RESEARCH ARTICLE

Active Learning Module for Protein Structure Analysis Using Novel Enzymes

Jessica I. Kelz[1,§], Gemma R. Takahashi[2,§], Fatemeh Safizadeh[1], Vesta Farahmand[1], Marquise G. Crosby[2], Jose L. Uribe[1], Suhn H. Kim[1], Marc A. Sprague-Piercy[2], Elizabeth M. Diessner[1], Brenna Norton-Baker[1], Steven M. Damo[3], Rachel W. Martin[1,2,*], Pavan Kadandale[2,*]

[1]Department of Chemistry, University of California, Irvine, CA, USA

[2]Department of Molecular Biology and Biochemistry, University of California, Irvine, CA, USA

[3]Department of Life and Physical Sciences, Fisk University, Nashville, TN, USA

"§" equal contribution

"*" corresponding authors

Received: 15 August 2021

Accepted: 11 January 2022

Published: 0 Month 2022

10.35459/tbp.2021.000209

ABSTRACT

A major challenge for science educators is teaching foundational concepts while introducing their students to current research. Here we describe an active learning module developed to teach protein structure fundamentals while supporting ongoing research in enzyme discovery. It can be readily implemented in both entry-level and upper-division college biochemistry or biophysics courses. Preactivity lectures introduced fundamentals of protein secondary structure and provided context for the research projects, and a homework assignment familiarized students

with 3-dimensional visualization of biomolecules with UCSF Chimera, a free protein structure viewer. The activity is an online survey in which students compare structure elements in papain, a well-characterized cysteine protease from *Carica papaya*, to novel homologous proteases identified from the genomes of an extremophilic microbe (*Halanaerobium praevalens*) and 2 carnivorous plants (*Drosera capensis* and *Cephalotus follicularis*). Students were then able to identify, with varying levels of accuracy, a number of structural features in cysteine proteases that could expedite the identification of novel or biochemically interesting cysteine proteases for experimental validation in a university laboratory. Student responses to a postactivity survey were largely positive and constructive, describing points in the activity that could be improved and indicating that the activity was an engaging way to learn about protein structure.
Keywords: protein structure prediction, enzyme, biochemistry, active learning, undergraduate

I. INTRODUCTION

The Protein Data Bank (1) contains more than 174 000 structures of biomolecules as of early 2021, and familiarity with protein structures is necessary for understanding the literature in many subfields of biology. Experimentally, protein structures are generally solved by X-ray crystallography, nuclear magnetic resonance spectroscopy, cryogenic electron microscopy, or, for complex molecular assemblies, a combination thereof. Advances in experimental methodology, including automated data collection at synchrotron beamlines, improved nuclear magnetic resonance instrumentation, and the “resolution revolution” in cryogenic electron microscopy have greatly accelerated the pace of protein structure determination studies. As this methodology becomes easier to use, familiarity with protein structures has become an essential competency needed for many types of biological research. Being able to visualize the relevant molecular structures improves mechanistic understanding of enzyme activity, protein–protein

interactions, and regulation of biological processes such as transcription and translation. Connecting protein structure to function has been identified by the American Society for Biochemistry and Molecular Biology as 1 of 5 foundational concepts in molecular biology education, and learning how to relate the primary sequence to 3-dimensional (3D) structure is a prerequisite for the associated learning goals (2).

Learning to interpret protein structures is therefore one of the fundamental tasks of a student in an introductory biochemistry course. This topic is traditionally considered difficult, and analysis of semantic distance between fields shows that molecular biology and biochemistry are culturally isolated from other disciplines (3). Therefore, a large corpus of field-specific language must be learned starting in the introductory classes, even without considering the information-packed graphical symbology used to express chemical structures. Examples in textbooks and lectures, not to mention the current literature, interchangeably switch between different representations of the same molecules depending on the features being emphasized. Representations in which all atoms are shown are generally eschewed because the distracting level of atomic detail obscures the overall fold and key structural motifs and makes it difficult to locate functional residues without prior knowledge. Space-filling models are useful for building intuition about molecular shape and, with appropriate color coding, surface properties such as charge and hydrophobicity, but they do not allow visualization of the protein interior.

Ribbon or licorice diagrams that omit side chains and individual atoms and represent α-helical and β-strand secondary structure elements as coiled helices or flat ribbons, respectively, highlight the 3D organization of the protein. These diagrams were first systematized by Jane Richardson in 1981 (4), although similar drawings had already appeared in individual structural biology papers. Although every introductory biochemistry textbook has a concise explanation of

these diagrams, we recommend Richardson's original review to students who are interested in structural biology: various structural motifs are clearly explained, numerous instructive examples of structural motifs are presented, and the beautiful hand-drawn diagrams highlight the human effort that went into developing this highly efficient representation scheme. Computer programs for automating the production of ribbon diagrams soon followed (5, 6), and modern Protein Data Bank (PDB) structure viewers, such as UCSF Chimera (7), PyMOL, version 1.8 (Schrödinger LLC, New York, NY), and Visual Molecular Dynamics (8), use these representations as one of the standard settings. Several such programs are available online for free and are relatively easy to install and use. Here we take advantage of these tools to have students apply their recently gained knowledge about protein structure to an enzyme discovery project with the use of structures predicted from genomic data.

This activity is linked to an ongoing project in the lab of RWM, where a major research goal is the discovery of novel enzymes from genome and transcriptome data, in particular from carnivorous plants. These plants have adapted to grow in nutrient-poor environments by obtaining much of their nitrogen from protein in insect prey (9). Carnivorous plants are expected to have a variety of proteases with different activities, because they rely on these enzymes for digestion as well as the more typical functions of plant proteases: cellular housekeeping, defense against insects and pathogens, and hydrolysis of seed storage proteins. In the Venus flytrap (*Dionaea muscipula*), expression of at least 1 digestive protease is upregulated in response to prey stimuli (10). As expected, the genomes of the Cape sundew (*Drosera capensis*) (11) and the Albany pitcher plant (*Cephalotus follicularis*) (12) have yielded many new proteases—so many, that the main problem is choosing appropriate targets for experimental investigation. In general, determination of experimental structures is a bottleneck for enzyme discovery from nucleic acid

sequencing data. Advances in sequencing methodology have outstripped even the rapid pace of development in structural biology methods, in part because of the difficulties inherent in sample preparation. Preparing protein samples of sufficient quantity and purity for structural studies is time consuming and expensive and requires extensive training and experience, as does interpretation of the data. Performing these experiments is impractical for every putative enzyme discovered from a genome or transcriptome. Therefore, we use structural models derived from sequence data with protein structure prediction tools such as Rosetta (13, 14) and I-TASSER (15). Although the predicted structures do not capture every detail, particularly when considering side chain conformations, we find that they are highly reliable for predicting the overall folds of enzymes belonging to well-known structural classes, including the cysteine proteases used in this activity. This capability was illustrated by the crystal structure of a cysteine protease from *D. muscipula* (16), which was solved after we predicted its structure (17). Our predicted structure has excellent overall agreement with the experimental one and captures all of the functionally important features of the active site. Results such as this, as well as ongoing validation efforts such as the CASP competition (18), provide evidence that structures predicted in this manner are sufficient to verify functional folds and active sites for well-known enzyme classes. With recent machine learning–based advances in protein structure prediction such as AlphaFold (19) and RoseTTAFold (20), it is now possible to obtain large numbers of predicted structures for members of an enzyme class of interest, such that the activity can be updated frequently or tailored to fit the theme of a particular class.

Predicting structures en masse for enzymes discovered from genomic data provides a foundation for predicting which proteins will have functional differences from well-characterized members of the same enzyme class; however, examination of the structures and prediction of

functionality is not easily automated. Some features, such as extra domains, are apparent from the sequence alone and could be detected with standard software tools. Others are more subtle and require examination by a human with some training in protein structure analysis. For instance, even relatively small occluding loops can dramatically alter substrate specificity by partially blocking the active site cleft, and these cannot necessarily be identified in sequence space because they interact with the active site cleft in 3 dimensions. Fortunately, given an appropriate reference protein, undergraduate biochemistry students can learn to identify such features relatively quickly in the context of a class activity. Here we describe such an active learning module for students in an undergraduate biochemistry class. Students received training in protein sequence and structure analysis and then worked individually to identify similarities and differences between papain, a well-characterized plant cysteine protease, and a novel protein from either *D. capensis*, *C. follicularis*, or the extremophilic microbe *Halanaerobium praevalens* (21).

II. SCIENTIFIC AND PEDAGOGICAL BACKGROUND

A major challenge in teaching protein structure interpretation is that the connection between the intermolecular forces holding proteins together and the 3D structures that result is abstract. Furthermore, many students enter introductory biochemistry with limited 3D visualization skills, such that practicing a task that requires manipulating protein structures in a virtual 3D environment is helpful. The examples presented in introductory textbooks are often selected to present a wide range of different structural motifs, which provides a good overview of existing structures but can come across as disconnected. Here we introduce a particular enzyme class, cysteine proteases (MEROPS family C1) (22), and invite students to look for relatively subtle structural differences. We selected cysteine proteases because there are a large number of

characterized structures for this enzyme class in the PDB, making structure prediction very useful for determining overall folds and relative domain orientations. At the same time, there are no shortage of newly discovered and uncharacterized cysteine proteases, because many plants have multiple paralogs of these common defensive proteins (23, 24), of which only a few have been studied in detail. *D. capensis* has 44 cysteine proteases (17), which we have previously modeled and categorized according to the classification scheme of Richau et al. (23), whereas *C. follicularis* has at least 16 (12). Our protein set consisted of the 16 novel papain-like cysteine proteases from *C. follicularis*, matched with 17 cysteine proteases from *D. capensis*, whose structural features had already been examined by the Martin group. One additional cysteine protease from the extremophilic microbe, *H. praevalens*, was also included to assess the robustness of this characterization method when examining proteins that are less closely related. Each student was assigned a unique protease from this set of 34, and all students used the crystal structure of papain from *Carica papaya* [UniProt ID, PAPA1_CARPA; PDB ID, 9PAP] (25) as a reference protein to compare structural features. The main objectives of this class activity were to introduce students to the basics of protein structure, to help them examine and manipulate protein structures in a virtual 3D environment, and to provide an opportunity to participate in a live enzyme discovery research project.

Our active learning module was motivated by the success of Course-based Undergraduate Research Experiences (CUREs), which have numerous benefits for students, including making research experiences more equitably available to all students (26), increasing scientific affect (27), improving scientific skills (28), and increasing student retention (29). Furthermore, participation in CUREs early in their university experience improved the odds of students graduating with a science, technology, engineering, or mathematics degree and improved student

GPAs when they graduated (30). Shorter term gains from CUREs included improved content knowledge, increased probability of pursuing longer term, apprenticeship-based research experiences before graduation (29, 31), and abrogation of some so called "achievement gaps" for minoritized students (32). Traditionally, CUREs have been implemented either in lab courses or in the lab sections of theory courses. CURE courses often have limited enrollment and are usually available only to upper-division students. However, a variety of research-based active learning activities have recently been developed, some of which also include opportunities for students to contribute to community resources (33) or citizen science initiatives (34). A major objective of this activity is to provide an introduction to an active research project very early in the undergraduate experience. Given the numerous benefits of exposing students to research experiences, we sought to create a shorter research experience on the basis of our enzyme discovery research, embedded within a lecture course typically taken by first-year undergraduates.

Aside from the educational benefits of the class activity itself, this experience gives students an opportunity to learn about ongoing research at their university. It also helps them see their instructors as scientists, as well as teachers, and provides an opening for interested students to join a research group as early as their first year at university. Over the last few years, a total of 12 undergraduates (including 3 coauthors on this paper) have joined the authors' enzyme discovery efforts by independent study (course credit for research), summer research programs after performing various early versions of this activity, or both. We have found that this type of activity enables recruitment of students at an earlier career stage, compared with the more typical situation in which upper-division students join labs either as part of a formalized capstone course or after being exposed to research topics in more specialized classes. In the event that not every

student who is interested in performing follow-up research can be accommodated because of space or enrollment constraints, which can happen after announcing the opportunity to a large class, it is useful to have a list of other faculty who offer undergraduate research experiences. In the future, we also plan to develop a full CURE course based on this type of research, which would make it possible for more students to participate in an extended study of novel enzymes and potentially become coauthors on a publication.

As a pilot for the large class, we first performed the activity by Zoom with undergraduate students in Chem341L (Physical Chemistry Lab), an upper-division course at Fisk University, a private historically Black university in Nashville, TN (October 2020). There is precedent for sophisticated protein structure activities in upper-division biophysical courses such as this. For example, undergraduate students assigned to solve the crystal structure of a small protein from its electron density map were very successful even without knowledge of the protein sequence, modeling ambiguous residues using chemical knowledge to identify local interactions, and in some cases producing a better result than the original structure (35). Other activities have focused on the use of molecular dynamics tools to teach structure visualization, ligand interactions (36), and noncovalent interresidue interactions (37).

In this activity, graduate students taught a lesson introducing protein structure concepts in general and important structural features of proteases in particular. The lecture material focused on secondary and tertiary protein structure, with examples of types of secondary structures found in globular proteins as well as the importance of intrinsically disordered proteins. An informal and highly interactive class discussion also took place around current protease projects in the lab of RWM, including the carnivorous plant proteins in this dataset, as well as the SARS-CoV-2 main protease (M^{pro}), which served as a transition into the hands-on activity. The goal of the

activity was to help students solidify their knowledge and exercise what they learned from the lecture, using their new insight to help discover novel structural features in papain-like protein structures. Because of the small class size (9 students) and the students' relatively advanced knowledge of molecular structure, each student was able to examine multiple structures and compare notes about different protein features, including pro-sequences, granulin domains, and differing degrees of active site cohesion. Three-dimensional–printed structures of selected proteins were provided, because there is evidence that examining 3D-printed models of protein structures helps students build accurate mental models of protein structure (38).

To incorporate this module into a large lecture course, we created a shorter version that we implemented in 2 sections of a lower division biochemistry course. This class had a large enrollment (356 students in one section and 252 students in the other section) and was required for all students in several majors, including Biology, Pharmaceutical Sciences, Nursing, and Public Health. The course is taught as a one-quarter survey course of major concepts in biochemistry, including amino acid properties and protein structure and function.

In the rest of this paper, we describe the design of lecture materials and the cysteine protease survey and discuss the results of the activity and its assessment, which we hope will be useful for other biochemistry educators. The survey materials and the protein models used are provided in the Supplementary Material.

III. MATERIALS AND METHODS

A. Protein sequences and structural models

Sequence alignments were performed with Clustal Omega (39), with settings for gap open penalty = 10.0 and gap extension penalty = 0.05, hydrophilic residues = GPSNDQERK, and the BLOSUM weight matrix. For the *D. capensis* proteases, the presence and position of a

signal sequence flagging the protein for secretion was predicted by SignalP 4.1 (40, 41). An initial model was created for each complete sequence by the Robetta (13) implementation of Rosetta (14). Any residues not present in the mature protein were removed, disulfide bonds identified by homology to papain were added, and the protonation states of active site residues were fixed to their literature values. Each corrected structure was then equilibrated in explicit solvent under periodic boundary conditions in NAMD (42) by the CHARMM22 forcefield (43) with the CMAP correction (44) and the TIP3P model for water (45; after this minimization, each structure was simulated at 293K for 500 ps, with the final conformation retained for subsequent analysis. The published structure of papain (PDB ID: 9PAP) (25) was used as the initial starting model (after removal of heteroatoms and protonation by REDUCE) (46), and similarly equilibrated before use as a reference.

For the proteases from *C. follicularis* and *H. praevalens* retrieved from UniProt (47) (Supplemental Table S1), structure prediction was performed by I-TASSER (15). Signal sequences were not removed from these proteins, to leave them as a point of discussion for the class activity.

The sequence alignments, minimal quality control (e.g., removal of proteins lacking the active site residues), and molecular modeling were performed by the research team in preparation for the activity; students were provided with sequences and structural models for their proteins.

B. The cysteine protease survey

The cysteine protease survey was designed to guide students through the process of comparing a novel cysteine protease structure to that of papain in UCSF Chimera. Questions identified characteristics like various secondary structure locations, blocked active sites, and

relative lengths of N- and C-termini. The full survey can be found in the Supplementary Material.

C. Postactivity survey

After completion of the activity, students were asked to answer a questionnaire about their experience. The survey was administered in Canvas as a regular weekly activity for the class. The questions were: "1. In how many classes at UCI (prior to this one) did you have the opportunity to apply the concepts you were learning about in class to a research project? 2. Please tell us what you liked best about the project. 3. Please tell us what you liked least about the project. 4. Do you agree or disagree with the following statement: This research project helped me understand protein structure-function better. 5. Do you agree or disagree with the following statement: This research project should continue to be a part of this course. 6. How can this research project be improved?"

IV. RESULTS AND DISCUSSION

A. Preactivity training

During the class period before the protease discovery activity, a general introduction to protein structure was presented. The concepts of primary, secondary, and tertiary structures were introduced, along with a primer on interpreting ribbon diagrams. Examples are shown in Figure 1.

Before the in-class exercise, an introductory lecture on cysteine protease discovery was presented, taking approximately 20 min. This lecture was delivered by a graduate student directly involved in the research and began with a description of the motivation for discovering new cysteine proteases. Examples presented included finding highly specific proteases to cleave expression tags or break down proteins into smaller peptides for bottom-up proteomics and, on

the other hand, finding very general proteases to break down biofilms and cleave protease-resistant aggregates such as amyloid fibrils. The overall workflow of the project was summarized, emphasizing the large number of proteases discovered from the *D. capensis* genome and how molecular modeling could help narrow down the targets chosen for experimental characterization. The graduate researcher also explained how the students' responses would be used by the group: their answers regarding which proteins have features that are significantly different from papain's were aggregated and used in the manner of crowdsourcing data. Because 509 students completed the activity and there were only 34 unique proteins, each protein was subject to independent analysis from multiple participants. Although

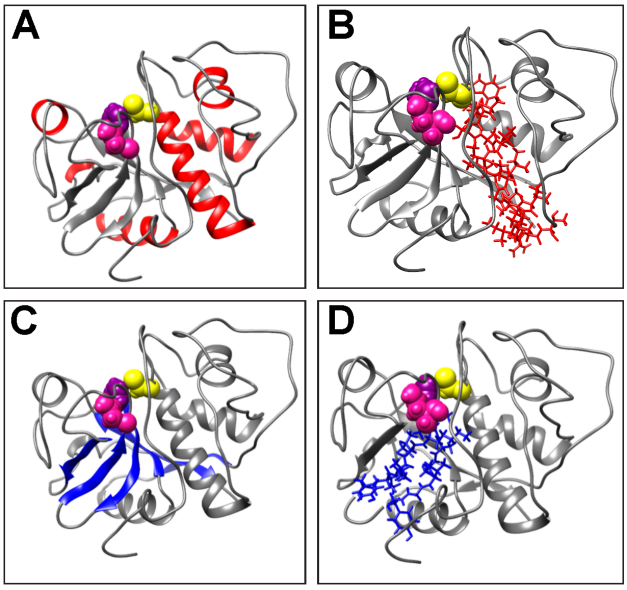

Fig 1. Papain secondary structure examples presented in presurvey lecture. (A) All α-helices (red) displayed as ribbons. (B) One α-helix (red) displayed with all atoms shown as stick models. (C) All β-strands (blue) displayed as ribbons. (D) Two β-strands (blue) displayed with all atoms shown as stick models.

students were allowed to work together in small groups, each student was randomly assigned a different protein, so it is likely that most of the observations of a given protein were independent.

This method enabled the research team to identify potentially interesting target proteins that multiple observers indicated had significant differences from the reference protein.

Finally, some examples of *D. capensis* cysteine proteases with functional features different from papain's were shown. During the initial training, it was pointed out that although the correlation between structure and function is not perfectly predictable, enzymes that are structurally very similar are likely to be functionally similar as well. Therefore, enzymes that structurally resemble papain are likely to have similar activity to this well-characterized protease, whereas enzymes with notable differences of the types described in the background lecture are more likely to provide novel functionality. In future versions of the activity, we plan to ask students specifically whether their assigned protein is a good candidate for further study and to explain their reasoning.

The example proteases are shown in Figure 2. The first, aspain, has an unusual active site configuration with an aspartic acid taking the place of the canonical asparagine and a large occluding loop partially blocking the active site, potentially modulating substrate specificity. The second, DCAP_6097, has a C-terminal granulin domain, which is common in proteases that cleave storage proteins during seed sprouting. Both contain examples of structures students may encounter when studying novel papain-like proteases. Students were also instructed in how to compare aligned sequences and locate particular amino acid residues on the protein structure. Overall, the background material took one full 50-minute class period, with a second class period devoted to the active learning activity. Students were then allowed 2 extra days to work on the survey before having to submit their responses; this arrangement provided some flexibility, but more than half of the responses were received by the end of the designated activity day. In total, students were given about 5 d to complete the activity.

B. In-class exercise

To provide practical experience comparing structurally related proteins, we assigned each student a protein model from our set of predicted structures, which they were instructed to compare to papain. Every student was given 2 PDB files to download: the reference papain structure and the predicted structure of a novel protein. An example is shown in Figure 3. The structure of papain (Fig 3A) and the model of the novel protein DCAP_4793 (Fig 3B) are very similar in overall fold, and differences are difficult to determine when examining them separately. However, overlaying them (Fig 3C) reveals some potentially functionally relevant differences. The region labeled 1 shows the difference in length of 2 β-strands and the loop connecting them: both the strands and the loop are longer in DCAP_4793 than in papain. The area labeled 2 shows a short α-helix in DCAP_4793 that is absent in papain. Both proteins have a long helix ending in the area labeled 3, but it is longer in papain than in DCAP_4793. Differences in backbone position of the long loops are also observed (e.g., in the region labeled 4), but these are considered to be a result of variable dynamics in these structural elements rather than persistent, meaningful differences. Discussion of which of these structural differences are likely to be functionally relevant was arguably the most difficult part of the exercise, and at the same time led to valuable conversations about the types of judgement calls made by structural biologists and how protein structure can serve as a starting point for hypotheses about function.

C. Detection of novel protease features

Not all protease features were interpreted in the same way; some were correctly identified by most participants, whereas others received mixed responses of varying accuracy. Students did, for example, correctly match most large α-helices to those in papain (Fig 4A,D) but often struggled to identify partially or fully blocked active sites (Fig 4C,D). Furthermore, more

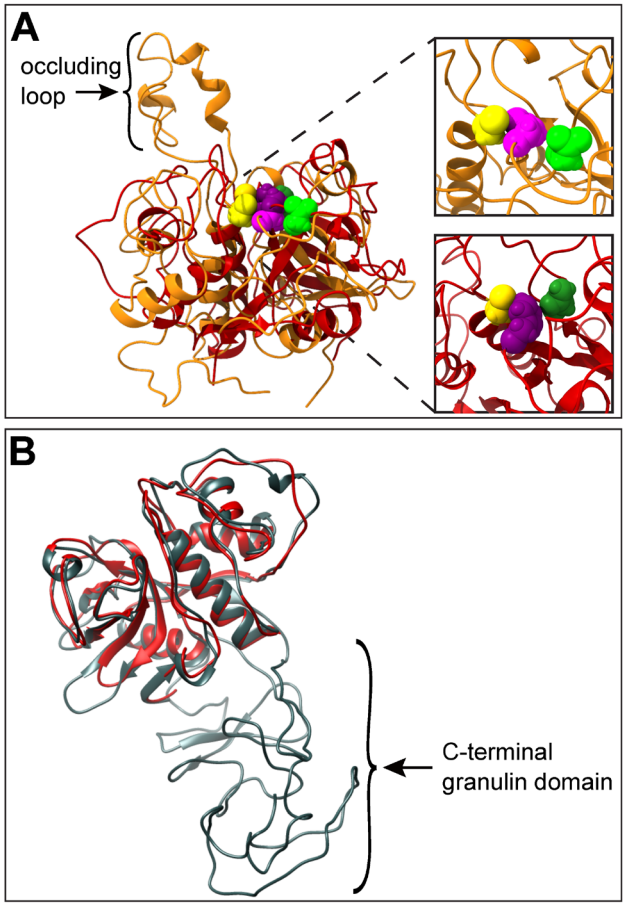

Fig 2. Example cysteine proteases, aligned with papain (red), presented to students before taking the in-class survey. (A) Aspain: DCAP_3968 (orange). Aspain's unusual active site (top inset) replaces the typical asparagine (dark green) of papain (bottom inset) with aspartic acid (lime green). Its occluding loop, which partially blocks the active site, is indicated by an arrow. Other active site residues: cysteine, gold/yellow; histidine, purple/magenta for papain and aspain, respectively. (B) DCAP_6097 (dark grey). DCAP_6097's C-terminal granulin domain, indicated by an arrow, extends well beyond the rest of the papain-aligned structure.

ambiguous structural features, like papain's small sixth α-helix (Fig 4B,D), were identified with mixed levels of success. Representative data for several of these questions are shown in Figure 4: Q3: "Is there an α-helix on your structure that lines up with the first α-helix in papain? (yes/no)"; Q4.5: "Is there an α-helix on your structure that lines up with the sixth α-helix in papain? (yes/no)"; Q13: "Do you see a feature that is partially or fully blocking the active site? (yes/no)";

Q17: “What differences does your protein have when compared to papain that were either not fully captured or not addressed at all in earlier questions? (free response)”. For DCAP_5945 (Fig 4A), Q3 and Q13 were answered accurately, because this protein does have an α-helix that matches papain’s first α-helix and does not appear to have a blocked active site. In the free response to Q17, most students also suggested the presence of DCAP_5945’s granulin domain, describing a much longer sequence and extra secondary structure elements. DCAP_5945’s Q17 response bar shows that a number of students responded with some identifying description of this granulin domain (yes they did or no they did not). These responses demonstrate what students did very well: identify large structural features that were clearly explained in presurvey presentations. Other questions, however, did not receive such consistent answers. Papain’s sixth α-helix is an example of a more ambiguous structural feature, whose presence or absence in other proteins is subject to interpretation. For example, DCAP_6547 (Fig 4B) does contain an α-helix near papain’s sixth α-helix, but a lack of overlapping residues and some variation in local

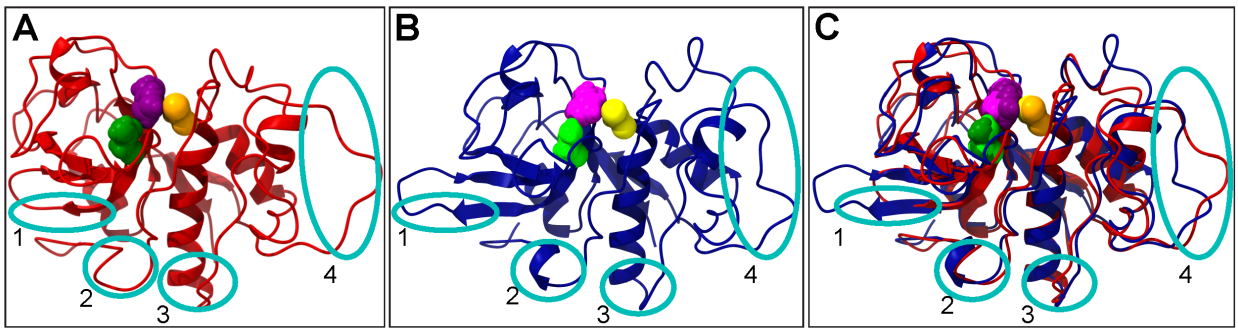

Fig 3. Comparison of the reference papain structure to a molecular model of a new protein, DCAP_4793. (A) The papain structure is shown in red. Circled areas (cyan) highlight differences in compared with DCAP_4793. (B) The molecular model for DCAP_4793, generated with Rosetta, is shown in blue. (C) An overlay of the 2 proteins in panels A and B highlights similarities and differences described in the main text. The active site residues in both proteins are shown as space-filling models with color codes as follows: cysteine, gold/yellow; histidine. purple/magenta; asparagine, dark/lime green for papain and DCAP_4793, respectively.

torsion angles make it difficult to judge whether these are truly aligned; in this case, both "yes" and "no" are reasonable answers to Q4.5. Additionally, most students did not recognize a large N-terminal pro-sequence blocking the active site in many proteins, answering "no" to Q13; this can be seen in the responses given in Figure 4B and D. When viewing the accuracy of student responses as a whole (Fig 4D), clear differences emerge between questions. Question Q3 was answered with relatively high levels of accuracy, whereas Q4.5 received responses of mixed accuracy, although several proteins had no unambiguously correct answer. In contrast to the largely accurate responses to Q3 and Q4.5, in Q13, students were able to identify active sites that were not blocked with good accuracy but did have difficulty identifying blocked active sites, which suggests that more instruction should be given on this topic in future implementations of the activity.

Discrepancies may have a number of causes, including the inherent difficulty of capturing snapshots of certain dynamic protein features (e.g., very short α-helices or flexible termini), differences in survey interpretation, and use of structural cues, rather than Chimera's predictive software for secondary structure identification. For example, the ambiguous alignment of papain's sixth α-helix in several proteins (Fig 4B,D) is likely a result of the torsion angle cutoff used to define true α-helices in Chimera; despite the clear visual alignment of these coil-like structures, part or all of their residues may not be considered α-helical in nature. These results speak to the importance of both clarity in what is being asked of participants, as well as emphasis on natural variation of the structural patterns they are asked to characterize. For many of these features, however, different responses are simply a result of varied, but equally valid interpretations of ambiguous data. This kind of harmless variance contributes to the strength of crowdsourced studies and allows researchers to note potentially mobile or disordered regions.

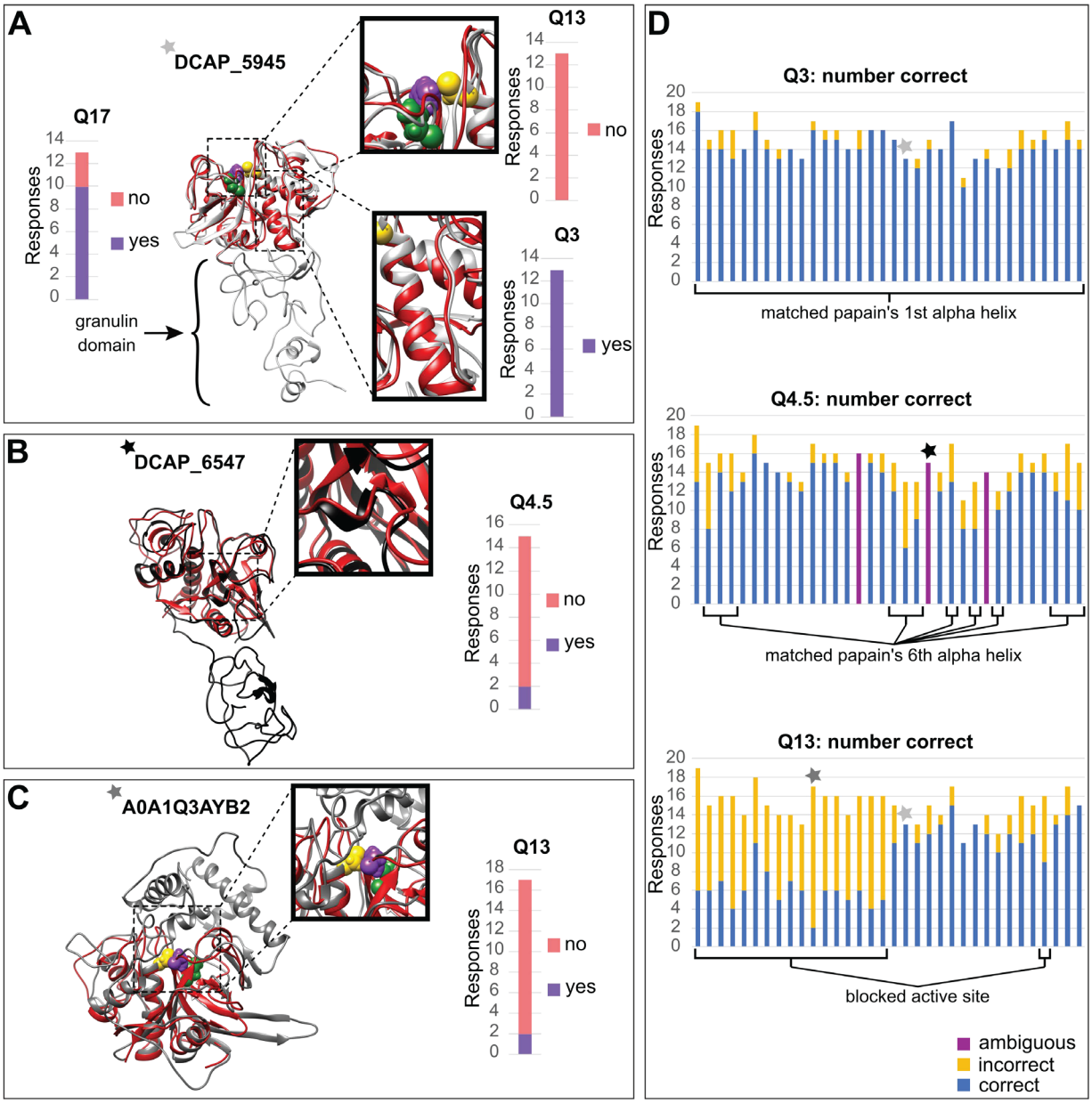

Fig 4. Example survey questions and student responses using proteins aligned to papain (red). (A) Examples of accurate and informative student responses using DCAP_5945 (light grey). (B) Example of ambiguity in student responses with DCAP_6547 (black). (C) Example of inaccuracy in student responses with *C. follicularis* protein A0A1Q3AYB2 (dark grey). (D) Accuracy of all student responses to Q3, Q4.5, and Q13. All 34 proteins are shown in each panel, and those whose examples appear in panels A, B, and C are indicated by colored stars (DCAP_5945, light grey; DCAP_6547, black; A0A1Q3AYB2, dark grey). Black brackets below each graph show which subsets of proteins contain the feature in question.

Consequently, future iterations will work to refine the organization and clarity of presurvey presentations and survey questions, without biasing students' answers. Another modification that could improve students' experience as well as help the instructors identify points of confusion would be to ask students to explain their answers regarding whether particular structural features are present or whether their assigned protein is different from papain or not.

On the research side, student answers will be used by the research group in aggregate. The approach we are using relies on a crowdsourcing model, where multiple students answer questions about each protein independently. Using the data effectively therefore depends on the observation that there is only 1 right and many possible wrong answers, such that the consensus is more likely to be correct than any one answer chosen from the class. This methodology was first introduced by Francis Galton in 1907 (48) and later elaborated for anthropological studies where the reliability of individual informants is unknown (49). Modern versions have been used to solve a variety of problems in fields ranging from engineering and computer science to text analysis (50, 51).

Table 1. In how many classes at UCI (prior to this one) did you have the opportunity to apply the concepts you were learning about in class to a research project?

0, *n* (%)	**1, *n* (%)**	**2, *n* (%)**	**3, *n* (%)**	**4, *n* (%)**	**≥5, *n* (%)**
243 (67.7)	54 (15.0)	35 (9.8)	17 (4.7)	4 (1.1)	6 (1.7)

Therefore, proteins that have been identified by several students as having novel features can be selected for further investigation, whereas those that are agreed to be similar to the reference protein do not merit further scrutiny. Proteins that generate an unusually high level of disagreement may also be of interest, both from the standpoint of improving the instruction and

because they may have interesting features that were not captured by the survey questions (which are made up in advance of detailed examination of the novel proteins). Of course, this strategy is vulnerable to systematic errors if everyone in the class shares a common misconception, making the quality of the instructional materials critical for the research outcome as well as for the students. Because the student results are used in aggregate, the students will be acknowledged as a group in the publication (e.g., Bio98, Winter 2020). However, students who are interested in further participation in enzyme discovery research are offered the opportunity to sign up for research credits. So far, 7 undergraduates have become coauthors on related projects by this mechanism. In our experience, the students recruited in this way are at an earlier stage in their degree program and are more likely to belong to historically underrepresented demographic groups compared with those identified by more traditional methods.

To encourage open discussion and to minimize stress from having to produce correct descriptions of sometimes ambiguous results, this activity was graded only for participation: full credit was granted for submitting a screenshot of the assigned protein model. After completion of the activity, students were given feedback en masse in a class presentation by the graduate student researchers. The “correct” or expert answers referred to in Figure 4 were generated by having 2 experienced undergraduate researchers (with at least 6 mo of experience with protein structure analysis) answer the questions independently. Conflicting answers were then adjudicated by a graduate student. To provide feedback within 1 wk and to be consistent with how we envision using these data for research in the future, this time-consuming process was initially performed only for a subset of enzymes for which several students described features worthy of further investigation. The full set of answers presented in part in Figure 4D was generated later, to assess which aspects of our training module could be improved. The examples

chosen for the follow-up presentation included 1 protein that did not appear to be significantly different from papain and several that had novel features. For example, proteins with occluding loops, granulin domains, pro-sequences, and extra or missing secondary structure elements were shown and the relevant features pointed out. Other instructors may prefer to give each student personalized feedback, although this requires a tradeoff between using new, research-relevant examples and the research team being able to complete the analysis of every protein quickly enough to provide feedback to the students while the activity is fresh in their minds.

D. Student experience assessment

Students' responses to the questions about their experience with the activity ($N = 359$) are summarized in the tables. Results are not mutually exclusive because multiple features were coded from each answer where applicable. Therefore, the number of responses in each category does not necessarily add up to 100%. Table 1 shows in how many classes students were given the opportunity to apply concepts learned in class to a research project. Most students had never performed a similar activity in a class before, although some reported as many as 3 such experiences. Table 2 summarizes the most common responses given for what students liked best about the project. The most common responses cited the interactivity of the activity, seeing how concepts learned in class applied to real-world examples, and having the opportunity to contribute to an ongoing research project. Many students mentioned applying their knowledge to a real-world problem (25.9%) or knowing their work would contribute to an active research project (25.1%) (e.g., "I really enjoyed putting what I have learned to use! It really motivated me to work hard on that assignment and to pay attention in lecture, as I knew it had pertinent information I would need." Others focused on the interactive format of the exercise (22.8%) and the ability to view the proteins in 3D, examine them from different angles, and correlate

Table 2. Please tell us what you liked best about the project (topics from free response).

Real world, *n* (%)	**Research, *n* (%)**	**3D, *n* (%)**	**Interactive, *n* (%)**	**Understand better, *n* (%)**	**Chimera, *n* (%)**	**Instruction, *n* (%)**	**Fun, *n* (%)**
93 (25.9)	90 (25.1)	86 (24.0)	82 (22.8)	59 (16.4)	55 (15.3)	38 (10.6)	30 (8.4)

sequence with structural features (24.0%), none of which are possible with a picture in a textbook. “What I like the most about this project is that I got to look at the protein in 3D, and it is very interesting. On the textbook or online, the protein are always 2D and we cannot spin it around to see its structure.” Some students specifically stated that doing the activity helped them understand protein features (16.4%), and others indicated that it was fun (8.4%). Roughly 15% cited using the UCSF Chimera software as one of their favorite aspects of the project, with several of them explaining that they enjoyed learning a tool that is used by researchers working on protein structures. “I loved the program Chimera and how easy it was to visualize the protein. It was very interesting to compare the different structures to each other based on their sequencing. I felt like a real scientist” and “I liked actually getting to use software that professionals use! It was also nice to apply my own knowledge on something useful, it makes me remember what I’m learning more effortlessly and I enjoy it.” Around 11% mentioned some aspect of the instruction as among their favorite features, including the topic lectures by the instructor or graduate students or the survey activity itself.

Table 3 summarizes the most common responses given for what students liked least about the project. The most common responses focused on some aspect of the instructions being confusing or hard to follow (33.4%), or difficulty or frustration with the Chimera software (17.8%), although many also said that they got used to the software with practice. “What I liked least about the project was that the instructions were not always clear. While doing the survey

during lecture time, I found myself confused by the instructions and I feel that affected the responses I submitted into the survey." "I did not like having to download Chimera and go through that entire process for only a one time use." "Getting used to using Chimera was my least favorite part, but it was also part of the learning experience." "It was somewhat tough to get acquainted with the program in the beginning, but practice over the week helped with this." Some students thought that the activity was rushed and they would have preferred either more class time or more time to work with their groups (3.1%). A few students did not like the open-ended nature of the assignment given that it is part of a live research project. Some were concerned about possibly providing incorrect information for the project (1.4%) or frustrated about not finding out the correct answer at the end (1.7%). "I didn't like how stressful it was to think about how it could affect real research if we got a part incorrect." "The right answer is

Table 3. Please tell us what you liked least about the project (topics from free response).

Instructions, *n* (%)	**Chimera, *n* (%)**	**Blank, *n* (%)**	**Survey design, *n* (%)**	**Rushed, *n* (%)**	**No right answer, *n* (%)**	**Stressed, *n* (%)**
120 (33.4)	64 (17.8)	57 (15.9)	32 (8.9)	11 (3.1)	6 (1.7)	5 (1.4)

Table 4. How can this research project be improved (topics from free response)?

Clearer instructions, *n* (%)	**More feedback, *n* (%)**	**Chimera video or demo, *n* (%)**
116 (32.3)	28 (7.8)	27 (7.5)

not known." "I wish I could have been able to look at other proteins gathered from the project to see what they looked like." However, the total number of negative responses to participation in an active research project with no known answer (3.1%) were far outweighed by the positive ones described above (25.9% "real world experience" + 25.1% "research" = 51.0%). About 16%

of respondents specifically stated that they did not have a least favorite part or that they liked everything about the exercise (blank responses were not included in this category). The least liked aspects of the project included finding the instructions for using Chimera and comparing the 2 proteins confusing and feeling rushed to complete the activity during the class time. The most commonly given suggestions for improvement focused on making the instructions more clear (Table 4). Another idea that was mentioned frequently was to allow students to analyze their proteins as a team. Finally, one student commented that the activity was difficult because of color blindness, which is a useful reminder that instructions for changing the default colors in Chimera should be specifically discussed in the future. Overall, students indicated that this activity helped them understand protein structure and function (Table 5) and should continue to be a part of this course (Table 6). As such, future iterations of this activity will implement suggestions described in Table 4 to make it a more engaging and informational part of their curricula.

V. CONCLUSION

This interactive exercise is adaptable for use in both smaller, upper-division and larger introductory biochemistry courses and can serve as an early exposure to current research projects; it could also be repeated after additional training with more advanced material. It enables students to use fundamental knowledge of protein secondary structures and motifs gained from lectures to build new skills actively that are essential for more advanced study and participation in research on structural biology and protein function. Student feedback after participation in the in-class activity was generally positive. In particular, students indicated that the potential for the work conducted in class to affect real-world research benefited their short-term engagement with the material and bolstered their sense of the value of investing in learning

the information long-term. Criticism was primarily centered on actionable areas of improvement, such as providing more detailed instructions for using the software tools. We expect that future iterations will further benefit from tempering student expectations about the process and continuing to improve clarity in both the presentations and survey by conducting a separate analysis of how interpretations could lead to inconsistent answers. Increased participation and further development in this type of pedagogical tool will serve not only to improve students' educational experience, but also expedite the pipeline for discovering new enzymes that are worthy of experimental validation, a particularly relevant activity in light of recent developments in protein structure prediction. A full description of how the crowdsourced data are used to help streamline the enzyme discovery process will be the topic of a forthcoming publication. Equally importantly, we find that this activity serves as a mechanism to recruit undergraduate researchers at an earlier career stage.

Table 5. Do you agree or disagree with the following statement: This research project helped me understand protein structure/function better (choose one).

Strongly agree,	**Agree,**	**Mildly agree,**	**Mildly disagree,**	**Disagree,**	**Strongly disagree,**
***n* (%)**	***n* (%)**	***n* (%)**	***n* (%)**	***n* (%)**	***n* (%)**
72 (20.1)	170 (47.4)	96 (26.7)	11 (3.1)	5 (1.4)	5 (1.4)

Table 6. Do you agree or disagree with the following statement: This research project should continue to be a part of this course (choose one).

Strongly agree,	**Agree,**	**Mildly agree,**	**Mildly disagree,**	**Disagree,**	**Strongly disagree,**
***n* (%)**	***n* (%)**	***n* (%)**	***n* (%)**	***n* (%)**	***n* (%)**
95 (26.5)	158 (44.0)	71 (19.8)	15 (4.2)	7 (1.9)	13 (3.6)

VI. IRB STATEMENT

This work, which is classified as exempt (research involving normal education practices in an established educational setting), was carried out in accordance with the standards established by the University of California, Irvine, Institutional Review Board (UCI IRB protocol 264).

SUPPLEMENTAL MATERIAL

The full text of the cysteine protease survey activity, the assessment survey, and a ZIP file containing the PDB files used in this exercise are available at: https://doi.org/10.35459/tbp.2021.000209.s1 and https://doi.org/10.35459/tbp.2021.000209.s2.

AUTHOR CONTRIBUTIONS

RWM and PK designed the course module concept. GRT, JIK, MGC, MAS-P, JLU, SMD, RWM, and PK designed and taught the protein structure lecture material and activity training and JIK created the 3D-printed structural models. SMD and PK performed the course design and taught the classes. SHK, MAS-P, JIK, GRT, MGC, BN-B, and RWM designed the cysteine protease survey activity. JIK, PK, and RWM designed the assessment survey. RWM analyzed the assessment survey data. FS, VF, GRT, JIK, and JLU analyzed the cysteine protease survey data and provided the expert answers for each protein. JLU, FS, VF, GRT, MGC, EMD, and JIK provided protein models and structure visualizations for the course materials and the manuscript. RWM, GRT, JIK and PK wrote the manuscript.

ACKNOWLEDGMENTS

This research was supported by National Science Foundation (NSF) award DMR-2002837 to RWM and D. J. Tobias and National Aeronautics and Space Administration award 80NSSC20K0620 to RWM and C. T. Butts. MAS-P was supported by the Howard Hughes

Medical Institute Gilliam Fellowship for Advanced Study. MGC was supported by National Institutes of Health award R25 GM055246 MBRS to the Initiative for Maximizing Student Development program at the University of California, Irvine. JIK and BN-B acknowledge support from the NSF Graduate Research Fellowship Program. This material is based on work supported by the NSF Graduate Research Fellowship under grant DGE-1321846. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the NSF. The funders had no role in the design and conduct of the study; in the collection, analysis, and interpretation of the data; and in the preparation, review, or approval of the manuscript. We thank Carter Butts for advice about survey design. Most importantly, we gratefully acknowledge the hard work and helpful input of the students in Bio98, Winter 2021 (UC Irvine), and in Chem341L, Fall 2020 (Fisk University).

REFERENCES

1. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res* 28:235–242.

2. American Society for Biochemistry and Molecular Biology. 2021. Structure and function: macromolecular structure determines function and regulation. Accessed 12 August 2021. https://www.asbmb.org/education/core-concept-teaching-strategies/foundational-concepts/structure-function.

3. Vilhena, D. A., J. G. Foster, M. Rosvall, J. D. West, J. Evans, and C. T. Bergstrom. 2014. Finding cultural holes: how structure and culture diverge in networks of scholarly communication. *Social Sci* 1:221–238.

4. Richardson, J. S. 1981. The anatomy and taxonomy of protein structure. *Adv Protein Chem* 34:167–339.

5. Lesk, A. M., and K. D. Hardman. 1982. Computer-generated schematic diagrams of protein structures. *Science* 216:539–540.

6. Carson, M., and C. E. Bugg. 1986. Algorithm for ribbon models of proteins. *J Mol Graph* 4:121–122.

7. Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 25:1605–1612.

8. Humphrey, W., W. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J Mol Graph* 14:33–38.

9. Pavlovic, A., M. Krausko, and L. Adamec. 2016. A carnivorous sundew plant prefers protein over chitin as a source of nitrogen from its traps. *Plant Physiol Biochem* 104:11–16.

10. Libiaková, M., K. Floková, O. Novák, L. Slováková, and L. Pavlovičc. 2014. Abundance of cysteine endopeptidase dionain in digestive fluid of Venus flytrap (*Dionaea muscipula* Ellis) is regulated by different stimuli from prey through jasmonates. *PLoS One* 9:e104424.

11. Butts, C. T., J. C. Bierma, and R. W. Martin. 2016. Novel proteases from the genome of the carnivorous plant *Drosera capensis*: structural prediction and comparative analysis. *Proteins Struct Fund Bioinforma* 84:1517–1533.

12. Fukushima, K., X. Fang, D. Alvarez-Ponce, H. Cai, L. Carretero-Paulet, C. Chen, T.-H. Chang, K. M. Farr, T. Fujita, Y. Hiwatashi, Y. Hoshi, T. Imai, M. Kasahara, P. Librado, L. Mao, H. Mori, T. Nishiyama, M. Nozawa, G. Pálfalvi, S. T. Pollard, J. Rozas, A. Sánchez-Gracia, D. Sankoff, T. F. Shibata, S. Shigenobu, N. Sumikawa, T. Uzawa, M. Xie, C. Zheng, D. D. Pollock, V. A. Albert, S. Li, and M. Hasebe. 2017. Genome of the pitcher plant *Cephalotus* reveals genetic changes associated with carnivory. *Nat Ecol Evol* 1:0059.

13. Kim, D. E., D. Chivian, and D. Baker. 2004. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* 32:W526–W531.

14. Raman, S., R. Vernon, J. Thompson, M. Tyka, R. Sadreyev, J. Pei, D. Kim, E. Kellogg, F. DiMaio, O. Lange, L. Kinch, W. Sheffler, B.-H. Kim, R. Das, N. V. Grishin, and D. Baker. 2009. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* 77:89–99.

15. Yang, J., R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang. 2015. The I-TASSER suite: protein structure and function prediction. *Nat Methods* 12:7–8.

16. Risor, M. W., L. R. Thomsen, K. W. Sanggaard, T. A. Nielsen, I. B. Thøgersen, M. V. Lukassen, L. Rossen, I. Garcia-Ferrer, T. Guevara, C. Scavenius, E. Meinjohanns, F. X. Gomis-Rüth, and J. J. Enghild. 2016. Enzymatic and structural characterization of the major endopeptidase in the Venus flytrap digestion fluid. *J Biol Chem* 291:2271–2287.

17. Butts, C. T., X. Zhang, J. E. Kelly, K. W. Roskamp, M. H. Unhelkar, J. A. Freites, S. Tahir, and R. W. Martin. 2016. Sequence comparison, molecular modeling, and network analysis predict structural diversity in cysteine proteases from the Cape sundew, *Drosera capensis*. *Comput Struct Biotechnol J* 14:271–282.

18. Moult, J., K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano. 2018. Critical assessment of methods of protein structure prediction (CASP)—round XII. *Proteins Struct Funct Bioinforma* 86:7–15.

19. Senior, A. W., R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* 577:706–710.

20. Baek, M., F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, and D. Baker. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373:871–876.

21. Ivanova, N., J. Sikorski, O. Chertkov, M. Nolan, S. Lucas, N. Hammon, S. Deshpande, J.-F. Cheng, R. Tapia, C. Han, L. Goodwin, S. Pitluck, M. Huntemann, K. Liolios, I. Pagani, K. Mavromatis, G. Ovchinikova, A. Pati, A. Chen, K. Palaniappan, M. Land, L. Hauser, E.-M. Brambilla, K. P. Kannan, M. Rohde, B. J. Tindall, M. Göker, J. C. Detter, T. Woyke, J. Bristow, J. A. Eisen, V. Markowitz, P. Hugenholtz, N. C. Kyrpides, H.-P. Klenk, and A. Lapidus. 2011. Complete genome sequence of the extremely halophilic *Halanaerobium praevalens* type strain (GSL). *Stand Genomic Sci* 4:312–321.

22. Rawlings, N. D., A. J. Barrett, P. D. Thomas, X. Huang, A. Bateman, and R. D. Finn. 2018. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res* 46:D624–D632.

23. Richau, K. H., F. Kaschani, M. Verdoes, T. C. Pansuriya, S. Niessen, K. Stüber, T. Colby, H. S. Overkleeft, M. Bogyo, and R. A. L. van der Hoorn. 2012. Subclassification and bio-chemical analysis of plant papain-like cysteine proteases displays subfamily-specific characteristics. *Plant Physiol* 158:1583–1599.

24. Misas-Villamil, J. C., R. A. L. van der Hoorn, and G. Doehlemann. 2016. Papain-like cysteine proteases as hubs in plant immunity. *New Phytol* 212:902–907.

25. Kamphuis, I. G., K. H. Kalk, M. B. Swarte, and J. Drenth. 1984. Structure of papain refined at 1.65 Å resolution. *J Mol Biol* 179:233–256.

26. Bangera, G., and S. E. Brownell. 2014. Course-based undergraduate research experiences can make scientific research more inclusive. *CBE Life Sci Educ* 13:602–606.

27. Delventhal, R., and J. Steinhauer. 2020. A course-based undergraduate research experience examining neurodegeneration in *Drosophila melanogaster* teaches students to think, communicate, and perform like scientists. *PLoS ONE* 15:1–22.

28. Junge, B., C. Quiñones, J. Kakietek, D. Teodorescu, and P. Marsteller. 2010. Promoting undergraduate interest, preparedness, and professional pursuit in the sciences: an outcomes evaluation of the SURE program at Emory University. *CBE Life Sci Educ* 9:119–132.

29. Jordan, T. C., S. H. Burnett, S. Carson, S. M. Caruso, K. Clase, R. J. DeJong, J. J. Dennehy, D. R. Denver, D. Dunbar, S. C. R. Elgin, A. M. Findley, C. R. Gissendanner, U. P. Golebiewska, N. Guild, G. A. Hartzog, W. H. Grillo, G. P. Hollowell, L. E. Hughes, A. Johnson, R. A. King, L. O. Lewis, W. Li, F. Rosenzweig, M. R. Rubin, M. S. Saha, J. Sandoz, C. D. Shaffer, B. Taylor, L. Temple, E. Vazquez, V. C. Ware, L. P. Barker, K. W. Bradley, D. Jacobs-Sera, W. H. Pope, D. A. Russell, S. G. Cresawn, D. Lopatto, C. P. Bailey, and G. F. Hatfull. 2014. A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. *mBio* 5:e01051-13.

30. Rodenbusch, S. E., P. R. Hernandez, S. L. Simmons, and E. L. Dolan. 2016. Early engagement in course-based research increases graduation rates and completion of science, engineering, and mathematics degrees. *CBE Life Sci Educ* 15:ar20.

31. Siritunga, D., M. Montero-Rojas, K. Carrero, G. Toro, A. Vélez, and F. A. Carrero-Martínez. 2011. Culturally relevant inquiry-based laboratory module Implementations in upper-division genetics and cell biology teaching laboratories. *CBE Life Sci Educ* 10:287–297.

32. Hurst-Kennedy, J., M. Saum, C. Achat-Mendes, A. D'Costa, E. Javazon, S. Katzman, E. Ricks, and A. Barrera. 2001. The impact of a semester-long, cell culture and fluorescence microscopy CURE on learning and attitudes in an underrepresented STEM student population. *J Microbiol Biol Educ* 21:45.

33. Haas, K. L., J. M. Heemstra, M. H. Medema, and L. K. Charkoudian. 2018. Collaborating with undergraduates to contribute to biochemistry community resources. *Biochemistry* 57:383–389.

34. Bliese, S. L., M. Berta, and M. Lieberman. 2020. Involving students in the distributed pharmaceutical analysis laboratory: a citizen-science project to evaluate global medicine quality. *J Chem Educ* 97:3976–3983.

35. Horowitz, S., P. Koldewey, and J. C. Bardwell. 2014. Undergraduates improve upon published crystal structure in class assignment. *Biochem Mol Biol Educ* 42:398–404.

36. Hati, S., and S. Bhattacharyya. 2016. Incorporating modeling and simulations in undergraduate biophysical chemistry course to promote understanding of structure-dynamics-function relationships in proteins. *Biochem Mol Biol Educ* 44:140–159.

37. Justino, G. C., C. P. Nascimento, and M. C. Justino. 2021. Molecular dynamics simulations and analysis for bioinformatics undergraduate students. *Biochem Mol Biol Educ* 49:570–582.

38. Howell, M. E., C. S. Booth, S. M. Sikich, T. Helikar, K. van Dijk, R. L. Roston, and B. A. Couch. 2020. Interactive learning modules with 3D printed models improve student understanding of protein structure-function relationships. *Biochem Mol Biol Educ* 48:356–368.

39. Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539–539.

40. Petersen, T. N., S. Brunak, G. von Heijne, and H. Nielsen. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 8:785–786.

41. Nielsen, H. 2017. Predicting Secretory Proteins with SignalP. *Protein Function Prediction: Methods and Protocols*. 59–73. Methods in Molecular Biology. New York, NY: Springer.

42. Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten. 2005. Scalable molecular dynamics with NAMD. *J Comput Chem* 26:1781–1802.

43. MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102:3586–3616.

44. MacKerell, A. D., M. Feig, and C. L. Brooks. 2004. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* 25:1400–1415.

45. Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79:926.

46. Word, J. M., S. C. Lovell, J. S. Richardson, and D. C. Richardson. 1999. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol* 285:1735–1747.

47. The UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 49:D480–D489.

48. Galton, F. 1907. Vox populi (the wisdom of crowds). *Nature* 75:450–451.

49. Romney, A. K., S. C. Weller, and W. H. Batchelder. 1986. Culture as consensus: a theory of culture and informant accuracy. *Am Anthropol* 88:313–338.

50. Barbier, G., R. Zafarani, H. Gao, G. Fung, and H. Liu. 2012. Maximizing benefits from crowdsourced data. *Comput Math Organ Theory* 18:257–279.

51. Li, G., J. Wang, Y. Zheng, and M. J. Franklin. 2016. Crowdsourced data management: a survey. *IEEE Trans Knowl Data Eng* 28:2296–2319.