Education Corner

BASIL, RCSB Protein Data Bank, and the NSF

By Paul A. Craig and Bonnie L. Hall

This year, the National Science Foundation (NSF) is celebrating its 75th anniversary. NSF support was essential in the original development of BASIL (Biochemistry Authentic Scientific Inquiry Lab). Ongoing NSF support over the past ten years has enabled the BASIL community to grow in numbers and in collaboration with other teacher/scholar teams who are seeking to change undergraduate biochemistry education. At the same time, NSF support has also provided support for our most critical online resource, the RCSB Protein Data Bank, which has always provided us with the structures that we study and, increasingly, is providing us with the tools that our students use to explore these structures and predict their function.

History of BASIL

BASIL began as a research project led by Paul Craig (Rochester Institute of Technology) and Herbert J. Bernstein (Dowling College) when our students developed the ProMOL plugin for PyMOL[1] and began using it to explore proteins of unknown function in the PDB, which had primarily been generated by the Structural Genomics Initiative [2,3]. Brett Hanson and Charlie Westin designed ProMOL to search for enzyme active sites in protein structures using about thirty annotated motifs from the Mechanism and Catalytic Site Atlas (M-CSA) [4]. As we developed ProMOL, we used hundreds of structures from the RCSB PDB to create a library of well-annotated enzyme active site templates. We then discovered that the Structural Genomics Initiative had deposited >4000 protein structures that lacked functional annotation. The undergraduate research students were able to suggest functions for these proteins by using their active site library, along with results from BLAST[5], Pfam[6], and DALI[7]. They then began testing these predicted functions in our research lab. Figure 1 illustrates an example of an active site alignment done in ProMOL. Similar alignments were found for more than 50 proteins of unknown function in the RCSB PDB[8].

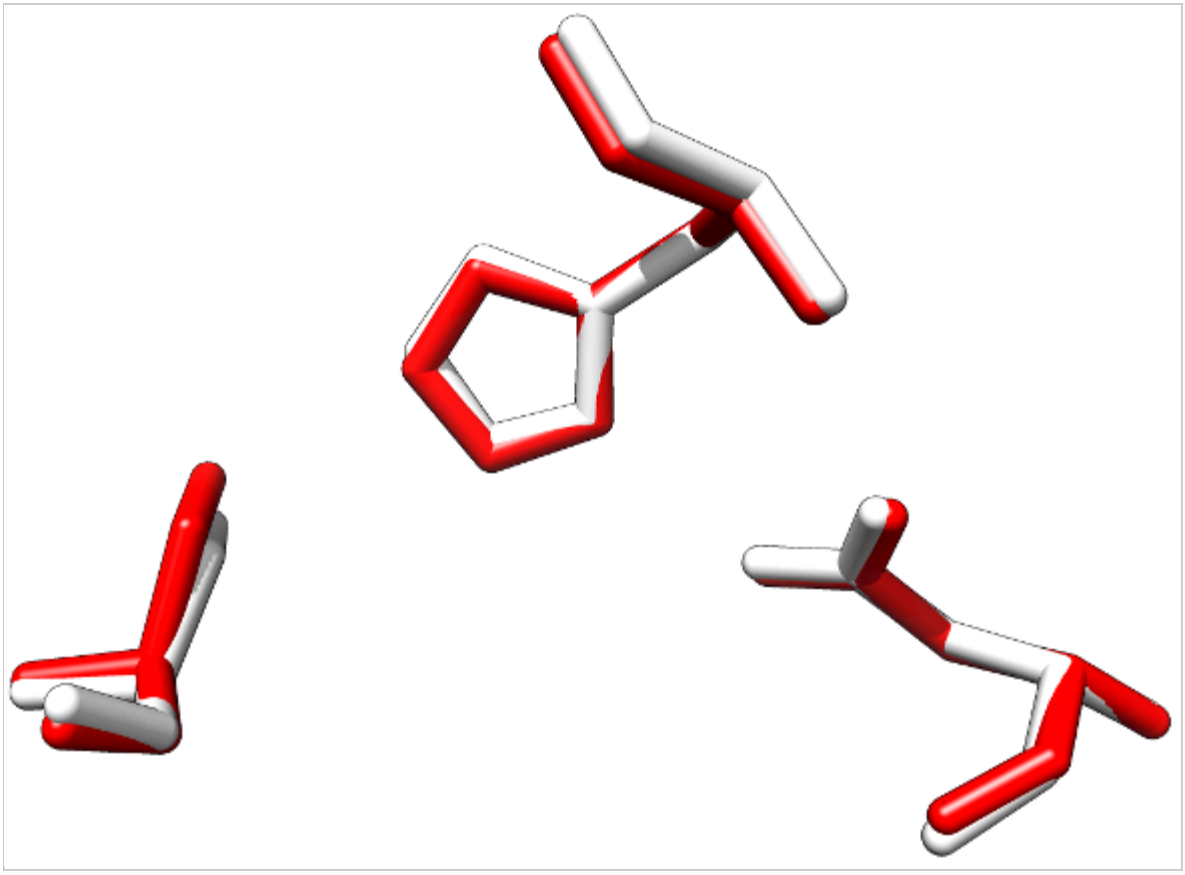

Figure 1. Alignment for a serine protease with ProMOL and PyMOL. Alignment of PDB entry 1AFQ (bovine gamma chymotrypsin; the query in red) with a motif template based on 1A0J (in white) a trypsin structure from Atlantic salmon. Three residues from 1AFQ (His 57, Asp 102 and Ser 195) aligned with the three homologous residues from 1A0J (His 57, Asp 102 and Ser 195).

After six or seven years of funding from NIH, our program officer encouraged us to move our project to the Division of Undergraduate Education (DUE) at NSF. The DUE focuses on how to improve STEM education and engage more undergraduate students in STEM. Although we had missed the deadline, NSF allowed us to submit an off-cycle proposal to the IUSE (Improving Undergraduate STEM Education) program. We did not receive funding, but did receive excellent feedback that led to funding the following year. NSF support since then has enabled us to develop our initial project into the BASIL curriculum and to validate its impact on student learning[9–11]. We are currently offering workshops (listed on the BASIL website) to help people adopt the BASIL curriculum and to promote valid assessment of student learning using the BASIL CURE (Course-based Undergraduate Research Experience).

The BASIL Curriculum

The BASIL curriculum includes computational methods to predict protein functions and wet lab methods to confirm the predicted functions. The curriculum allows students to develop skills in reading scientific literature, collaborating, and generating and testing hypotheses; in doing so they transition to thinking and identifying as scientists.This approach of asking students to learn science the way scientists “do” science is well-aligned with the mission of the NSF DUE.

The BASIL curriculum uses the structures and resources of the RCSB Protein Data Bank extensively. In fact, the A in our logo (designed by a medical illustration student at RIT) is based on a beautifully symmetric PDB structure. In the current BASIL curriculum students download the sequence (in FASTA format) and structure (in legacy PDB file format) of their target proteins and follow the curriculum described in the Figure 2 flowchart. BASIL still

includes an enzyme active site alignment comparison to an RCSB PDB template structure, although students now use a different tool instead of ProMOL.

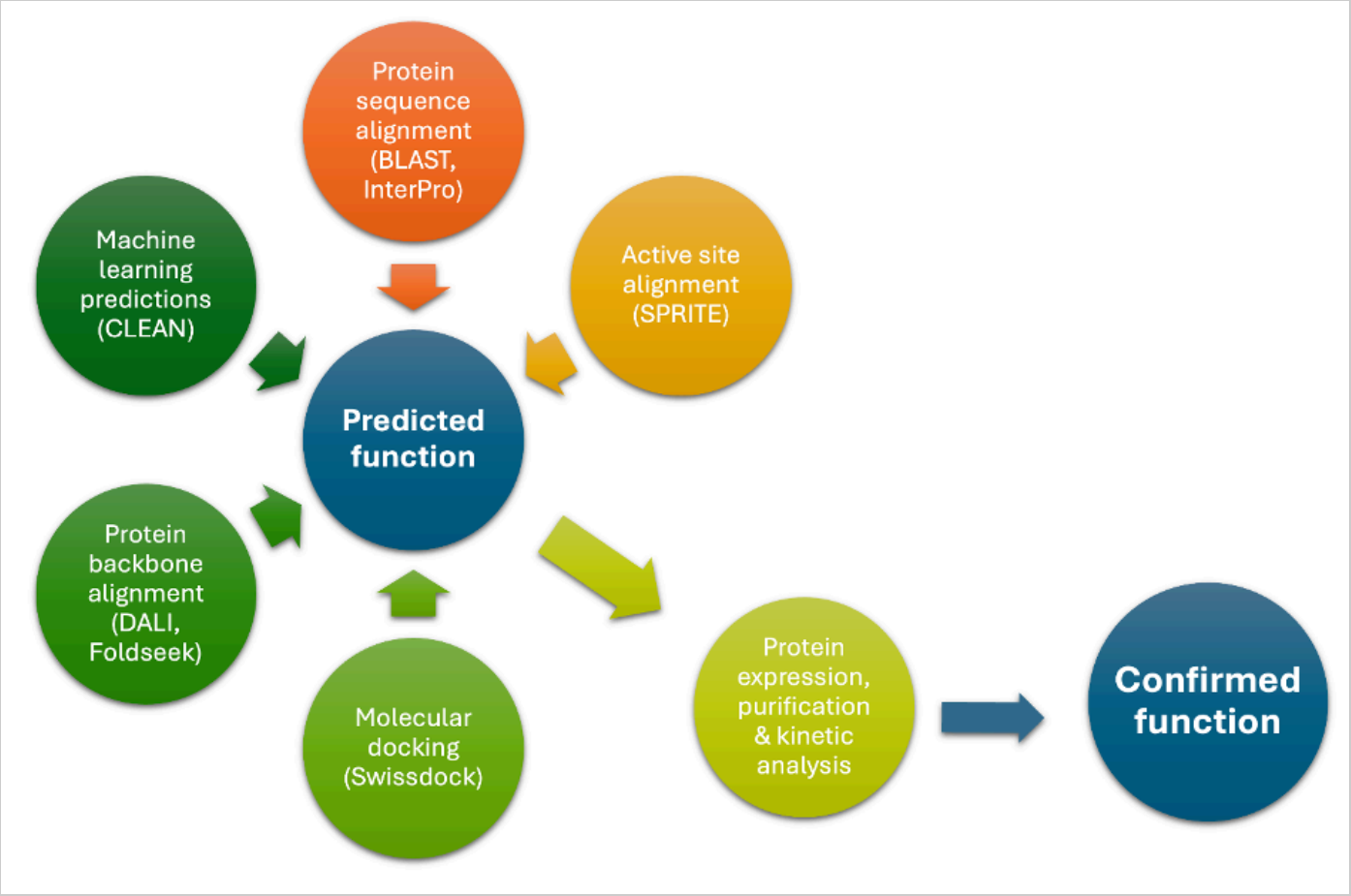

Figure 2. A flowchart describing the BASIL approach to protein function prediction.

Students work with proteins that have known structures but lack functional annotation. A search of the RCSB PDB with the term “unknown function” revealed about 4000 hits when we started this project in 2012, with many more proteins now without an assigned function due to the rapid increase in computationally-solved structures. Continuous NSF support of this public protein repository allows any student at any university to undertake this kind of research project. Initial versions of the BASIL curriculum used a plugin for PyMOL called ProMOL [1], that enabled users to compare a query protein against a library of several hundred enzyme active sites based on the Mechanism and Catalytic Site Atlas [4]. Initial searches revealed about a dozen of these proteins of unknown function as potential serine hydrolases, so we chose that as the starting point for BASIL. We found that plasmids containing the genes for most of these proteins were available at the DNASU Plasmid Repository. One member of the BASIL core team, Mike Pikaart (Hope College), worked with DNASU to create the BASIL starter pack, a collection of ten proteins that can be ordered for minimal cost. Details of the clones and their associated PDB files are included in Table 1.

Table 1. The BASIL Starter Kit contains plasmids which can be used to overexpress the indicated proteins. It is available for purchase at modest cost from DNASU.

Clone ID	Species	PDB ID

LiCD00311532	*Listeria innocua*	3DS8
RrCD00335772	*Cupriavidus pinatubonensis* JMP134	3B7F
BsCD00370437	*Bacillus subtilis* subsp. subtilis str. 168	3CBW
SaCD00432683	*Staphylococcus aureus* subsp. aureus Mu50	3H04
EfCD00450400	*Enterococcus faecalis* V583	3L1W
BsCD00531324	*Bacillus subtilis*	2O14
UnCD00534390	Unknown	3FEQ
EfCD00584424	*Enterococcus faecalis* V583	2QRU
UnCD00663579	Unknown	4DIU
CpCD00696668	*Chitinophaga pinensis* DSM 2588	4Q7Q

The BASIL curriculum was originally designed to have students begin with the computational modules, formulate a hypothesis about the function of the protein and then move to the wet lab to express, purify and characterize the protein. The current computational modules include sequence analysis with BLAST and Interpro [12], active site alignment with SPRITE [13], a full backbone alignment with Dali [7], and docking with Autodock Vina on the SwissDock website [14]. Recently we have developed modules based on two AI tools: CLEAN [15], which predicts the Enzyme Commission Class for the protein based on its sequence, and Foldseek [16], which uses a unique algorithm to perform rapid full backbone alignments of submitted structures.

BASIL instructors are encouraged to adapt the curriculum to work on their campuses and a number of implementations other than the original approach described in the preceding paragraph have been successful [17]:

1. Some faculty begin with the wet lab and integrate the computational materials over the course of the term.
2. At one campus, the computational modules and wet lab modules are taught in separate courses. At the ten-week point in the fourteen-week semester, the students from the two courses are brought together to discuss their findings and they work together on a poster presentation for the end of the semester.
3. One BASIL instructor uses the computational modules only in a graduate level enzymology course.
4. Faculty have used BASIL in a fully online environment.

During the pandemic a number of colleagues contacted the BASIL team and implemented the computational modules as their biochemistry lab course as a fully online response to required emergency remote instruction; others had their students complete all of the modules during the pandemic. The computational modules

transitioned readily to fully online implementation. Students completed the wet labs by working with existing data (SDS-PAGE gels, protein assays, enzyme activity assays) and we found they achieved most of the desired learning objectives [18]. Arthur Sikora (Nova Southeastern University) is developing a fully virtual BASIL curriculum that is currently being tested at Nova and Ursinus College.

The BASIL curriculum is fully open source. In addition, instructors can request access to resources that include teaching guidelines, answers to assessments, explanations of the data analysis, and alternative protocols if they lack access to an instrument. For example, we provide alternatives to sonication for cell lysis, and methods for making your own auto-induction medium in order to control costs.

BASIL and the PDB

The BASIL curriculum currently has 11 modules, six of which utilize the PDB extensively. Some modules require data from the PDB in order to characterize a protein of interest, while others rely on tools that are based on the extensive amount of data about protein structure that is available at the PDB. BASIL modules relying on the PDB include:

1. CLEAN: Requires input of the amino acid sequence of the protein of interest in a FASTA format, which for many proteins can be found on the PDB site (even those without a PDB ID assigned, if Computed Structure Models are included). The predictive model for CLEAN was trained using a collection of proteins that include proteins from the PDB.
2. DALI/FOLDSEEK: Compares the structure of the protein of interest to a collection of PDB proteins (either the full PDB collections or a curated PDB25 subset of proteins). Also requires a protein structure input, submitted either by PDB ID or as a .pdb file (can be downloaded from the PDB site for many proteins, especially if Computed Structure Models are included).
3. InterPro: Accesses information from the PDB (and other databases) to predict protein domains/clans/families. Requires input of the amino acid sequence of the protein of interest in a FASTA format.
4. SPRITE: Compares the protein of interest to a collection of proteins annotated from X-ray crystallographic structures archived in the PDB. Also requires input of either a PDB ID or an upload of a .pdb file (either can be obtained from the PDB).
5. SWISSDOCK: Requires a .pdb structure file, either using a PDB ID or from an uploaded .pdb file. Utilizes PDB structures to validate and improve docking algorithms. The PDB also provides a collection of potential ligands that can be utilized.
6. ENZYME ASSAYS: Requires identification of a specific ligand, with the PDB being one database that can provide ideas for identifying a relevant ligand.

As BASIL has grown, we have focused on making our resources fully accessible to all institutions. One of our major steps has been to move from requiring users to install software to using fully online web applications so that students and faculty can access them easily. The team of software engineers and designers at the RCSB PDB have developed multiple resources that are helpful in curriculum development and implementation.

1. The Mol* viewer enables all BASIL users to fully explore their structures in methods that were formerly only available with standalone programs like PyMOL and ChimeraX. This includes examination of docking results generated by SwissDock.
2. The Sequence Annotations Viewer enables students to directly compare the sequence with the structure in the integrated Mol* structure view window.

3. Students download the FASTA sequences of proteins for exploration on BLAST, InterPro and CLEAN. They also use PDB IDs with Dali and Foldseek.
4. The Advanced Search function enables students to find potential substrates for docking exercises once they have assigned a predicted enzyme commission class.

Ongoing support for RCSB PDB provided by the NSF, NIH, and DOE has been essential for the development and maintenance of the BASIL curriculum. Faculty and students benefit from access to this free digital data resource, with the BASIL curriculum being just one example of how NSF support has enhanced STEM education.

Student Publications Based on the BASIL Curriculum

Examples of student work using the BASIL curriculum can be found online. One set of student work is housed on Proteopedia (Table 2), where students curate their protein function identification work. Two examples of Proteopedia pages of student work are provided. Students also present their work at a variety of conferences, with abstracts provided for 5 such presentations (Table 3).

Table 2. Student generated resources housed on Proteopedia.

PDB ID	**Topic**	**Report**
3HDT	This structure is described as a putative kinase and we have shown it has limited activity as a cytidylate kinase	Proteopedia
3R8E	A novel glucose kinase	Proteopedia

Table 3. Student presentations at national scientific conferences.

PDB ID	**Topic**	**Meeting Abstract**
	Novel carboxylesterase protein function	ASBMB/DiscoverBMB 2023
2O14 3H04	GDSL lipase/esterase family alpha/beta hydrolase family	ASBMB/DiscoverBMB 2024
3R8E	Novel glucose Kinase	ASBMB/DiscoverBMB 2023
1ZBS 3DNU	Predicted to be an N-acetyl-glucosamine (NAG) kinase. Predicted to be a toxin-antitoxin type 2 protein that functions as a serine-threonine kinase	Experimental Biology, 2020
	General presentation of multiple putative kinases	Experimental Biology, 2019

Building and Maintaining the BASIL Community

The BASIL Curriculum has two distinct groups of users. One group comprises faculty exploring biochemistry laboratory curriculum ideas. The curriculum is available free of charge, including 11 student modules, instructor resource documents, and assessment questions for each module. There is a BASIL users Slack channel, where faculty support each other with curriculum adoption, troubleshooting, and updates. The other user group for the BASIL curriculum comprises students being asked to use one or more modules in a course. Students tend to access only the modules page, downloading modules when directed to by an instructor. Students are not given access to instructor resources and are not members of the Slack channel. A Level 1 IUSE grant from the NSF supported the initial development of the BASIL curriculum modules. Ongoing NSF support has allowed the BASIL curriculum to be shared with 140 campuses; portions or all of the curriculum have been adopted at >50 of them. It has also supported assessing and refining the effectiveness of the BASIL curriculum, with the goal of improving STEM education and enhancing scientific workforce development.

As BASIL has grown, the NSF requirement for a sustainability plan has helped shape how the community is maintained. A steering committee now focuses on high level issues and long term goals, with input from a core team of BASIL faculty. BASIL adopters with an interest in the management of BASIL are invited to join the core team, helping sustain BASIL in the long-term. Two committees support instructors, one focused on onboarding for new users and one focused on supporting all instructors (including workshops about the various computational tools, assessment of student work, and other emerging topics). An assessment committee focuses on both assessing student work and evaluating the effectiveness of the BASIL curriculum as a tool in biochemistry education. A modules development committee makes sure the curriculum is updated regularly, develops new modules, and identifies modules needing to be retired. A data management team helps with curating data and with maintaining the BASIL website, so the resources remain freely available.

A Network of Support, The Fabric of Undergraduate Research

The relationship between BASIL and the NSF is critical for the BASIL community. Were the PDB not available for any length of time, faculty using the BASIL curriculum would not be able to implement it. PDB structures are used in all of our computational modules and many of the wet lab modules. Members of the BASIL community in turn work with the PDB on providing feedback about various tools and developments that would be of use to the academic community. BASIL students and faculty have participated in summer PDB workshops at at Rutgers, twice in person and once online. The support of the NSF for both BASIL and the PDB has been essential, as neither would have flourished without that support. The support of the NSF IUSE program has driven the growth and development of the BASIL curriculum and community.

With our current NSF support for BASIL, we continue to build the community by providing open source materials. This includes virtual workshops throughout the year, as well as the more intensive BASIL week offerings each summer. Planning is underway for an online 10th anniversary BASIL celebration from June 9-13, 2025, bringing together faculty to improve STEM education and the BASIL curriculum (https://www.basilbiochem.org/basil-week-2025). Expansion of the BASIL curriculum is underway, including a potential new module focused on Molecular Dynamics simulations to help assign protein functions. Community members are also working on Python scripting for some of the computational modules and data analysis for the wet lab modules. There may soon be enough computational modules to develop a full semester computational version of BASIL. Individuals are welcome to contribute their skills and interests to BASIL, including developing methods for studying non-hydrolase enzyme classes, assigning functions for non-enzymatic proteins, and the development of novel modules that focus on fluorescence, surface plasmon resonance, and the synthesis of novel substrates.

References

1. B. Hanson, C. Westin, M. Rosa, A. Grier, M. Osipovitch, M. L. MacDonald, *et al.* Estimation of protein function using template-based alignment of enzyme active sites. (2014) *BMC Bioinformatics*. **15**, 87.
2. M. R. Chance, A. R. Bresnick, S. K. Burley, J. S. Jiang, C. D. Lima, A. Sali, *et al.* (2002) Structural genomics: A pipeline for providing structures for the biologist. Cold Spring Harbor Lab,.
3. H. M. Berman, J. D. Westbrook, M. J. Gabanyi, W. Tao, R. Shah, A. Kouranov, *et al.* The protein structure initiative structural genomics knowledgebase. (2009) *Nucleic Acids Research*. **37**, D365–D368.
4. A. J. M. Ribeiro, G. L. Holliday, N. Furnham, J. D. Tyzack, K. Ferris, J. M. Thornton Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. (2018) *Nucleic Acids Res*. **46**, D618–D623.
5. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. (1997) *Nucl. Acids Res.* **25**, 3389–3402.
6. J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, *et al.* Pfam: The protein families database in 2021. (2021) *Nucleic Acids Research*. **49**, D412–D419.
7. L. Holm Dali server: structural unification of protein families. (2022) *Nucleic Acids Research*. **50**, W210–W215.
8. T. McKay, K. Hart, A. Horn, H. Kessler, G. Dodge, K. Bardhi, *et al.* Annotation of Proteins of Unknown Function: Initial Enzyme Results. (2015) *J Struct Funct Genomics*. **16**, 43–54.
9. S. M. Irby, N. J. Pelaez, T. R. Anderson How to Identify the Research Abilities That Instructors Anticipate Students Will Develop in a Biochemistry Course-Based Undergraduate Research Experience (CURE). (2018) *CBE—Life Sciences Education*. **17**, es4.
10. S. M. Irby, N. J. Pelaez, T. R. Anderson Anticipated learning outcomes for a biochemistry course-based undergraduate research experience aimed at predicting protein function from structure: Implications for assessment design. (2018) *Biochemistry and Molecular Biology Education*. **46**, 478–492.
11. S. M. Irby, N. J. Pelaez, T. R. Anderson Student Perceptions of Their Gains in Course-Based Undergraduate Research Abilities Identified as the Anticipated Learning Outcomes for a Biochemistry CURE. (2020) *J. Chem. Educ.* **97**, 56–65.
12. T. Paysan-Lafosse, M. Blum, S. Chuguransky, T. Grego, B. L. Pinto, G. A. Salazar, *et al.* InterPro in 2022. (2023) *Nucleic Acids Res*. **51**, D418–D427.
13. N. Nadzirin, E. J. Gardiner, P. Willett, P. J. Artymiuk, M. Firdaus-Raih SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures. (2012) *Nucleic Acids Research*. **40**, W380–W386.
14. M. Bugnon, U. F. Röhrig, M. Goullieux, M. A. S. Perez, A. Daina, O. Michielin, *et al.* SwissDock 2024: major enhancements for small-molecule docking with Attracting Cavities and AutoDock Vina. (2024) *Nucleic Acids Research*. **52**, W324–W332.
15. T. Yu, H. Cui, J. C. Li, Y. Luo, G. Jiang, H. Zhao Enzyme function prediction using contrastive learning. (2023) *Science*. **379**, 1358–1363.
16. M. van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, J. Lee, C. L. M. Gilchrist, *et al.* Fast and accurate protein structure search with Foldseek. (2024) *Nat Biotechnol*. **42**, 243–246.

17. R. Roberts, B. Hall, C. Daubner, A. Goodman, M. Pikaart, A. Sikora, *et al.* Flexible Implementation of the BASIL CURE. (2019) *Biochemistry and Molecular Biology Education*. **47**, 498–505.
18. A. Sikora, S. M. Irby, B. L. Hall, S. A. Mills, J. R. Koeppe, M. J. Pikaart, *et al.* Responses to the COVID-19 Pandemic by the Biochemistry Authentic Scientific Inquiry Lab (BASIL) CURE Consortium: Reflections and a Case Study on the Switch to Remote Learning. (2020) *Journal of Chemical Education*. **97**, 3455–3462.

Paul A. Craig is a professor of chemistry at the Rochester Institute of Technology since 1993. His research spans computational biochemistry, biochemistry education, and protein function prediction. He has been working with the BASIL team since 2015.

Bonnie L. Hall is a professor of Chemistry at Grand View University. She is interested in protein function, including predicting new functions, enzyme engineering, and machine learning approaches. She has been with the BASIL team since 2018.

Join us!

If you are interested in learning more about BASIL and the BASIL curriculum

1. Explore the community and the curricular modules online at basilbiochem.org.
2. Request access to instructor resources for BASIL
3. Follow BASIL online at LinkedIn
4. Join our Slack channel

5. Contact us at basilbiochem@gmail.com