Translating Molecular Science Innovations into Biotechnology Solutions

A UIDP Workshop Report

May 14-15, 2024
Washington, DC

This material is based upon work supported by the National Science Foundation (NSF) under award #2419731. Report finalized on January 31, 2025.

Translating Molecular Science Innovations into Biotechnology Solutions

Executive Summary and Principal Workshop Findings

A highly interactive, National Science Foundation (NSF) funded UIDP workshop convened more than 50 representatives from academic, government, industry, and nonprofit sectors on May 14-15, 2024, in Washington, DC. The academic participants were primarily members of the funded teams for the first two cycles (MFB1 and MFB2) of the Molecular Foundations for Biotechnology (MFB) initiative that has evolved into a powerful collaboration between the National Science Foundation (four directorates with NSF-Chemistry [CHE] and NSF-Molecular and Cellular Biosciences [MCB] as the leads) and the National Human Genome Research Institute [NHGRI] of the National Institutes of Health. MFB1 focuses on protein interactions with small molecules, and MFB2 focuses on machine learning (ML) methods on the function of biomolecular systems.

The workshop's goals were to explore the translation potential of the currently funded MFB work and to identify frontier areas of research in the sphere of discovery science of paramount importance to the pharma/biotech industry and particularly well suited for the multidisciplinary research that the MFB collaborative initiative is well positioned to catalyze.

The two-day workshop was formulated to activate cross-pollination of ideas by featuring presentations from all three sectors: (1) short talks from all funded MFB1 and MFB2 teams, (2) high-level visionary statements from federal government representatives (including leaders of three of the eight science and technology directorates, including Mathematical and Physical Sciences [MPS], Biology [BIO], and Technology, Innovation and Partnerships [TIP]) and MFB-specific statements from leaders of the CHE and MCB divisions at NSF and NHGRI at the NIH, and (3) priority area/challenge articulations from the thought leaders representing pharma/biotechnology sector participants (including Pfizer, Novartis, Sanofi, Eli Lilly, Amgen, Merck, Biogen, Halo, Ono Pharma, and Nexo). Attendees representing private foundations (Chan Zuckerberg Initiative) or venture capital interests (Google Ventures/BioPharma Discovery and Fundomo) also presented their perspectives.

Several key areas for follow-on discussion were identified, including:

I. Innovations in Target Identification

Target identification remains a key element of biomedical science, in which innovations in fundamental science can be game changers, opening new areas of research. This has been seen particularly in activity-based protein proteomics, an area identified in the research discussions. There is still a need for **innovations in tools of chemical biology, bioinformatics, and systems biology to identify biomolecular targets,** for example, by comparing "omics" patterns associated with normal homeostasis versus those associated with dysfunctional biology. In terms of biological regulation/signaling, there remains a need to develop **new tools that allow the investigator to read individual proteoforms**, with attention to all post-translational modifications, ideally with spatial and temporal resolution.

Membrane proteins and receptors and RNAs, including micro-RNA and other non-coding RNAs, are important and less well-characterized areas of biomacromolecular space to explore, both for specific modulations of function and to better define the roles and structures of these biomacromolecules.

II. Innovations in Target Analysis/Dynamics

Participants called out several challenges in this area, particularly the notion of **developing methods to identify inducible pockets and to study proteins and RNA as dynamic ensembles** rather than as static structures. With the advent and continued development of cryo-electron microscopy and modern computational methods, including the rapid expansion of AI/machine learning-based protein structure prediction/design tools such as AlphaFold and RoseTTAFold, there are real opportunities to improve methods to observe conformational ensembles.

In addition to considering an ensemble of conformers for a targeted protein, investigators are also challenged to consider the protein in its often-crowded cellular context. Potential directions include working toward **innovative methods of studying protein structure and function** in such crowded protein environments, and also modeling these environments.

Related to these issues, participants encouraged federal government representatives to continue working to provide access to *state-of-the-art biophysical facilities*, perhaps in partnership with industrial co-sponsors, to ensure broad accessibility to such biophysical tools.

III. Novel Modalities to Manipulate Biology

Today, there is a sea change in mechanisms by which cellular biology can be re-routed using **heterobifunctional small molecules that serve as proximity inducers**. While most of these proximity inducers involve two unrelated ligands for two different targets connected via linker, there is also great interest in heterobifunctional molecules that do not require a linker, so-called molecular glues. The novel modalities that take such a chemical approach to manipulate biology, including proteolysis-targeting chimeras (PROTACs), lysosome-targeting chimeras (LYTACs), phosphorylation targeting chimeras (PhosTACs), regulated induced proximity targeting chimeras (RIPTACs), and deubiquitination (DUBTACs), hold great promise for offering biomedical scientists entirely new ways to approach biological targets.

At the two extremes, these agents range from heterobifunctionals that tag a specific protein for targeted protein degradation (PROTACs and LYTACS) to those that do just the opposite and increase the concentration of an

Key Workshop Findings

The participants noted the need for the federal government to continue investment in biophysical facilities, noting this could be pursued in partnership with private enterprise. Other needed investments include:

1. Chemical biology, bioinformatics, and systems biology tools to identify biomolecular targets.
2. Tools to read individual proteoforms, with attention to all post-translational modifications, ideally with spatial and temporal resolution.
3. Exploration of less well-characterized areas of biomacromolecular space, such as membrane proteins and receptors and RNAs.
4. Use of heterobifunctional small molecules to re-route cellular biology.
5. Innovative approaches to delivery of a therapeutic agent to a specific targeted site.
6. More chemical transformations that can be performed in biological systems.
7. Standard formats and repositories for data storage, including a structural database that would allow the rational design of PROTACs-like heterobifunctionals.
8. A “Human Proteome Project,” focused on fully characterizing all PTMs that occur in expressing the approximately 20,000 genes in the human genome.

essential protein by telegraphing that protein for DUBTACs. While several of these approaches have already led to clinical candidates (particularly PROTACs), there is much room for maturation of this important cutting-edge area; for example, obtaining biological structures of the bound heterobifunctionals (cryo-EM could make a difference here) and building a theoretical/computational chemistry arm for the field of proximity inducers.

IV. Delivery

Pharma/biotech sector participants emphasized the challenge associated with **delivery of a therapeutic agent to a specific targeted site**. This is a complex problem that involves multiple scales, from targeting particular physiological loci (e.g., crossing the blood-brain barrier) to controlling/limiting organ distribution or specificity for a discrete tissue type or cell type, to crossing the cell membrane, to targeting particularly sub-cellular compartments (e.g., the cell nucleus or the mitochondria). While antibody-drug conjugates (ADCs) are well developed in this area, there is room for improvement and innovation in this and related approaches.

For example, regulated induced proximity targeting chimeras (RIPTACs) target a specific cell type with an antibody, nanobody, or related moiety, and in so doing, also target an essential cell cycle protein (perhaps through proximity-induced high-affinity binding), leading to a specifically targeted lethal chemical biology outcome.

Additionally, antibody-sIRNA-conjugates (ARCs) offer the potential to deliver a specific gene-silencing siRNA molecule to a specific cell type. Research challenges here include **better, more reproducible siRNA-protein conjugating chemistry** and considering **innovative approaches to protecting the siRNA cargo and releasing that cargo in the targeted site.**

This dovetails with a major discussion on the important role of lipid vesicles, liposomes, or alternative carriers for sensitive drug cargos, such as mRNA vaccine cargos. Great enthusiasm was expressed for investment in the **fundamental chemistry and biology that underlie such successful carrier mechanisms,** particularly for cargo like mRNA that cannot be too generally immunogenic but must safely be delivered to the targeted cell nuclei for translation into the antigenic protein or protein fragment to induce the specific targeted immune response desired.

V. Bioorthogonal Chemistry

Notwithstanding the recent 2022 Nobel Prize in Chemistry for bioorthogonal chemistry,[1] participants expressed the need for the discovery and development of a much larger array of **chemical transformations that can be performed in biological systems**, e.g., largely aqueous systems with a high concentration of thiol/thiolate nucleophiles (glutathione), variations in pH, salt concentration, and redox potential. Such tools are useful for targeting specific cell-surface glycoproteins or receptors, for example, and may allow for many more *in vivo* self-assembly and selection approaches in chemical biology, the surface of which is only now beginning to be scratched.

[1] The Nobel Prize in Chemistry 2022. NobelPrize.org. Nobel Prize Outreach AB 2025. Tue. 14 Jan 2025. <https://www.nobelprize.org/prizes/chemistry/2022/summary/

VI. Big Data, Data-Sharing Repositories, Data Analytics, and AI and Machine Learning

Big data and the need for both **widely accepted formats and repositories for data storage** arose as an important sub-theme of the workshop. On the one hand, the success of the Protein Data Bank (PDB) was underscored as the foundation upon which protein structure prediction programs such as AlfaFold and RoseTTAFold are built. Alternatively, participants expressed the need for similarly rich databases for the structure and function of other biomacromolecules, such as nucleic acids (particularly RNAs),[2] lipids (e.g., liposomes), and even vaccine adjuvants. Having a **structural database that would allow the rational design of PROTACs-like heterobifunctionals** was seen as aspirational, as almost no such structures have been solved and released.

One focus where the big data challenge looms largest is in the area of the **"Human Proteome Project,"** consisting of fully characterizing what many estimate to be some one million or so actually expressed proteoforms when considering all post-translational modifications (PTMs) that occur in expressing the approximately 20,000 genes in the human genome. This problem is perhaps two orders of magnitude more complex than the completed Human Genome Project. And while efforts have been initiated toward a corresponding Human Proteome Project, there may well be room for U.S. federal agencies, academics, private funders, and industrial chemists to take on a much larger role in what will need to be a major, well-coordinated international effort to be successful. The data challenges for such an effort are significant, and yet such an initiative could have major implications for bringing the full weight of innovations in omics measurements to the level of "smart, readable, personalized fingerprints" to guide future efforts in molecular personalized medicine.

Biomedical Long-Term Broader Impacts

Fundamentally new findings in the molecular sciences have the potential to vastly accelerate biotechnology development. With private sector co-investment and partnership in pursuit of mutually relevant research, federal agencies can identify funding priorities in areas with the most promise to yield high-reward results.

The leaders from pharmaceutical and biotechnology companies articulated several clear priorities in terms of challenge areas in biomedical science, specifically:

I. neurodegenerative indications;
II. cardiovascular health;
III. oncology/cancer biology; and
IV. immune system-related indications.

Importantly, there was specific emphasis on more attention to women's health care, an area that is historically under-resourced. This includes cancer, endometriosis, menopause/hormonal balance, birth and fertility, and post-partum depression.

High-level themes emerging from the workshop included the critical role of data (storage, analysis, and access), as well as tools and platforms that can leverage AI and machine learning to enhance discovery. Pooled data (perhaps using privacy-preserving encryption technologies) from both industry and academia is needed to create the training datasets to harvest the full potential of AI. Industry participants noted that this would require forming a consortium of companies and academics to power

[2] NIH's Nucleic Acid Database provides some of the features described here. https://pmc.ncbi.nlm.nih.gov/articles/PMC3964972/

the data, similar to the PDB-enabled AlphaFold and other structure-prediction options. If enabled, AI and ML could specifically be applied to improve predictions of antibody-antigen interactions or protein-protein complex interactions. Near-term steps to accomplish this would include:

1. Identify the structure and functionality for pooling data in the framework of a central academic consortium, including the process required for companies to blind data on targets so what is captured is narrowed to the local area of interactions (this is to ameliorate concerns about intellectual property). Data on binding affinity should also be included, using a common technique to characterize it.
2. Determine if companies would be open and interested in joining the consortium.
3. Engage a data science company to structurally capture the local interactions/hotspots that could be used to train the AI or ML model.
4. Launch a pilot to determine if future complex interactions can be predicted with a limited test set that includes strong and weak interacting proteins.

Appendix A: Workshop Participants	
Universities and Research Institutions	David Baker, University of Washington /Institute for Protein Design Abhishek Chatterjee, Boston College Phil Cole, Harvard University Sean Cutler, University of California Riverside Peter Dorhout, Iowa State University Taekjip Ha, Boston Children's Hospital Sagar Khare, Rutgers University Crystal Leach, New York University Jared Lewis, Indiana University Sarah Maurer, Central Connecticut State University Daniel Nomura, University of California, Berkeley Shu-Bing Qian, Cornell University Matthew Realff, Georgia Institute of Technology Herman Sintim, Purdue University Joanna Slusky, University of Kansas Ian Wheeldon, University of California, Riverside
Industry Representatives	Sanjeev Ahuja, Merck Adam Gilbert, Pfizer Rishi Jain, Novartis Pharmaceuticals- Biomedical Research Joel Kaar, Amgen Matt LaMarche, Sanofi Kevin Leland, Halo Cures, Incorporated Hiroshi Ochiai, Ono Pharma USA Andy Phillips, Nexo Therapeutics Christalyn Rhodes, Eli Lilly and Company Laura Silvian, Biogen Hidehiro Suzuki, Ono Pharma USA
Other Representatives	Jonah Cool, Chan Zuckerberg Initiative Rajeev Surati, Fundomo Wendy Young, Google Venture/BioPharma Discovery
NSF and NIH Observers	Catalina Achim, NSF David Berkowitz, NSF Denise Caldwell, NSF Christine Chow, NSF Sorin Draghici, NSF Erwin Gianchandani, NSF Theresa Good, NSF Lin He, NSF Manju Hingorani, NSF P. Shing Ho, NSF

	Carolyn Hutter, NIH National Human Genome Research Institute John Jewett, NSF Tingyu Li, NSF Martin Liu, NSF Susan Marqusee, NSF Pumtiwitt McCarthy, NSF Kenneth Moloy, NSF Ishita Mukerji, NSF Waleed Nasser, NSF Ian Nova, NIH National Human Genome Research Institute Jon Rainier, NSF David Tennenhouse, NSF Allen Walker, NSF
Organizing Team	Anthony Boccanfuso, UIDP Mike Brizek, UIDP Natalie Brown, UIDP Morgan Jones-King, UIDP

Appendix B: Award Abstracts

Developing Next-Generation Approaches to Targeted Protein Degradation
Daniel Nomura and Thomas Maimone, University of California, Berkeley
Sept. 1, 2021-August 31, 2024

Synopsis
Our research in this proposal aims to overcome major challenges in realizing the full potential of targeted protein degradation approaches. We currently have a dearth of E3 ligase recruiters and we need to identify new small-molecule recruiter and permissive E3 ligase pairs that can be exploited in either heterobifunctional or molecular glue-based degraders. In this proposal, we have developed next-generation approaches to targeted protein degradation (TPD) to specifically degrade disease targets for therapeutic applications through (1) recruiting core components of the ubiquitin-proteasome machinery with E2 ubiquitin conjugating enzymes and Cullin adaptor proteins; and (2) discovering rational design and discovery principles for molecular glue degraders. We have discovered covalent recruiters against essential components of the ubiquitin proteasome machinery, including E2 UBE2D and CUL4 and CUL1 adaptor proteins DDB1 and SKP1 that we demonstrate can be used in PROTAC applications. We have also discovered a covalent degradative handle that targets the E3 ligase RNF126 that can be transplanted across many different protein-targeting ligands to develop molecular glue degraders of their targets. We have also discovered a second covalent degradative handle that targets the E3 ligase DCAF16 that can also be used for rational design of molecular glue degraders.

Significance
Our findings in expanding the scope of new recruiters that can be used in PROTAC applications that exploit core and essential components of the ubiquitin-proteasome system and in discovering rational design strategies to develop molecular glue degraders help overcome some of the major bottlenecks in the targeted protein degradation field. The findings from this proposal will accelerate the progress of developing degraders against undruggable protein targets and in translating these next-generation therapies into the clinic.

Deciphering the Logic of PTM Crosstalk via Novel Chemical Technology: Histones and Beyond
Philip Cole, Harvard University, and Ben Garcia, Washington University in St. Louis
Sept. 1, 2021-August 31, 2024

Synopsis
We have developed a new method to decipher the logic of how post-translational modifications (PTMs) influence each other in cellular proteins. Over the past thirty years, there has been an explosion in the identification of modifications on cellular proteins using mass spectrometry and allied approaches giving rise to incredible molecular diversity. In but a few of these cases, it is understood how these PTMs, which often occur in clusters, influence each other within the same protein molecule. By combining the power of a designed enzyme, bar-coded tags, and mass spectrometry, we can now quantitatively analyze PTM crosstalk in intact histone H3 tails. We have designed a mutant, circular sortase transpeptidase enzyme (cW11) that efficiently cuts and pastes bar-coded tags onto the C-terminus of histone H3 proteins isolated from mammalian cells. This specific tandem mass tag labeling has been used in multiplex middle-down mass spectrometry to quantify changes in the co-occurrence of PTMs. Using this method, we were able to quantify nearly 200 distinct proteoforms that include combinations of Lys methylation, Lys acetylation, Lys propionylation, Ser/Thr phosphorylation, and Arg methylation.

We have compared the histone H3 PTM patterns that result from cell treatment with two distinct histone deacetylase inhibitors, the HDAC1-3 selective inhibitor MS-275 and the CoREST complex selective inhibitor corin. These inhibitor treatments induce a wide range of changes in histone H3 PTMs that may influence cell growth and gene expression. We have also applied cW11 sortase to streamline the preparation of modified nucleosomes, including those that contain asymmetric H3 tails, which has facilitated the structural and functional analysis of eraser enzyme interactions with chromatin.

Significance
Our generation of cW11 sortase has enhanced both the production of designer nucleosomes and revealed new insights into crosstalk between histone modifications isolated from mammalian cells. From a translational perspective, this work can aid in the structure-based discovery of inhibitors of histone writer, eraser, and reader proteins. Our technology will also be useful in the pharmacological analysis of histone enzyme inhibitor action in cells.

Bioorthogonal Chemistries Targeting 5-hydroxytryptophan for Biological Discovery and Biologics Development
Abhishek Chatterjee, Tim van Opijnen, and Eranthie Weerapana, Boston College
Sept. 1, 2021-Aug. 31, 2025

Synopsis
We have developed a new class of bioorthogonal conjugation reactions targeted to the noncanonical amino acid 5-hydroxytryptophan (5HTP) for two enabling applications. First, we have created a fully genetically encoded system to tag and purify newly synthesized proteins in pathogenic bacteria, which involves: (1) an inducible biosynthetic pathway to generate 5HTP in cells from tryptophan, (2) stochastic incorporation of the biosynthesized 5HTP into newly synthesized proteins using an engineered tryptophanyl-tRNA synthetase, and (3) the use of our 5HTP-selective bioconjugation chemistry to tag and purify the 5HTP-labeled proteins for proteomic characterization. This platform was used to investigate the proteomic changes in pathogenic bacteria during infection. Additionally, we have used the 5HTP-selective chemistry to generate therapeutically relevant protein and virus conjugates. Furthermore, the orthogonality between our 5HTP-labeling chemistry and strain-promoted azide-alkyne cycloaddition has enabled the synthesis of novel therapeutic protein conjugates distinctly functionalized with two different cargoes.

Significance
The ability to precisely modify proteins at predefined sites has numerous powerful applications in both biotechnology and basic research. Our technology provides a general method for achieving site-specific protein modification by: (A) co-translationally introducing the novel unnatural amino acid 5-hydroxytryptophan at desired sites within proteins expressed in live cells, and (B) developing reactions that can bioorthogonally label the 5-hydroxytryptophan residue on the resulting proteins. We have demonstrated the utility of this strategy by creating therapeutically relevant site-specific antibody-drug conjugates and developing a technology to map changes in protein expression in pathogenic bacteria in during host invasion

Collaborative Research: MFB: Ultra-Fast Development of Portable Small Molecule Sensor-Actuators
Tim Whitehead, University of Colorado, Boulder; Ian Wheeldon, UC Riverside; Sean Cutler, UC Riverside; Francis Peterson, Medical College of Wisconsin
Sept. 1, 2021-Aug. 30, 2024

Synopsis
Plants sense the stress hormone abscisic acid (ABA) using chemical-induced dimerization (CID) modules, including the receptor PYR1 and HAB1, a protein phosphatase inhibited by ligand-activated PYR1. This system is unique because of the relative ease with which ligand recognition can be reprogrammed; this property stems from two interconnected biochemical properties: affinity amplification and dimerization. PYR1's phosphatase partner acts analogously to a co-receptor, boosting apparent ligand-binding affinity up to ~100-fold, allowing μM-affinity receptor-ligand interactions to be sensed with overall nM sensitivities. Using this scaffold, we previously developed multiple agrochemical-regulated CID modules, including PYR1MANDI, designed to enable agrochemical control of plant drought response pathways. For our MFP project, we have focused on improving this platform by developing a systematic understanding of its chemical scope using high-throughput screens for new sensors. We used computationally guided mutagenesis to develop libraries of PYR1 variants harboring binding-site mutations and screened them to identify hundreds of small-molecule-regulated CID modules. Cheminformatic analyses of the hits indicate a broad chemical scope for structurally diverse drug-like small molecules. We've used these new CID modules to engineer plants and microbes with chemically regulated phenotypes, such as sensing environmental toxins, and to deepen our understanding of PYR1/ligand interactions using structural biology. This MFB project is a collaboration between Tim Whitehead of the Department of Chemical and Biological Engineering at the University of Colorado, Boulder, Sean Cutler and Ian Wheeldon at the University of California, Riverside, and Francis Peterson at the Medical College of Wisconsin.

Significance
The development of methods to screen large antibody libraries in vitro using phage display and yeast display ushered in a new era of high-throughput biology. These open-source libraries of natural or synthetic antibody sequences were transformative because they enabled small and medium-sized research labs to develop custom antibodies against new targets for their basic and applied research problems. Our MFP project is creating analogous systems for developing chemical biosensors and CID systems regulated by user-specified ligands.

MFB: Deep-Learning Enabled Structure Prediction and Design of Protein-DNA Assemblies
David Baker and Frank DiMaio, University of Washington; Barry Stoddard, Fred Hutchinson Cancer Center
Sept. 1, 2022-Aug. 31, 2025

Synopsis
Sequence-specific DNA-binding proteins (DBPs) play critical roles in biology and biotechnology, and there has been considerable interest in the engineering of DBPs with new or altered specificities for genome editing and other applications. While there has been some success in reprogramming naturally occurring DBPs using selection methods, the computational design of new DBPs that recognize arbitrary target sites remains an outstanding challenge. We describe a computational method for the design of small DBPs that recognize specific target sequences through interactions with bases in the major groove, and employ this method in conjunction with experimental screening to generate binders for five distinct DNA targets. These binders exhibit specificity closely matching the computational models for the target DNA sequences at as many as six base positions and affinities as low as 30–100 nM. The crystal structure of a designed DBP-target site complex is in close agreement with the design model, highlighting the accuracy of the design method. The designed DBPs function in both Escherichia coli and mammalian cells to repress and activate transcription of neighboring genes. Our method is a substantial step towards a

general route to small, and hence readily deliverable, sequence-specific DBPs for gene regulation and editing.

Collaborative Research: MFB: Integrating deep learning and high-throughput experimentation to rapidly navigate protein fitness landscapes for non-native enzyme catalysis
Philip Romero, Duke University; Anthony Gitter, University of Wisconsin, Madison; Jared Lewis, Indiana University
Nov. 1, 2022-Oct. 31, 2025

Synopsis
Our research seeks to develop a transformative framework to engineer enzymes that catalyze non-natural and difficult chemical reactions. Our approach leverages advances in artificial intelligence, molecular modeling, and high-throughput experimentation to rapidly navigate the sequence-function landscape for novel biocatalysts. These studies will contribute to our understanding of enzyme function and the capacity of algorithms to design complex macromolecules, and also advance the theoretical foundations of deep learning and optimization. This work will enable rapid and reliable design of biocatalysts to perform chemo-, site- and stereoselective reactions that remain intractable using synthetic organic methodology. It will also result in publicly available computational tools and experimental methods that other researchers can readily leverage to engineer diverse classes of proteins and biomolecules.

To date, we have developed a new computational framework, Mutational Effect Transfer Learning (METL), for guiding protein engineering that integrates biophysical modeling and artificial intelligence. METL trains a transformer-based artificial intelligence model on large-scale biophysical simulations of proteins. This initial training imbues the model with biophysical knowledge, which enables it to predict the functional consequences of protein sequence mutations from small amounts of experimental data.

In parallel with these efforts, we developed a sequence binning procedure that maps enzyme activity data obtained from standard analytical methods to sequences obtained via long-read next-generation sequencing. Sequence-function data obtained for variants of the enzyme SadX generated during a directed evolution campaign aimed at improving the non-native azidase activity of this enzyme are now being used to train machine learning models customized for this experimental design. Because each sequence variant is associated with a categorical activity bin instead of a quantitative score, we adapted machine learning training procedures to use ordinal loss functions that operate on ranks across bins. The models nominated beneficial mutations for further experimental characterization.

Significance
The ML-guided protein engineering strategies outlined in this proposal will improve our ability to rapidly and reliably engineer custom biocatalysts, which will transform the field of chemical synthesis to produce fine chemicals, polymers, natural products, and pharmaceuticals under environmentally friendly “green” conditions. These capabilities will strengthen the United States’ bioeconomy and secure its position as a global power for the next century. The open-source METL software and models can be reused and adapted by others for a wide variety of protein engineering applications, even beyond biocatalysis.

MFB: Targeting the Dark Proteome by Machine-learning-guided Protein Design
Sagar Khare, Adam Gormley, and Guillaume Lamoureux, Rutgers University
Sept. 1, 2022-Aug. 31, 2025

Synopsis
We are developing targeted protein editors guided by novel machine learning (ML) approaches. The ability to precisely edit genomes has transformed modern biotechnology and medicine; however, technology for the in-situ precision editing of proteins—the workhorses of biology—has been lacking. Our research team is developing and testing protein design approaches for highly selective new editor proteins that contain a binding domain and an enzymatic domain to ensure the selective recognition and modification of chosen target proteins that have intrinsically disordered regions.

MFB: NSF-BSF: Data-Adaptive and Metamorphosis Machine Learning Architectures for Generative Protein Design of Metal Biosensors
Joanna Slusky, University of Kansas; Rachel Kolodny and Margarita Osadchy, Haifa University
Sept. 1, 2022-Aug. 30, 2025

Synopsis
Outer membrane proteins, which are almost exclusively in the outer membrane of gram-negative bacteria, have extreme sequence diversity. This is due to extensive bacterial evolutionary diversification. In addition, due to the constraints of membrane protein insertion, all outer membrane proteins have the same fold, which is a β-barrel. The result is a large number of sequences and extensive sequence diversity for this particular fold. The sequence diversity and structural homogeneity of membrane β-barrels make them well-suited for generative design. The relative diversity and homogeneity of membrane β-barrels position β-barrel generative design to potentially identify sequence space that has not been sampled by evolution but can still yield this same fold. A better understanding of such dark evolutionary sequence space may allow for improved protein design of other folds as well. To use membrane β-barrels for generative design, we need extended sequence sets of proteins and improved structural features for those sequence sets. We have created datasets for improved training, and we have developed computational methods to evaluate the sequences we generate. We have used these datasets and evaluation methods for a variety of machine-learning architectures.

Significance
Membrane β-barrels are ideal for the design of useful biomolecules, especially sensors and enzymes. Membrane β-barrels are well-suited for biosensing because of their shape and location. The variety of radii available to β-barrels, and the ability to proscribe chemical moieties on the interior of the barrel, allow for the binding of a wide variety of analytes with high specificity. Moreover, because these proteins span membranes, the act of binding necessarily provides a measurable output of change in conductance across the membrane. β-barrels also make excellent enzymes. The cylindrical shape of β-barrels makes them excellent targets for enzyme design. Of all the protein families that exhibit barrel structures, 68% are found to be various kinds of enzymes. The barrel shape allows the creation of a cavity of defined size that can sequester substrates from the aqueous environment.

Deciphering RNA-based regulatory logic with interpretable machine learning
Shu-Bing Qian, Cornell University, and Oded Regev, New York University
Sept. 1, 2022-August 31, 2025

Synopsis
RNA transcripts contain multiple, complex, and overlapping codes that dictate their biochemical processing. Two RNA processing mechanisms, RNA splicing and 5'UTR regulation, play key roles in the

fundamental transfer of information from DNA to functional RNA and protein products. Understanding the regulatory logic of these RNA-based codes is required for the rational design of RNA transcripts in biotechnology. Despite decades of genetics, biochemistry, and bioinformatics research, understanding RNA-based regulatory logic – that is, which RNA features dictate the processing for any given input sequence – remains elusive. More recent attempts to apply "off-the-shelf" machine learning methods to limited datasets were not designed to explain how they arrive at their predictions and do not provide insights into the underlying regulatory logic. We propose designing and deploying "interpretable-by-design" machine learning algorithms trained on data obtained from massively parallel reporter assays (MPRAs).

Application 1: A comprehensive understanding of the splicing code in determining exon skipping. The determination of which parts of the transcript are exons depends on a complex code known as the splicing code. How exactly an exon's sequence determines whether it would be included or not is unknown. To address this challenge, we use an MPRA dataset consisting of over 350,000 constructs and analyze it using an interpretable neural network model. In preliminary work, we were able to achieve high accuracy and derive multiple insights into splicing code.

Application 2: A comprehensive understanding of the role of the 5'UTR code in translation initiation and RNA stability. Precise control of protein synthesis by engineering sequence elements in 5'UTR remains a fundamental challenge. To accelerate our understanding of cis-regulatory code embedded in 5'UTR, we develop MPRA from a synthetic mRNA library composed of over one million 5'UTR variants. Our results expose diverse sequence features of 5'UTR in controlling mRNA translatability and stability.

Significance
This project will demonstrate that interpretable machine learning can be used to decipher RNA-based regulatory logic, a critical step forward for basic research with direct and generalizable implications for biotechnology applications. Given its remarkable flexibility, efficient delivery, and proven safety, mRNA has become a powerful and versatile therapeutic platform. Since mRNA therapeutics rely on the translation machinery in target cells, a better understanding of the regulatory logic between cis-sequence elements and trans-acting factors will facilitate rational design of mRNAs with optimal translation potential, high stability, and low immunogenicity.

Accelerating the Discovery of Novel Liposome Formations with Origins-of-Life Insights, Laboratory Automation, and Machine Learning
Joshua Schrier, Fordham University; Sarah Maurer, Central Connecticut State University
October 2022-September 2025

Synopsis
Liposomes have many applications, including for human health and synthetic biology, and can be formulated from a large array of lipids, both synthetic and natural. By developing machine learning and autonomous research methods, we propose that liposome formulation can be optimized for stability and functionality. More speculatively, developing artificial protocells with simple components would not only inform our knowledge about how life evolved, but also enable the creation of engineered abiotic biochemical systems. Traditional approaches to chemical evolution have been biased by considering a "best guess" for starting conditions and reactants based on extant organisms and considered only a relatively limited number of chemical inputs (< 10 reactants) to tame combinatorial complexity. In this project, the investigators use a combination of laboratory automation and machine-learning-guided experimentation to obtain datasets and statistical baselines, needed to test algorithms for exploring and optimizing these complex, non-ideal mixtures. We have developed an automation protocol using an

OpenTrons OT2 liquid handler (Post-doc Ekosso, CCSU) and automated microscopy, including automated micrograph processing (Post-doc Liu, Fordham), and are in round 2 of our first active learning campaign to minimize lipid concentrations for liposome formation. Even in these very early stages, our method has found liposomes at <5 mM total amphiphile concentration. We have also developed a method for generating liposomes from archaeal lipids, to serve as a complement positive control for our existing bacterial liposomes.

Significance
Automating this workflow, including microscopy, has had a profound impact on the person-hours required to evaluate a large number of compositions. While we are still in the early phases of the ML method development and testing, we expect to minimize the concentrations needed for liposome formation, which has implications for improving the economics, stability, and functionality of liposomes for a range of applications.

Novel Graph Neural Networks to Understand, Predict, and Design Allosteric Transcription Factors
Corey Wilson, Yao Zie, and Matthew Realff, Georgia Tech
Sept. 1, 2022-August 31, 2025

Synopsis
Protein allostery is an important protein function that enables communication between different parts of a functional protein that are widely separated. Our lack of understanding of the mechanism of allostery prevents scientists and engineers from designing this critically important function. We are addressing how to predict and design allosteric pathways in the LacI transcription factor. We are constructing complete datasets of one position at a time mutations from wild type and engineered LacI that have repressor and anti-repressor function, respectively. The genotype to phenotype mapping captures performance data in the form of the dynamic range of expression as a function of the mutated sequences from the underlying transcription factor. These data sets will allow us to use machine learning to fine-tune predictions from general sequence to function deep learning tools to predict functional and non-functional protein sequences with higher accuracy. We will be able to identify positions within the protein sequence that appear to change the function of the protein that are not directly related to DNA or ligand binding and, hence, candidates for allosteric communication. The identification of key positions and residue identities will narrow the search space for the design of new allosteric pathways and, hence, enable new transcription factors to be engineered.

Significance
Our overall research goal is to engineer intelligent microbes. This requires that we create microbes with logic, memory and communication features that are under our control. This control allows us to program communities of microbes, such as those that comprise the microbiome in the human gut, to deliver therapeutics or undertake diagnostic tasks. It enables programming of biosecurity function directly into the microbes to protect from accidental or nefarious release of pathogens and protection of intellectual property. The control of metabolic pathways to produce new pharmaceuticals and other chemicals and fuels is supported by programmable gene regulation and can significantly improve overall production rates and titers. Transcription factors are at the heart of this programming and allostery is the mechanism to translate our input signals into actions at the genetic level. Expanding the types of signals that we can use and the range of engineered transcription factors under our control will allow for the expansion of the functionality of transcriptional programming.

Appendix C: Workshop Agenda

Tuesday, May 14, 2024

Time	Session
8:30 a.m.-5 p.m.	**Registration and Check-In**
9-9:10 a.m.	**Welcome** *Denise Caldwell and Susan Marqusee, NSF*
9:10-10 a.m.	**Workshop Charge and Introduction to MFB** *David Berkowitz, NSF; Carolyn Hutter, NIH; Theresa Good, NSF*
10-10:20 a.m.	**Break**
10:20-10:40 a.m.	**Awardee Presentation: UC Berkely** MFB: Developing Next-Generation Approaches to Targeted Protein Degradation *Daniel Nomura, UC Berkeley*
10:40-11 a.m.	**Awardee Presentation: Harvard University** Collaborative Research | Deciphering the Logic of PTM Crosstalk via Novel Chemical Technology: Histones and Beyond *Philip Cole, Harvard Medical School*
11-11:20 a.m.	**Awardee Presentation: Boston College** Bioorthogonal Chemistries Targeting 5-hydroxytryptophan for Biological Discovery and Biologics Development *Abhishek Chatterjee, Boston College*
11:20-11:40 a.m.	**Awardee Presentation: Medical College of Wisconsin, University of Colorado at Boulder, and UC Riverside** Collaborative Research | Ultra-Fast Development of Portable Small Molecule Sensor-Actuators *Sean Cutler and Ian Wheeldon, UC Riverside*
11:40 a.m.-1 p.m.	**Lunch and Speaker Presentation** *Taekjip Ha, Boston Children's Hospital, Harvard Medical School, and Howard Hughes Medical Institute*
1-1:20 p.m.	**Awardee Presentation: University of Washington** MFB: Deep-Learning Enabled Structure Prediction and Design of Protein-DNA Assemblies *David Baker, University of Washington*
1:20-1:40 p.m.	**Awardee Presentation: Indiana University, Morgridge Institute for Research, Inc., and University of Wisconsin-Madison** Collaborative Research | Integrating Deep Learning and High-throughput Experimentation to Rapidly Navigate Protein Fitness Landscapes for Non-native Enzyme Catalysis *Jared Lewis, Indiana University*
1:40-2 p.m.	**Awardee Presentation: Rutgers University** Targeting the Dark Proteome by Machine-learning-guided Protein Design *Sagar Khare, Rutgers University*

2-2:20 p.m.	**Awardee Presentation: University of Kansas** NSF-BSF: Data-adaptive and Metamorphosis Machine Learning Architectures for Generative Protein Design of Metal Biosensors *Joanna Slusky, University of Kansas*
2:20-2:40 p.m.	**Break**
2:40-3 p.m.	**Awardee Presentation: Cornell University and New York University** Collaborative Research | Deciphering RNA-based Regulatory Logic with Interpretable Machine Learning *Shu-Bing Qian, Cornell University*
3-3:20 p.m.	**Awardee Presentation: Fordham University** Accelerating the Discovery of Novel Liposome Formations with Origins-of-Life Insights, Laboratory Automation, and Machine Learning *Sarah Maurer, Central Connecticut State University*
3:20-3:40 p.m.	**Awardee Presentation: Georgia Tech** Novel Graph Neural Networks to Understand, Predict, and Design Allosteric Transcription Factors *Matthew Realff, Georgia Tech*
3:40-4 p.m.	**Break**
4-4:45 p.m.	**Looking forward with NSF TIP** *Waleed Nasser and Erwin Gianchandani, NSF*
4:45-5:15 p.m.	**Report Outs, Key Takeaways, and Pulse of the Participants**
5:15-6:15 p.m.	**Reception**
Wednesday, May 15, 2024	
Time	**Session**
8-8:30 a.m.	**Day 1 Recap**
8:30-9 a.m.	**Chemistry Workshop 2022 Update/Case Study** *Ken Moloy, Program Director, NSF-CHE*
9-10 a.m.	**Future Directions in Molecular Biotechnology: Corporate Perspectives**
10-10:15 a.m.	**Group Discussion Charge**
10:15-10:30 a.m.	**Break**
10:30-11:30 a.m.	**Small Group Discussions Led by Industry Representatives**
11:30 a.m.-12 p.m.	**Report Out, Brainstorming on Future Research Challenges, and Next Steps**
12 p.m.	**Workshop Concludes**