Hannah K. Wayment-Steele
wayment@brandeis.edu | (650) 781-2862
www.hwaymentsteele.github.io

Education

Ph.D., Chemistry, Stanford University 2016 – 2021
Advisors: Rhiju Das, Vijay Pande
M.Phil., Chemistry, Cambridge University 2015 – 2016
Advisor: Daan Frenkel
B.A., Chemistry and Mathematics double major, Music minor, Pomona College 2011 – 2015

Honors and Awards

Jane Coffin Childs Postdoctoral Fellowship 2022
Award for Outstanding Graduate Research, *ACS PHYS division & J. Chem. Phys.* 2021
Chemical Computing Group Excellence Award, *ACS COMP division* 2021
Joseph R. McMicking Award, *Stanford Chemistry Department* 2021
NSF Graduate Research Fellowship 2016
Churchill Scholarship, *Sir Winston Churchill Foundation of the USA* 2015
John Stauffer Prize for Academic Merit in the Sciences*, *Pomona College* 2015
Beckman Scholar 2014
Goldwater Scholar 2014

**Awarded to one STEM graduate annually who exhibits the highest academic promise*

Research

Jane Coffin Childs Postdoctoral Fellow, Brandeis University 2022 – present
Advisor: Dorothee Kern
- Developed deep-learning-based approaches to predict multiple conformational states in proteins
- Created benchmarks of nuclear magnetic resonance (NMR) measurements of dynamics in proteins and integrated with state-of-the-art deep-learning approaches, including AlphaFold2 and language models
- Advised students in projects on protein language model interpretability

Visiting Faculty Researcher, Google Brain Oct. 2022 – Apr. 2023
Host: Lucy Colwell
- Consulted on projects using deep learning in structural biology and biotechnology

Postdoctoral Fellow, Wyss Institute, Harvard Medical School 2021 – 2022
Advisor: William Shih
- Developed novel assays for ultra-sensitive biomolecule detection

Graduate research, Stanford University 2016 – 2021
- Developed improved algorithms for RNA thermodynamic prediction using statistical mechanics to link high-throughput experiment and machine learning
- Created biophysical models for RNA degradation, applied methods to design experimentally validated model mRNA therapeutics with improved shelf lives
- Linked dynamical systems theory and unsupervised machine learning frameworks to create improved analysis tools for molecular dynamics simulations of proteins

Graduate research, Cambridge University 2015 – 2016
- Improved understanding of DNA nanomaterial nucleation and assembly via molecular modelling

Peer-Reviewed Publications (*Equal contributions)

Wayment-Steele, H.K.*, Ojoawo, A.*, Otten, R., Apitz, J.M., Pitsawong, W., Ovchinnikov, S., Colwell, L.J., Kern, D. “Prediction of multiple conformational states via sequence clustering and AlphaFold2”. (2023) *Nature* (In Press).

Wayment-Steele, H. K.*, Kladwang, W.*, Watkins, A. M.*, Kim, D. S.*, Tunguz, B.*, ... Das, R. (2022) Deep learning models for predicting RNA degradation via dual crowdsourcing. *Nature Machine Intelligence* (4) 1174-84.

Wayment-Steele, H.K., Kladwang, W., Strom, A. I., Becka, A., Lee, J., Treuille, A., Eterna Participants, Das, R. (2022). RNA secondary structure packages evaluated and improved by high-throughput experiments. *Nature Methods* (19) 1234-42.

Leppek, K.*, Byeon, G.W.*, Kladwang, W.*, Wayment-Steele, H. K.*, Kerr, C. H.*, ... Barna, M., Das, R. (2022) Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nature Communications* (13) 1536.

Andreasson, J. O., Gotrik, M. R., Wu, M. J., Wayment-Steele, H. K., Kladwang, W., Portela, F., Wellington-Oguri, R., Eterna Participants, Das, R., Greenleaf, W. J. (2022). Crowdsourced RNA design discovers diverse, reversible, efficient, self-contained molecular sensors. *Proceedings of the National Academy of Sciences* (119) 18.

Wayment-Steele, H.K., Kim, D.S., Choe, C.A., Nicol, J.J., Wellington-Oguri, R., Sperberg, R.A.P., Huang, P., Eterna Participants, Das, R. (2021). Theoretical basis for stabilizing messenger RNA through secondary structure design. *Nucleic Acids Research*, 48(18), 10604-10617.

Kostrz, D., Wayment-Steele, H. K., Wang, J. L., Follenfant, M., Pande, V. S., Strick, T. R., Gosse, C. (2019). A modular DNA scaffold to study protein–protein interactions at single-molecule resolution. *Nature Nanotechnology*, 14(10), 988-993.

Wayment-Steele, H. K., Pande, V. S. (2018). Variational encoding of protein dynamics benefits from maximizing latent autocorrelation. *The Journal of Chemical Physics*, 149(21), 216101.

Hernandez, C. X.*, Wayment-Steele, H. K.*, Sultan, M. M.*, Husic, B. E., Pande, V. S. (2018). Variational Encoding of Complex Dynamics. *Physical Review E*, 97(6), 062412.

Sultan, M. M., Wayment-Steele, H. K., Pande, V. S. (2018). Transferable neural networks for enhanced sampling of protein dynamics. *Journal of Chemical Theory and Computation*, 14(4), 1887-1894.

Husic, B. E., McKiernan, K. A., Wayment-Steele, H. K., Sultan, M. M., Pande, V.S. (2018) A minimum variance clustering approach produces robust and interpretable coarse-grained models. *Journal of Chemical Theory and Computation*, 14(2), 1071-1082.

Wayment-Steele, H. K., Frenkel, D., Reinhardt, A. (2017) Investigating the role of boundary bricks in DNA brick self-assembly. *Soft Matter* (2017) 13, 1670-1680.

Agnarsson, B., Wayment-Steele, H. K., Höök, F., Kunze, A. Monitoring of single and double lipid membrane formation with high spatiotemporal resolution using evanescent light scattering microscopy. (2016) *Nanoscale* (8), 19219-19223.

Wayment-Steele, H. K., Jing, Y., Swann, M. J., Johnson L. E., Agnarsson, B., Johal, M. S., Kunze, A. (2016) Effects of Al^{3+} on phosphocholine and phosphoglycerol containing solid supported lipid bilayers. *Langmuir* 32:7, 1771–1781.

Wayment-Steele, H.K., Johnson L. E., Tian, F., Dixon, M. C., Benz, L., Johal, M. S. "Monitoring N3 Dye Adsorption and Desorption on TiO2 surfaces: A combined QCM-D and XPS study." *ACS Applied Materials & Interfaces* (2014) 6, 9093-9099.

Tian, F., Cerro, A.M., Mosier, A. M., Wayment-Steele, H. K., Shine, R. S., Park, A., Webster, E. R., Johnson, L. E., Johal, M. S., Benz, L. (2014) "Surface and Stability characterization of a nanoporous ZIF-8 thin film". *Journal of Physical Chemistry C* 118, 14449-14456.

Invited reviews & book chapters

Wayment-Steele, H.K., Das, R. Learning RNA structure prediction from crowd-designed RNAs. *Nature Methods* 19, 1181–1182 (2022).

Wayment-Steele, H. K., Wu, M., Gotrik, M., Das, R. (2019). Evaluating riboswitch optimality. *Methods in Enzymology*, 623, 417-450.

Teaching

Instructor, Biomolecular NMR course, Swedish NMR Centre, Univ. Gothenburg Sep. 2023

Co-instructor, Topics in Genomics (BIOL 4013) Fall 2020
Co-instructor: Gloria Regisford, Biology Department, Prairie View A & M Univ.
45 students
- Invited and hosted 7 visiting speakers from historically underrepresented backgrounds
- Developed a novel final project using Eterna as a platform for students to create puzzles based on RNA molecules relevant to health and disease
- Developed interactive class activities on reading scientific literature and introduction to principles in RNA biophysics
- Coordinated interactive career information sessions with volunteer graduate students

Teaching assistant, Macromolecules (BIOC 241) Fall 2020
Instructors: Rhiju Das, Pehr Harbury, Biochemistry Dept., Stanford Univ.
40 students
- Developed and led interactive virtual course sessions once a week
- Held office hours, assisted in grading

Teaching assistant, Thermodynamics (CHEM 175) Spring 2017
Instructors: Bianxiao Cui, Will Pfalzgraff, Dept. of Chemistry, Stanford Univ.
60 students
- Assisted instructors in developing a new computational lab section for the class
- Helped develop lab handouts, code bases and exercises
- Assisted in running and delivering lectures for three weekly lab sections
- Developed homework and exam material, held office hours, assisted in grading

Teaching assistant, Statistical Mechanics (CHEM 171) Winter 2017
Instructor: Tom Markland, Dept. of Chemistry, Stanford Univ.
60 students
- Prepared and delivered weekly hour-long discussion sections to clarify concepts in statistical mechanics
- Developed corresponding activities for discussion sections
- Developed homework and exam material, held office hours, assisted in grading

Teaching assistant, Accelerated Chemical Principles (CHEM 31X) Fall 2016
Instructors: W. E. Moerner, Charlie Cox, Dept. of Chemistry, Stanford Univ.
150 students
- Directed two weekly experimental lab sections of 15 students each
- Assisted in writing and developing homework, quizzes, exam materials, held office hours, assisted in grading

Supervisor, Nanotechnology Doctoral Training Centre Michaelmas 2015
Cavendish Labs, Cambridge University
20 students
- Met weekly with groups of first-year PhD students to supervise a theoretical chemistry lab exercise
- Developed course handouts and code for data analysis

Mentorship, Outreach

Program advisor, Undergraduate and DEI education, Nucleate accelerator 2021-2022
- Advised undergraduate and PhD students on initiatives and grants to engage students in biotech-related research.

Program mentor, Center for Genetically Encoded Materials REU, UC Berkeley Summer 2021
- Mentored 2 undergraduates in projects creating an online database of ribosome studies.

Outreach, Eterna Project, Stanford University 2019-2021
- Mentored PVAMU undergraduates in directed reading projects related to RNA vaccines, graduate school application preparation
- Assisted in writing grants for outreach initiatives
- Assisted in science communication, social media presence, hiring

Program mentor, Stanford Summer Research Program Summer 2020
- Mentored 1 undergraduate in a data science project interpreting machine learning models for RNA structure prediction

Mentor for Teaching assistants, Department of Chemistry, Stanford University 2017-2018
- Met monthly with 4 graduate teaching assistants per quarter (12 in total) to discuss teaching strategies and help guide goal-setting for teaching
- Each quarter, ran classroom observation sessions, coordinated teaching evaluations, and summarized and discussed feedback with each mentee
- Helped plan, coordinate and run weeklong Chemistry department TA training orientation at start of fall quarter
- Developed and ran training sessions on effective teaching strategies and grading

Professional Service

General Chair, "Machine Learning for Structural Biology" workshop at NeurIPS 2023
- Led writing and submission of workshop proposal to NeurIPS (35% acceptance rate)
- Established a novel program working with the journal *PRX Life* to create an option for authors to publish their contributions, with publication fees waived
- Managed selection of invited speakers and fundraising to support travel grants for underrepresented students

Reviewer 2019-present
- Nature Methods, Nature Communications, PLOS Comp. Bio., Nucleic Acids Research, Vaccines, and more

Organizer, NSF Protein Folding Consortium Conference, Berkeley, CA Spring 2017

Senator, Stanford Chemistry Student-Hosted Colloquium Committee 2016 – 2018

Invited talks

HHMI award lecture, Quantitative Biology program, Brandeis University. *"Learning the languages of life (Or, why is AlphaFold like ChatGPT?)"* October 2023

Prairie View A & M University, remote. *"Intro to RNA structure."* every Fall, 2021-2023

CASP special interest group for ensembles, remote. *"Predicting multiple conformational states using AlphaFold2 and clustering."* June 7, 2023

Boston Protein Modeling and Design Club, Cambridge, MA. *"Understanding (and discovering?) fold-switching proteins."* April 12, 2023

Machine Learning for Proteins, remote. *"Understanding fold-switching proteins using AlphaFold2 and sequence clustering."* April 28, 2023

Relay therapeutics, Cambridge, MA. *"Predicting multiple conformational states by combining AlphaFold2 and sequence clustering."* Jan. 17, 2023

Pomona College Chemistry Dept, Claremont, CA. *"Inferring RNA structure and stability via high-throughput experiment."* July 22, 2022

Inceptive Nucleics, remote. *"Inferring RNA structure and stability via high-throughput experiment."* April 13, 2022

Schrödinger Multiscale modelling for biotherapeutics symposium, remote. *"Improving the Stability of mRNA therapeutics through biophysics, machine learning, and crowdsourcing."* May 13, 2021

TEDx Washington High, Fremont, CA. *"Designing stabilized vaccines with community science."* May 1, 2021

Center for HIV-1 Studies Annual Workshop, remote. *"Inferring RNA ensembles via high-throughput data."* April 5, 2021

IEEE Silicon Valley Chapter, Information Theory Society. *"Improving the stability of mRNA therapeutics."* March 24, 2021

Conference Presentations (Contributed)

"Computational Aspects of Biomolecular NMR" Gordon Research Conference, Mt. Snow, VT. *"Have protein language models learned dynamics? Evaluating with a large-scale benchmark of NMR relaxation data."* June 20, 2023

Machine Learning in Structural Biology workshop, Neural Information Processing Systems conference, New Orleans, LA. *"Predicting conformational landscapes of known and putative fold-switching proteins using AlphaFold2"* Dec 3, 2022

International Conference on Intelligent Systems for Molecular Biology (ISMB). *"Improving RNA structure prediction with high-throughput crowdsourced data."* July 13, 2020

Media engagement

Nature. *"Remarkable AI tool designs mRNA vaccines that are more potent and stable."* May 2, 2023

Fifty Years Podcast. *"Screening for Enhanced RNA Vaccines with Kathrin Leppek, Gun Woo Byeon, and Hannah Wayment-Steele."* October 14, 2021

National Geographic. *"Future COVID-19 vaccines might not have to be kept so cold."* April 13, 2021

Patent Applications

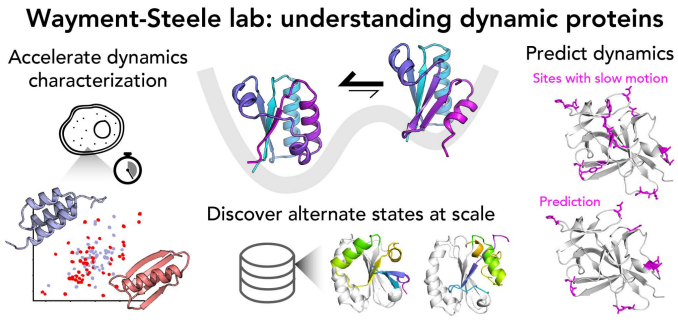
H. K. Wayment-Steele, E. Sharma, R. Das, W. Greenleaf. 63/245,744, “Systems and Methods to Determine Nucleic Acid Thermodynamics and Uses thereof”, Sep. 17, 2021.

R. Das, H. K. Wayment-Steele. PCT/US2021/040026, “Systems and Methods to Enhance RNA Stability and Translation and Uses Thereof”, July 1, 2021.

R. Das, C. A. Choe, H. K. Wayment-Steele, W. Kladwang, 17/364,890, “Systems and Methods to Enhance RNA Stability and Translation and Uses Thereof”, June 30, 2021.

Predicting and discovering protein dynamics

How does a biological sequence encode not just one, but multiple conformational states? This is the question that first drew me to biophysics, and it continues to drive my research interests. In many areas of biology, researchers have shown that proteins are exquisitely tuned to occupy multiple structures to perform their functions, including signaling, catalysis, transport, and more. The ability to accurately predict protein dynamics – multiple conformations, their probabilities, and the rates of transitioning between them – is the next grand challenge for structural biology. Given the sheer scale of biology and the difficulty of studying proteins in motion, we are far from a comprehensive understanding of dynamic proteins. Electron microscopy and tomography (ET) have not reached the time or spatial resolution needed to characterize domain-level dynamics, preventing access to critical details[1]. For instance, in the signaling proteins critical across human health, entire outcomes are determined by the presence or absence of a phosphate group[2] – just five atoms. In contrast, NMR spectroscopy offers superb sensitivity to dynamics spanning from picoseconds to hours motions, yet currently, the experiments and analysis are time-consuming. Faster means of understanding dynamics would help us better learn the roles and evolution of protein dynamics, would significantly accelerate therapeutic discovery efforts, and would allow us to design enzymes with the same finesse as nature for applications from breaking down neuro-degenerative plaques to capturing carbon.

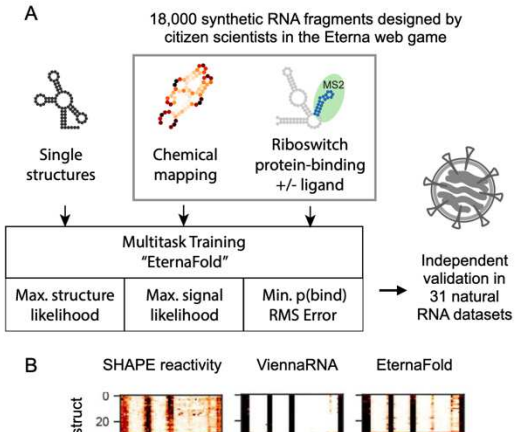


Through my proposed research, I envision that in future years we will be able to experimentally determine the populations of multiple states and kinetics of any protein, even *in vivo*, in the span of an afternoon. When combined with emerging powers of cryo-ET, these methods will transform the space- and time- resolution at which we can understand molecular systems in cells. In its first years, my lab will (1) develop methods to increase the throughput of NMR by an order of magnitude, (2) predict kinetics leveraging large-scale collections of NMR observables, and (3) identify proteins related to disease with alternate states as potential therapeutic targets.

Prior research. Pursuing the question of how a single biological sequence encodes information about multiple structures, while striving to use the most cutting-edge techniques available and make impact where possible, has led me to gain unique range of expertise in topics spanning molecular simulation, dynamic programming, NMR spectroscopy, large language models, and even mRNA vaccine design.

Improving interpretation of molecular dynamics simulations with deep learning. Molecular dynamics (MD) simulations offer the potential to understand atomistic details of protein dynamics. As MD simulations grow in length- and time-scales, new theoretical and statistical methods were needed to parse the resulting data. I connected a classic framework from unsupervised learning, the variational autoencoder, to dynamical systems theory to create an improved approach for extracting long-timescale processes from MD data[3,4] and demonstrated our approach to incorporate the autocorrelation of processes in the learning function was the key element[5]. However, the outsized ratio of theory to experimental data left me wishing for a more data-rich problem in biomolecular ensembles. I found this exactly in a different macromolecule: RNA.

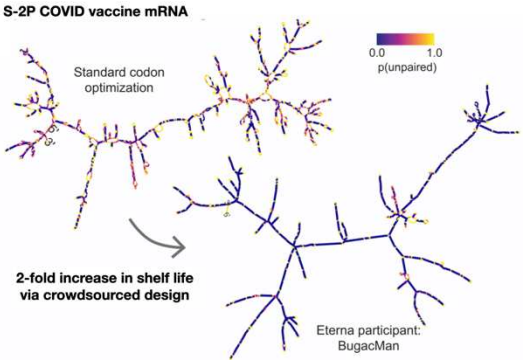
Advancing RNA structure prediction and design. The computer-aided design of RNA molecules is increasingly widespread in a variety of synthetic biology and medical applications. A variety of algorithms based on dynamic programming and increasingly sophisticated techniques have been in development for the last 50 years, yet it was understood in the field that they were insufficient at predicting thermodynamics-based observables beyond a single most likely structure. Working with Rhiju Das, we hypothesized that signal to improve one of these models lay in datasets of synthetic RNAs whose structure characterized via the online RNA design game Eterna. I developed a statistical



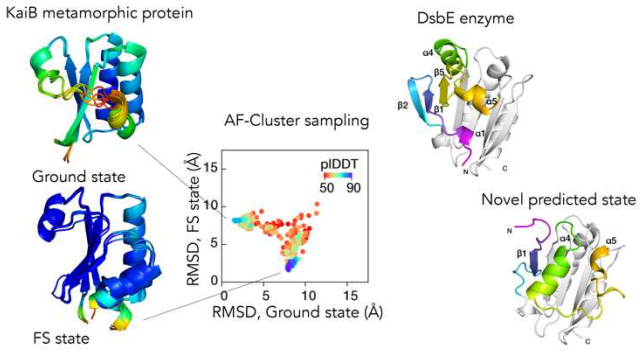
Multi-task training on thousands of RNA molecules created an improved thermodynamic framework to predict RNA ensemble-averaged observables. B. Example prediction of unpaired probabilities of synthetic test set RNAs in ViennaRNA, the most widely-used RNA folding package, and EternaFold.

mechanical framework to update nearest-neighbor thermodynamic parameters based on 1) predicting unpaired probabilities of individual nucleotides through chemical mapping data, and 2) predicting riboswitch activity using large-scale riboswitch datasets[6] (Fig. 2A). The resulting model, dubbed EternaFold, demonstrated improved performance that generalized to diverse independent datasets of molecules including complete viral genomes probed in vivo and synthetic designs modeling mRNA vaccines.[7]

Designing shelf-stable mRNA therapeutics with biophysically-principled ML. The field of RNA design emerged as poised to address one of the greatest global health challenges delivered by COVID-19: the development of refrigerator-stable mRNA vaccines to enable their more equitable distribution. Vaccines based on messenger RNA (mRNA) held immense promise, yet concerns persisted regarding their thermostability. A largely unexplored strategy to reduce mRNA hydrolysis is to design mRNAs that form double-stranded regions, which are protected from in-line cleavage and enzymatic degradation, while coding for the same proteins. The amount of stabilization that this strategy can deliver and the most effective algorithmic approach to achieve stabilization were poorly understood. I developed a framework relating the hydrolysis rate of a RNA molecule to readily-computed base-pair probabilities, and demonstrated that many mRNA targets we tested were predicted to gain at least 2-fold increase in half-life through computational sequence design (Fig. 3), while maintaining wide diversity in morphologies.[8] Our team launched a crowdsourced challenge to solicit diverse RNA designs, and experimentally probed their degradation. I distilled features of these data in a machine learning model that was used to guide a stochastic mRNA design algorithm. This resulted in sequences that had a 2.5-fold increase in shelf life over conventional methods, and increased protein expression over conventional designs, even in model vaccines formulations from Pfizer.[9] We used these data to launch a crowdsourced machine-learning challenge, in which over 1600 teams collaboratively created highly accurate deep learning models for predicting RNA degradation.[10]



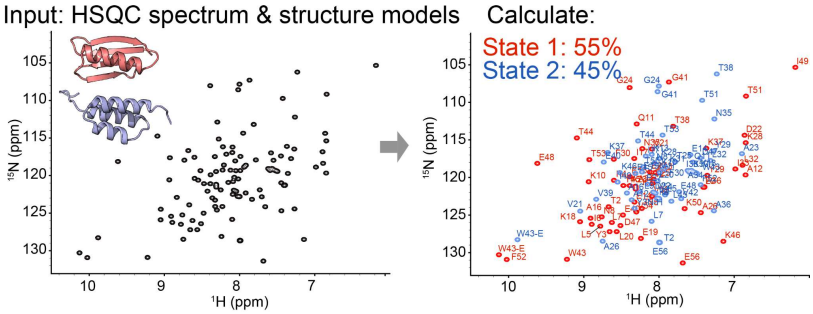
Adapting AlphaFold2 to predict alternate protein conformational states. Following my PhD, I wished to return to thinking about multiple conformational states of proteins. Because AF2 uses information in evolutionary couplings to make structure predictions, I hypothesized that it ought to be able to predict multiple conformations of proteins that contain strong evolutionary pressure for multiple states. I reasoned that multiple sets of couplings for multiple conformations ought to be most obvious in “metamorphic” proteins, proteins which completely rearrange their secondary structures as part of their function. I demonstrated that by simply clustering sequences from a multiple sequence alignment (MSA) and using those clusters as input to AF2, we could predict both conformations for the metamorphic proteins KaiB, RfaH, and Mad2.[11] The few metamorphic proteins that are known have been discovered by chance, yet they play essential roles from prokaryotes to humans. I was curious if we could use our method to screen for new alternate conformational states. Indeed, applying our method AF-Cluster to proteins with no known alternate state, we predicted a novel conformation for a secreted oxidoreductase from *Mycobacterium tuberculosis*. My research indicates that for the first time, we might be able to systematically identify fold-switching proteins across any organism of interest. This will unlock a deep new understanding of protein functions, allostery, and potentials for more complex rational design of multi-state proteins.



Large-scale benchmarks of slow protein dynamics. I am convinced that we are poised to be able to qualitatively predict multiple conformations of proteins, and that the missing ingredient for achieving quantitative prediction lies in the lack of large-scale data on dynamics. I developed a standardized benchmark of NMR relaxation datasets comprising 101 proteins and am currently testing if fine-tuning large language models to predict which amino acids are exchange-broadened across thousands of proteins in the Biological Magnetic Resonance Bank is sufficient to already have predictive power in predicting which residues have slow motion in relaxation experiments. This lays the groundwork for future directions detailed more in my proposed research.

Proposed Research.

Aim I. Increasing NMR throughput by unlocking "HSQC-only" mode. NMR provides exquisite sensitivity to thermodynamics and kinetics, and much information is present in the workhorse $^{1}H^{15}N$ HSQC spectrum: how many folded states are present and their populations. However, assigning a HSQC currently requires multiple further triple-resonance experiments that extends the collection time to many days, with further weeks often required to analyze. Given the ever-improving accuracy of structure prediction methods and increasing scope of structure-related deep learning, do we still need triple resonance experiments to gain useful information from HSQCs? An "HSQC-only" paradigm would let us rapidly accelerate data collection while in no way detracting from more detailed NMR characterization where desired. In conjunction with inevitable improvements in automated protein expression[12] and shorter pulse sequences for kinetics experiments (CPMG, CEST), within the decade I predict one scientist will be able to collect populations and kinetics of ~10 proteins in a week. This is readily extensible to *in vivo* NMR, currently primarily limited by the time needed for triple resonance assignments[13], as well as non-equilibrium experiments. "HSQC-only" NMR would be viable if we could accurately predict chemical shifts from structure models.



Aim I outcome: immediately calculate states and populations from HSQC.

Although the problem of predicting two numbers (the N and H backbone chemical shifts) per residue given a 3D structure prior might seem like a simple task, highly accurate empirical chemical shift prediction has eluded the protein NMR field since initial work on the problem in the 1970's[14]. Accurate chemical shifts can be calculated using ab initio techniques for some small proteins, but ab initio calculations quickly become intractable. Chemical shifts are extremely sensitive to neighboring electronic effects, such as backbone nitrogen lone pairs or the presence of distant sulfur atoms, and existing efforts have not proven sensitive enough to enable accurate assignment from N and H chemical shifts alone[15].

To address this problem, I intentionally will *not* start by testing deep learning architectures myself but will take a new approach: I will conduct a crowdsourced deep learning challenge to rapidly solicit and test approaches. I estimate it would take any single academic group 5-10 years to make the headway that my lab will enable in 6 months of a crowdsourced competition. Indeed, introducing the seemingly simple problem of predicting a HSQC spectrum to non-NMR experts is a key outcome of this project. I hypothesize that chemical shift prediction will be very analogous to other tasks that deep learning has gained traction on such as predicting binding pockets[16] and ddGs of point mutations[17], yet even if one challenge does not create models with sufficient predictive power, posing the question as a worthy challenge will hopefully attract the attention of the best minds to eventually solve the problem, much as how AlphaFold was a result of researchers at DeepMind becoming familiar with the protein folding problem[18] through community competition-oriented initiatives like FoldIt[19] and CASP. The key steps to successfully running this challenge include: 1) designing the challenge –metrics for evaluation, train/test splits, explanations, supplemental featurizations, available codebases – so that non-NMR experts have sufficient grasp of the problem to be productive, 2) creating sufficient data to train from, and 3) designing stringent blind tests that evaluate our end goal.

1) *Crowdsourced challenge design.* I have significant experience communicating complex tasks in biomolecular prediction to non-experts, most recently by contributing to running a crowdsourced deep learning challenge in collaboration with researchers at Kaggle[10]. This task was predicting RNA degradation, a problem that also appeared to our team that it would take years for an academic lab to make headway on. The competition, which lasted a total of 3 months, resulted in models that achieved an accuracy where 41% of predictions were within experimental error[10]. The deep learning community surrounding the competition hypothesized and tested featurizations and architectures from across natural language processing and computer vision that it would have taken PhD students years to identify, code up, and test. I also have broader experience in setting up blind challenges through the Eterna platform[6-9], which interacting with amateur citizen scientists and machine learners from a variety of backgrounds.

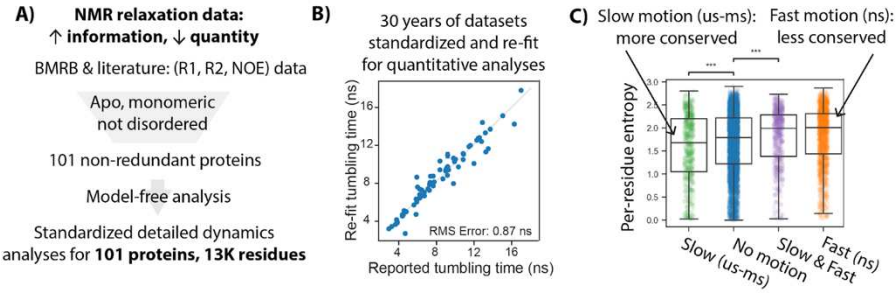
2) *Data.* Previous empirical chemical shift prediction work using the BMRB has used subsets of all available chemical shifts by limiting to proteins for which experimental 3D structures could be confidently matched. We will expand the usable data to the entirety of chemical shifts deposited in the BMRB (~12,000) by positing that AlphaFold2 models are sufficiently accurate for these purposes. We can augment these data with small molecule chemical shifts and/or quantum mechanical-based calculations. I have previously curated data at the scale of the

entire BMRB (see Aim II), which gives me preparation to understand existing challenges with BMRB data and ideas for how to improve its usability.

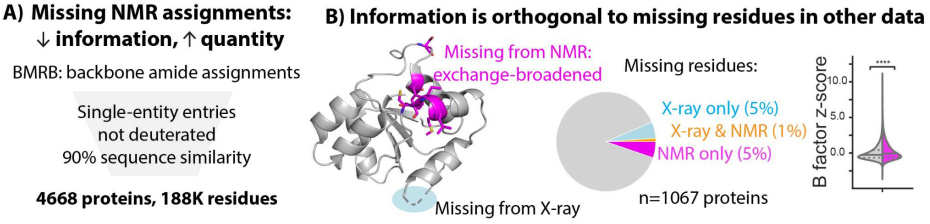
3) *Designing suitable tests.* My lab will create a diverse set of blind tests from new data collected in our own lab as well as collect unpublished datasets from other NMR labs. Furthermore, our post-competition analysis of what kinds of chemical environments proved easiest and most difficult to predict will be key to using the results of this competition to advance continued algorithms and data collection.

The winning models from the crowd-sourced challenge in phase I will serve as core modules for methods my lab will develop to automatically deconvolve and assign HSQC spectra for multiple states in conjunction with prior structure models provided by AF2 or other sampling methods. We will test our methods by designing mutations to switch populations within protein families with two folded states such as the G_A/G_B proteins[20] and KaiB^{RS}, a novel system we introduced in ref. 11, which will be ideal for testing multistate design due to its monomer-monomer transition, small size, and clear signal in HSQCs. We continue to expand our fitting and design methods to more complex systems. This will be transformative for rapidly obtaining information on multistate proteins to facilitate multistate design. As a fallback, if N and H shift prediction alone is not yet accurate enough, we will explore using sparse triple resonance experiments for our continued efforts.

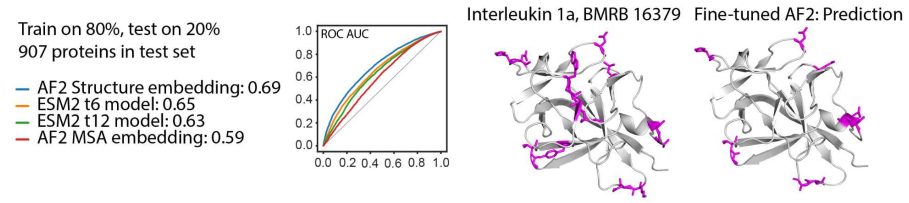
Aim II: De novo prediction of protein motions by learning from large-scale NMR datasets. Despite considerable interest in the ability to predict protein conformational changes on the timescales of microseconds to milliseconds (μs-ms), which are often critically important to biological function, a missing ingredient has been a lack of large-scale "ground truth" data of protein dynamics. Researchers have been trying to improve prediction of multiple conformations via MD simulations for decades, yet these comparisons have been limited to a handful of model proteins.



Where can we find "ground truth" measurements for μs-ms dynamics on the scale of thousands of proteins? Dynamics manifest themselves in all structural biology measurements, but NMR relaxation experiments uniquely give a direct readout of dynamics. I compiled a database of 101 non-redundant datasets of NMR relaxation experiments (R1, R2, NOE), all re-fit using the same assumptions[17]. This represents the first large-scale collection of dynamics data, and is amenable to start using in machine learning to develop methods to predict directly from sequence, which residues are undergoing fast or slow motions. Strikingly, this dataset revealed a finding about the evolutionary conservation of dynamics: residues exhibiting only slow motion (us-ms timescale) were more evolutionarily conserved than static residues, which in turn were more conserved than residues undergoing any sort of fast motion (Fig. 4).

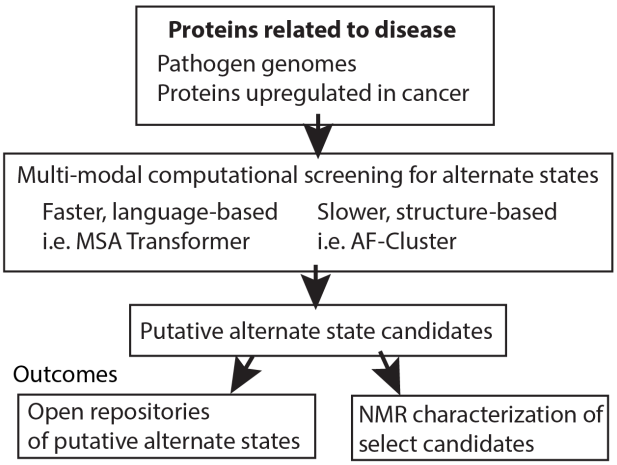


I then realized there were markers of μs-ms dynamics hiding in plain sight. I compiled a benchmark (Fig. 5A) consisting of residues missing from $^{1}H^{15}N$ HSQCs due to exchange broadening. These residues missing from NMR have low overlap with missing residues in X-ray and Cryo-EM data (Fig. 5B), highlighting that the signal from these NMR-derived data on thousands of proteins represent an orthogonal set of data for deep learning efforts to predict dynamics. Initial fine-tuning tests reveal that AF2 and even the unsupervised language model ESM already contain some capacity to predict such exchange-broadening (Fig. 5C). These results show that there are sequence patterns that we can leverage to predict μs-ms protein dynamics, and they exist at the scale of thousands. The next step for using these to train



and evaluate all-atom models in predicting kinetic lifetimes is modeling the conformations and actual mechanistic transitions underlying these data. 1) We will use enhanced sampling molecular dynamics to add bias at residues with exchange broadening. We will first evaluate conformations sampled in systems where end states of a diverse set of transition types are known, in collaboration with Paul Robustelli (Dartmouth). This will generate atomic-level training data for these systems representing transitions at the μs-ms timescale, and provide populations of states via the deposited bias. We will determine a computational scheme that enables this MD at the scale of the BMRB and make these sampled data publicly available. 2) We will use these all-atom conformations as training data for a Boltzmann generator-type framework that explicitly includes barrier heights to account for kinetics. CASP prediction challenges will provide an independent test of methods: CASP15 in 2022 had 9 pilot targets with multiple conformational states[21] and will doubtless continue to expand initiatives for multiple conformations.

Aim III: Discovering novel druggable conformations relevant to understudied diseases. The advent of methods such as AF2 has enabled many databases of predicted protein structures. However, developing therapeutics requires understanding a target's alternate states and kinetics.[22,23] I previously used AF-Cluster[11] to identify a novel alternate state for a pathogenic protein: a secreted oxidoreductase from *M. tuberculosis*, which we verified indeed occupies two stably folded states. My lab will screen proteins implicated in disease for alternate conformations to provide a more complete set of models of their conformational states. I will maintain my collaboration with Dr. Gloria Regisford and Dr. Ashley Oyewole Andrea at Prairie View A&M University, to engage them and their students in identifying overlooked diseases and pathogens to direct our efforts. We can start immediately with existing methods including AF-Cluster and apply improved methods from Aim II. We will use computational pipelines to screen structures for putative drug-binding sites[24] and will characterize candidates in conjunction with fragment libraries with NMR[25].



Future directions. With our advances in Aim I to rapidly characterize dynamic proteins, improved prediction methods from Aim II, and discovering dynamic proteins in novel contexts from Aim III, there are many more questions my lab will be poised to tackle. For instance, can we map all the ways dynamics evolve in protein families, and can we apply these rules to create designer enzymes? Can we predict the structures and actions of proteins far from equilibrium? Our work will ultimately help us reach a deep and general understanding of how life encodes and uses motion.