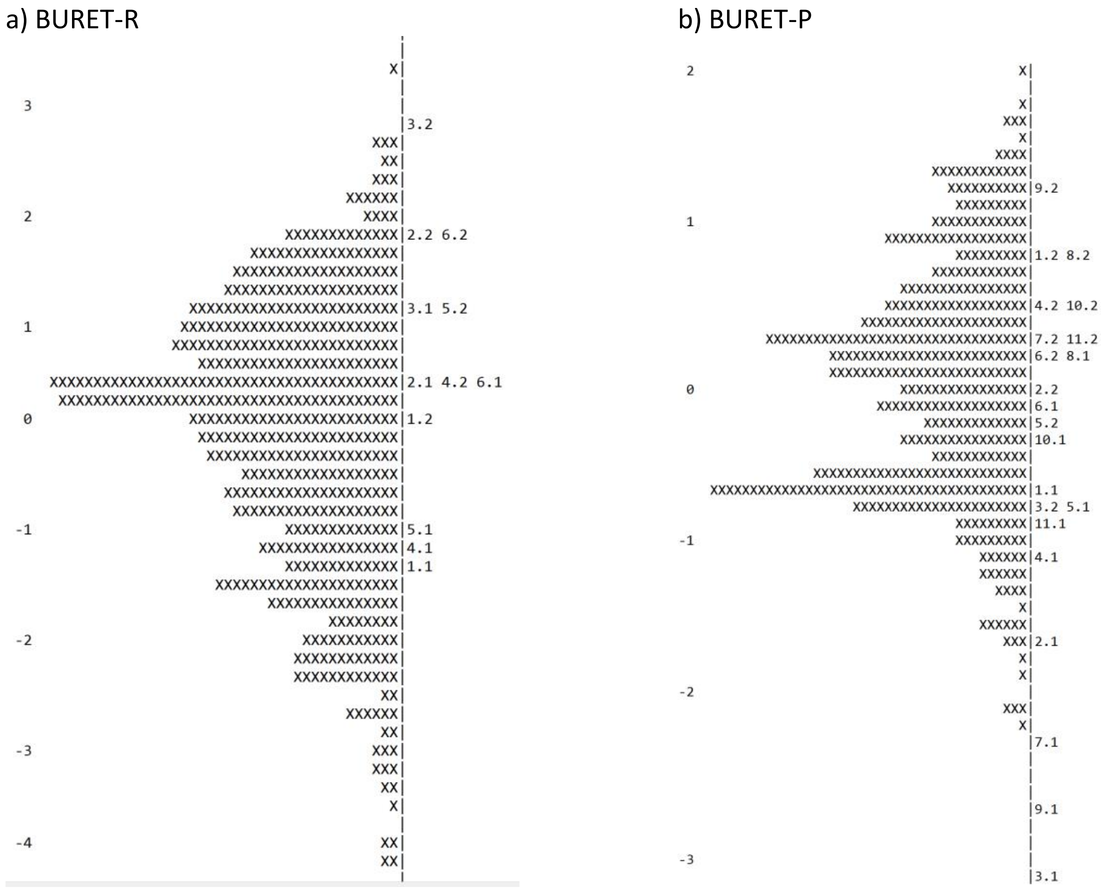


Figure 3. Wright maps for the BURET instruments

Two other variables that are highly correlated with increased research experience are year in school and whether the research experience was part of a course. As previously mentioned, most of the novice researchers in the sample were enrolled in a CURE, while most of the advanced researchers were participating in a URE in a faculty lab and had previously completed a CURE. To determine which of these variables was the best predictor of score on the BURET instruments, factorial ANOVAs were run using year in school, semesters of research experience, and URE/CURE as the independent variables. For the BURET-R, only URE/CURE was a significant predictor ($p < 0.05$). For the BURET-P, only semesters of research experience was a significant predictor ($p < 0.05$). No interaction terms were significant in either case. I was unable to look at whether there were differential effects for students who identified as a URM because too few of these students with at least 2 semesters of research experience were recruited. Future work will be needed to investigate this aspect of the instruments.

Comparison of the BURET-R and BURET-P

Overall, largely similar results were obtained from the BURET-R and the BURET-P instruments, which were generally administered several weeks apart. Student scores on the BURET-R and BURET-P instruments were significantly correlated with one another ($r = 0.4$, $p < 0.01$). The items on which students tended to excel or struggle were similar across the two instruments, with some variations based on the exact relationships between the items assessed and the specific prompt or interview questions being answered. Targeted questions asked during the poster presentations generally elicited more specific information than the broader reflective prompts, resulting in more items being coded when assessing poster presentations. Poster presentations were also much longer than the written responses to BURET-R; on average, written responses were 248 words in length, while poster presentations (including answers to questions) were 1,682 words in length. In general, students scored higher on BURET-P (average score = 3.1) than on BURET-R (average score = 2.3). This can also be seen by looking at individual participants; 85% of the participants scored higher on their poster presentations than on their written responses.

<u>Research Question 2</u>

What do the BURET instruments reveal about what undergraduates understand about research and what they are still learning at different stages of their undergraduate careers?

The BURET instruments provided information about the progress students made on each of the BURET Indicators as they gained in research experience. The following sections describe the characteristics of student progression along each Indicator using the KI framework, including undergraduate student performance on each item and the items that most differentiate novice and advanced study participants. Items are grouped by which Indicator they are most closely associated with to provide a more holistic picture of each primary area of assessment (Table 7).

Table 7. BURET-P items grouped by most related BURET Indicator

Indicator	Items
1	Placing their work in a broader scientific context
	Placing their work in a broader societal context
	Incorporating references to previous work
	Integrating additional content knowledge
2	Providing rationale for an experimental design choice
	Addressing limitations of an experimental design choice
	Comparing alternatives to an experimental design choice
	Number of experimental design choices with some rationale
3	Interpreting their data
	Analyzing sources of error and uncertainty
4	Proposing next steps for the project

Indicator 1: Communicating Significance

The first BURET Indicator assesses how well students can communicate the significance of their specific project to the overarching research questions of the laboratory and the broader scientific field. Three of the four items corresponding to this Indicator for the BURET-P instrument showed statistically significant differences between novice and advanced students. Advanced students demonstrated a more sophisticated understanding of their project's scientific context ($p < 0.05$), referred to previous work more often ($p < 0.05$), and integrated more content knowledge into their presentations ($p < 0.05$) when compared to less experienced students (Table 6). An analysis of students' responses to the BURET-R instrument also provided evidence that they develop in their ability to place their project into a broader context (Table 5).

For example, advanced students often demonstrated a more integrated understanding of scientific context by explaining the current state of the field or how their research might affect projects in other labs. One student who received a score of 4 stated, "I was ... working on investigating ... the mechanical properties of polycarbonate urethane. Our research is particularly relevant to joint implants and joint replacements, ... the current industry standard polymer is called ultra-high molecular wave polyethylene. ... Polycarbonate urethane or PCU is being pioneered as a new material. ... But it's pretty new so we're still doing research on the very mechanical properties and how it will react to being in the body and in an ionic environment where there is salts and stuff like that, that can affect its microstructure." In this response, the student clearly connects their work on the mechanical properties of PCU to the broader field of material science, particularly in the area of artificial joints.

A student would receive a 2 on the "Incorporating references to previous work" item by clearly referring to previous research but failing to explicitly link that research to their experimental design or compare it to their own results. In contrast, one student stated that, "There'd been, not a consensus, but almost every single study that we had read previously looking for these heavy metals in chocolate, but also in other candy, had focused on the cocoa, then being the source and maybe mentioned other possible sources in passing." The student then compared this body of previous work with their own work, which found a possible alternate source of heavy metals, resulting in a score of 4.

Additionally, advanced students scored higher on providing context by integrating more additional content knowledge into their presentation and answers. Additional content knowledge was defined as "exhibiting scientific content knowledge beyond what is required to describe the project." Students received a 2 by simply providing some additional clarification, or a 4 by providing multiple examples or extensive discussions of relevant information. It should be noted that this does not directly measure the content knowledge of a student, but rather the extent to which students have *integrated* that content knowledge into discussions of their research.

In contrast to the findings for placing work into a scientific context, both novice and advanced students integrated their work into a *societal* context at a high level. To receive a score of 4, a student needed to explicitly connect their work to a specific societal need or make an explicit statement about the possible use of their results. For example, "[While] removing lead from the gasoline solves our problem of potential lead contamination, … the problem with that is that, when you replace the lead with aromatics, ethylbenzene and toluene are particularly toxic to organisms, as well. And since modern day highways and roads are designed to funnel water off of the road into the surroundings, aquatic toxicity and environmental damage is a big problem with gasoline. So, what we wanted to do was determine the toxicity of ethylbenzene and another aromatic known as toluene, which are the two most common ones, and see how toxic they are to aquatic life." Students in both groups generally scored well on this item. However, the overall trend for the "big picture" items is that advanced students were more able to place their work into a larger context than the novice students.

Indicator 2: Justifying the Experimental Design

The second BURET Indicator was assessed with three items that focused on how students discussed their experimental design choices, which were defined as approaches, strategies, techniques, or other decisions made during the experimental design process. When asked to provide a rationale for an experimental design choice, the difference between novice and advanced student responses on the BURET-R instrument was highly significant ($p < 0.001$). Although more advanced students generally scored higher than novices on the BURET-P instrument on providing rationale, addressing limitations, and comparing alternatives to their experimental design choices, none of these differences were statistically significant.

Examples of experimental design choices varied broadly, from why the research group chose to study a certain topic to the specific instruments used to collect raw data. In the BURET-P interview protocol, students were asked *why* they made a given design choice *instead of* something else, and they were also asked about the *limitations* of that choice. In general, both novice and advanced students scored highly on providing a rationale for an experimental design choice related to their project for the BURET-P instrument; over half of the students scored 4 or higher, which requires a clear description of the design choice *and* an explicit rationale that demonstrates domain-specific content knowledge. For example, "We chose to use micro plasma atomic emissions spectroscopy because of its wide dynamic range. While there were many other instruments that would have worked similarly well, but not within this large range. And we were very uncertain as to whether we were over diluting or under diluting our samples…. We only had rough EPA guidelines to kind of guide our choices." The marginally higher average score for advanced students compared to novices was not significant. However, advanced students did explain a greater number of their decisions than novices. To reflect this, an item was included that simply counted the number of design choices for which the student provided some rationale. This number was significantly higher ($p < 0.05$) for advanced students, reflecting the greater detail in which they described their experimental design.

Students were less proficient at discussing the limitations of experimental design choices. A representative response is, "Limitations, well the biggest thing is when trees die they fall down and just sometimes you can't tell that they were there at all," which received a 3 for identifying a reasonable limitation of their fieldwork but not integrating in domain-specific content knowledge. However, some students were able to discuss limitations more fluently; for example, the following excerpt scored a 4: "The limitations of that technique are that bringing it under PBS, which is phosphate buffered saline, only mimics the ionic concentrations. It doesn't mimic the chemical function. So [what] we'd like to do for further research is hydrate it in [inaudible], which … mimics in vivo synovial fluid." Both novice and advanced students showed moderate levels of sophistication on the "comparing alternatives" item but rarely scored as high as 4, for which they needed to make a clear comparison between their choice and the alternative, explaining why their choice was superior. For example, "We decided to use MPAS instead of graphite furnace atomic absorption spectroscopy, even though both measure lead very well. Because MPAS has a larger dynamic range, and we were very uncertain as to the concentration we were gonna get."

Although the BURET-R rubric did not have items for scoring discussions of limitations or alternatives on a 0-5 point scale, prompt responses were coded for these items as present or absent. Overall, these numbers were low; only 13% of students alluded to limitations of their design choices, and nobody compared their design choices to alternatives. Moreover, there did not seem to be a clear difference between novice and advanced students in whether they chose to discuss any limitations. Because of these factors, the presence or absence of these items were not used in the overall assessment.

Indicator 3: Interpreting the Data

Items for the third BURET Indicator measured the extent to which students were able to analyze and interpret data in order to construct explanations and models that are relevant to their research question. The data interpretation item for BURET-P focused on constructing explanations that demonstrated domain-specific content knowledge; advanced students were significantly more likely ($p < 0.05$) to score higher on this item. This is consistent with results from the BURET-R instrument, on which advanced students scored higher on describing or interpreting their data analysis ($p < 0.01$).

For example, "And we found that with lower concentrations of silver, we get the same amount of silver conductivity" scored a 2 because there was a clear statement about the experimental results, but no additional comments were made about the data or their conclusions. A score of 4 required students to *explain* what they observed: "We stained the plates, which contained the cellulose media, with Congo red, which is a dye that binds to cellulase. So what that allowed us to do is once we washed the excess dye away, we got results that looked like this: the bacterial colonies that didn't produce any cellulase show no halo, and the whole plate is red, because the cellulose is still there, the dye binds, it's all still there. The ones that you see here have a halo of white, are positive results. They produce cellulase, and we know that because around the bacterial colony, is a halo where the cellulose has been degraded, and the dye doesn't bind." This student describes the underlying mechanism of the assay, explaining what is happening on a molecular and cellular level to justify their interpretation.

While scores for data interpretation were generally relatively high, students performed less well on analyzing sources of error and uncertainty. Most students identified a clear potential source of error or expressed skepticism about their results, but less than half of the students elaborated on their answer or connected that source of error to either their experimental design or their conclusions. A more complete response might explain how the experiment was designed to control for possible sources of error. For example, "And also, to avoid error we wanted to use NMR. First, we dissolve our wristbands using deuterated chloroform, and then running that through NMR, and seeing if there are any errors that we can possibly encounter for contamination. We just wanted to make sure the wristbands were mostly silicon. We had a positive control and negative control in just the chemical that we tested." However, most students did not discuss sources of error at this level and there were no significant differences between novice and advanced students.

Indicator 4: Proposing Future Investigations

A single assessment item aligned with the final BURET Indicator measures the extent to which students are able to generate hypotheses and plan future experiments relevant to their research question and in response to their analysis and interpretation of data. This item primarily evaluates the rationale given with next steps for the research project proposed by the student. Interestingly, advanced students did not score significantly higher on this item than novice students for either the BURET-R or -P instrument.

Students who received a 2 on this item typically suggested "more"-based continuations of their work with no rationale: more trials, more substrates, more different temperatures, and so on. In contrast, students who received a 4 would include a rationale that integrates domain-specific content knowledge, for example: "In the future we hope to perform confocal microscopy to determine the depth of infiltration, that's another common problem with current scaffolds is that they'll grow in an x-y plane and spread out in a nice flat layer, but they don't go into the bi-layer membrane. So that's what we're hoping to get with these fiber mats later, when you spin onto a mesh collector plate you get these really nice nodes, and we're hoping that cells could easily fit into those pores and infiltrate deeper into the membrane." Most students fell in between these two points; over 50% of participants scored a 3 on this item.

In earlier versions of the rubric, there were items to specifically assess the hypotheses students made about their next steps and whether they discuss the potential impact of this proposed future work. However, so few students suggested any hypotheses (4%) that the final version of the coding protocol only noted the presence or absence of a hypothesis. Similarly, while many students vaguely alluded to ways in which their proposed future work might be impactful, only a few (11%) addressed this explicitly, and this too was marked only as present or absent in the final version of the coding rubric. Overall, neither of these items distinguished between novice and advanced undergraduates, so they were not included in the final analysis of student performance.

DISCUSSION AND CONCLUSIONS

Two novel instruments for assessing how undergraduate researchers grow in their understanding of scientific research are introduced in this study. These instruments assess student discussions of their own research project, complementing previously published instruments that assess the ability of undergraduates to answer questions about completely different research scenarios (e.g., Harsh, 2016) or specific components of the research process like experimental design (e.g., Deane et al., 2014). The BURET instruments use the Knowledge Integration framework to evaluate how undergraduates develop an integrated understanding of the different components of research and the scientific practices and content of their projects. They can be applied to different types of research situations and across various scientific disciplines, in contrast to most existing tools. The instruments are able to distinguish between students at different levels of research experience, and evidence is presented for their validity and reliability. Excerpts from the rich datasets of student responses provide a detailed picture of how students progress in their understanding of research and help to identify specific areas where they need more support to fully develop as researchers.

Novice undergraduate students require more guidance to place their research into a larger scientific context

The largest differences between novice and advanced undergraduates involved providing a scientific context for their work. Even though faculty report that undergraduates are sometimes given key papers to read when starting in a lab, this work shows that their understanding of the connection between their experimental work and the broader scientific context is often weak. The faculty interviews I conducted suggest that, at least in some research groups, minimal emphasis is placed on teaching novice undergraduates the scientific context of their research projects. Multiple faculty members singled out the first Indicator as important but "hard in some cases for undergrads, they don't necessarily see the big picture at this time." Several faculty members also mentioned that having their undergraduates read the literature was a weak point in their mentoring. The lower priority given to these areas by faculty mentors, particularly for novice students, may help explain why there is such an increase in performance once students have been participating in research for at least two semesters. Reading the scientific literature has been shown to be challenging for novice students, but these skills develop over time as they work with their graduate student mentors to read more papers (Nelms & Segura-Totten, 2019). It is suggested that mentors use published approaches for teaching students to read the literature (Hoskins et al., 2011; Krontiris-Litowitz, 2013; Sato et al., 2014) to help their undergraduates understand the scientific context of their project more rapidly.

In contrast, students at all levels performed well on providing an integrated societal context for their work, and more advanced students did not receive higher average scores on this item. The ability to discuss the broader impacts of a research project is a valued skill, with some institutions offering courses explicitly aimed at training students in this area (Heath et al., 2014; MacFadden, 2009). In two of the CUREs included in this study, students developed research questions, often addressing a societal issue of interest to them, and as a result, they could fluently discuss the societal relevance of their project. Because novice students were strong on this item, there was little growth with more research experience.

Support is needed for beginning undergraduate researchers to better justify their experimental design and interpret their data

Experimental design is central to all research experiences. Previous attempts to assess gains in experimental design ability during scientific research experiences showed a general trend that participation in a CURE or URE improves student reasoning in this area (Dasgupta et al., 2014; Harsh et al., 2017; Harsh, 2016; Shanks et al., 2017; Sirum & Humburg, 2011). However, identifying the limitations of an experimental design has been found to be a weak point, even for graduate students (Gilmore et al., 2015). In this work, it was found that both novice and advanced students scored relatively high on their ability to rationalize experimental design choices. It was also found that more advanced students recognized that rationalizing experimental design, including providing the limitations of and alternatives to their experimental design choices, is an important component of talking about their research. The difference between novice and advanced students' rationalizations of experimental design choices for BURET-R is highly significant ($p < 0.001$). Similarly, advanced students were more likely to include limitations and alternatives as components of their presentation of their research.

The general trends observed for experimental design also hold for data interpretation; students generally performed well on giving straightforward interpretations of their data but were less likely to provide a richer description unless specifically prompted. Scores on the combined data analysis and interpretation items on both instruments were relatively high, with advanced students scoring significantly higher than novice students. This is consistent with other studies showing that data interpretation skills correlate with increased research experience (Harsh et al., 2017; White et al., 2011). In contrast, one of the lowest scoring items for both novice and advanced students was their ability to identify and discuss potential sources of error in their work. Students may deliberately focus on more positive aspects of their project, or the low scores may reveal a genuine deficit among undergraduates, who have been shown to struggle with critically analyzing experimental designs (Varela et al., 2005; White et al., 2011). This study suggests that students may benefit from targeted interventions in these areas throughout their undergraduate career.

Novice and advanced students were equally proficient at proposing future work for their projects

The advanced and novice students in the study sample were equally successful at proposing next steps for their research projects. This is surprising, because when faculty were asked what specifically they look for as signs of progress in their undergraduate researchers, many focused on day-to-day independence, including "thinking about what's next, what would be the next experiment after this one." The faculty interviewed by Laursen et al. (2010) also identified taking initiative, making decisions, and acting independently as markers of student progress.

A concept from the literature that is closely related to the item on proposing future work is that of iteration, as students scored higher when the proposed work was linked in some way to their most recent results. Authentic research is an iterative process, where the data from one experiment helps inform the next. Some have suggested that iteration is an essential part of an undergraduate research experience (Auchincloss et al., 2014), and efforts have been made to explicitly include iteration in CUREs (Light et al., 2019). Although there are instruments that measure whether a student perceives iteration to be a part of their research experience (Corwin et al., 2015), to my knowledge, there are no instruments that assess student proficiency in proposing next steps for an ongoing research project.

It was anticipated that advanced students would be more experienced at proposing future experiments and would therefore be able to more fluently discuss them in their written responses and poster presentations. Although this was not reflected in the average scores, it was observed that only advanced students received the highest possible score for proposing next steps on either the BURET-R or BURET-P instrument. Additionally, most of the advanced graduate students who were interviewed during the development of the instrument (see Methods), scored at the highest level on the BURET-P for this item.

One potential explanation for the discrepancy between expectations and observed results for undergraduate researchers on average is that many of the advanced undergraduate presenters were weeks away from graduation. These students were likely in the process of concluding their research and were not planning longer-term directions of the project. As a result, their scores on proposing future work might be lower than if they had been interviewed earlier. In contrast, many novice students were in a one-semester CURE in which they were explicitly instructed to talk about future work as part of their poster presentation. Their relative success in this area suggests that, contrary to faculty expectations, even novice students can be expected to propose the next steps of their research project, and this expectation should be more explicitly integrated into UREs.

The BURET instruments apply to a range of scientific disciplines

One advantage of the BURET instruments is that they can be applied to projects spanning a wide range of scientific fields, which is rare for instruments of this type, as many previously published instruments that evaluate undergraduate research are discipline specific. The BURET instruments attempt to account for discipline-specific knowledge without being restrictive. For example, for a higher score of 4 on the BURET data interpretation item, a student must explain what they observed in a way that demonstrates domain-specific content knowledge relevant to their research project. Because content knowledge for a diverse sampling of undergraduate researchers can be from a variety of disciplines, it has previously been difficult to measure with existing instruments that utilize a single hypothetical scenario. This work demonstrates that such knowledge can generally be identified using the BURET instruments. For example, all of the excerpts in Table 8 scored a 4 on data interpretation except the atmospheric science passage, which scored a 3 because domain-specific content knowledge was vaguely alluded to instead of explicitly stated. My collaborators and I envision the BURET instruments being used by educational researchers to monitor student progress in a unified way across UREs and CUREs in different disciplines.

Table 8. Excerpts from poster presentation transcripts: Interpretation of observed results across various scientific disciplines

Scientific Discipline	Excerpt from Student Poster Presentation
Biochemistry	My interpretation of these results is that the R-pal is utilizing the thiosulfate to grow and produce ammonia, so that's the main takeaway of this experiment and that if we took out thiosulfate and replaced it with another electron donor then they would grow with those electrons donated from that.
Inorganic Chemistry	What I've done here is I've synthesized a magnet that targets the lanthanide that has a strongly axial crystal field, but also a radical bridge, and this works very well because the 2,2'-bipyrimidine, that is substituted with chlorines, is a very weak epineural donor and so the crystal field becomes more axial because you have such a weak epineural donor even though you still have a radical lanthanide bridge.
Materials Science	But the decrease is that prevention of growth that I was talking about, [due to] the charge neutralization of the bromide ions on the ends of the surfactant. So, if the surfactant is more packed, no more gaps are available for precipitation to occur, and so you can't grow any nanorods per se. All you're gonna be left with is a bunch of spherical nanoparticles, no growth curve. So that's the reason for this decrease.
Ecology	So specifically for the Cottonwood Creek, we found that there may be a correlation between copper and rainfall depending on drought year. And we think this is because the rice farms are nearby, and when they use copper as a pesticide, it leaches into groundwater and surface water. And this may be increased by rainfall, whatever, there's a lot of rainfall that can wash away a lot.
Atmospheric Science	What is shown here is the VOC reactivity to show it's relatively constant, and then the NO_x concentrations, and the ozone concentrations. So the NO_x decreases from weekday to weekend because there are less giant trucks driving. Then this is showing that ozone decreases, but it doesn't really decrease that much, it's basically the same.

Limitations

Self-selection bias, a well-known limitation of undergraduate research studies, points out that those who participate are likely to be among the most highly motivated and high performing students.

Selection bias is expected to be minimal in this case, as approximately 70% of chemistry majors, who make up the majority of the sample, participate in undergraduate research, giving them an opportunity to participate in the poster session from which the participants were recruited.

Non-uniform experimental conditions. Data were necessarily collected from a variety of CURE and URE contexts, such that some students typed their responses, while others submitted hand-written responses, leading to differences in length of response. I have attempted to counteract these issues by designing the BURET instruments specifically to deeply probe the quality of the responses.

Preliminary CURE Experience. The most common trajectory for undergraduates in the department of chemistry at this institution is to take a CURE prior to starting a URE in a faculty research group. At some institutions, students may start a URE without any prior CURE experience or enroll in a CURE concurrently with or after participating in a URE. Because it was found that BURET-R scores appear to be sensitive to the type of research experience, more work will be needed for its use in different universities and for comparisons across different sequencing of CURE and URE experiences.

Implications

The BURET-R and BURET-P instruments have been used to characterize the progression of student expertise and reveal weaknesses in the learning outcomes of the undergraduate researchers. While time-intensive to code, I envision the use of the BURET instruments to be highly valuable in the contexts of future mentoring and undergraduate research studies. The BURET-R instrument is more generally applicable, as it can be quickly administered. For assessing poster presentations in a variety of educational contexts, the corresponding BURET-P instrument can provide a more detailed picture of student knowledge integration. These instruments offer a method of assessing student learning in relationship with students' own research projects. Because the focus is on the student's project and not on answering questions about a hypothetical scenario, the instruments are authentic and can be used across scientific disciplines.

Moreover, the BURET instruments provide an informal, low-stakes method for mentors to check on the progression of their students. Research mentors can regularly observe students setting up and analyzing the results of experiments, but they often have fewer opportunities to probe how their undergraduate students think about the research project more broadly. Additionally, the act of responding to the BURET-R prompts is itself a useful opportunity for the student to reflect on their project, which may not be a regular feature of their research experience. Similarly, answering the BURET-P protocol questions is an inherently useful activity, as it can help students to strengthen their poster talks and provide practice taking questions from the audience.

It should be noted that these instruments were designed to probe what students spontaneously discuss when asked to talk about their research project. Thus, it is suggested that students not receive a copy of the rubric beforehand, as it would be easy to "game" the assessment to receive a much higher score. The BURET instruments should also not be used as a summative assessment of what a student knows, because students likely know much more than what they choose to talk about in their written responses or poster presentations. Instead, the data should be used as a way to start a conversation with the student about what they might consider highlighting more in their research discussions and how to better turn what they know into an integrated narrative about their project.

At a departmental or institutional level, the BURET instruments can be used at regular intervals to assess how well a particular research experience is supporting student learning as they progress from novice to advanced researchers. The BURET instruments complement self-report survey data by enabling educational researchers to directly measure student learning with respect to knowledge and skills that are critical for their development as scientists. In the event that certain BURET Indicators are of greater importance with a particular student group, specific probes, like the interview questions in the BURET-P instrument, can be used to further explore student thinking for different components of the research process. Both of the BURET instruments can be used to provide students with feedback about their strengths and knowledge gaps with respect to the research project they are working on in a CURE or URE. These instruments can also be used to compare different research experiences, providing individual CUREs or UREs with information about the areas in which students need additional instruction or training from their research mentors.

CONCLUSION

The research presented in this dissertation has a variety of implications for how we conceptualize and facilitate the development of expertise in our students. Knowing more about how expertise manifests itself, both in the classroom and the research lab, can provide us with targets toward which we can attempt to guide our students. Expertise requires a broad base of conceptual knowledge, problem-solving skills to apply that knowledge, and the ability to integrate knowledge into a coherent understanding of one's scientific subfield. The work described in the preceding chapters characterizes novice and expert behaviors and suggests effective methods for helping students develop into scientific practitioners.

In Chapter 1, student problem solving was examined at a high level of detail, specifically the ability to solve complex predict-the-product problems in organic chemistry. The overall approach taken by different students was surprisingly consistent, although the more expert-like graduate students did tend to consider multiple possible pathways more often than the sophomore undergraduates. However, certain details did distinguish more and less successful problem solvers at the same level of experience. In particular, explicitly identifying functional groups by name was strongly associated with greater success, especially "compound" functional groups like acetal and diene. It is proposed based on these results that assigning a name to a given structure better activates relevant knowledge about the common reactive pathways for a certain functional group.

More generally, an important first step for problem solving is taking stock of the information given by the problem and linking it to connected knowledge. This step can be extremely rapid for an expert practitioner, to the point where they may not even realize they are doing it. Students who observe an expert solving a problem may then conclude that such preliminary steps are not necessary, or that they are only required for "dumb" people. It is therefore suggested that when instructors model the problem-solving process, they attempt to slow down this initial step and explain clearly which features are important to notice and what information should be associated with a given aspect of the problem. It would be reasonable to require students to write down this important information for untimed assignments like problem sets, in the hope that this sort of training would develop into a much faster routine over time.

Solving organic chemistry problems requires not just problem-solving skills, but also a broad range of conceptual knowledge, the amount of which can often be overwhelming for students. The ability to readily assimilate new information by adding it to existing cognitive structures allows an expert to learn much more efficiently than a novice, who may view the same material as an incomprehensible mass of seemingly disconnected facts and ideas. To help generate the appropriate mental architecture, it has been proposed that students be allowed to explore data to gain familiarity with which details are most pertinent to the target concepts (Schwartz & Bransford, 1998). In particular, the presence of contrasting cases that differ by one important variable is suggested to be beneficial for a student's ability to grasp the underlying principles when presented later through a lecture or text. The Preparation for Future Learning (PFL) lessons described in Chapter 2 were designed with this goal in mind.

Overall, PFL lessons were shown to be an effective and engaging method for students to interact with the material and develop a conceptual understanding of key topics in organic chemistry. Feedback on the lessons showed that this was generally an enjoyable and memorable experience for the students, and quizzes given at the end of the lesson showed that students seemed to initially grasp the primary learning goals. While students who attended these lessons did not outperform their classmates on relevant questions from the subsequent midterm, they did score significantly higher on corresponding questions from the final, at least for the acid-base lesson. Based on these results, it is proposed that students are equally able to learn topics like acid-base trends from a PFL lesson, a traditional lesson, or a lecture. However, the *persistence* of that learning seems to be greatest for the PFL lesson. A common issue in teaching is that students seem to forget everything they learned the moment the test is over. The research presented here suggests that properly preparing students to learn core concepts may be one way to circumvent this outcome.

As students advance in their education and engage in undergraduate research, they are confronted with a substantially different set of problems. Although the form is often the same as what they have seen in classes (e.g., how do I make that?, what happened when I mixed those?), the problems encountered in research are much more open-ended and are unlikely to have one single optimal solution. This makes the assessment of student progress as a researcher a difficult task. It seems inappropriate to test for open-ended problem-solving abilities using well-defined questions with a clear answer. Complicating matters further is that research results in very specialized knowledge specific to a given project, unlike coursework, in which all students are expected to master the same set of material. Previous attempts to measure aptitude in facets of research like experimental design have generally led to instruments in which students are asked questions about a standardized scenario in the approximate field of their research (e.g., Dasgupta et al., 2014). However, answering analytical questions about the effects of a certain chemical on the development of frogs (J. A. Harsh, 2016a) may have little to do with the expertise gained by conducting research on, for example, femtosecond spectroscopy or synthesis of functionalized carbon nanotubes. It was based on this type of discrepancy that our research team and I set out to develop instruments to assess how well students integrated domain-specific knowledge into discussions of their *own* research project. These instruments were shown to be valid and reliable, and they offer a fundamentally different alternative to existing measures.

The instruments described in Chapter 3 are intended to be used primarily as formative assessment, giving students some idea about where they are with respect to an expert-like understanding of their own work. Student reflections on their project, either through written journal entries or through discussions of a prepared poster, can be analyzed through the lens of the BURET instruments. The results of such an analysis can then be used to start a conversation about how the student might better integrate their knowledge into a more complete and coherent internal narrative regarding their project. It is recommended that undergraduate researchers reflect more frequently about their research, giving both their mentors and themselves more opportunity to assess progress. The items on which they are assessed can vary based on what the research mentor thinks is most appropriate for their student's current stage of development, whether it be "big picture" questions, knowledge of alternative methods, or the ability to propose future work.

It is hoped that the work described in this dissertation will convince readers to be more deliberate about their pedagogical decisions, both in the classroom and in the research laboratory. Suggestions on developing conceptual knowledge, problem-solving abilities, and research aptitudes are given throughout, along with specific methods for assessing whether students are gaining the appropriate types of expertise. Students will generally gain some level of expertise simply by having more experience, but with the correct pedagogical choices, instructors can more efficiently and effectively put students on a path toward success.

REFERENCES

Airey, J., & Linder, C. (2009). A disciplinary discourse perspective on university science learning: Achieving fluency in a critical constellation of modes. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *46*(1), 27–49.

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist*, *48*(2), 87–113.

Auchincloss, L. C., Laursen, S. L., Branchaw, J. L., Eagan, K., Graham, M., Hanauer, D. I., Lawrie, G., McLinn, C. M., Pelaez, N., & Rowland, S. (2014). *Assessment of course-based undergraduate research experiences: A meeting report*. Am Soc Cell Biol.

Bhattacharyya, G. (2014). Trials and tribulations: Student approaches and difficulties with proposing mechanisms using the electron-pushing formalism. *Chemistry Education Research and Practice*, *15*(4), 594–609.

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

Blumer, L. S., & Beck, C. W. (2019). Laboratory courses with guided-inquiry modules improve scientific reasoning and experimental design skills for the least-prepared undergraduate students. *CBE—Life Sciences Education*, *18*(1), ar2.

Bode, N. E., & Flynn, A. B. (2016). Strategies of successful synthesis solutions: Mapping, mechanisms, and more. *Journal of Chemical Education*, *93*, 593–604.

Bodner, G. M. (2003). Problem solving: The difference between what we do and what we tell students to do. *University Chemistry Education*, *7*(2), 37–45.

Bond, T. G., & Fox, C. M. (2007). Fundamental measurement in the human sciences. *Chicago, IL: Institute for Objective Measurement*.

Bonous-Hammarth, M. (2000). Pathways to success: Affirming opportunities for science, mathematics, and engineering majors. *Journal of Negro Education*, 92–111.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *The Journal of General Psychology*, *49*(2), 229–240.

Bowen, C. W. (1994). Think-aloud methods in chemistry education: Understanding student thinking. *Journal of Chemical Education*, *71*(3), 184–190.

Brando, B. (2019). *Study behaviors, problem-solving, and exam design in organic chemistry* [Doctoral Dissertation]. University of California, Berkeley.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000a). How experts differ from novices. In *How people learn: Brain, mind, experience, and school* (Expanded, pp. 31–50). National Academy Press.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000b). *How people learn* (Vol. 11). Washington, DC: National academy press.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000c). Learning and transfer. In *How people learn: Brain, mind, experience, and school* (Expanded, pp. 51–78). National Academy Press.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, *24*, 61–100.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32–42.

Bunce, D. M., Gabel, D. L., & Samuel, J. V. (1991). Enhancing chemistry problem-solving achievement using problem categorization. *Journal of Research in Science Teaching*, *28*(6), 505–521.

Butz, A. R., & Branchaw, J. L. (2020). Entering Research Learning Assessment (ERLA): Validity evidence for an instrument to measure undergraduate and graduate research trainee development. *CBE—Life Sciences Education*, *19*(2), ar18.

Carlone, H. B., & Johnson, A. (2007). Understanding the science experiences of successful women of color: Science identity as an analytic lens. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *44*(8), 1187–1218.

Carpi, A., Ronan, D. M., Falconer, H. M., & Lents, N. H. (2017). Cultivating minority scientists: Undergraduate research increases self-efficacy and career ambitions for underrepresented students in STEM. *Journal of Research in Science Teaching*, *54*(2), 169–194.

Cartrette, D. P., & Bodner, G. M. (2010). Non-mathematical problem solving in organic chemistry. *Journal of Research in Science Teaching*, *47*(6), 643–660.

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

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

Chang, M. J., Eagan, M. K., Lin, M. H., & Hurtado, S. (2011). Considering the impact of racial stigmas and science identity: Persistence among biomedical and behavioral science aspirants. *The Journal of Higher Education*, *82*(5), 564–596.

Chang, M. J., Sharkness, J., Hurtado, S., & Newman, C. B. (2014). What matters in college for retaining aspiring scientists and engineers from underrepresented racial groups. *Journal of Research in Science Teaching*, *51*(5), 555–580.

Charters, E. (2003). The use of think-aloud methods in qualitative research: An introduction to think-aloud methods. *Brock Education Journal*, *12*(2), 68–82.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152.

Coil, D., Wenderoth, M. P., Cunningham, M., & Dirks, C. (2010). Teaching the process of science: Faculty perceptions and an effective methodology. *CBE—Life Sciences Education*, *9*(4), 524–535.

Corwin, L. A., Graham, M. J., & Dolan, E. L. (2015). Modeling course-based undergraduate research experiences: An agenda for future research and evaluation. *CBE—Life Sciences Education*, *14*(1), es1.

Corwin, L. A., Runyon, C. R., Ghanem, E., Sandy, M., Clark, G., Palmer, G. C., Reichler, S., Rodenbusch, S. E., & Dolan, E. L. (2018). Effects of discovery, iteration, and collaboration in laboratory courses on undergraduates' research career intentions fully mediated by student ownership. *CBE—Life Sciences Education*, *17*(2), ar20.

Corwin, L. A., Runyon, C., Robinson, A., & Dolan, E. L. (2015). The laboratory course assessment survey: A tool to measure three dimensions of research-course design. *CBE—Life Sciences Education*, *14*(4), ar37.

Crawford, G. L., & Kloepper, K. D. (2019). Exit interviews: Laboratory assessment incorporating written and oral communication. *Journal of Chemical Education*, *96*(5), 880–887.

Cruz-Ramírez de Arellano, D., & Towns, M. H. (2014). Students' understanding of alkyl halide reactions in undergraduate organic chemistry. *Chemistry Education Research and Practice*, *15*(4), 501–515.

Dasgupta, A. P., Anderson, T. R., & Pelaez, N. (2014). Development and Validation of a Rubric for Diagnosing Students' Experimental Design Knowledge and Difficulties. *Cell Biology Education*, *13*(2), 265–284. https://doi.org/10.1187/cbe.13-09-0192

Dasgupta, Annwesa P., Anderson, T. R., & Pelaez, N. (2014). Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. *CBE—Life Sciences Education*, *13*(2), 265–284.

Dasgupta, Annwesa P., Anderson, T. R., & Pelaez, N. J. (2016). Development of the neuron assessment for measuring biology students' use of experimental design concepts and representations. *CBE—Life Sciences Education*, *15*(2), ar10.

Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2014). Development of the biological experimental design concept inventory (BEDCI). *CBE—Life Sciences Education*, *13*(3), 540–551.

DeCocq, V., & Bhattacharyya, G. (2019). TMI (Too much information)! Effects of given information on organic chemistry students' approaches to solving mechanism tasks. *Chemistry Education Research and Practice*, *20*(1), 213–228.

deGroot, A. D. (1965). *Thought and Choice in Chess*. Mouton.

Doğan, A., & Kaya, O. N. (2009). Poster sessions as an authentic assessment approach in an open-Ended University general chemistry laboratory. *Procedia-Social and Behavioral Sciences*, *1*(1), 829–833.

Egan, D. E., & Schwartz, B. J. (1979). Chunking in recall of symbolic drawings. *Memory & Cognition*, *7*(2), 149–158.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251.

Ericsson, K. A., & Simon, H. A. (1993). Introduction and Summary (Ch. 1). In *Protocol analysis: Verbal reports as data* (Revised, pp. 1–62). MIT Press.

Estrada, M., Burnett, M., Campbell, A. G., Campbell, P. B., Denetclaw, W. F., Gutiérrez, C. G., Hurtado, S., John, G. H., Matsui, J., & McGee, R. (2016). Improving underrepresented minority student persistence in STEM. *CBE—Life Sciences Education*, *15*(3), es5.

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

Finkenstaedt-Quinn, S. A., Watts, F. M., Petterson, M. N., Archer, S. R., Snyder-White, E. P., & Shultz, G. V. (2020). Exploring student thinking about addition reactions. *Journal of Chemical Education*, *97*(7), 1852–1862.

Flynn, A. B. (2014). How do students work through organic synthesis learning activities? *Chemistry Education Research and Practice*, *15*(4), 747–762.

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

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A description of think aloud method and protocol analysis. *Qualitative Health Research*, *3*(4), 430–441.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415.

Galloway, K. R., Leung, M. W., & Flynn, A. B. (2019). Patterns of reactions: A card sort task to investigate students' organization of organic chemistry reactions. *Chemistry Education Research and Practice*, *20*(1), 30–52.

Gilmore, J., Vieyra, M., Timmerman, B., Feldon, D., & Maher, M. (2015). The Relationship between Undergraduate Research Participation and Subsequent Research Performance of Early Career STEM Graduate Students. *The Journal of Higher Education*, *86*(6), 834–863. https://doi.org/10.1353/jhe.2015.0031

Gin, L. E., Rowland, A. A., Steinwand, B., Bruno, J., & Corwin, L. A. (2018). Students who fail to achieve predefined research goals may still experience many positive outcomes as a result of CURE participation. *CBE—Life Sciences Education*, *17*(4), ar57.

Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *CBE—Life Sciences Education*, *11*(4), 364–377.

Graulich, N., & Schween, M. (2018). Concept-oriented task design: Making purposeful case comparisons in organic chemistry. *Journal of Chemical Education*, *95*, 376–383.
Griffeth, N., Batista, N., Grosso, T., Arianna, G., Bhatia, R., Boukerche, F., Crispi, N., Fuller, N., Gauza, P., & Kingsbury, L. (2015). An Undergraduate Research Experience Studying Ras and Ras Mutants. *IEEE Transactions on Education*, *59*(2), 91–97.
Grove, N. P., Cooper, M. M., & Cox, E. L. (2012). Does mechanistic thinking improve student success in organic chemistry? *Journal of Chemical Education*, *89*(7), 850–853.
Grove, N. P., Cooper, M. M., & Rush, K. M. (2012). Decorating with arrows: Toward the development of representational competence in organic chemistry. *Journal of Chemical Education*, *89*, 844–849.
Harackiewicz, J. M., & Hulleman, C. S. (2010). The importance of interest: The role of achievement goals and task values in promoting the development of interest. *Social and Personality Psychology Compass*, *4*(1), 42–52.
Harsh, J. A. (2016a). Designing performance-based measures to assess the scientific thinking skills of chemistry undergraduate researchers. *Chemistry Education Research and Practice*, *17*, 808–817.
Harsh, J. A. (2016b). Designing performance-based measures to assess the scientific thinking skills of chemistry undergraduate researchers. *Chemistry Education Research and Practice*, *17*(4), 808–817.
Harsh, J., Esteb, J. J., & Maltese, A. V. (2017). Evaluating the development of chemistry undergraduate researchers' scientific thinking skills using performance-data: First findings from the performance assessment of undergraduate research (PURE) instrument. *Chemistry Education Research and Practice*, *18*(3), 472–485. https://doi.org/10.1039/C6RP00222F
Heath, K. D., Bagley, E., Berkey, A. J. M., Birlenbach, D. M., Carr-Markell, M. K., Crawford, J. W., Duennes, M. A., Han, J. O., Haus, M. J., Hellert, S. M., Holmes, C. J., Mommer, B. C., Ossler, J., Peery, R., Powers, L., Scholes, D. R., Silliman, C. A., Stein, L. R., & Wesseln, C. J. (2014). Amplify the Signal: Graduate Training in Broader Impacts of Scientific Research. *BioScience*, *64*(6), 517–523. https://doi.org/10.1093/biosci/biu051
Hernandez, P. R., Woodcock, A., Estrada, M., & Schultz, P. W. (2018). Undergraduate research experiences broaden diversity in the scientific workforce. *BioScience*, *68*(3), 204–211.
Hinsley, D. A., Hayes, J. R., & Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In *Cognitive Processes in Comprehension* (pp. 89–106). Lawrence Erlbaum Associates, Inc.
Hogan, K., & Maglienti, M. (2001). Comparing the epistemological underpinnings of students' and scientists' reasoning about conclusions. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *38*(6), 663–687.
Hoskins, S. G., Lopatto, D., & Stevens, L. M. (2011). The CREATE approach to primary literature shifts undergraduates' self-assessed ability to read and analyze journal articles, attitudes about science, and epistemological beliefs. *CBE—Life Sciences Education*, *10*(4), 368–378.
Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526.
Killpack, T. L., & Fulmer, S. M. (2018). Development of a Tool to Assess Interrelated Experimental Design in Introductory Biology. *Journal of Microbiology & Biology Education*, *19*(3).
Kraft, A., Strickland, A. M., & Bhattacharyya, G. (2010). Reasonable reasoning: Multi-variate problem-solving in organic chemistry. *Chemistry Education Research and Practice*, *11*(4), 281–292.
Krieter, F. E., Julius, R. W., Tanner, K. D., Bush, S. D., & Scott, G. E. (2016). Thinking like a chemist: Development of a chemistry card-sorting task to probe conceptual expertise. *Journal of Chemical Education*, *93*(5), 811–820.

Krim, J. S., Coté, L. E., Schwartz, R. S., Stone, E. M., Cleeves, J. J., Barry, K. J., Burgess, W., Buxner, S. R., Gerton, J. M., & Horvath, L. (2019). Models and Impacts of Science Research Experiences: A Review of the Literature of CUREs, UREs, and TREs. *CBE—Life Sciences Education*, *18*(4), ar65.

Krontiris-Litowitz, J. (2013). Using Primary Literature to Teach Science Literacy to Introductory Biology Students. *Journal of Microbiology & Biology Education*, *14*(1), 66–77. https://doi.org/10.1128/jmbe.v14i1.538

Laungani, R., Tanner, C., Brooks, T. D., Clement, B., Clouse, M., Doyle, E., Dworak, S., Elder, B., Marley, K., & Schofield, B. (2018). Finding some good in an invasive species: Introduction and assessment of a novel CURE to improve experimental design in undergraduate biology classrooms. *Journal of Microbiology & Biology Education*, *19*(2).

Laursen, S., Hunter, A.-B., Seymour, E., Thiry, H., & Melton, G. (2010). *Undergraduate research in the sciences: Engaging students in real science*. John Wiley & Sons.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge university press.

Light, C. J., Fegley, M., & Stamp, N. (2019). Emphasizing iterative practices for a sequential course-based undergraduate research experience in microbial biofilms. *FEMS Microbiology Letters*, *366*(23), fnaa001.

Lin, T.-J., Lin, T.-C., Potvin, P., & Tsai, C.-C. (2019). Research trends in science education from 2013 to 2017: A systematic content analysis of publications in selected journals. *International Journal of Science Education*, *41*(3), 367–387.

Linn, M. C. (1995). Designing computer learning environments for engineering and computer science: The scaffolded knowledge integration framework. *Journal of Science Education and Technology*, *4*(2), 103–126.

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

Linn, M., Eylon, B.-S., Kidron, A., Gerard, L., Toutkoushian, E., Ryoo, K., Bedell, K. D. B., Swearingen, A., Clark, D., & Virk, S. (2018). *Knowledge integration in the digital age: Trajectories, opportunities and future directions*. International Society of the Learning Sciences, Inc.[ISLS].

Linn, Marcia C., Palmer, E., Baranger, A., Gerard, E., & Stone, E. (2015). Undergraduate research experiences: Impacts and opportunities. *Science*, *347*(6222), 1261757.

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review*, *29*, 693–715.

MacFadden, B. J. (2009). Training the Next Generation of Scientists about Broader Impacts. *Social Epistemology*, *23*(3–4), 239–248. https://doi.org/10.1080/02691720903364100

Maltese, A. V., Harsh, J. A., & Svetina, D. (2015). Data visualization literacy: Investigating data interpretation along the novice—expert continuum. *Journal of College Science Teaching*, *45*(1), 84–90.

Mason, A., & Singh, C. (2011). Assessing expertise in introductory physics using categorization task. *Physical Review Special Topics - Physics Education Research*, *7*(2), 020110.

Mondisa, J.-L., & McComb, S. A. (2018). The role of social community and individual differences in minority mentoring programs. *Mentoring & Tutoring: Partnership in Learning*, *26*(1), 91–113.

National Academies of Sciences & Medicine. (2017). *Undergraduate research experiences for STEM students: Successes, challenges, and opportunities*. National Academies Press.

Nelms, A. A., & Segura-Totten, M. (2019). Expert–Novice Comparison Reveals Pedagogical Implications for Students’ Analysis of Primary Literature. *CBE—Life Sciences Education*, *18*(4), ar56. https://doi.org/10.1187/cbe.18-05-0077

Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning–a review of test instruments. *Educational Research and Evaluation*, *23*(3–4), 78–101.

Ortiz, N. A., Morton, T. R., Miles, M. L., & Roby, R. S. (2020). What About Us? Exploring the Challenges and Sources of Support Influencing Black Students' STEM Identity Development in Postsecondary Education. *The Journal of Negro Education*, *88*(3), 311–326.

Peteroy-Kelly, M. A., Marcello, M. R., Crispo, E., Buraei, Z., Strahs, D., Isaacson, M., Jaworski, L., Lopatto, D., & Zuzga, D. (2017). Participation in a year-long CURE embedded into major core genetics and cellular and molecular biology laboratory courses results in gains in foundational biological concepts and experimental design skills by novice undergraduate researchers. *Journal of Microbiology & Biology Education*, *18*(1).

Petterson, M. N., Watts, F. M., Snyder-White, E. P., Archer, S. R., Shultz, G. V., & Finkenstaedt-Quinn, S. A. (2020). Eliciting student thinking about acid–base reactions via app and paper–pencil based problem solving. *Chemistry Education Research and Practice*, *21*(3), 878–892.

Polya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press.

Remich, R., Naffziger-Hirsch, M. E., Gazley, J. L., & McGee, R. (2016). Scientific growth and identity development during a postbaccalaureate program: Results from a multisite qualitative study. *CBE—Life Sciences Education*, *15*(3), ar25.

Robnett, R. D., Chemers, M. M., & Zurbriggen, E. L. (2015). Longitudinal associations among undergraduates' research experience, self-efficacy, and identity. *Journal of Research in Science Teaching*, *52*(6), 847–867.

Rodemer, M., Eckhard, J., Graulich, N., & Bernholt, S. (2020). Decoding case comparisons in organic chemistry: Eye-tracking students' visual behavior. *Journal of Chemical Education*, *97*(10), 3530–3539.

Rodenbusch, S. E., Hernandez, P. R., Simmons, S. L., & Dolan, E. L. (2016). Early engagement in course-based research increases graduation rates and completion of science, engineering, and mathematics degrees. *CBE—Life Sciences Education*, *15*(2), ar20.

Ryoo, K., & Linn, M. C. (2012). Can dynamic visualizations improve middle school students' understanding of energy in photosynthesis? *Journal of Research in Science Teaching*, *49*(2), 218–243.

Sabers, D. S., Cushing, K. S., & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensional, and immediacy. *American Educational Research Journal*, *28*(1), 63–88.

Sadler, T. D., Burgin, S., McKinney, L., & Ponjuan, L. (2010). Learning science through research apprenticeships: A critical review of the literature. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *47*(3), 235–256.

Sato, B. K., Kadandale, P., He, W., Murata, P. M. N., Latif, Y., & Warschauer, M. (2014). Practice Makes Pretty Good: Assessment of Primary Literature Reading Abilities across Multiple Large-Enrollment Biology Laboratory Courses. *CBE—Life Sciences Education*, *13*(4), 677–686. https://doi.org/10.1187/cbe.14-02-0025

Schoenfeld, A. (1987). What's all the fuss about metacognition? In *Cognitive Science and Mathematics Education* (pp. 189–215).

Schooler, J. W. (2002). Verbalization produces a transfer inappropriate processing shift. *Applied Cognitive Psychology*, *16*, 989–997.

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, *122*(2), 166–183.

Schultz, P. W., Hernandez, P. R., Woodcock, A., Estrada, M., Chance, R. C., Aguilar, M., & Serpe, R. T. (2011). Patching the pipeline: Reducing educational disparities in the sciences through minority training programs. *Educational Evaluation and Policy Analysis*, *33*(1), 95–114.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction*, *16*(4), 475–522.

Schwartz, D. L., Bransford, J. D., & Sears, D. (2005). Efficiency and innovation in transfer. In *Transfer of Learning from a Modern Multidisciplinary Perspective* (pp. 1–51). Information Age Publishing.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*(2), 129–184.

Shanks, R. A., Robertson, C. L., Haygood, C. S., Herdliksa, A. M., Herdliska, H. R., & Lloyd, S. A. (2017). Measuring and Advancing Experimental Design Ability in an Introductory Course without Altering Existing Lab Curriculum †. *Journal of Microbiology & Biology Education*, *18*(1), 1–8. https://doi.org/10.1128/jmbe.v18i1.1194

Simon, H. A. (1980). Problem solving and education. In *Problem Solving and Education: Issues in Teaching and Research* (pp. 81–96). Erlbaum.

Sirum, K., & Humburg, J. (2011a). The Experimental Design Ability Test (EDAT). *Bioscene: Journal of College Biology Teaching*, *37*(1), 8–16.

Sirum, K., & Humburg, J. (2011b). The Experimental Design Ability Test (EDAT). *Bioscene: Journal of College Biology Teaching*, *37*(1), 8–16.

Snyder, J. L. (2000). An investigation of the knowledge structures of experts, intermediates and novices in physics. *International Journal of Science Education*, *22*(9), 979–992.

Stone, E. M. (2014). Guiding Students to Develop an Understanding of Scientific Inquiry: A Science Skills Approach to Instruction and Assessment. *Cell Biology Education*, *13*(1), 90–101. https://doi.org/10.1187/cbe-12-11-0198

Stratton, D. H. (2020). Types of instructional strategies and their effect on Preparation for Future Learning in differentiation. *International Journal of Educational Research*, *104*, 101691.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296.

Thiry, H., Laursen, S. L., & Hunter, A.-B. (2011). What experiences help students become scientists? A comparative study of research and other sources of personal and professional gains for STEM undergraduates. *The Journal of Higher Education*, *82*(4), 357–388.

Timmerman, B. E. C., Strickland, D. C., Johnson, R. L., & Payne, J. R. (2011). Development of a 'universal' rubric for assessing undergraduates' scientific reasoning skills using scientific writing. *Assessment & Evaluation in Higher Education*, *36*(5), 509–547.

Varela, M. F., Lutnesky, M. M., & Osgood, M. P. (2005). Assessment of student skills for critiquing published primary scientific literature using a primary trait analysis scale. *Microbiology Education*, *6*, 20.

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

Webber, D. M., & Flynn, A. B. (2018). How are students solving familiar and unfamiliar organic chemistry mechanism questions in a new curriculum? *Journal of Chemical Education*, *95*(9), 1451–1467.

Weinrich, M. L., & Sevian, H. (2017). Capturing students' abstraction while solving organic reaction mechanism problems across a semester. *Chemistry Education Research and Practice*, *18*(1), 169–190.

Wheatley, G. H. (1984). *Problem solving in school mathematics. MEPS Technical Report 84.01, School Mathematics and Science Center, Purdue University, West Lafayette, IN*.

White, B., Stains, M., Escriu-Sune, M., Medaglia, E., Rostamnjad, L., Chinn, C., & Sevian, H. (2011). A Novel Instrument for Assessing Students' Critical Thinking Abilities. *Journal of College Science Teaching*, *40*, 102–107.

White, H. B., Benore, M. A., Sumter, T. F., Caldwell, B. D., & Bell, E. (2013). What skills should students of undergraduate biochemistry and molecular biology programs have upon graduation? *Biochemistry and Molecular Biology Education*, *41*(5), 297–301.

White, R., & Gunstone, R. (2014). *Probing understanding*. Routledge.

Wiggins, G. (1998). *Educative Assessment. Designing Assessments To Inform and Improve Student Performance.* ERIC.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA press.

APPENDICES – CHAPTER 1

1.1. Interview protocol for think-aloud interviews

Questions about the Course

- Thanks so much for coming, I really appreciate your help.
- For the first part of the interview, I just have a couple background questions and then some general questions about your experiences in 12A and B.

1. Courses
    - Did you do the standard 4A/B, 12A/B sequence?
    - What other chemistry courses have you taken so far?
2. What is your year in school and intended major?
    - Do you have an intended subfield or area of interest?
    - Do you know what you're thinking of doing after?
    - Are you looking to join any labs, or have you already? Which ones?
        - Have you done any chemistry research prior to joining a lab here?
3. What was your study routine like for organic, both in terms of ongoing learning and exam prep?
    - What activities did you spend the most time on?
    - What activities were most helpful?
    - Were there things you did at first but stopped doing because you did not find them helpful?
    - Are there changes you would make if you did it again?
    - Did this vary between the two semesters?
4. What were your biggest struggles in organic chemistry?
    - Did the biggest issues change over the course of the year?
5. When you see a problem on an exam (let's say predict the products) and you draw a blank at first, what do you do next?
6. What advice would you give to students who are starting 12A in the fall?

Questions about Section

Next I have a few questions that pertain to the discussion section(s) you attended. (Only Q8 asked for students not attending ChemScholars)

7. What aspects of section did you find to be most and least useful? Please feel free to be as honest as possible; I am very interested in learning how to improve this section for future.
    - What changes would you make?
8. Did you attend any other sections regularly? If so, what was your experience there, and how did it differ from this section.
    - What did you find particularly helpful or unhelpful about the other section(s)?

Thinkaloud Portion

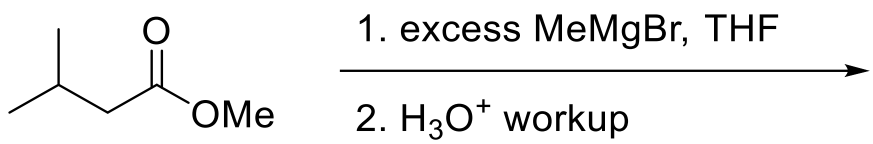
Part of what I'm trying to study is the detailed thought processes that go on in students' minds while they are working on solving typical organic chemistry problems. What I'm going to have you do is work through a few predict-the-products organic problems, and I want you to vocalize your thoughts as you have them, to the best of your ability. After you've finished working on the problem, let me know. If you are unsure of the answer, treat it like an exam and give your best guess.

Also, I want to say up front that you shouldn't take these problems as an indicator of how prepared you are for the final, even though they involve similar material. I can explain more about why afterwards.

We'll have a warm-up question to practice thinking aloud. Do you have any questions for me?

Training Problem
[Give paper with training problem:]

Predict the major organic product(s) of the following reactions. Please indicate stereochemistry where appropriate.

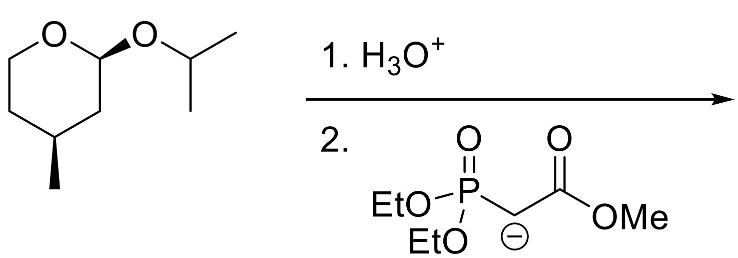


I want you to practice thinking aloud on this problem. I may prod you if you stop talking for too long. Also, I should point out that you're not *explaining* it to me or anybody else. We're trying to get at the best approximation of the thoughts you'd have if you were sitting alone, working on this problem without any cameras or stress or time constraints.

Great! For the remaining questions, I will allow you to finish working, and then I will have a number of follow-up questions. Some of those may involve asking you about a "hypothetical student answer" that you have not mentioned; these are asked to everyone and they do not indicate that your initial answer (or the hypothetical answer) is correct or incorrect.

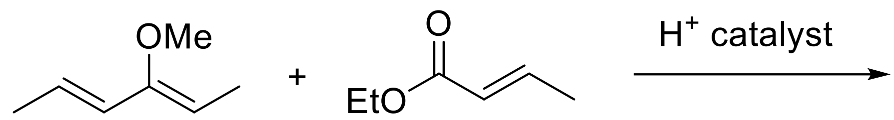
Problems
All questions will have the same instructions. [Each question presented sequentially on separate pieces of paper]

Problem 1
Predict the major organic product(s) of the following reactions.



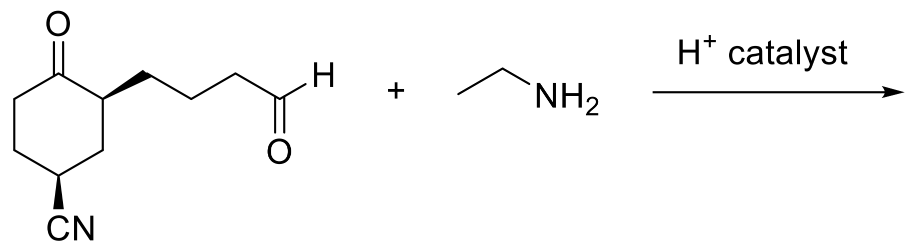
<u>Problem 2</u>

Predict the major organic product(s) of the following reaction.



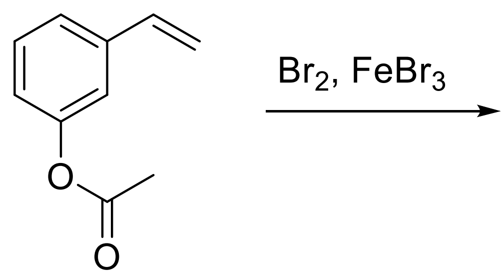
<u>Problem 3</u>

Predict the major organic product(s) of the following reaction.

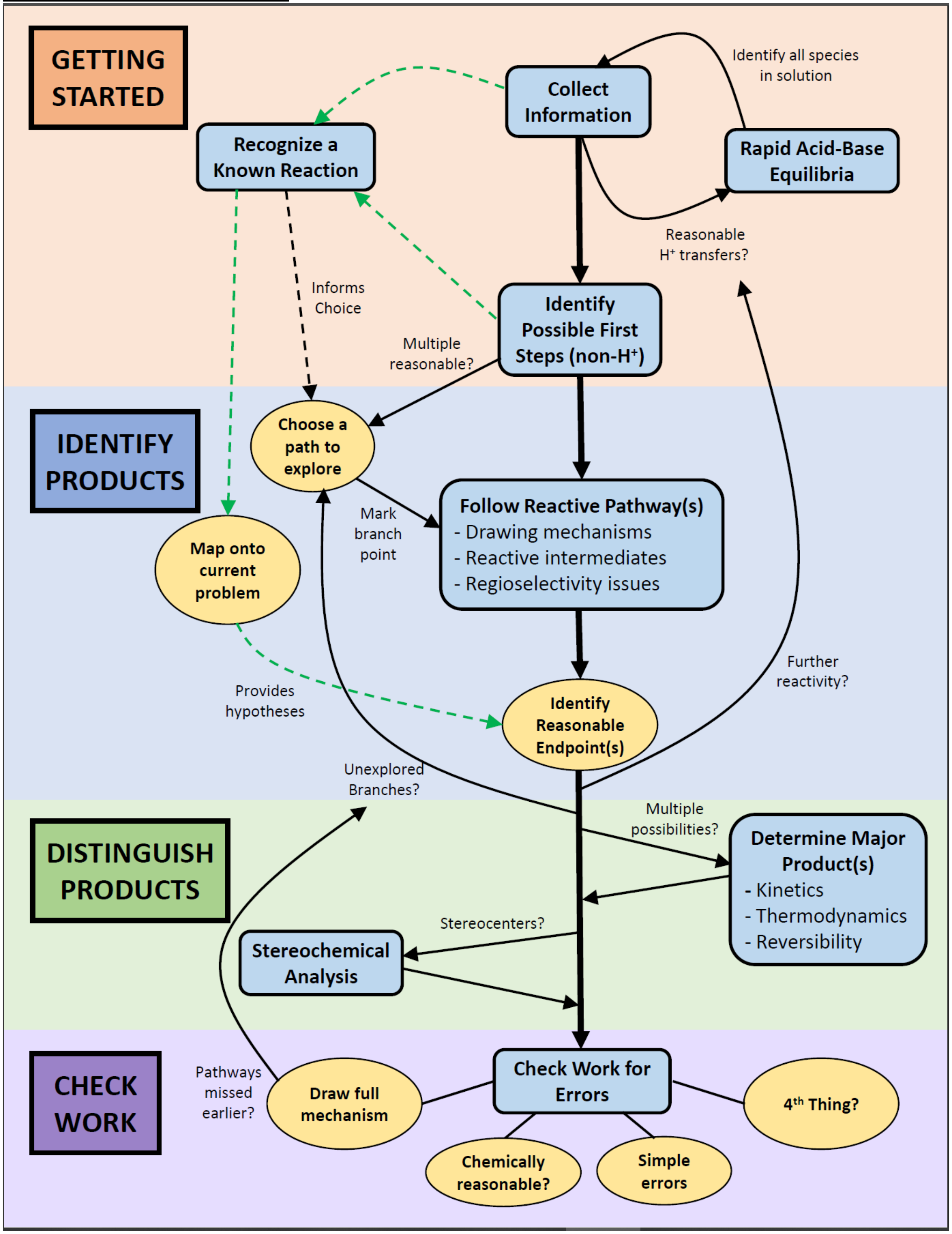


<u>Problem 4</u>

Predict the major organic product(s) of the following reaction.



1.2. First draft of the workflow



1.3. Advice for predicting reactivity (pre-study)

1. First Steps
    a. Identify the types of information available
    b. Determine what first steps are possible in this reaction mixture
    c. Choose one or more for further investigation
2. Identifying Products
    a. Following reaction pathways
        - Draw each step. On the paper. Not in your head.
        - Follow reactive intermediates (e.g., carbocations) until they run out of ways to make themselves more stable. There may be branching pathways.
        - Eventually: Recognize chunks of steps that often (but not always!!) go together. A subset of these are given names (e.g., "oxymercuration", "tautomerization", "aldol condensation"). For now, the chunks are small (many are just two steps), but this becomes more important as you learn reactions with longer mechanisms.
    b. For certain steps, pay attention to regioselectivity and stereochemistry (which products are possible, which are favored)
    c. Look for further reactivity
3. Distinguishing Products – Which are most likely to form?
    a. Key Questions
        i. What is the fastest reactive pathway? (kinetic considerations)
        ii. What is the most stable attainable product? (thermodynamic considerations)
        iii. Which steps are effectively irreversible under the given conditions? (does kinetics or thermodynamics dominate the result?)
        iv. What effects do various conditions (e.g., solvent, temperature) have on the product ratios?

1.4. Explanation sheet for workflow (used in focus groups)

1. **Collect Information** – Determine what information can be derived from the structures you are given. For example, where are the most acidic protons, most basic atoms, nucleophilic and electrophilic sites, functional groups, leaving groups, etc. It is highly recommended that you *write this information down*, since complex problems require you to hold a lot in your mind at once, and writing extends your working memory.

2. **Rapid Acid-Base Equilibria** – Do any rapid proton transfers, and determine what is present in solution at equilibrium. This will be the range of possible reactive species for the next step.

3. **Identify Possible First Steps (non-H^+)** – After the acid-base equilibrium, what are the possible elementary steps that could happen next? Often this will take the form of the strongest nucleophile attacking the strongest electrophile.
    a. **Determine Major Pathways** – If there are multiple possible first steps, narrow down potential pathways by determining whether one or more paths are likely to dominate

4. **Follow Reactive Pathways** – Draw an arrow-pushing mechanism for the elementary step chosen in the previous part. If the resulting intermediate is still reactive, "explore" this path by continuing a potential mechanism.
    a. **Identify Reasonable Endpoints** – At what point does the molecule seem to have no further possible reactivity? What neutral species can you isolate from this reaction, possibly after an aqueous workup?

5. **Determine Major Products** – If there are multiple reasonable products, which one is likely to be the major product (or products)? Is the reaction reversible? Will kinetics or thermodynamics dominate under the given conditions?

6. **Stereochemical Analysis** – If there are stereocenters in your final molecule, which of the possible stereoisomers will be formed, and will there a be a preference for some over others? Look back through the mechanism to determine at which point each stereocenter was formed, if they were not present at the beginning of the reaction.

7. **Check Your Work** – Are there alternate reactive pathways that you didn't explore? Can your final product react further under the given conditions? Did you make any other common mistakes like having a structure with five bonds to carbon? Did you make any drawing errors like losing a carbon?

A faster but less systematic way to approach the problem is to recognize a similarity between the current problem and a known reaction.

1. **Recognize Known Reaction** – What does this set of reagents and functional groups "typically do" when combined? Does it correspond to a reaction that you have specifically learned?

2. **Map onto Current Problem** – Draw out an example of a known reaction that you think applies to this problem. Map the atoms of the prototypical reaction onto the current substrate, and determine whether the same outcome is possible.

3. **Draw Full Mechanism** – Check that the known reaction is applicable in this case by walking through the mechanism and comparing it to the mechanism for the prototypical reaction. Is each step still likely to occur with the current substrate?

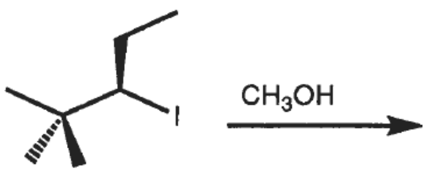
1.5. Workflow practice problems (used in focus groups)

Feedback Needed:

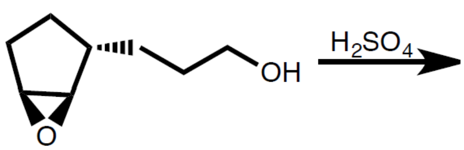
- Is the workflow understandable with minimal additional info?
  - What additional info is necessary to understand parts of the workflow?
- If you are unsure of an answer, does the workflow help organize your thinking?
- How could this workflow be improved, either in terms of usability or likelihood of generating reasonable answers?
- How can the workflow be pitched so that students will use it?

Predict the major organic products for the following reactions, using the workflow as a general guide to your process.

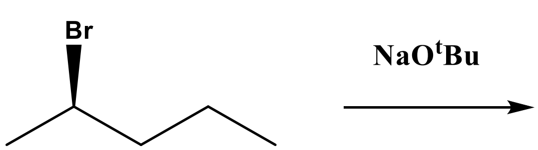
1.



2.

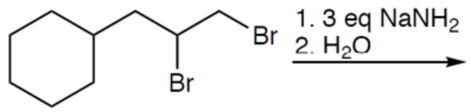


3.

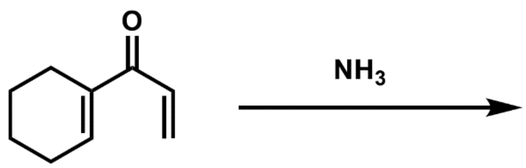


Predict the major organic products for the following reactions, using the workflow as a general guide to your process.

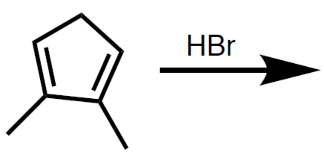
4.



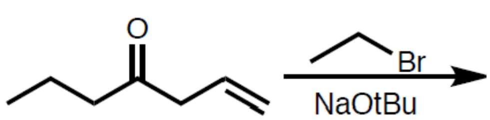
5.



6.



7.



Additional Feedback Needed:

- For reactions you recognize, does the "fast lane" seem appropriate?
- Can you think of additional prompting questions that are useful to think about when stuck? (E.g., What are the strongest nucleophile and electrophile in solution?)
- Could this workflow be simplified even more for when it's first rolled out (right after Midterm #2, partway through the SN1/SN2/E1/E2 section)?

1.6. Prompting questions to accompany workflow

FIRST STEPS

Collecting Information

- What is the potential function (acid, nucleophile, etc.) of each **functional group** or **reagent**?
- Is the reaction under acidic, basic, or neutral conditions?
- Which are the most **acidic** protons in the reaction?
    - Is one significantly (~5 pKa units) more acidic than the others, or are there multiple comparable protons?
- Where are the strongest **bases** in the reaction?
    - Are there multiple basic atoms of comparable strength?
- Where are the strongest **nucleophiles** in the reaction?
- Are there any good **electrophiles**?
- Are there any good **leaving groups**?
    - What types of atoms are they bonded to (e.g., secondary carbon)?
- Is there a **solvent** shown?
    - Is the solvent likely to be involved in the reaction?
- Is there a workup step shown?
- What other **conditions** are shown?
    - How many equivalents of each reagent are used?
    - Is a temperature or any other information given? What does that information tell you?
- Are there any particularly **unstable features**? (e.g., extremely strong acid, base, nucleophile, or electrophile, strained rings, O-O bonds, etc.)

Acid-Base Equilibria

- Where are the most acidic protons and most basic atoms in the reaction?
- Consider the strongest acid and the strongest base
    - Is a proton transfer likely based on **pKa values**?
    - Will that proton transfer be **reversible or irreversible**?
    - Are there other comparably acidic protons to consider (especially in reversible situations or if there is excess base)?
    - Are there other comparably basic atoms to consider (especially in reversible situations or if there is excess acid)?

Identifying Possible First Steps

- Where are the strongest **nucleophiles and electrophiles** in the reaction?
- Consider the strongest nucleophile and the best electrophile
    - Is the nucleophile strong enough to attack the electrophile?
        - If not, are there any good leaving groups? If so, could it leave spontaneously under the given conditions?
    - Are there other competing nucleophiles or electrophiles of comparable strength to consider (especially in a reversible situation)?
- Which of the ~10 **elementary steps** you've learned could possibly happen at this point?
    - Which of these are reversible under the given conditions?

IDENTIFYING POSSIBLE PRODUCTS

Determining Major Pathways

- Of the possible first steps, are some much more likely than others?
- Are there any steps that would be fast and **irreversible**?
- Are there multiple possible **regioisomers**?
    - Are any of them minor enough where that path doesn't need to be followed?

Following Reactive Pathways

- If there are familiar **reactive intermediates** (e.g., carbocation, oxyanion, enolate), what steps do they usually undergo to become less reactive?
- Are there multiple possible **regioisomers**?
    - Are any of them minor enough to ignore?

Identifying Reasonable Endpoint(s)

- Are there still **strong** acids, bases, nucleophiles, or electrophiles present?
    - If multiple **equivalents** are indicated, did more than one get used?
    - If a reagent is indicated as **catalytic**, did you regenerate it in your mechanism?
- Is the substrate **neutral**, or would it be neutral after aqueous workup?
- Could the proposed product **react further** under the reaction conditions?

DISTINGUISHING BETWEEN POSSIBLE PRODUCTS

Determining Major Products (if there are multiple possible ones)

- What is the fastest reactive pathway? (**kinetic** considerations)
- What is the most stable attainable product? (**thermodynamic** considerations)
- Are any steps effectively **irreversible** under the given conditions?
    - Are we generally under kinetic (irreversible) or thermodynamic (reversible) conditions?
- What effects do various **conditions** (e.g., solvent, temperature) have on the product ratios?

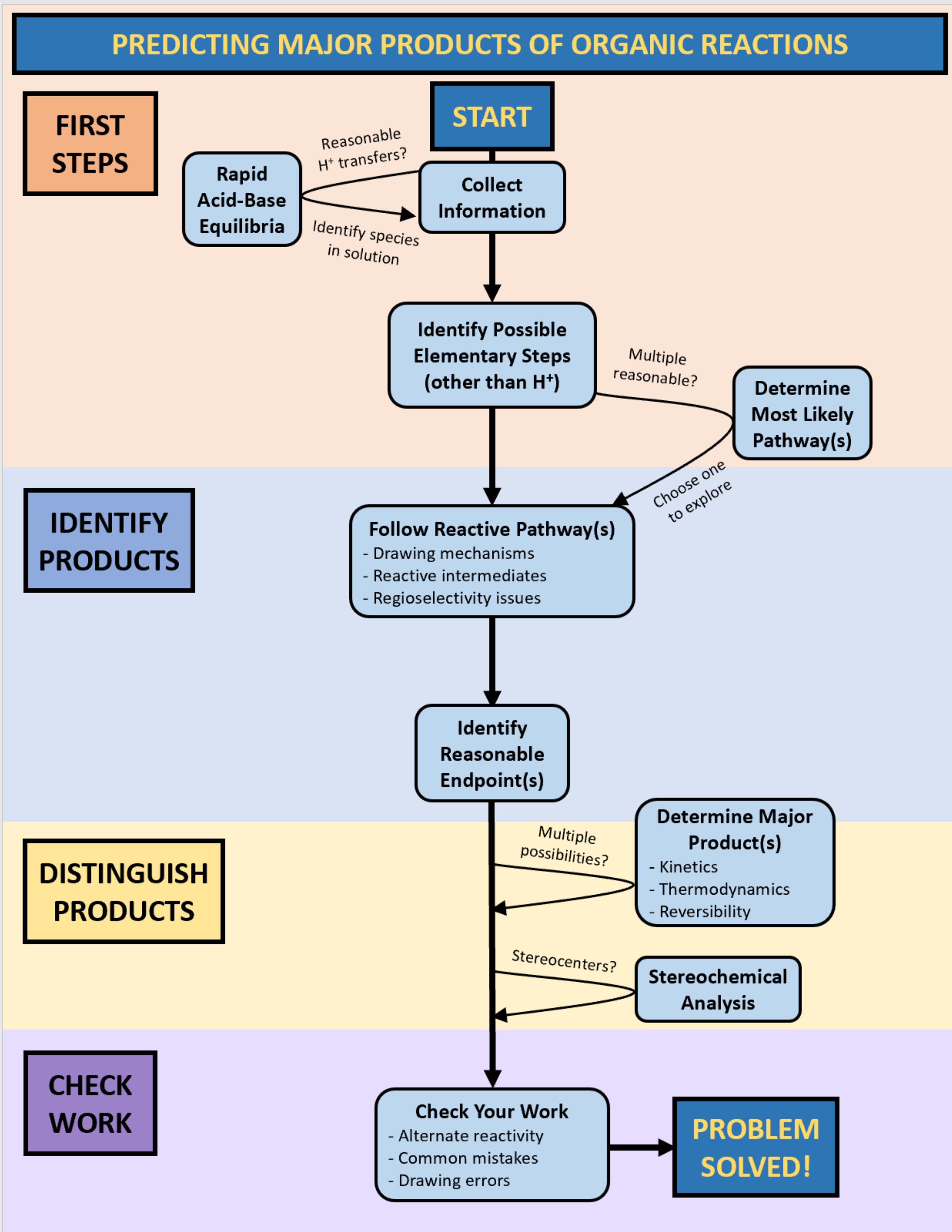
Stereochemical Analysis (if there are multiple possible stereoisomers)

- What stereochemistry is present in the starting molecules?
- **Which steps** in the mechanism involve creating or destroying stereocenters (or alkenes with stereochemistry)?
- For each new stereocenter formed, is one isomer heavily favored, slightly favored, or exactly as likely as the other?

CHECK YOUR WORK

- Look back through a complete **mechanism** for the proposed reaction
    - Are there **alternate branches** that might be competitive?
    - Can my proposed final product **react further** under the reaction conditions?
- Did I make any common mistakes?
    - Do I have 5 bonds to carbon or nitrogen?
    - Am I missing any formal charges?
- Did I make any drawing errors?
    - Did I accidentally add or lose any carbon atoms?
    - Do I need to add in stereochemistry?

1.7. Simplified workflow for introduction to students



1.8. Example student comments for each primary code

Code	Abbrev.	Example
Collect Information	CollInfo	"This group looks like a protecting group, which is like... and this is also an acetal, … and I know that these can be used to protect alcohols"
Acid-Base Equilibria	AcidBase	"Normally if it's just a carbonyl I would think this would protonate here, make it a reactive center"
Identify First Steps (non-H+)	IdentFirst	"Um, I am going to have the amine come in to the aldehyde because that's more reactive"
Follow Reactive Pathway	FRP	"We have this... this would be attacked by the negative charge on the carbonyl"
Recognize Similarity to Known Reaction	RecogRxn	"I was thinking that the HWE reaction, because of this step"
Map Onto Current Problem	MapTotal	"so that's going to create a double bond there to this guy, this will add to carbonyls"
Identify Reasonable Endpoints	IdentEnd	"Then our final product would be a 1 2 3 4 5 6-membered ring"
Propose Alternate Reactivity	PropAlt	"Or what also could possibly happen is if that leaves, then I could form this positive charge, which would be resonance stabilized"
Decide Between Pathways/ Products	Decide	"So I'm looking between these two, and it looks like, this one looks more reasonable because this looks like some alcohol that's a good leaving group"
Stereochemical Analysis	Stereo	"I want to say that this one is going to be opposite of that one, just because that would make the most sense stereochemically, but I don't think it matters which one goes up and which one goes down"
Assess Progress	AssessProg	"Okay, but the problem here is like, there's still a positive charge on the oxygen"
Check Work	CheckTotal	"I'm trying to think if this could attack here, that's a weird thing, oh I guess I can try, wait no, then that would form into a 4-membered ring so that probably shouldn't happen"

APPENDICES – Chapter 2

2.1. Additional contrasting cases lesson

In the third PFL lesson, the focus was on familiarizing students with how changes to a reaction can affect the outcome of that reaction. As discussed in the Introduction, Graulich and Schween (2018) suggest an approach in which students are given contrasting cases that differ by only a single variable, in order to help students focus on the meaningful features of the problem. This was accomplished by showing them contrasting cases like those in Figure 2.1.1. Some of the pairings involved reactivity they had already seen, and some incorporated novel reactions. Students were given the instructions and worked example shown in Figure 2.1.2.

Figure 2.1.1. Examples of contrasting cases

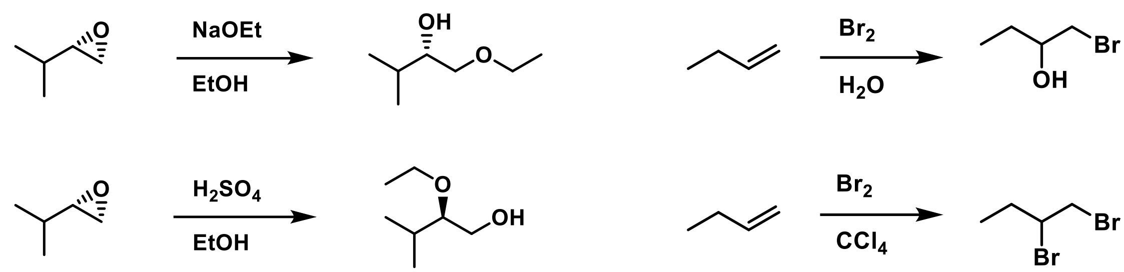
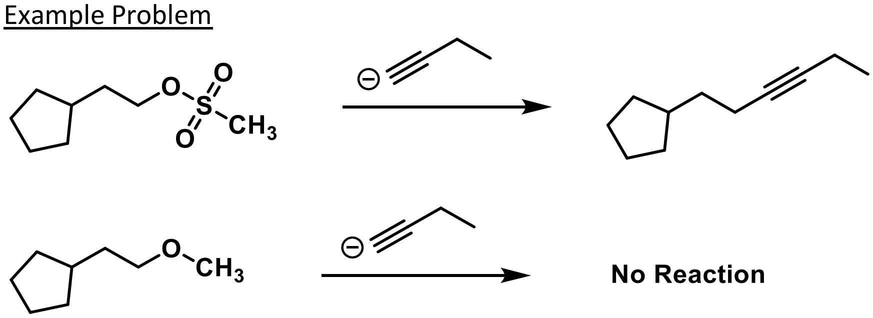


Figure 2.1.2. Instructions for the third PFL activity

Instructions: For each pair or set of reactions:

a) Explain why the outcome is different, focusing on features and chemical principles that account for that difference.
b) Identify key reaction features to look for when predicting the outcome of organic reactions.

Example Problem



a) The top reaction has a good leaving group (MsO^-), whereas the corresponding part of the bottom substrate (MeO^-) is a bad leaving group. The charge on MsO^- is heavily stabilized by resonance, whereas the charge on MeO^- is fairly unstable.
b) Leaving group ability: If you are doing an S_N2 (or E2) process, make sure your leaving group is a good leaving group. Good leaving groups generally have stabilized negative charges or are neutral (after leaving), and they are always very weak bases.

After students had time to explore and discuss the contrasting cases, the instructor went over the target concepts with the full group. The learning goals here are closely related to the predict-the-product workflow developed in Chapter 1. When students are first approaching a predict-the-product problem, they start by collecting information, including both information provided on the page and relevant knowledge that is immediately activated upon seeing some feature of the problem. The intent of this lesson is to help students get better at this crucial problem-solving step by highlighting to them what features they should be considering when thinking about organic reactivity.

This lesson took place in the later portion of Chem 12A, just before they started a unit on electrophilic addition. Students were given pre- and post-quizzes targeting the same key features to look for an immediate effect of participation in the lesson. No control lesson was offered, because only 9 students participated, due to a competing midterm review session for a physical chemistry course. Students scored significantly higher on the post-test than on the pre-test ($p < 0.01$). The gains were greater on the questions testing topics that had been covered as part of the lesson.

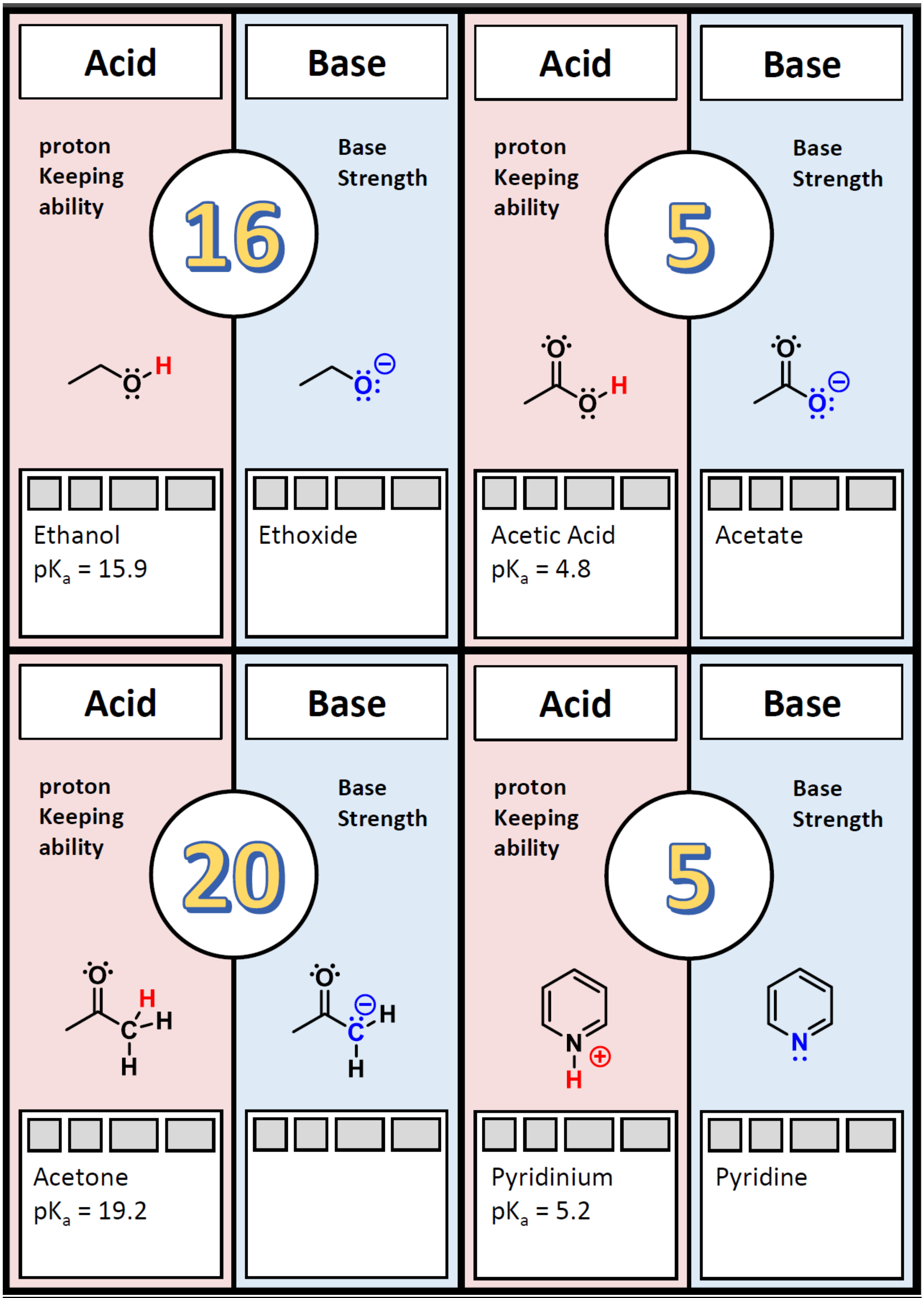
To generate a control group, the 18 most frequent ChemScholars attendees were split into those who attended this lesson and those who did not, resulting in 10 that didn't and 8 that did. Average attendance between the two groups was not significantly different. However, on predict-the-product exam questions, the experimental group did not score significantly higher than the comparison group or the rest of the class on any of the 15 questions investigated, using a stricter significance criterion of $p < 0.01$ due to multiple tests.

Overall, this was a lower quality lesson, made even less useful by the low turnout. The learning goals were slightly muddled, and the lesson was hampered by the fact that most of the information was either already understood by the students or too far advanced to be approachable. Teaching a cross-cutting concept like "what information do I need to be looking at to determine reactivity?" is potentially not well suited for a PFL lesson. As a result, this was left out of the main chapter to better focus on the successful lessons.

2.2. Implicit vs. explicit assessment

I originally went into this work with a more specific hypothesis, based on my understanding of the Preparation for Future Learning literature. Because PFL targets a specific type of transfer ("knowing with"), I proposed that students who participate in a PFL lesson may not do better on questions that explicitly test the content of that lesson ("knowing that"), but they will do better when that content is tested implicitly, as they "know with" their knowledge and bring it to mind when needed. However, splitting the exam questions into explicit and implicit questions did not yield differing results, so this hypothesis was not stated in the main text.

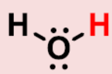
2.3. Full set of pK_a cards



Acid

proton Keeping ability

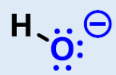
16



Water
pK_a = 15.7

Base

Base Strength

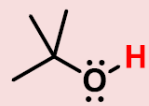


Hydroxide

Acid

proton Keeping ability

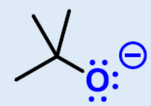
16



tert-Butanol
pK_a = 17

Base

Base Strength

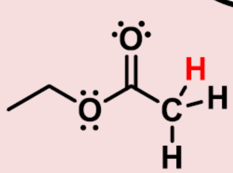


tert-Butoxide

Acid

proton Keeping ability

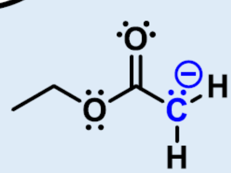
25



Ethyl acetate
pK_a = 25

Base

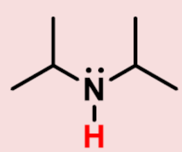
Base Strength



Acid

proton Keeping ability

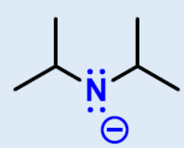
38



Diisopropyl-amine
pK_a = 38

Base

Base Strength

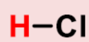


Diisopropyl-amide

Acid

proton Keeping ability

-7



Hydrochloric acid
pK_a = -7

Base

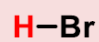
Base Strength

Chloride

Acid

proton Keeping ability

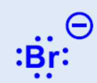
-7



Hydrobromic acid
pK_a = -8

Base

Base Strength

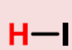


Bromide

Acid

proton Keeping ability

-7



Hydroiodic acid
pK_a = -10

Base

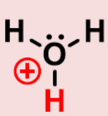
Base Strength

Iodide

Acid

proton Keeping ability

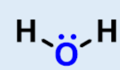
-2



Hydronium
pK_a = -1.7

Base

Base Strength

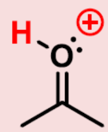


Water

Acid

proton Keeping ability

-7



pK$_a$ ≈ -7

Base

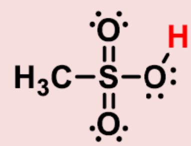
Base Strength

Acetone

Acid

proton Keeping ability

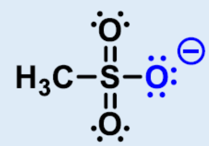
-2



Methane-sulfonic acid
pK$_a$ = -2.6

Base

Base Strength

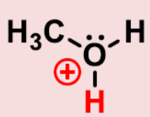


Mesylate

Acid

proton Keeping ability

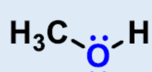
-2



pK$_a$ = -2.2

Base

Base Strength

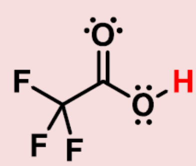


Methanol

Acid

proton Keeping ability

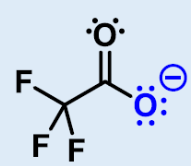
5



Trifluoro-acetic acid
pK$_a$ = -0.3

Base

Base Strength

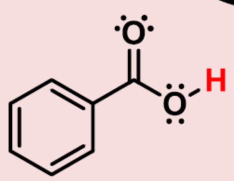


Trifluoro-acetate

Acid

proton Keeping ability

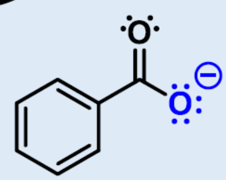
5



Benzoic Acid
pK_a = 4.2

Base

Base Strength

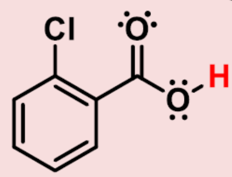


Benzoate

Acid

proton Keeping ability

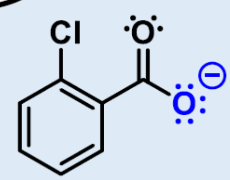
5



2-chloro-
benzoic acid
pK_a = 2.9

Base

Base Strength

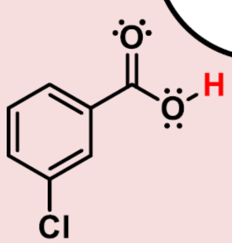


2-chloro-
benzoate

Acid

proton Keeping ability

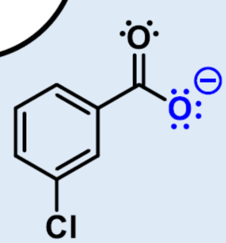
5



3-chloro-
benzoic acid
pK_a = 3.8

Base

Base Strength

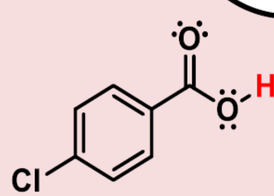


3-chloro-
benzoate

Acid

proton Keeping ability

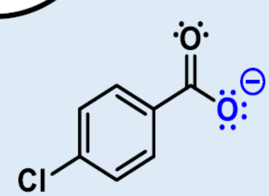
5



4-chloro-
benzoic acid
pK_a = 4.0

Base

Base Strength

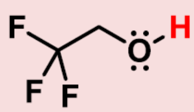


4-chloro-
benzoate

Acid

proton Keeping ability

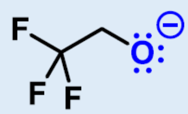
16



Trifluoro-
ethanol
pK_a = 12.5

Base

Base Strength

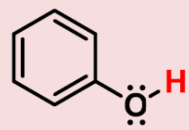


Trifluoro-
ethoxide

Acid

proton Keeping ability

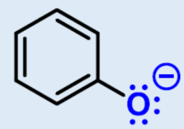
10



Phenol
pK_a = 10.0

Base

Base Strength

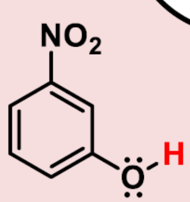


Phenoxide

Acid

proton Keeping ability

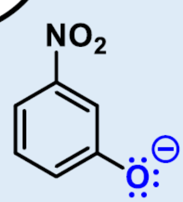
10



3-Nitro-
phenol
pK_a = 8.4

Base

Base Strength

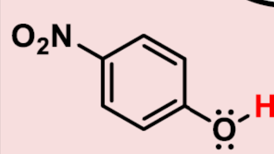


3-Nitro-
phenoxide

Acid

proton Keeping ability

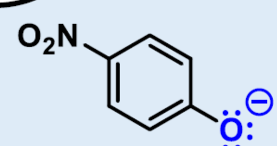
10



4-Nitro-
phenol
pK_a = 7.1

Base

Base Strength

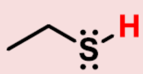


4-Nitro-
phenoxide

Acid

proton Keeping ability

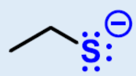
10



Ethanethiol
pK_a = 10.6

Base

Base Strength

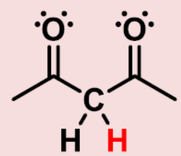


Ethane-thiolate

Acid

proton Keeping ability

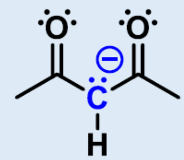
10



Pentane-2,4-dione
pK_a = 9.0

Base

Base Strength



Acid

proton Keeping ability

10

H–C≡N:

Hydrogen cyanide
pK_a = 9.4

Base

Base Strength

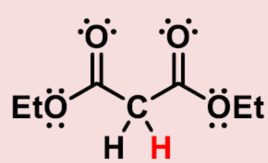
⊖:C≡N:

Cyanide

Acid

proton Keeping ability

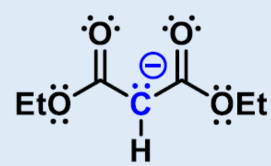
10



Diethyl-malonate
pK_a = 13

Base

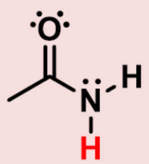
Base Strength



Acid

proton Keeping ability

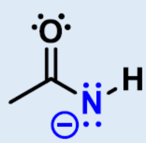
16



Acetamide
pK_a = 15.1

Base

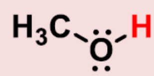
Base Strength



Acid

proton Keeping ability

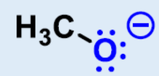
16



Methanol
pK_a = 15.5

Base

Base Strength

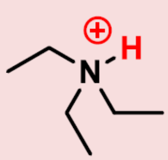


Methoxide

Acid

proton Keeping ability

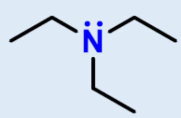
10



Triethyl-
ammonium
pK_a = 10.8

Base

Base Strength

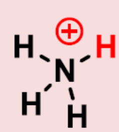


Triethylamine

Acid

proton Keeping ability

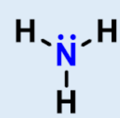
10



Ammonium
pK_a = 9.2

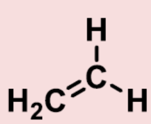
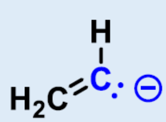
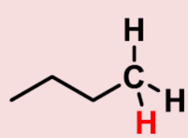
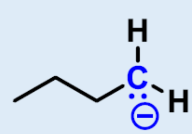
Base

Base Strength



Ammonia

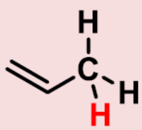
Acid	Base	Acid	Base
proton Keeping ability	Base Strength	proton Keeping ability	Base Strength
25		38	
H−C≡C−H	⊖:C≡C−H	H−H	:H⊖
Acetylene pK_a = 25	Acetylide	Hydrogen pK_a = 35	Hydride

Acid	Base	Acid	Base
proton Keeping ability	Base Strength	proton Keeping ability	Base Strength
44		50	
			
Ethylene pK_a ≈ 44		Butane pK_a ≈ 50	

Acid

proton Keeping ability

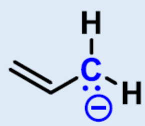
44



Propene
pK_a = 43

Base

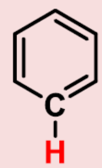
Base Strength



Acid

proton Keeping ability

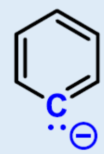
44



Benzene
pK_a = 43

Base

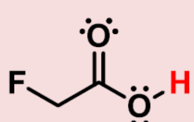
Base Strength



Acid

proton Keeping ability

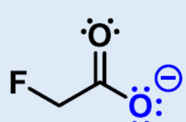
5



Fluoroacetic acid
pK_a = 2.7

Base

Base Strength

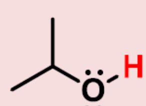


Fluoroacetate

Acid

proton Keeping ability

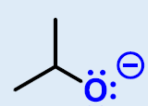
16



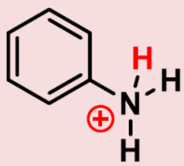
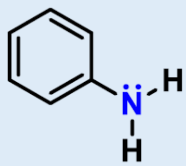
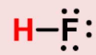
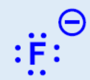
Isopropanol
pK_a = 16.5

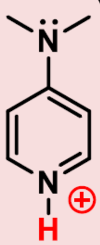
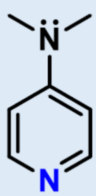
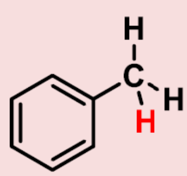
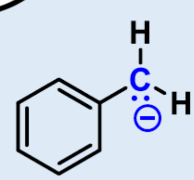
Base

Base Strength



Isopropoxide

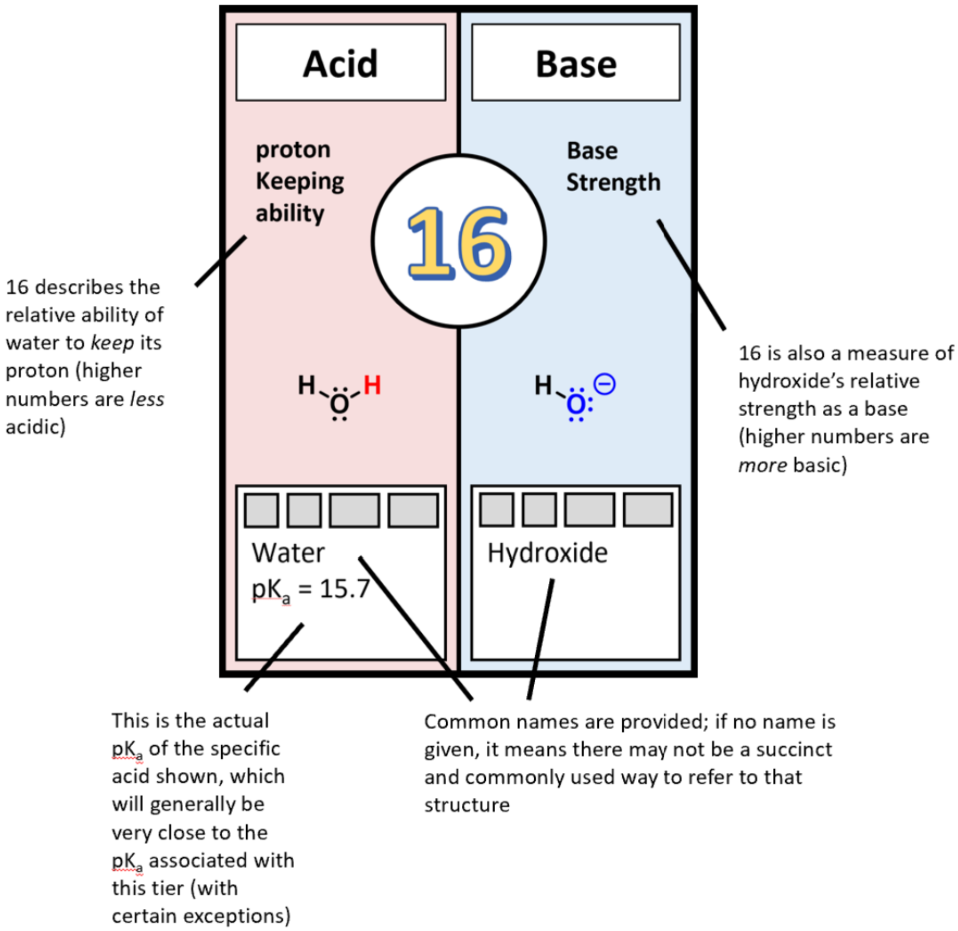
Acid	Base	Acid	Base
proton Keeping ability	Base Strength	proton Keeping ability	Base Strength
5		5	
			
pK_a = 4.6	Aniline	Hydrofluoric acid pK_a = 3.2	Fluoride

Acid	Base	Acid	Base
proton Keeping ability	Base Strength	proton Keeping ability	Base Strength
5		44	
			
pK_a = 9.6	DMAP (Dimethyl-aminopyridine	Toluene pK_a = 41	

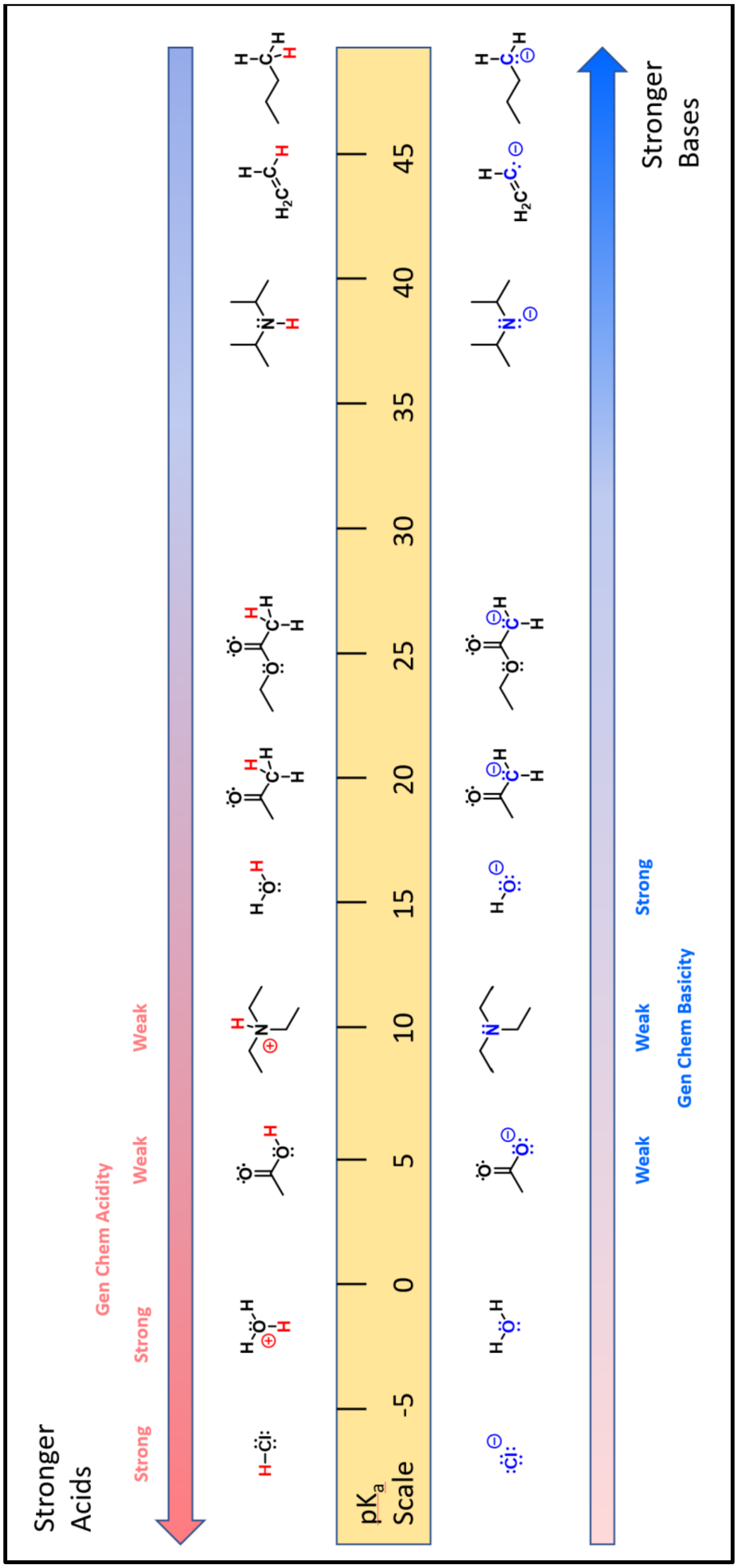
2.4. Explainer sheet for pK_a cards

Each card depicts one conjugate acid-base pair.
The central number corresponds to one of 10 "tiers" of acid-base pair strengths, which is equal to the pK_a for a typical acid at that tier.

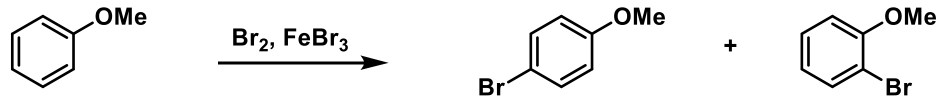
The tiers are: -7, -2, 5, 10, 16, 20, 25, 38, 44, and 50

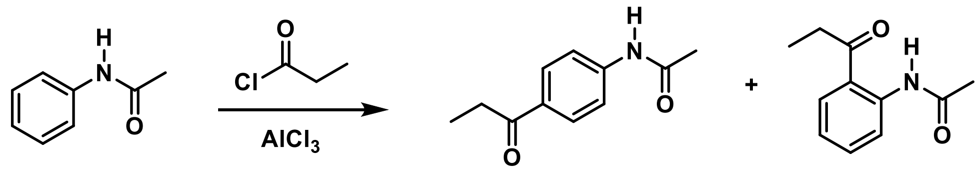


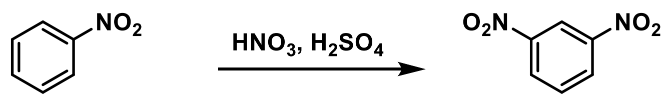
2.5. pK_a ruler

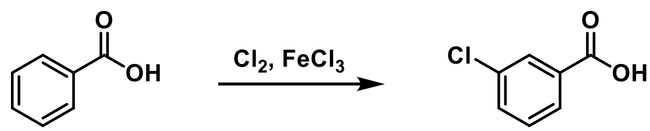

Stronger Acids
Strong
Strong
Gen Chem Acidity
Weak
Weak
pKa Scale
-5
0
5
10
15
20
25
30
35
40
45
Weak
Weak
Strong
Gen Chem Basicity
Stronger Bases

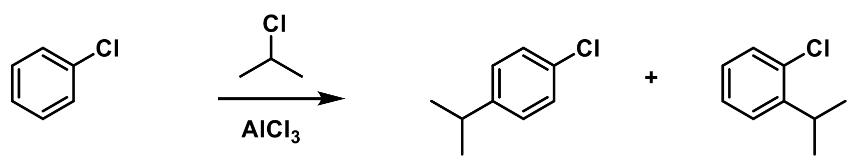
2.6. Complete dataset for S_EAr lesson

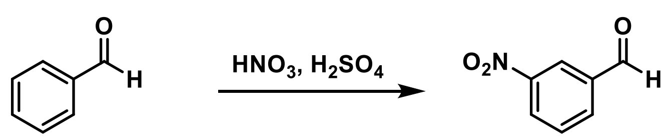


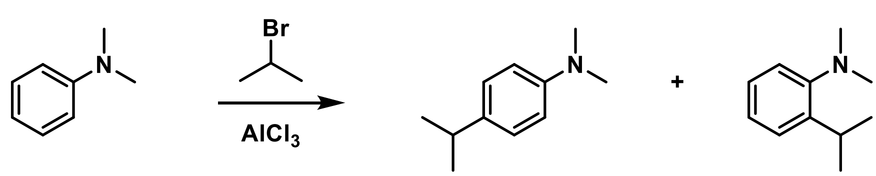


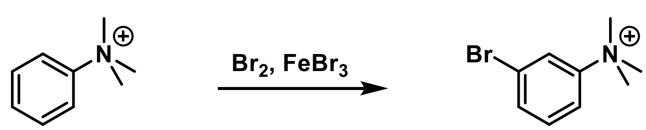


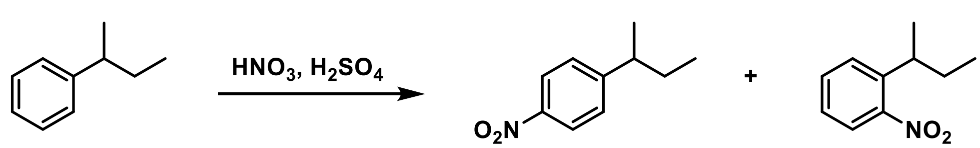


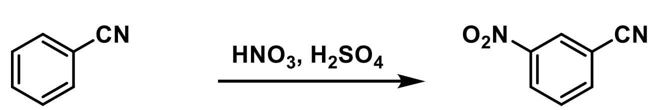


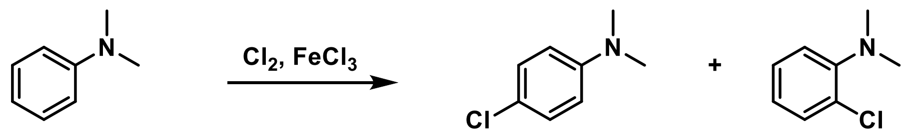


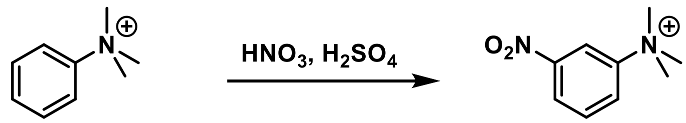


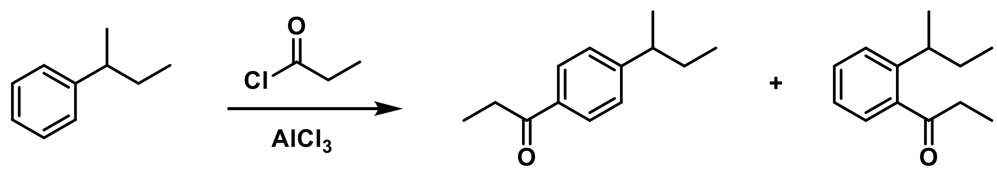


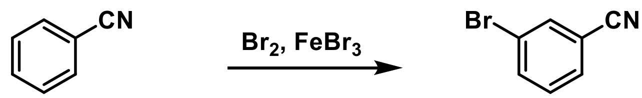


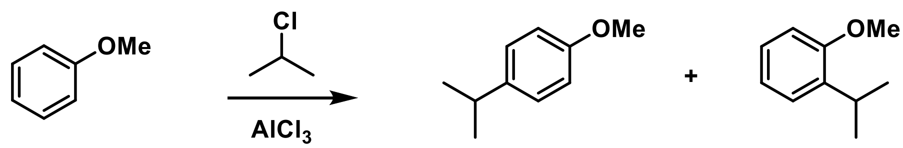


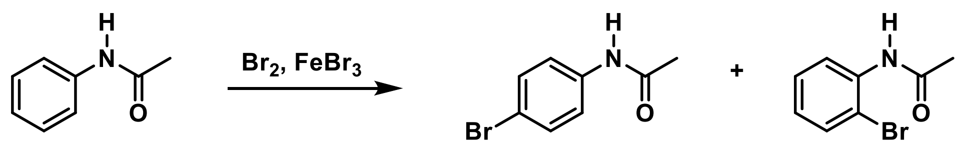

Cl2, FeCl3
N
Cl
+
N
Cl

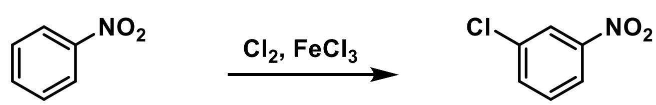

N
HNO3, H2SO4
O2N
N

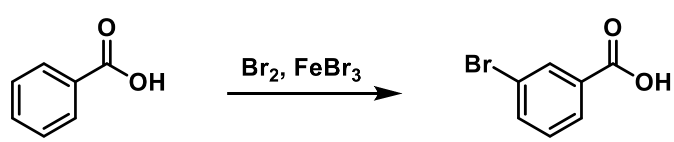

O
Cl
AlCl3
O
+
O

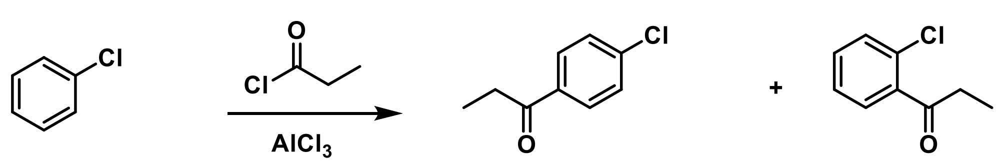

CN
Br2, FeBr3
Br
CN

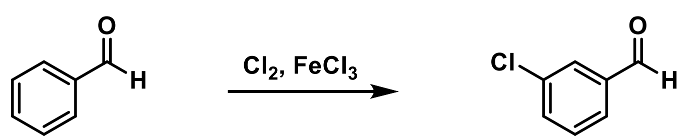

OMe
Cl
AlCl3
OMe
+
OMe


H
N
O
Br2, FeBr3
H
N
O
Br
+
H
N
O
Br


NO2
Cl2, FeCl3
Cl
NO2


O
OH
Br2, FeBr3
Br
O
OH

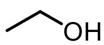

Cl
O
Cl
AlCl3
Cl
O
+
Cl
O


O
H
Cl2, FeCl3
Cl
O
H

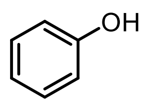
2.7. Acid-base lesson pre-test

For all questions, explain your reasoning in the box to the right. The explanations do not need to be long or use complete sentences, just a few key words on why you chose that answer.

1. For each pair of molecules, circle the molecule that is the stronger acid:



vs.

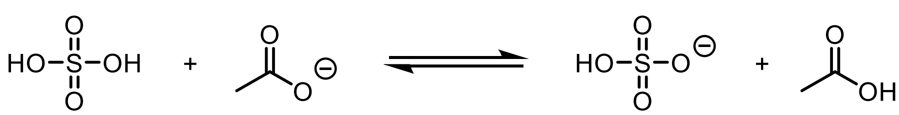


Molecule with a pK_a of 1 vs. Molecule with a pK_a of 5

2. Circle the compound that is the stronger base

$K^{+}\ OH^{-}$ vs. $Na^{+}\ NH_2^{-}$

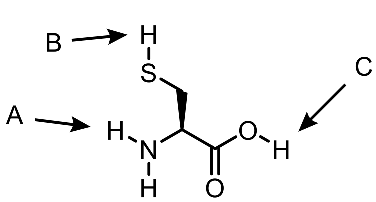
3. Which side of the following acid-base equilibrium would be favored?



   a. The left side
   b. The right side
   c. About equally favored

   If unsure, is there any additional information that would help you answer this question?

4. Rank the three indicated protons from most acidic to least acidic.



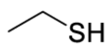
2.8. Practice problems for the control lesson

Problems:

1. Based on the pKa table and what you've learned about trends in acidity and basicity, estimate the pKas of the following compounds. Explain your reasoning by comparing the molecule of unknown pKa to a molecule of known pKa.

comparable molecule

A.

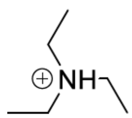


estimated pKa=

known pKa=

The unknown pKa will be greater than/less than/approximately equal to the pKa of the known compound because:

B.

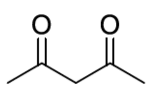


estimated pKa=

known pKa=

The unknown pKa will be greater than/less than/approximately equal to the pKa of the known compound because:

C.

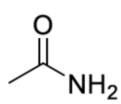


estimated pKa=

known pKa=

The unknown pKa will be greater than/less than/approximately equal to the pKa of the known compound because:

D.

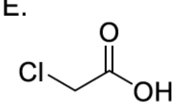


estimated pKa=

known pKa=

The unknown pKa will be greater than/less than/approximately equal to the pKa of the known compound because:

E.

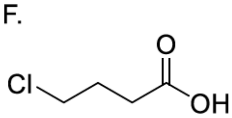


estimated pKa=

known pKa=

The unknown pKa will be greater than/less than/approximately equal to the pKa of the known compound because:

F.



estimated pKa=

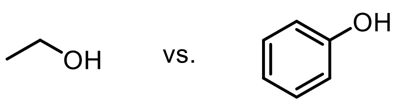
known pKa=

The unknown pKa will be greater than/less than/approximately equal to the pKa of the known compound because:

2.9. Acid-base lesson post-test

For all questions, explain your reasoning in the box to the right. The explanations do not need to be long or use complete sentences, just a few key words on why you chose that answer.

1. For each pair of molecules, circle the molecule that is the stronger acid:

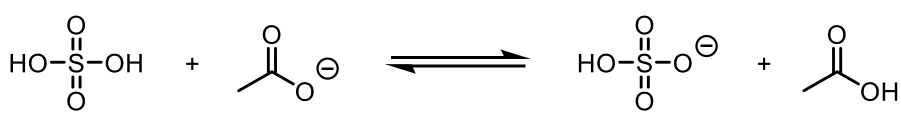


Molecule with a pK_a of 1	vs.	Molecule with a pK_a of 5

2. Circle the compound that is the stronger base

$K^{\oplus}\ {}^{\ominus}OH$ vs. $Na^{\oplus}\ {}^{\ominus}NH_2$

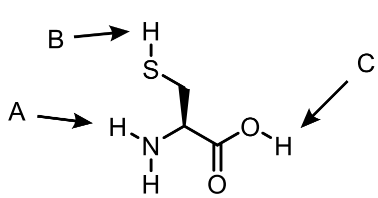
3. Which side of the following acid-base equilibrium would be favored?



   a. The left side
   b. The right side
   c. About equally favored

   If unsure, is there any additional information that would help you answer this question?

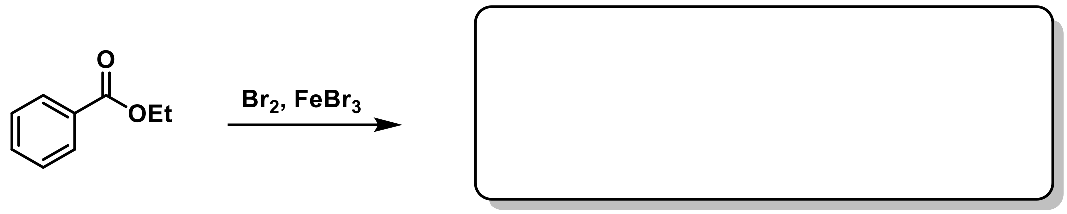
4. Rank the three indicated protons from most acidic to least acidic.



2.10. S_EAr lesson post-test

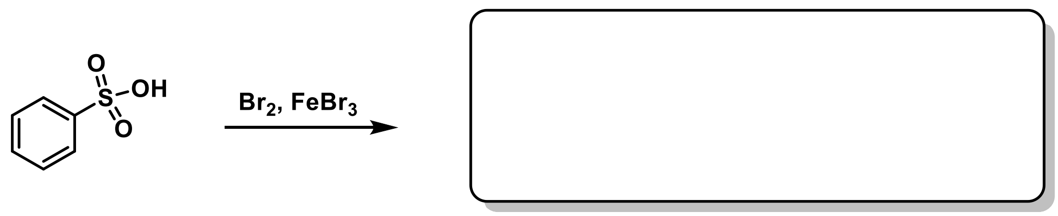
Predict the major organic product(s) of the following reactions. Assume only one equivalent of Br_2 is used. Provide a brief rationale (words or structures) for each answer.

1.



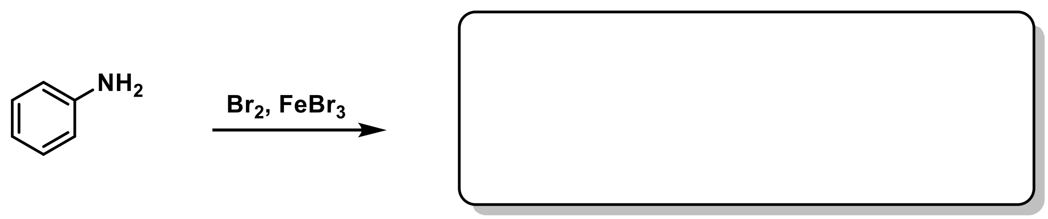
Rationale:

2.



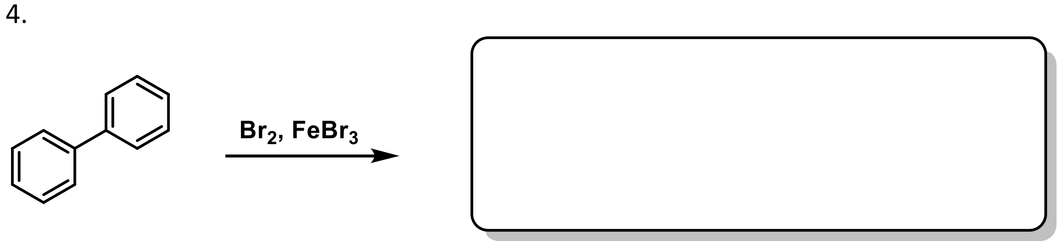
Rationale:

3.



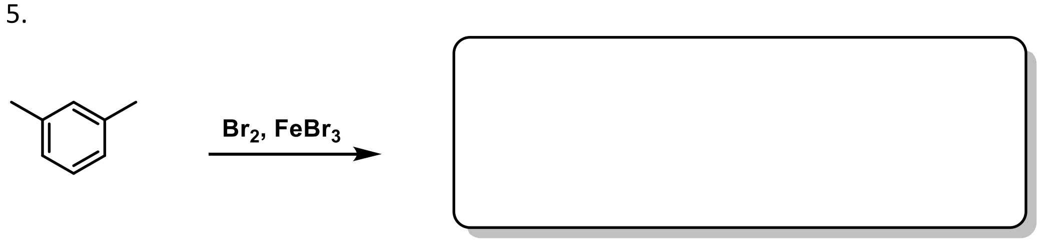
Rationale:

4.



Rationale:

5.



Rationale:

2.11. Observational note

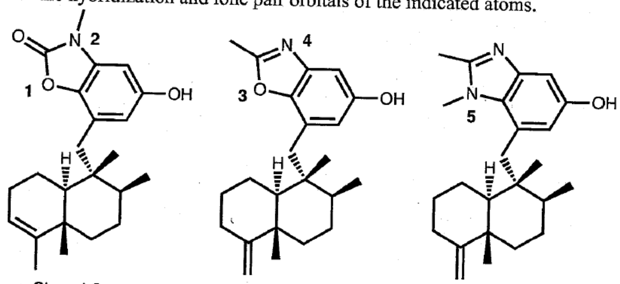
The observational data from the sections did not prove especially useful, so no results from this were included in the main text. However, there was one interesting thing that we did *not* hear. When piloting the S_EAr lesson, all three small groups spontaneously started referring to substituents as electron-donating or electron-withdrawing groups. However, they were assigning these labels purely on inductive effects, resulting in the wrong classification for most groups. As part of the lesson, the lead researcher went over the idea with each group that things can be withdrawing by induction but donating by resonance, and the resonance effects generally dominate if there is a pi system. However, in the S_EAr lesson reported on in this chapter, no groups were overheard misclassifying substituents. We do not know what caused this difference between cohorts.

2.12. Relevant exam questions (Acid-base lesson)

Midterm 1

2f. (Explicit)

e. Two other forms of Cinerol are found in nature. All three Cinerols are shown below. Fill in the table for the hybridization and lone pair orbitals of the indicated atoms.

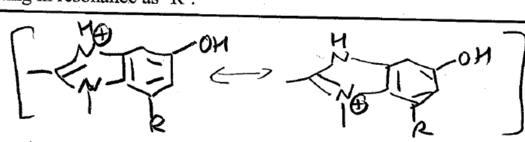


Cinerol G   Cinerol E   Cinerol A

	Oxygen **1**	Nitrogen **2**	Oxygen **3**	Nitrogen **4**	Nitrogen **5**
Atomic hybridization	sp2	sp2	sp2	sp2	sp2
1st Lone pair orbital	p	p	p	sp2	p
2nd Lone pair orbital (oxygen only)	sp2		sp2		

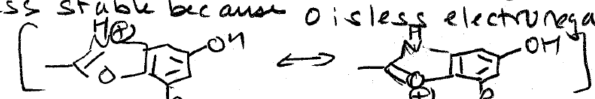
f. Which of the Cinerols in part e is most basic? Draw the protonated molecule and explain your selection. If resonance is a part of your explanation, draw the relevant resonance structures. You do not need to draw the entire structure in your resonance structures. You may abbreviate the parts of the molecule that are not participating in resonance as 'R'.

A is most basic

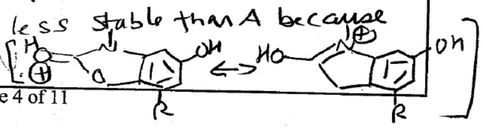


In A most important resonance structure shares ⊕ on the other N.

In Cinerol E, the ⊕ is shared on the oxygen which is less stable because O is less electronegative



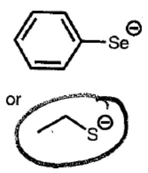
In Cinerol G, the oxygen is protonated & ⊕ is shared between N & O which is less stable than A because O is more electroneg than N



Page 4 of 11

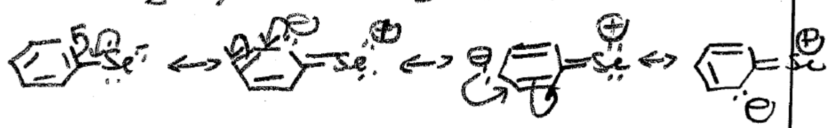
6a. (Explicit)

a. Circle the strongest **base**.



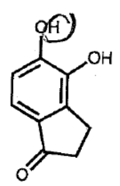
Explanation

Because sulfur is above se in periodic table. Therefore, S is smaller & stabilizes ⊖ less. In addition  is resonance stabilized, while  is not

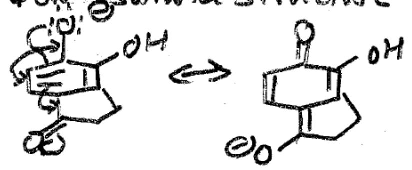


6b. (Explicit)

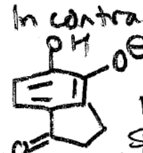
b. Circle the most **acidic** hydrogen.



Explanation The conjugate base is stabilized by resonance & one resonance structure has ⊖ on second O.



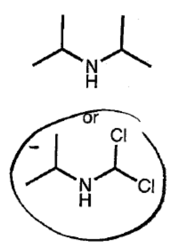
In contrast



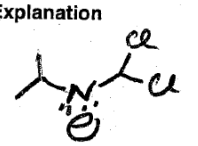
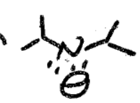
Does not have resonance structures that distribute ⊖ onto the 2nd O

6c. (Explicit)

c. Circle the strongest **acid**



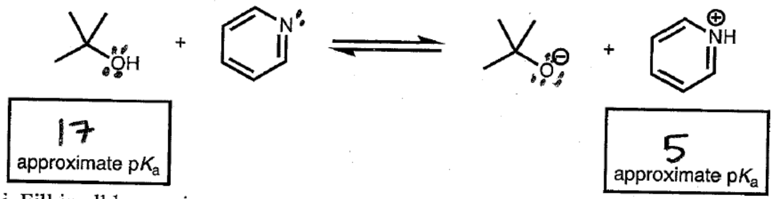
Explanation

 is more stable than 

because Cl are e⁻ withdrawing by induction while the CH_3 groups are e⁻ donating by induction

6d. (Explicit)

d. Consider the following equilibrium.



i. Fill in all lone pairs.
ii. Fill in approximate pK_a values in the boxes.
iii. Will the products or starting materials be favored in this equilibrium? Explain your answer.

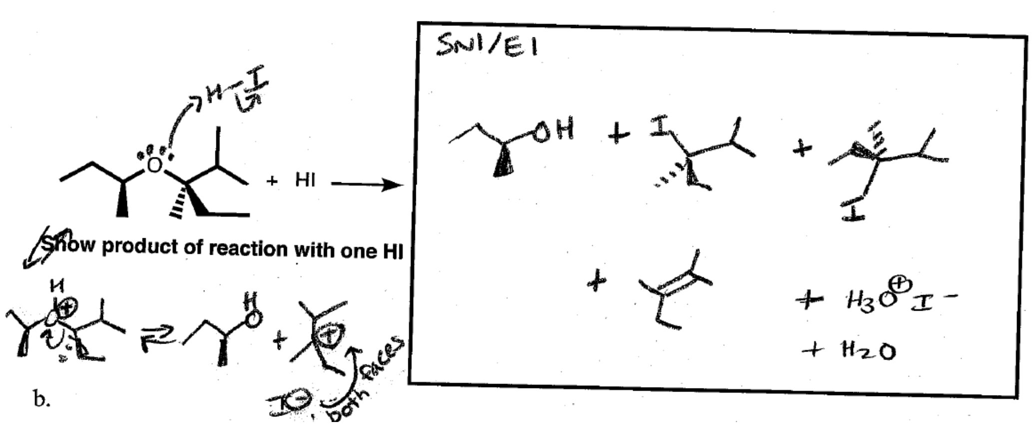
The starting materials are favored because the stronger conjugate base & conjugate acid are on the product side & are less stable than the conjugate acid & base on the starting material side.

Midterm 2

1a. (Implicit)

... part 1a, you will fill in the reagent.

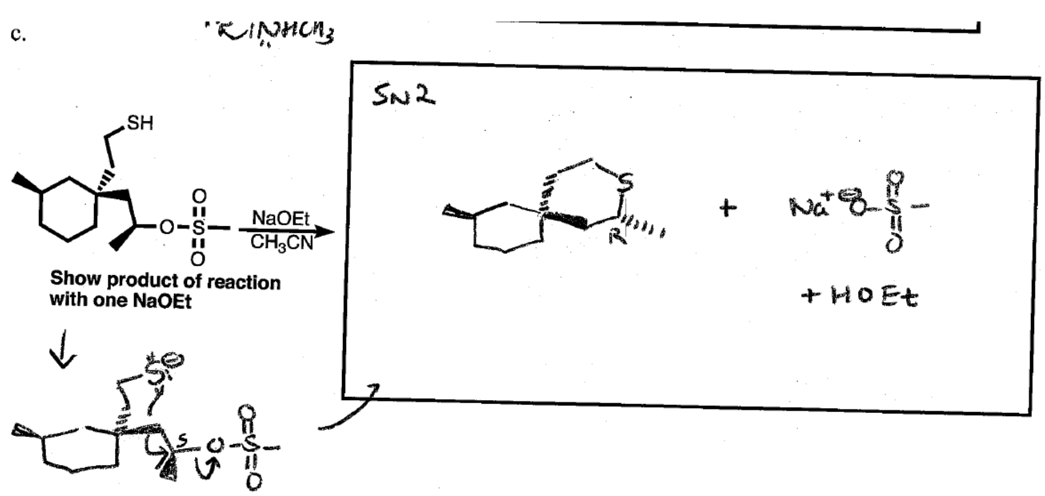
a.



b.

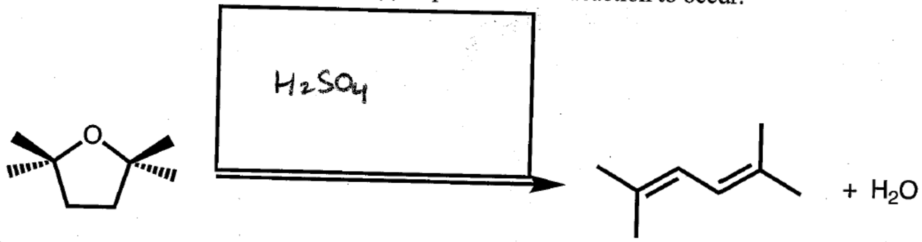
1c. (Implicit)

c.



1d. (Implicit)

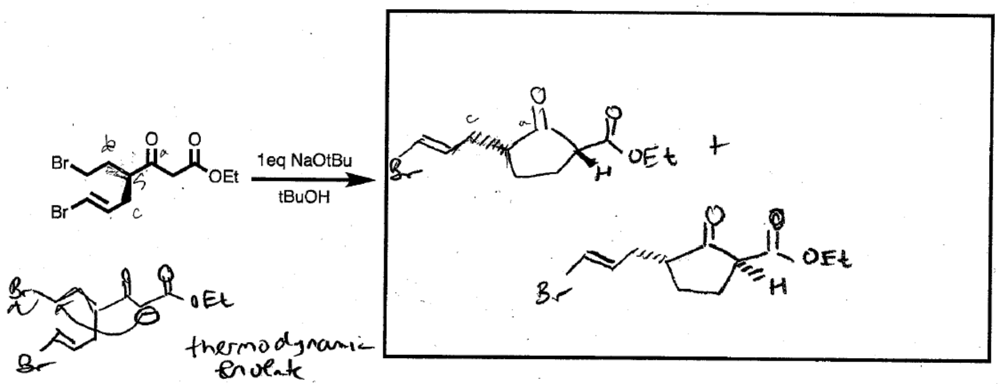
d. For this reaction, fill in the reagent(s) required for this reaction to occur.



Midterm 3

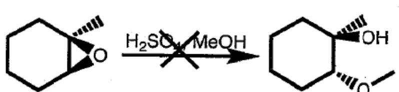
1d. (Implicit)

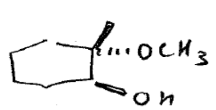
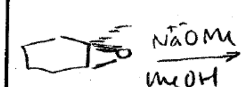
d.



3b. (Implicit)

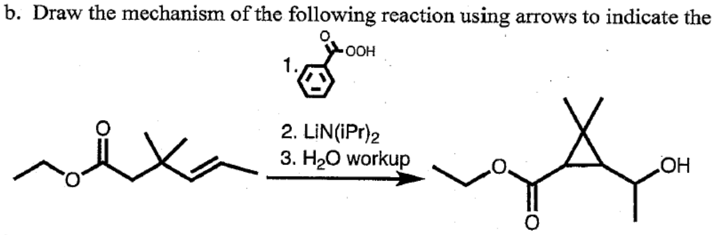
b.



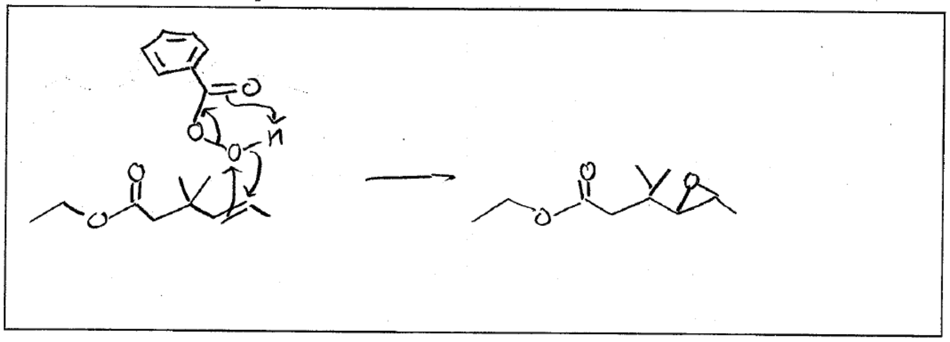
What product is actually made? (Draw structure or NR for no reaction)	Why was desired product not formed? (Explain in 1 sentence and include drawings of any relevant structures)	How could substrate OR reaction be changed to give desired product? Draw your revised reaction.
	Protonated epoxide will operat more substituted carbon because this carbon has more partial positive charge.	Use $^-OCH_3$ to give SN2 mechanism. 

4b2. (Implicit)

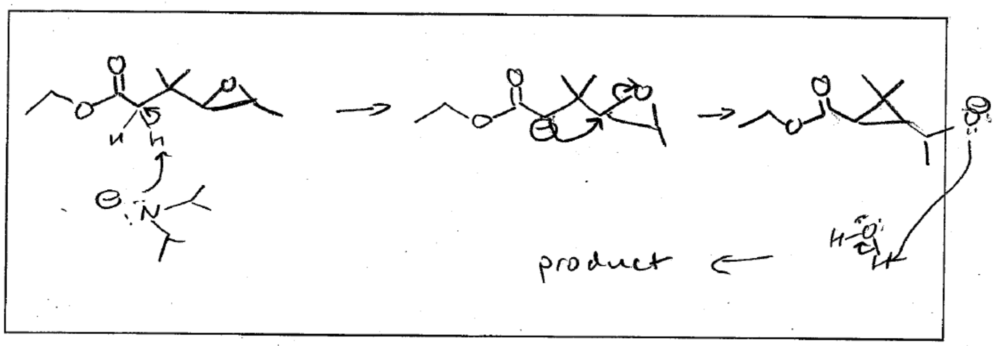
b. Draw the mechanism of the following reaction using arrows to indicate the flow of electrons.



i. Reaction indicated in step 1.



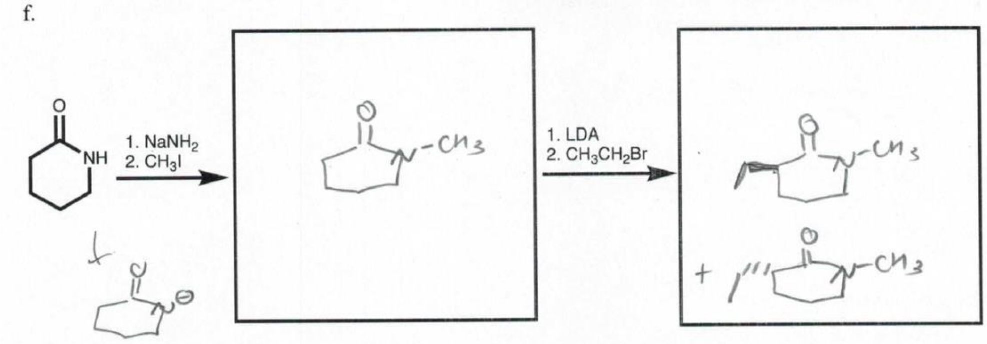
ii. Reaction indicated in steps 2 and 3.



Final

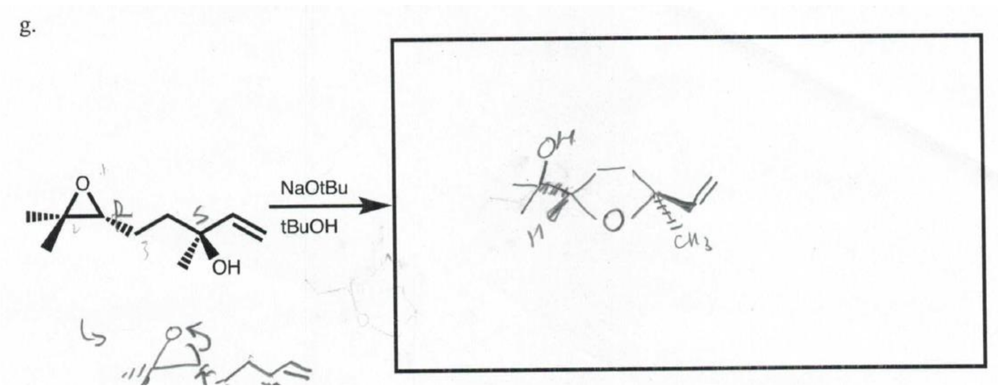
1f. (Implicit)

f.



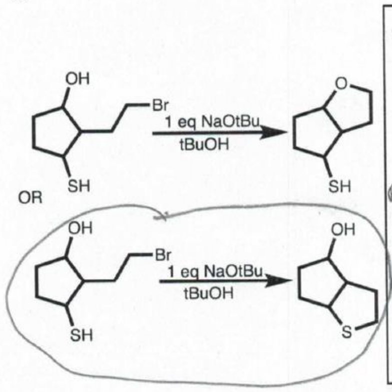
1g. (Implicit)

g.



2a. (Implicit)

a.

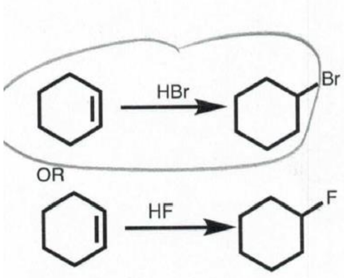


The 1st step is deprotonation. Only one eq of base is added & S-H is more acidic than O-H & therefore, all base will be used up deprotonating R-SH, which will then do an SN2 on alkyl bromide. R-SH is more acidic than R-OH because S is one row down in periodic table & larger size stabilizes negative charge in conjugate base.

Page 4 of 19

2d. (Implicit)

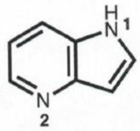
d.



Rate of reaction depends on strength of electrophile. H-Br is much more acidic than H F & therefore, a stronger electrophile. Transition state includes a partially broken H-X bond & more acidic H-X will be more stable in T.S, & rxn will go faster.

4a2. (Explicit)

4. (43 points) Consider the molecules below.
a.



	Hybridization	Orbital of lone pair
N1	sp2	p
N2	sp2	sp2

i. Fill in the table next to the molecule.
ii. Is this molecule aromatic? Explain your answer.

yes. 2n+2 e- n = 4 10π e-
cyclic, planar, fully conjugated

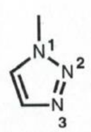
iii. Which nitrogen atom in this molecule is **most** basic? Explain your answer.

N2 is most basic, lone pair is sp2 & not involved in resonance making it less stable & more basic

N1 lone pair is part of aromatic system & is less basic

4c2. (Explicit)

c. Consider the molecule below.



	Hybridization	Orbital of lone pair
N1	sp2	p
N2	sp2	sp2
N3	sp2	sp2

i. Fill in the table next to the molecule.
ii. Is this molecule aromatic? Explain your answer.

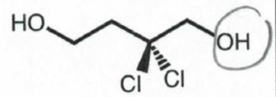
yes 6π e-, fully conjugated, cyclic, planar

iii. Which nitrogen atom in this molecule is **least** basic? Explain your answer.

N1 is least basic because it is part of aromatic system. Protonating at N1 would break aromaticity. Protonating at N2 & N3 would not effect aromaticity

4d. (Explicit)

d. Which is the most acidic proton in the molecule below? Explain your selection.

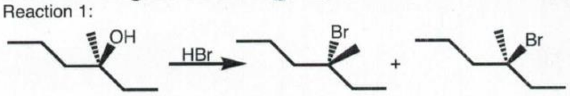


This OH is closer to the electron-withdrawing Cl groups. These stabilize conjugate base by induction.

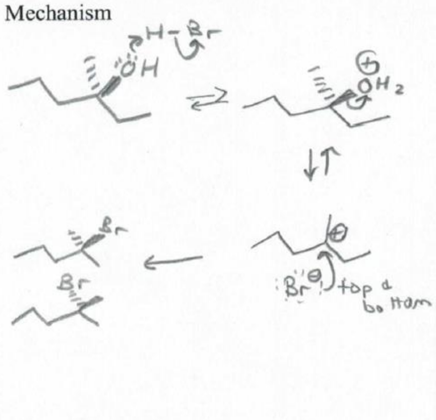
8a. (Implicit)

8. (28 points) The two reactions below produce different products. For each reaction, in box on the left, draw a mechanism of the reaction using arrows to show the flow of electrons. In the box on the right, draw a reaction coordinate diagram. Use the protonated alcohol as your starting material in the reaction coordinate diagram. In the diagram, label ΔG° for the reaction and ΔG‡ for the rate determining step.

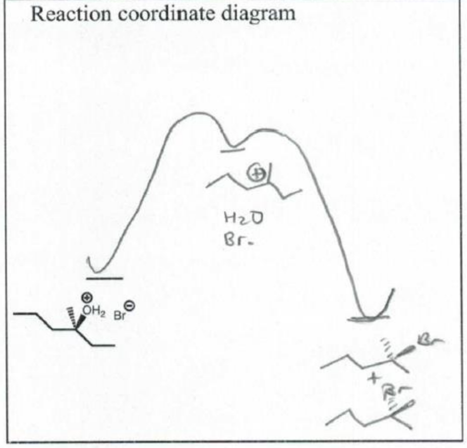
Reaction 1:



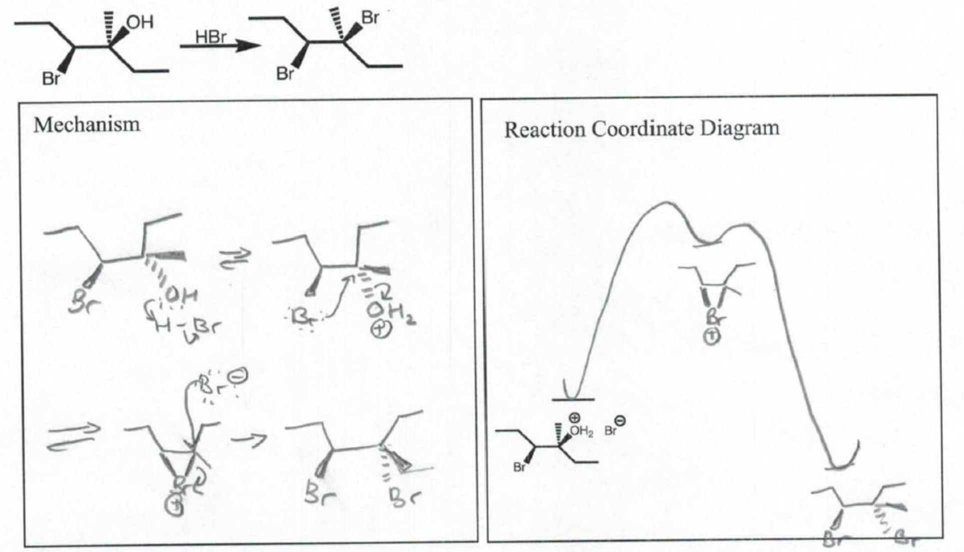
Mechanism



Reaction coordinate diagram

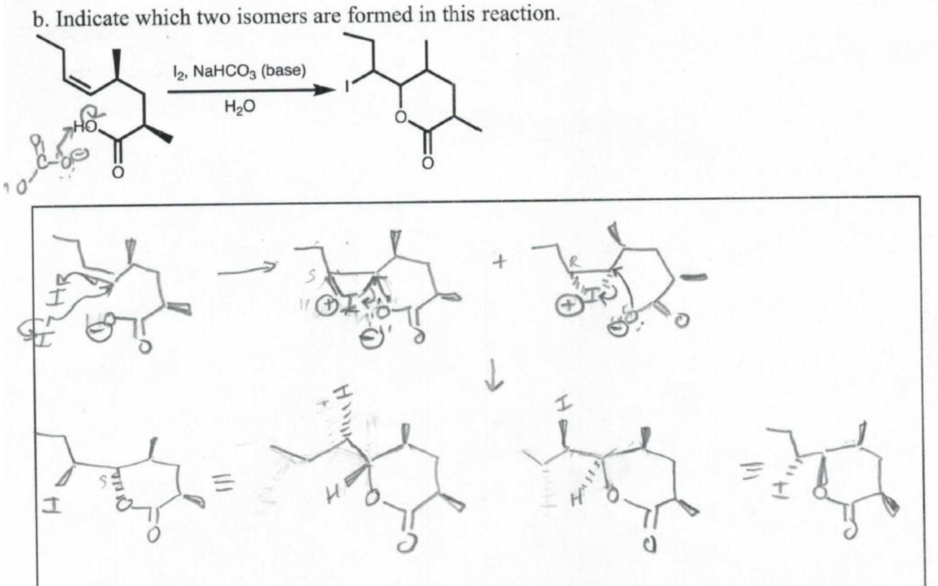


Reaction 2:



10b. (Implicit)

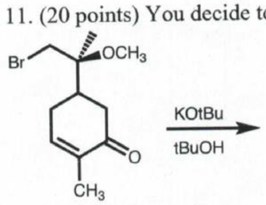
b. Indicate which two isomers are formed in this reaction.



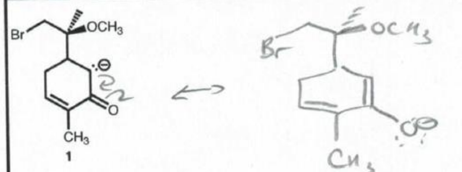
Page 16 of 19

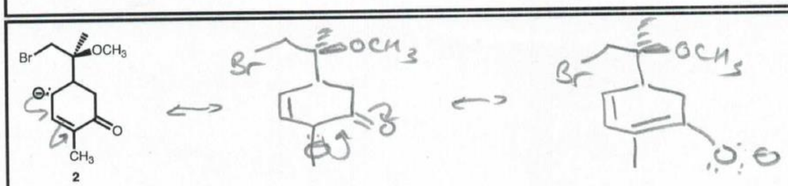
11a. (Implicit)

11. (20 points) You decide to perform the following reaction.



a. You expect one of two anions to form. Draw the resonance structures of each in the boxes below.



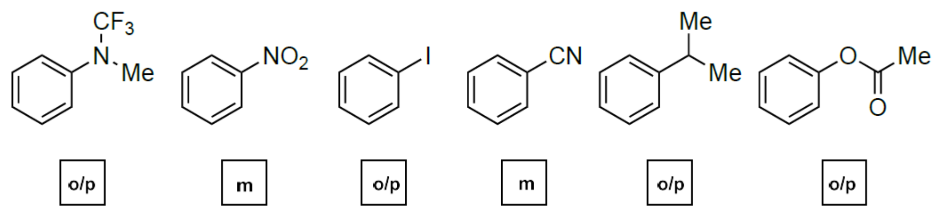


2.13. Relevant exam questions (S_EAr lesson)

Midterm 2

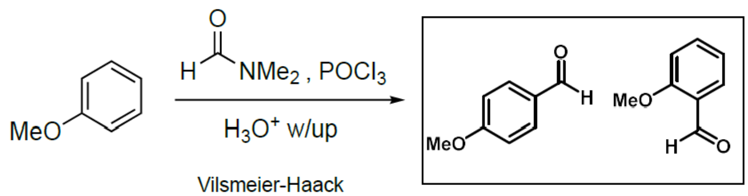
1a.

(A) (6 points) Indicate which of the following substituents would direct ortho/para (**o/p**) or meta (**m**) in the boxes provided.



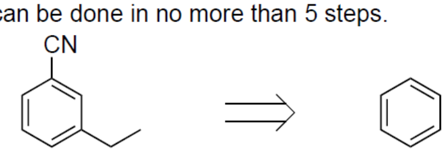
2a. 100% for the entire class

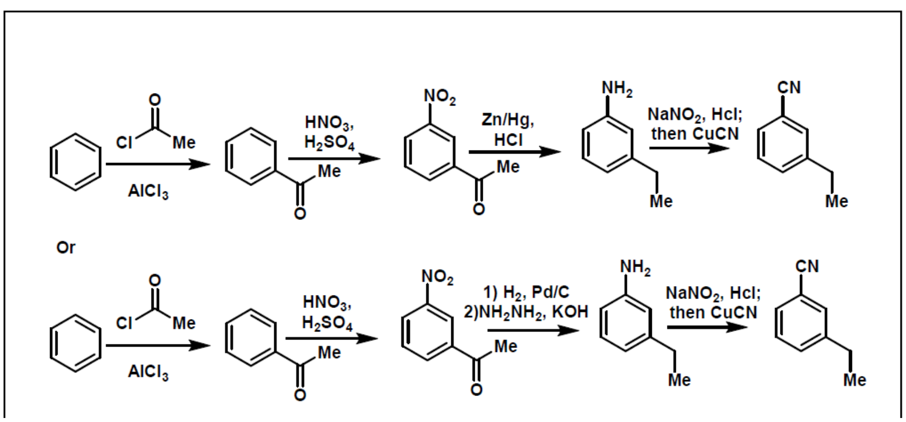
(A)



6.

Question 6: (16 points) Propose a reasonable synthesis of 3-ethylbenzonitrile starting from benzene and any inorganic or organic reagent with two or less carbon atoms. The shortest synthesis can be done in no more than 5 steps.

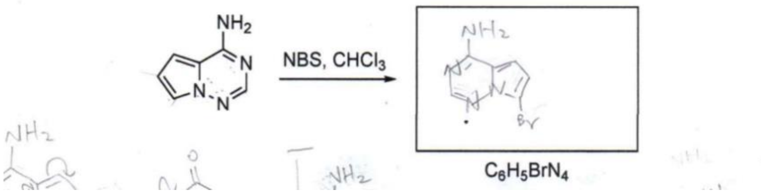




Final

5d.

(D) S_EAr bromination of pyrrolo[2,1-*f*][1,2,4]triazin-4-amine with NBS yields one product. Predict the structure and justify your answer with a clear drawing of resonance structures.



APPENDICES – Chapter 3

3.1. Sampling procedures for the BURET study

All of the students enrolled in the three target CUREs were invited to be part of this study in person by one of the researchers. A few weeks before their corresponding poster session, students were asked to respond to the two reflective prompts, and answers were collected from all students who agreed to participate. This included 135 members of the chemistry CURE, for a response rate of 58%, 11 students in the pre-service teacher CURE, for a response rate of 65%, and 9 participants from the transfer student CURE, for a response rate of 90%. All of the consenting students in the pre-service teacher and transfer student CUREs were then interviewed at the final poster session for those courses. For the chemistry CURE, a subset of 35 consenting students were interviewed at the final poster session. To choose a representative sample, students were stratified by major and prior research experience. After intentionally oversampling 7 URM students, a random sample of 28 was chosen from among the remaining 128 students. From this pool of CURE participants, 6 URM students and a random sample of 9 other students were chosen for further analysis. An additional 20 students for whom we had prompt responses but not poster session interviews were also randomly selected for analysis.

Additionally, all students presenting at one of the two target URE poster sessions were invited by email to participate in this study. For the chemistry poster session, 112 students were invited to participate and 66 consented, for a response rate of 59%. For the pre-service teacher poster session, 23 of the 26 students (88%) responded affirmatively to the invitation. Responses to the reflective prompts were collected via Qualtrics a few weeks prior to the poster sessions. All consenting students who provided answers to our reflective prompts (30 from the chemistry poster session and 23 from the pre-service teacher session) were interviewed at these poster sessions, using the same protocol. From this pool of URE participants, 6 URM students and a random sample of 24 other students were chosen for further analysis. An additional 15 students for whom we had prompt responses but not poster session interviews were also randomly selected for analysis. In total, the dataset we analyzed included 80 responses to reflective prompts and 55 poster session interviews.

3.2. BURET-R coding rubric

0 - Absent
1 - Partial concept or Irrelevant/Incorrect
2 - Isolated concept
3 - Partial link
4 - Basic link
5 - Complex link

CONTEXT Indicator 1: Place the research questions or goals of their laboratory and/or project in the context of the larger field.	
Score	Definition
0	- Does not explain goals of experiment or project
1	- Partial or unclear description of experiment and/or project goals
2	- States goal of experiment OR - States goal of project
3	- Clearly states goal of experiment AND - States goal of project (vagueness allowed - pretty low bar)

4	- Partial Link AND (- Explains how expt advances larger project OR - Explicit link of project to broader significance (scientific or societal))
5	- Partial Link AND (- Explains how expt advances larger project or the portion of the project they are working on) AND - Explicit link of project to broader significance (scientific or societal))
PREVIOUS WORK - CHECKBOXES (applicable = 1, not applicable = 0)	
PW	- Makes any reference to previous work relevant to current research, potentially vague
PW+	- Clearly describes results of previous work, with implied connection to current project OR - Clearly describes how previous work influenced current work

DESIGN CHOICES (RATIONALE)

Indicator 2: Justify their experimental design as appropriate for their research question and scientific content of their project.

Score	Definition
0	- Coder cannot identify any design choice discussed

1	- Partial or unclear description of one design choice OR - Design choice has obvious flaws
2	- Clear description of one design choice, but rationale is poor or absent
3	- Clear description of one design choice AND - Gives reasonable (sounding) rationale but vague, implied, or invokes little to no content knowledge
4	- Clear description of one design choice AND - Gives explicit rationale for choice of instrument or experiment that integrates (some) domain-specific content knowledge
5	- Above, but multiple distinct reasons for design choice are discussed
LIMITATIONS AND COMPARISONS - CHECKBOXES (applicable = 1, not applicable = 0)	
Limit	- Mentions a limitation
Limit+	- Elaborates on limitation, demonstrating content knowledge
Altern	- Mentions an alternative to design choice used
Comp	- Makes an explicit comparison to an alternative design choice

VARIABLES - WHAT IS BEING MEASURED, MANIPULATED, OR COMPARED? Indicator 2: Justify their experimental design as appropriate for their research question and scientific content of their project.	
Score	Definition

0	- Does not indicate what type of data is being collected or discuss any other relevant variables
1	- Isolated Concept but vague or implied (unclear what they are actually measuring, manipulating, comparing) OR - Basic instrument verification on a standard
2	- Clearly identifies what is being measured (raw OR analyzed)
	OR - Clearly identifies one or more variables being manipulated, compared, or held constant
3	- Isolated Concept, AND (- Provides basic rationale for choice of variables and/or range being investigated OR - Gives details on how or to what extent the variables are manipulated) OR - Basic link (4)2, but rationale or predictions are vague or questionable
4	- Clearly identifies what is being measured (raw OR analyzed) AND - Clearly states one or more variables being manipulated, controlled or compared AND (- Provides rationale (clear, but slightly generic okay) for why manipulated variables would affect measurements/output OR - Provides reasonable prediction of how manipulated variables will affect output)
5	- Basic Link AND - Rationale and/or predictions are strong and integrate content knowledge

DATA MANIPULATION Indicator 3: Analyze and interpret data in order to construct explanations and models that are relevant to their research question.	
Score	Definition
0	- Does not describe any analysis of raw data
1	- States that no data analysis was performed OR - States that results are inconclusive with no elaboration
2	- States a procedure for analyzing or manipulating data with no elaboration
3	- Links raw data to analyzed results, but discussion of data or analysis method/procedure is vague
4	- Clearly links raw data to analyzed results, including (clear) description of the analysis process
5	- Above, plus discusses at least one assumption or consequential decision made during analysis
ERROR ANALYSIS - CHECKBOXES (applicable = 1, not applicable = 0)	
Stats	- States method of statistical analysis
Stats+	- Elaborates on statistical analysis
Human	- States that human error is an issue, or describes error source that is essentially human/user error
Error	- States a reasonable source of error that doesn't reduce to human error
Skep	- Expresses skepticism of *plausible* results

DATA INTERPRETATION

Indicator 3: Analyze and interpret data in order to construct explanations and models that are relevant to their research question.

Score	Definition
0	- Does not describe results OR - Has not collected data yet
1	- Appears to misinterpret data/results OR - Unclear how conclusion is supported by results OR - Implies data interpretation but does not sufficiently describe
2	- Summarizes results without interpretation OR - Pre-packaged conclusion OR - States an interpretation with no connection to data
3	- Summarizes results and links to content knowledge or compares to expectations, but vague or minimal insights
4	- Gives plausible explanation for results (or compares results to expectations in a way) that integrates clear content knowledge
5	- Above, but integrates extensive content knowledge OR - Discusses alternate interpretations

NEXT STEPS Indicator 4: Generate hypotheses and plan future experiments in response to their analysis and interpretation of data and research question.	
Score	Definition
0	- Did not respond to prompt
1	- Completely different goals for future work with no/minimal relationship to current work
2	- Simple quantitative extension, modification, or new experiment with no or poor rationale OR - “Continue with the plan” OR - Repeat experiment with simple issue fixed OR - Very different direction/goals for future work, but some reasonable rationale
3	- Simple quantitative extension with good rationale OR - Modification or new experiment with credible but vague rationale OR - Repeat experiment after difficult-to-predict issue fixed (troubleshooting), link to content knowledge is vague or absent
4	- Modification or new experiment with clear rationale that integrates content knowledge OR - Next steps conditional on a particular outcome of current experiment (including some content knowledge-based rationale) OR

	- Troubleshoot with content knowledge-based rationale for change
5	- Multiple basic links (4), at least one of which is not a borderline Partial Link (3) OR - Conditional next steps for multiple possible outcomes (including content knowledge-based rationale)
NEXT STEPS IMPACT - CHECKBOXES (applicable = 1, not applicable = 0)	
Imp	- Mentions or vaguely states in next steps section how proposed next steps relate to scientific or societal impact
Imp+	- Explicit discussion in next steps section of how proposed next steps would have a scientific or societal impact

INTEGRATION OF (ADDITIONAL) CONTENT KNOWLEDGE	
Score	Definition
0	- Response does not integrate any scientific content knowledge beyond what is necessary to describe the experiment
2	- Below, but weaker
3	- Exhibits content knowledge beyond what is required to describe experiment
5	- Exhibits *extensive* content knowledge beyond what is required to describe experiment
CHECKBOX (applicable = 1, not applicable = 0)	
Inc	- Makes one or more factually incorrect statements, or makes clear errors in data analysis or interpretation

3.3. BURET-P coding rubric

0 - Absent
1 - Partial concept or Irrelevant/Incorrect
2 - Isolated concept
3 - Partial link
4 - Basic link
5 - Complex link

CONTEXT - SCIENTIFIC IMPORTANCE Indicator 1: Place the research questions or goals of their laboratory and/or project in the context of the larger field.	
Score	Definition
0	- Does not explain goals project
1	- Partial or unclear description of project goals
2	- Clearly states goal of project OR - States a very limited scientific application of their work
3	- States a general area of science that their work contributes to OR - Vague or implied version of below Should some content knowledge be required here?
4	- Discusses how future projects (by other labs) might be affected by current project OR - Suggests new research paths or projects that could be based on this work OR

	- Provides sufficient background for reader to understand current state of field
5	- Basic Link (4) with two out of three of the criteria present - Two different scientific contexts explained for the project - both at Basic Links

CONTEXT - SOCIETAL IMPACT OR APPLICATION

Indicator 1: Place the research questions or goals of their laboratory and/or project in the context of the larger field.

Score	Definition
0	- Does not explain goals of project
1	- Only discusses “personal” goals, but does not mention a societally-relevant topic
2	- Collecting data with no further connection to societal importance OR - Reader can infer societal importance or application of the data collected (i.e. mentions a societally-relevant topic (like semiconductor or cancer))
3	- Implies societal importance OR - Vague statement about the possible benefits or use of results
4	- Explicitly connects project to specific societal need OR - Explicit statement about the possible benefits or use of results Accurate content knowledge and coherent argument should be present. However, exact mechanism of connection does not need to be stated.

5	- Explicit comparison between current project goals and existing solutions to those problems. Exact mechanism of connection does need to be stated. OR - Explicit and specific statement about the possible benefits or use of results, including statement of existing societal issue or need

DESIGN CHOICES (RATIONALE) Indicator 2: Justify their experimental design as appropriate for their research question and scientific content of their project.	
Score	Definition
0	- Coder cannot identify any design choice discussed
1	- Partial or unclear description of design choice
2	- Clear description of one design choice, but rationale is poor or absent
3	- Clear description of one design choice AND - Gives reasonable (sounding) rationale but vague, implied, or invokes little to no content knowledge
4	- Clear description of one design choice AND - Gives explicit rationale that integrates domain-specific content knowledge
5	- Above, but multiple distinct reasons for design choice are discussed AND - Strong evidence of extensive content knowledge that supports their choices
RATIONALE CHECKBOXES (applicable = 1, not applicable = 0)	

Resour	- States that the design choice was made because that's the resources they had available
Flaw	- Design choice has major flaws or is based on poor/questionable assumptions

DESIGN CHOICES (LIMITATIONS)

Indicator 2: Justify their experimental design as appropriate for their research question and scientific content of their project.

Score	Definition
0	- Does not discuss any limitations of design choice
1	- Vague statement of a very generic limitation or logistical issue OR - Vague or implied description of a thoughtful limitation - implied in the description of results
2	- Clear statement of a very generic limitation or logistical issue OR - Vague description of a thoughtful limitation
3	- One or more thoughtful limitations mentioned, but content knowledge only implied
4	- Gives at least one explicit limitation that integrates domain-specific content knowledge
5	- Above AND (- Discusses how limitations affect conclusions OR - Discusses how limitation was addressed, minimized, avoided, etc.)

DESIGN CHOICES (COMPARISONS)	
Indicator 2: Justify their experimental design as appropriate for their research question and scientific content of their project.	
Score	Definition
0	- Does not mention any alternative design choices
1	- Mentions the fact that there are alternatives, but doesn't mention what these are.
2	- Mentions specific alternative, but no comparison OR - Compares to alternative because alternative is, in their opinion "not possible"
3	- Compares design choice to an alternative, but is somewhat vague or implied OR - Compares to alternative because alternative is, in their opinion "not possible", plus *why* it wouldn't be possible
4	- Comparison to an alternative design choice on a single facet with a clear statement of difference or advantage or reason to use one or the other
5	- Clear comparison to an alternative design choice on more than one facet - Suggested Addition: 3 or more Basic Links (4)

DATA INTERPRETATION

Indicator 3: Analyze and interpret data in order to construct explanations and models that are relevant to their research question.	
Score	Definition
0	- Does not describe results OR - Has not collected data yet
1	- Unclear how conclusion is supported by results OR - Implies data interpretation but does not sufficiently describe
2	- Summarizes results without interpretation OR - Pre-packaged conclusion
3	- Summarizes results and links to content knowledge or compares to expectations, but vague or minimal insights
4	- Gives plausible explanation for results (or compares results to expectations in a way) that integrates clear content knowledge
5	- Above AND (- Integrates extensive content knowledge OR - Discusses alternate interpretations OR - Discusses several significant/large sets of results that build on each other, leading to extensive discussion of data interpretation)
DATA INTERPRETATION CHECKBOXES (applicable = 1, not applicable = 0)	
MisInt	- Appears to misinterpret data/results

CONFIDENCE/ERROR ANALYSIS

Indicator 3: Analyze and interpret data in order to construct explanations and models that are relevant to their research question.

Score	Definition
0	- Does not identify any potential sources of error
1	- States that the experiment (or a large part of it) “didn’t work” without any elaboration as to why OR - Describes confidence in the ability of methods to answer the RQ
2	- Identifies a clear “error” in what was done OR - Vague reference to limitation of method/technique when discussing confidence in results OR - vague ‘doubts’ about data
3	- Identifies potential sources of error that are less “obvious” OR - Clear reference to limitation of method/technique when discussing confidence in results
4	- Clearly identifies potential reasonable source(s) of error AND - Mentions how these connect to at least *one* of the following: 1. Research questions 2. Experimental design (current or future) 3. Their conclusions

5	- Clearly identifies multiple distinct potential reasonable source(s) of error at the level of a Basic Link
ERROR CHECKBOXES (applicable = 1, not applicable = 0)	
Stats	- States method of statistical analysis
Stats+	- Elaborates on statistical analysis
Human	- States that human error is an issue, or describes error source that is essentially human/user error
Skep	- Expresses skepticism of *plausible* results

NEXT STEPS

Indicator 4: Generate hypotheses and plan future experiments in response to their analysis and interpretation of data and research question.

Score	Definition
0	- Does not discuss any potential future work
1	- Completely different goals for future work with no/minimal relationship to current work OR - Implies that they will "continue with the plan" but does not sufficiently describe
2	- Simple quantitative extension, modification, or new experiment with no or poor rationale OR - "Continue with the plan" OR - Repeat experiment with simple issue fixed

3	- Simple quantitative extension with good rationale OR - Modification or new experiment with credible but vague rationale OR - Repeat experiment after difficult-to-predict issue fixed (troubleshooting), link to content knowledge is vague or absent
4	- Modification, troubleshooting, or new experiment with clear rationale that integrates content knowledge
5	- Multiple Basic Links (4), at least one of which is not a borderline Partial Link (3) OR (- Basic Link AND - Explicitly links new choices to the *results* of current work)
NEXT STEPS - CHECKBOXES (applicable = 1, not applicable = 0)	
Imp	- Mentions or vaguely states how proposed next steps relate to scientific or societal impact
Imp+	- Explicit discussion of how proposed next steps would have a scientific or societal impact
Hyp	- States prediction about outcome of future work
Hyp+	- Prediction about outcomes of future work that is well-supported

PREVIOUS WORK Indicator 1: Place the research questions or goals of their laboratory and/or project in the context of the larger field.	
Score	Definition

0	- Does not mention any prior work
1	- Vague references to “other studies” without any specific designs/results or clear specification of how this informs part of project
2	- Clear reference to previous work, but no stated connection to current work OR - Vague reference to previous work with connection to current project
3	- Clear description of previous design or results AND - Vague connection to/influence on current work OR - Vague comparison b/w old and new design or results
4	- Summarizes previous work (specific design or results) and explicitly states how it connects to/influenced current work OR compares to current results
5	- Above AND (- Explanation of how current work is different or novel OR - Attempts to interpret sim/diff between current and previous results)
PREVIOUS WORK CHECKBOXES (applicable = 1, not applicable = 0)	
Own	- Refers to previous work, but only from members of own lab

INTEGRATION OF (ADDITIONAL) CONTENT KNOWLEDGE

Score	Definition
0	- Response does not integrate any scientific content knowledge beyond what is necessary to describe the project
2	- Below, but weaker
3	- Exhibits scientific content knowledge beyond what is required to describe project
4	- Exhibits *extensive* scientific content knowledge beyond what is required to describe project
5	- Multiple instances of above
CHECKBOX (applicable = 1, not applicable = 0)	
Inc	- Makes one or more factually incorrect statements