Towards Molecular Dynamics as a Tool for Assessing Protein Designs for Stability and Function

Christian Phillips

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2024

Committee:

Dave A. C. Beck

Joshua C. Vaughan

Program Authorized to Offer Degree:

Chemistry

University of Washington

Abstract

Towards Molecular Dynamics as a Tool for Assessing Protein Designs for Stability and Function

Christian Phillips

Chair of the Supervisory Committee:
Dave A.C. Beck
Chemical Engineering

Modern computational resources have revolutionized the way scientists understand the sequence-structure relationship. Combinations of AlphaFold2 predictions and bespoke machine learning models can generate variance in protein sequences targeting a desired characteristic. To understand and ground the success of these models, molecular dynamics simulations can be used to screen proposed mutants for desired characteristics and function. In this study, molecular dynamics simulations are used to validate outputs from NOMELT, a large language model targeting protein thermostability, and propose a novel, computationally designed thermostable red-emitting fluorescent protein. Demonstrated by established molecular dynamics campaigns used for assessing protein stability and thoughtful structural analysis for two use cases, NOMELT is capable of increasing the melting temperature of a given protein sequence while maintaining complex protein structure and function.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF EQUATIONS

ACKNOWLEDGEMENTS

I would like to thank my family and friends for all their support and patience during my entire college career. I could not have done it without my parents, the support of Pfaendtner Research Group members, and encouragement from my friends.

Chapter 1. INTRODUCTION

The proteins in this study have a rich history with a connection to the University of Washington. In this chapter, I will describe a brief history of protein design in general, how it relates to the proteins in this study, and the tools that I use to study them.

1.1 PROTEIN DESIGN

Proteins are the complex machinery of life that perform a variety of crucial and unique functions throughout nature. From the simplest prokaryotes to the human proteome with 20,000+ proteins, these proteins are constructed using the same 20 amino acids. The protein sequence, otherwise known as the primary structure, dictates the 3D shape of a protein–composed of secondary, tertiary, and quaternary structure–that is tied directly to the protein's function.[1,2] Four billion years of evolution has led to an incomprehensibly vast collection of proteins that are unique to a family of organisms, homologous proteins that perform the same function in very different organisms, even intrinsically disordered proteins, a growing family of proteins that are rapidly changing how we understand the sequence-structure relationship.[1,3,4] Rarely can scientists replicate the efficiency, specificity, and unique functions that nature has been doing "effortlessly" for millennia via proteins.[5,6] We, as humans, simply do not understand enough about the sequence-function and structure-function relationship of nature's universal protein language.

This is not for lack of trying. Nobel laureate Christian Anfinsen posited the sequence-function relationship in 1973 after studying the denaturization and subsequent refolding of ribonuclease A.[2] Since then, researchers and engineers have shown that small changes in primary structure can have profound effects on protein function, leading to mutated variants that function

differently from the wildtype starting sequence.[7–11] These studies are often time- and resource-intensive, require extensive knowledge of the reaction pathway or biological function, and are often limited to a researcher's chemical intuition and imagination. Until recently, directed evolution campaigns, circular permutation mutagenesis, and other wet-lab techniques were the most efficient way to study the sequence-structure relationship.[12–14]

In recent decades, there has been an explosion of resources for computational researchers tackling this structure-sequence problem on all fronts. For example, RCSB's Protein Data Bank is an open-source database with over 200,000 3D protein structures and counting; UniProtKB contains over 60 million protein sequences; and the National Institute of Health's GenBank has nearly 250 million genetic sequences, many of which encode for primary structure of proteins.[15–17] These databases, along with exponential increases in computation speed and increasing access to high-performance computational resources, have revolutionized our approach to understanding the sequence-structure relationship. Combining these computational resources and wealth of open-source data has led to programs like Google's AlphaFold (AF2), an ML model that claims to predict a protein's structure from a sequence.[18] This represents a great leap forward to unraveling the mysteries of nature's nearly universal protein sequence language but is certainly not the last word in *de novo* protein design. AF2 is simply one of many powerful tools that can be used in tandem to bolster our understanding of how changes in protein sequence can be leveraged for the next generation of materials, medicines, catalysts, and carbon sequestration.

1.2 Molecular Dynamics

Among the most powerful computational methods to help us understand the structure-function relationship is molecular dynamics (MD) simulations. On the surface, this technique can visualize protein conformational changes and drug binding events, sample thermodynamics ensembles, and

much more with atomistic resolution. Under the hood of all MD codes lies a simple Newtonian integrator leveraging classical physics, statistical mechanics, and Monte Carlo algorithms.[19–21]

MD is a powerful tool that allows for the 3D visualization of proteins in an interpretable way but has many shortcomings that can be augmented or improved on with other techniques. The primary shortcoming of MD experiments are the timescales that can be reasonably sampled by modern computational resources, traditionally limited by compute power available to researchers.[22] After the forces applied on each atom–a function of bond lengths, atomic charges, dihedral angles, and more–the 3D location of each atom is updated each timestep, meaning computation time increases as atoms are added to a system.[19,20] Modern MD simulations of modest size can achieve speeds over 100 nanoseconds per day, representing a small fraction of the minutes to hours some biological processes can proceed over. Simply put, atomistic-level detail of a biological function comes at the cost of being able to visualize snippets of a reaction or the chance of observing a rare event during a standard, unbiased MD simulation.[22,23]

While a thorough overview of MD is beyond the scope of this thesis,[19,20,24] there are many techniques and innovations that circumvent the classic sampling issue and take advantage of increasingly powerful computational resources. Enhanced sampling techniques, such as metadynamics and all related acronyms, are non-equilibrium sampling techniques that do thermodynamic work on a biological system to determine the most thermodynamically favorable configuration or conformation.[25] After reweighting to account for the added bias, a Gibbs free energy surface can be generated, and the most stable, thermodynamically favorable state can be calculated and assumed occupied on long enough timescales.[26] Modern MD engines and plugins such as GROMACS, LAMMPS, OpenMM, and many more are optimized to run on parallelized computer clusters, graphics processing units, and even bespoke computers for MD simulations.[27]

While these advancements have enabled researchers to study larger systems on previously inaccessible timescales, adequate sampling of the conformational space and properly parameterized force fields are still paramount for MD studies.[28–30]

1.3 NOMELT

An acronym for Neural Optimization for Melting-temperature Enabled by Leveraging Translation, NOMELT is a deep-learning model developed by colleague and Beck Research Lab alumnus Evan Komp.[31] NOMELT is an encoder-decoder large language model (LLM) trained on 4.3 million pairs of homologous low- and high-temperature prokaryotic protein sequences. This approach is novel–leveraging known trends in the distribution of amino acids for both mesophilic and thermophilic proteins[32,33]–and is designed to take in one mesophilic protein sequence, evaluate a library of proposed mutations, and return the most likely, thermostable variant of the input sequence and predicted AF2 structure. Where optimization and *de novo* designs of thermostable proteins are usually unique design problems in themselves, NOMELT demonstrates it is possible to translate from mesophilic to thermophilic sequences, similar to natural language processors like Google Translate that can recover meaning and syntactical information from one language to another.[31,34] Based on the ProtT5 foundational protein language model, NOMELT has an organized understanding of protein sequences found throughout nature at different evolutionary stages.[18,31,35]

More relevant to the focus of this thesis, however, is the training data used for NOMELT where optimal growth temperature (OGT) is a proxy for melting temperature. This well-established metric for living organisms allows for a broad, comprehensive dataset with enough datapoints necessary to train LLMs.[33,36] Using OGT as a stand-in for melting temperature operates under the assumption that to survive, an organism must have fully functional, therefore

properly folded, proteins to perform biological functions. The fluorescent proteins studied in this thesis have a very exact definition of a folded, functional state related to a critical biological function and can be quantified with computational methods.[37–40] Furthermore, a link between the OGT of an organism and the temperature range of fluorescence activity of proteins from that organism has been posited in literature, making OGT an especially relevant metric.[38,41] This will be elaborated on further in later chapters, including both applications and interpretations of outputs and use cases.

1.4 Fluorescent Proteins

The first fluorescent protein (FP) discovered by humans, known colloquially as green fluorescent protein (avGFP), was isolated from *Aqueorea victoria,* a bioluminescent jellyfish that lives in the waters off the west coast of North America.[42] Researchers Osamu Shimomura, Martin Chalfie, and Roger Tsien discovered avGFP and aequorin, a blue bioluminescent protein, from the "squeezate" of thousands of *A. victoria* jellyfish at the University of Washington's Friday Harbor Laboratory in 1962.[41–43] Both avGFP and aequorin are used in a biological example of Förster resonance energy transfer (FRET) to produce flashes of blue light on the release of Ca^{2+} ions.[42] It was not until the DNA encoding for avGFP was cloned by Doug Prasher in 1992 that the possible applications and special properties of avGFP were realized. [44]

An 11-strand β-barrel protein, avGFP is capable of fluorescence without any cofactors or substrate. Centered in the β-barrel motif is a chromophore that forms autocatalytically with dissolved molecular oxygen that is responsible for fluorescence process (**Figure 1**).[39,45] Cyclization of residues Ser65-Tyr66-Gly67 form a highly conjugated π-system in a small molecule, known as the chromophore, capable of absorbing and emitting red-shifted light in the visible spectrum (**Figure 1c)**.[10,41,42] The robust hydrogen-bonding system protects a concerted

proton transfer from the chromophore to a coordinated water molecule in the center of the protein upon excitation of light, giving off 507nm green light.[46] Crucially, avGFP isolated directly from *A. victoria* has a low brightness, as well as limited fluorescence above 30°C.[10,47]

The first examples of protein design of FPs addressed these shortcomings, paving the way for the imaging experiments FPs are applied in today. The brightness of avGFP was improved after researchers took inspiration from an FP found in *Renilla reniformis*; substituting a serine for a threonine at the 65th amino acid (S65T) nearly doubled the brightness and significantly reduced maturation time.[47] This is an example of a direct mutation of the chromophore's parent residues resulting in significant changes to the photophysical properties of FPs. Not long after the S65T mutation of avGFP came another substitution that enhanced the thermal properties of the brighter, mutant protein. Performing a proline substitution (S147P) allowed for reliable fluorescence at the optimum growth temperature of mammals, 37°C.[10] This mutation of a residue geometrically adjacent to the chromophore had little effect on the photophysical properties of the FP, yet it profoundly impacted the practicality of GFP and added thermal stability to an already stable protein scaffold. Modern examples of FP protein engineering operate under this same principle, screening huge libraries of single-point mutations or substitutions looking for an improvement on a desired feature.[48,49] These exhaustive approaches have yielded over 1,000 different variants over the past 25 years, each one a small combination of mutations different from a handful of biological examples or any of their familial, upstream mutants.[41,50] Development of green and yellowish-green FPs has plateaued after several hundred variants have been engineered to serve specific purposes, giving microscopists numerous options for nearly any imaging experiment.[41,43]

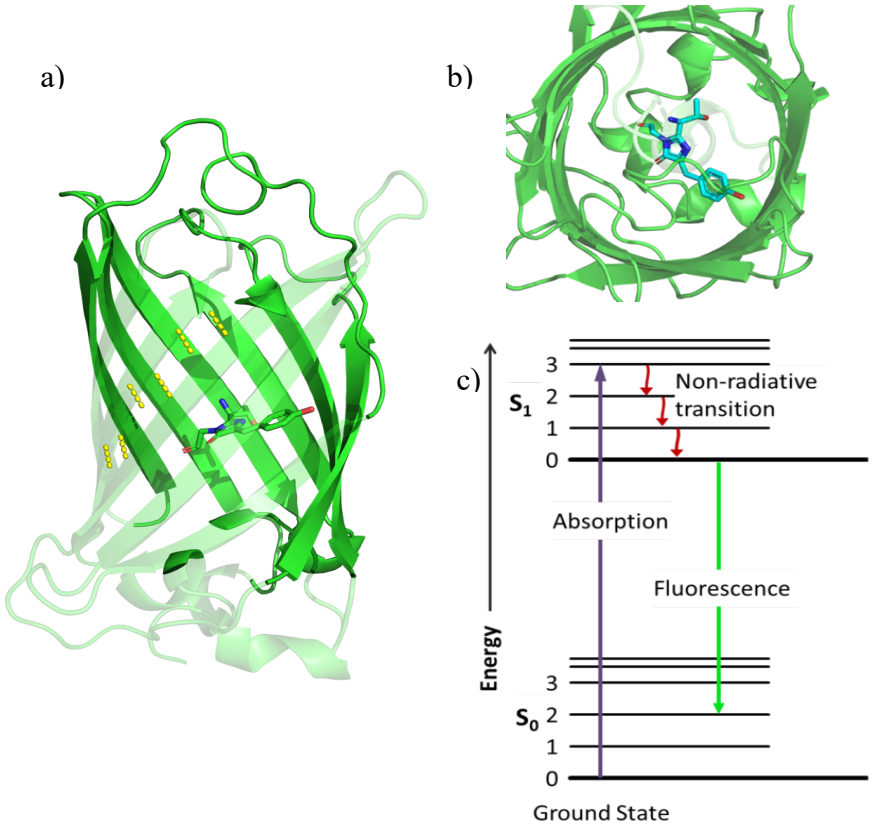


Figure 1: Features of avGFP Facilitating Fluorescence. (PDB:1EMB)

(a) Cutaway of avGFP with select interactions shown in one strand of β-barrel. (b) Top-down view of avGFP showing β-barrel cavity, chromophore in blue. (c) Jablonski diagram depicting vibronic spectrum and radiative action of fluorescence courtesy of Creative Commons.

Over the next several years, many other green-emitting FPs (GFP) were isolated from other *Aequorea* species, crustaceans, lancelets, and Anthozoa reef corals. The first example of red-emitting FPs (RFP), DsRed, was discovered and isolated from non-bioluminescent Anthozoa corals in 1999 and represents the starting point for most RFPs used today, including the popular "mFruit" series.[51–53] Another example of naturally occurring RFP, equaRFP, is found in *Entacmaea quadricolor,* a sea anemone commonly found in the Indo-Pacific.[54] These two FPs and their downstream variants represent the majority of RFPs used today.[50] With few examples found

naturally and less inspiration from nature, the development of RFPs is considerably slower than their GFP counterparts.[41,55] This has left experimentalists with underdeveloped, less specialized RFPs that are crucial for modern microscopy experiments involving extremophiles or harsh experimental conditions, including imaging or DNA fluorescence *in situ* hybridization (FISH) experiments.[56–58]

FPs have revolutionized many disciplines of science, but most relevant to this project are biomedical imaging, tracking of protein-protein interactions and dynamics, and FISH. These methods use photons emitted from FPs as a signal, making FP properties such as quantum yield (a ratio of photons emitted per photons used to excite, think efficiency) and brightness critical for achieving an adequate signal-to-noise ratio.[41,56,57] RFPs are ideal for deep mammalian tissue imaging (>100μm tissue depth) primarily because <650nm wavelength light is strongly absorbed by hemoglobin in the blood; more photons emitted by the FP that reach a sensor, producing a stronger signal and a higher-resolution image.[41,57,59] Many wildtype FPs have limited fluorescence at mammalian OGT, 37°C, hence early examples of thermostable avGFP variants.[10] Interestingly, RFPs tend to maintain fluorescence at higher temperatures, 37°C and above.[41,50] This has been attributed to the difference in OGT from *A. victoria*, native to the cool waters of the eastern Pacific, and OGT of stony corals and anemones found in the Indo-Pacific that have historically yielded RFPs.[52] While stable enough for most *in vivo* experiments, FISH experiments denature FP-tagged DNA probes complementary to target DNA regions at temperatures upwards of 70°C and harsh chemical environments, necessitating an especially durable and robust FP.[56] FISH is capable of tracking intra-nuclear movement of DNA, such as mitosis and DNA replication, but is becoming increasingly popular for clinical diagnoses and research of hereditary diseases.[58] For these reasons,

the goal of this thesis is to propose and validate a thermostable variant of an existing monomeric RFP sequence with machine learning models and MD simulation campaigns.

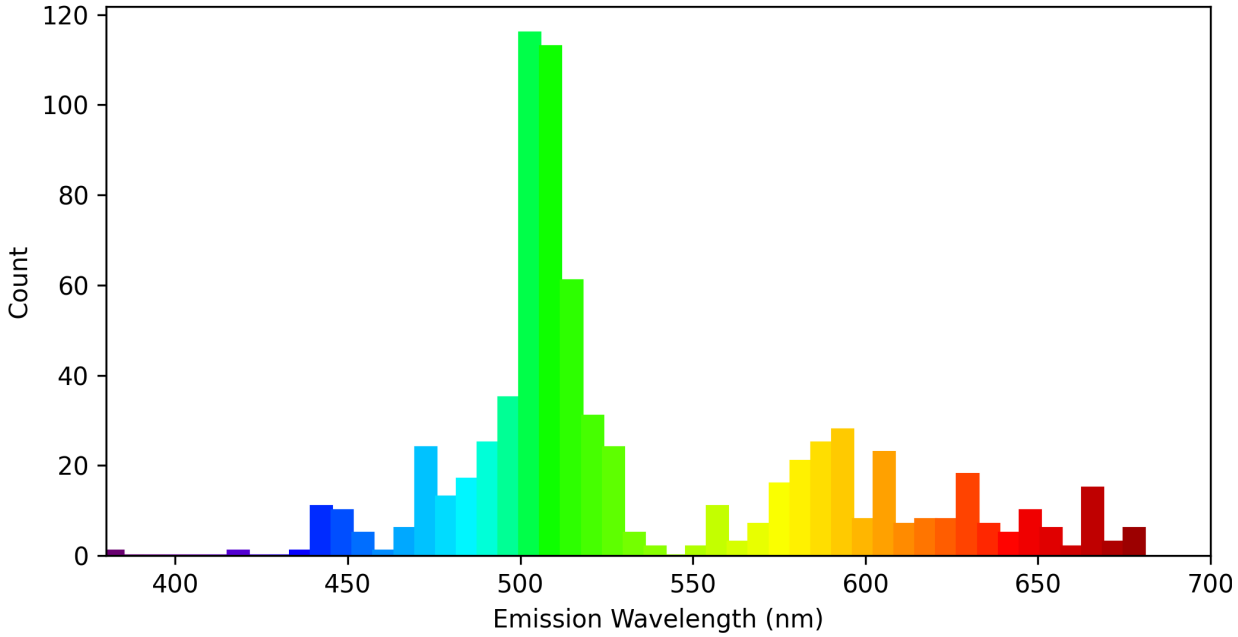


Figure 2: Non-comprehensive Histogram of FP Emission Wavelengths.[50]

Histogram with bins colored by primary emission wavelength for FPs found in open-source database, FPBase.

Chapter 2. NOMELT AND THE ENGRAILED HOMEODOMAIN

To validate the success and provide a ground truth for the NOMELT model, I performed an MD simulation campaign to thoroughly and quantitatively rank wildtype input sequences, intermediate and final sequence predictions from NOMELT, and thermophilic variants of the proposed in literature. To do this, I chose a small, well-studied, highly conserved (among eukaryotes) protein sequence that is used to transcribe DNA known as the engrailed homeodomain (ENH).

2.1 GROUND TRUTH AND ENH

Establishing a ground truth is critical for the development, validation, and practical usage of any machine learning model. Without validation of output data from *any* computational model, one cannot be certain that the model is accurately representing or reproducing real-life phenomena. A tangential example is the development of force fields used in MD simulations; vast libraries of atomic interaction parameters in different environments are approximated with quantum chemical calculations and benchmarked with simple, well-understood "toy systems" to ensure the MD simulations can accurately reproduce real phenomena.[28–30] To prove the hypothesis that NOMELT can accurately and reliably reproduce thermophilic homologs of an input sequence, I performed a large ensemble of MD simulations over a range of temperatures for wildtype proteins, previously engineered variants, and NOMELT output sequences. While ground truths are traditionally experimental wet lab results, probing the melting temperature of a single proposed variant with computational tools is more efficient, more time effective, and less prone to experimental issues that come with expressing new proteins.[60,61]

To understand NOMELT's output and quantify the model's success, we used the engrailed homeodomain (ENH) as an input sequence and test case. Facilitating gene regulation and

transcription of DNA during development, ENH is a highly conserved protein among vertebrates, albeit with a low sequence similarity.[62–64] Despite this, ENH homologs have a similar length (~60aa) and identical structural motifs known as a helix-turn-helix, or 3 α-helix bundle.[63] The third α-helix and N-terminus of ENH embed into the major and minor grooves of DNA, respectively, to "read" DNA. This low sequence similarity yet identical function is logical; only four base pairs are found in the DNA double helix, and the combination of sequences that form the functional helix-turn-helix is vast.[62,63] Simply put, the large sequence space of proteins capable of reading DNA and folding into a 3D shape complementary to DNA's double helix makes ENH nearly ubiquitous in vertebrates and beyond.[62,63] This includes some organisms known as extremophiles, capable of surviving at temperatures upwards of 90°C and extreme pressures found on geothermal vents, volcanic hot springs, etc.

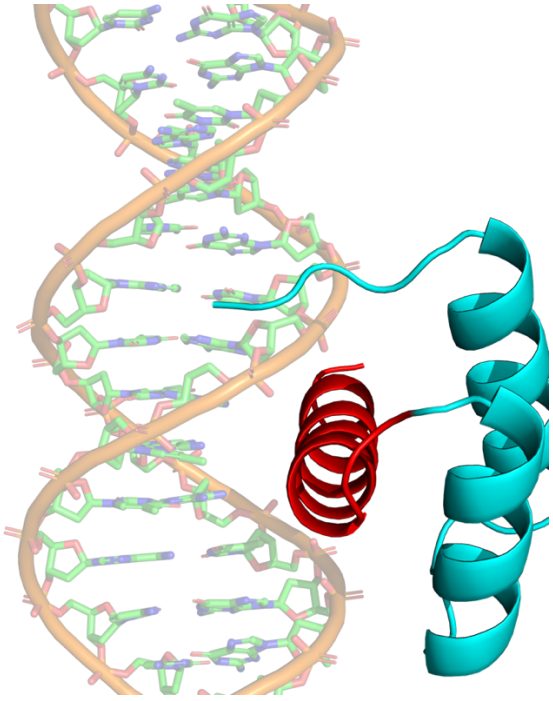


Figure 3: ENH Bound to DNA (PDB:3HDD[64]).

ENH (cartoon) interacting with DNA (sticks) with recognition helix (red) embedded in major groove of DNA. N-terminus (cyan) embedded in the minor groove of DNA.

Circling back to **Chapter 1**, these thermophiles, no matter how simple, must have proteins that remain folded and retain their function at an elevated OGT. Homologous proteins from thermophiles have a different distribution of amino acids constituting the primary structure compared to mesophiles yet similar 3D configurations to perform similar biological functions.[32,36] Clearly there is a relationship between primary structure and thermostability of a protein. ENH makes an excellent ground truth for NOMELT for two reasons: a large combination of sequences is capable of producing the simple three-helix bundle–giving NOMELT a sufficiently large permutation space to test and explore possible thermostable homologs–and many well-established examples of successful thermostable ENH variants in literature.[65,66] While ENH is not the most advanced or novel use case of NOMELT, proposing and validating a thermostable variant of a model system is an excellent system for showcasing NOMELT's translation ability and benchmarking against other techniques with the same goal: engineer a thermostable variant of a protein found in nature.

2.2 Computational Workflow and Details

Not only has ENH been subject to several thermostabilizing efforts, ENH has also been the focus of many computational studies probing unfolding and refolding of denatured proteins. There exists a wealth of papers using MD simulations to probe unfolding and refolding pathways.[24,67–70] Given ample peer-reviewed work and previous lab experience with this protein, I utilized an established computational protocol to quantify melting temperatures.[68–71]

The melting temperature of wildtype ENH is 56°C and is coveted by early computational studies for its relatively fast folding time, making long timescale simulations on older computational resources possible.[67,72] MD campaigns as recent as the late 00s report total

simulation time on the μs timescale with dated hardware, meaning a magnitude or two increase in computational time is possible with contemporary compute resources.[42,43,48] While there are competing definitions of melting points for proteins, the computational definition of melting point for this study of ENH is complete, global denaturation.

No matter how it is defined, the melting point of a protein is an ensemble property; there exist many valid unfolded or partially unfolded configurations and unfolding pathways varying in timescales between each configuration.[73,74] Because of this, melting point is impossible to accurately quantify with a single MD simulation. Even a small ensemble of independent MD simulations over a broad range of temperatures will not capture all the possible unfolded states and pathways. To overcome this, many replicas–independent simulations with randomly generated starting velocities and no memory of previous states–over a range of temperatures, from folded and stable to unfolded and dynamic, need to be performed.[68,71,74] This ensures that intermediate, metastable configurations during the many hypothesized and potentially undiscovered unfolding pathways are explored and weighted appropriately for this ensemble property. The collective variable (CV), or metric, used to quantify unfolding in this study is root-mean square deviation (RMSD) of the 3D protein structure to a reference PDB structure and is shown in equation 2-1. The 3D positions of each core residue's heavy atom's center of mass are subtracted from a reference structure, meaning a high RMSD reflects a different, i.e. unfolded, structure.

$$d(X, X_t) = \sqrt{\sum_i \sum_{\alpha}^{x,y,z} \frac{\omega_i}{\sum_j \omega_j} (X_{i,\alpha} - com_{\alpha}(X) - X_{ti,\alpha} + com_{\alpha}(X_t))^2} \quad (2\text{-}1)$$

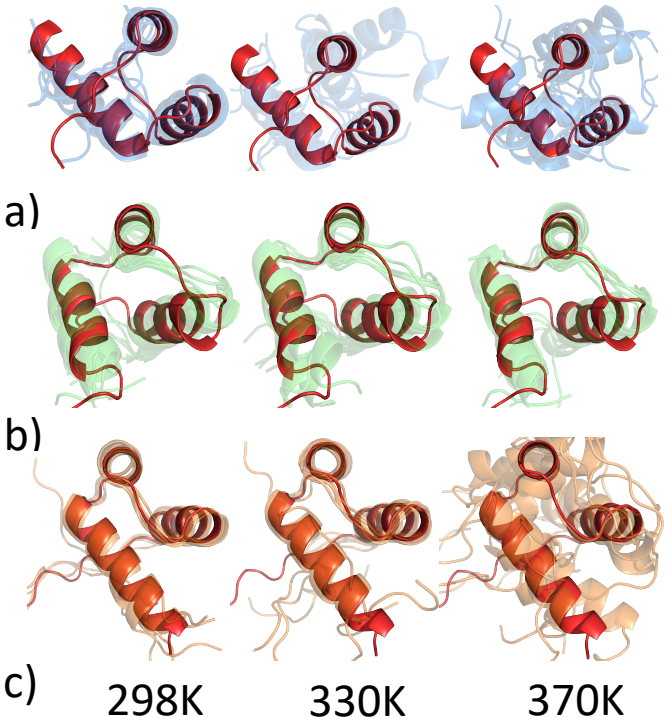
The simulation ensemble for each protein in this study was 5 replicas per temperature for 1μs each in 10K increments from 298K-370K, resulting in 125μs of total simulation time. The proteins in this study are wildtype ENH (PDB: 1ENH),[63] the proposed NOMELT variant, and a previously engineered thermostable variant from Mayo et al. with a >99°C melting temp (UVF) (PDB:2P6J).[65] All simulations were performed with the CHARMM36m force field,[75] GROMACS

2022.5[76] patched with Plumed 2.0[77,78] on Nvidia a-40 GPUs housed at the University of Washington's Hyak computer cluster. C- and N-termini were capped with NME and ACE groups with Pymol, respectively.[79] Identically TIP3P solvated and charge-neutralized cubic simulation boxes with periodic boundary conditions were subject to steepest descent energy minimization, followed by 500ps of NPT equilibration with the Brendsen barostat[80] for each system. Production runs were run in the NVT ensemble with V-rescale thermostat for 1μs each, using the Verlet cutoff scheme[81] and Particle-Mesh Ewald scheme for coulombic interactions.[82] The NPT equilibrations were run before each replica, ensuring independent, parallelizable simulations with random starting velocities.

2.3 Results and Discussion

Over the course of each replica below the experimentally derived melting temperature for wildtype ENH and UVF, thermal fluctuations increase with increasing temperature, but the protein remains globally folded and retains its function. Thermal fluctuations are expected, and thermostable variants have been known to have more flexible and dynamic backbone structures, due to the stabilizing effect of repacking core residues at elevated temperatures.[67,69,70] Beyond the published melting temperatures, a dramatic increase in RMSD is expected as the protein loses all secondary structure and function. In the context of this study, simulations beyond this point are uninteresting; thermally denatured proteins are not expected to function, and the success of NOMELT is unaffected by RMSD values past the melting temperature. However, adequate sampling beyond the melting point is important to ensure proper MD simulation setup and parameters. At or near the published melting temperature for ENH, an inflection point in the ensemble averaged RMSD values is observed as the protein begins to globally denature and structures pass through numerous unfolding pathways. Increasing this inflection point for the NOMELT variant and by how many

degrees is the key metric being used to quantify the success of NOMELT translation abilities. Furthermore, it is crucial that this inflection point for the wildtype ENH be consistent with literature values. Therefore, performing 5 replicas per 10K from 298K to 370K for wildtype ENH, a previously engineered variant, and the NOMELT proposed sequence provides the following: adequate sampling of transient unfolding/refolding pathways at multiple temperatures and timescales, a well-studied control system that reliably unfolds at moderate temperatures and MD-accessible timescales that validates MD settings and parameters, and a sanity check in the UVF system that should remain folded over the course of every temperature tested here.



d)

>NOMELT: **D**KRPHTEF**SS**AQL**A**RLKREF**NEN**RYLTEVRRQ**Q**L**SS**ELGLNE**A**QIEIWFQNKRAKIKK
>1enh: --RPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKI----

Figure 4: Final Frame Structural Snapshots of Replicas at 298K, 330K, and 370K. (a) ENH, (b) UVF, (c) NOMELT variant, and (d) sequence alignment of 1ENH and NOMELT variant with mutations in bold. Final structures, semi-transparent, are aligned with crystal reference structures (red).

Considering the output from NOMELT, 2^{14} possible mutations were suggested from a BEAM search of the training set. That is 14 binary mutations leading to 16,384 possible homologous sequences, which is an intractable number of MD simulations to run. However, 10 rounds of NSGA-ii evolutionary optimization with 10 sequences per run could increase the objective function, mAF-min, by 73%. While this function is used to assess the stability of a protein (higher value is more stable), it cannot be interpreted as a 73% increase of the melting temperature from input to output sequence. To completely relate this objective function to real metrics, a ground truth must be established.

Traditionally experimentally derived, the use of MD simulations to provide a ground truth for NOMELT is the best option for this study, primarily because of the difficulty associated with synthesizing designed proteins.[13,14] While beyond the scope of this thesis, synthesis, expression, isolation, and purification of design proteins is a resource- and time-intensive process. Because of the wealth of computational studies of ENH and engineered homologs and my research lab's experience with this system, a large ensemble of parallelizable MD simulations offered the quickest route to estimating the melting temperature of a computationally derived protein sequence. Furthermore, providing evidence that MD simulations can be used to validate predictions from ML models such as NOMELT means there are alternative and more accessible methods to providing a ground truth for future projects.

After completion and trajectory post-processing, the RMSD values for each replica were calculated with MDAnalysis 2.3[83] and ensemble averaged per temperature complete with error bars. The reference structure used in RMSD calculations for wildtype ENH, UVF, and NOMELT sequence were energy-minimized PDB structures for ENH and UVF, and the AF2 predicted structure, respectively. Note that from **Equation 2-1**, the differences in *xyz* coordinates between

heavy atoms are summed, meaning RMSD calculations that do not have the same number of atoms cannot be accurately compared. Because of this, only positions of amino acids present in the 3-helix bundle motif of wildtype ENH and their counterparts are considered. Furthermore, this increases the signal-to-noise ratio, as C- and N-termini ends are "floppy" and dynamic, where RMSD values of the 3-helix bundle is more indicative of folded or unfolded structures. Consistent with the literature, only the alpha-carbons of amino acids 10-52 are considered in the RMSD calculations for each system.[67,69] Lastly, RMSD ensemble averages for each temperature are subtracted by the mean RMSD at 298K, hence the zero RMSD at 298K for each system. This ensures no RMSD values are artificially inflated from less precise reference structures and highlights the deviations from average values that I am trying to measure.

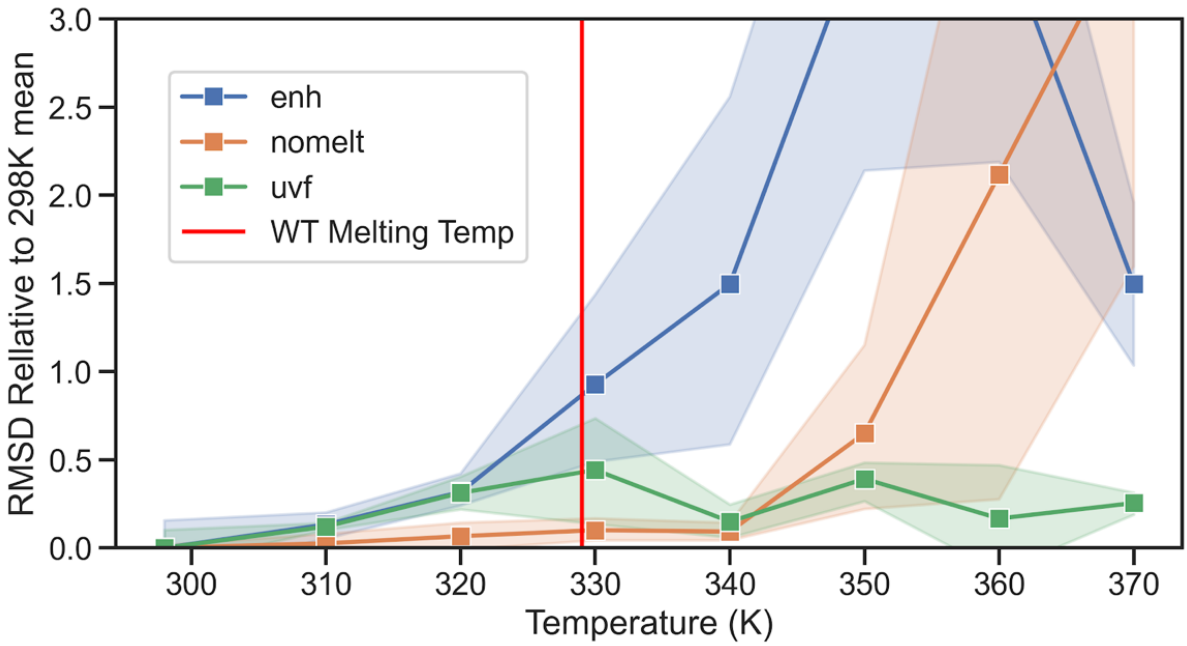


Figure 5 :Ensemble Averaged (n=5) RMSDs of ENH, UVF, and NOMELT. Normalized to 298K by subtracting values by average RMSD at 298K to remove effects of thermal fluctuations. Red line indicates experimentally derived melting temperature of ENH (56°C).

Considering the RMSD values in **Figure 5,** a sharp increase in RMSD is observed near ENH's literature melting point of 329K. This is indicative of most replicas beginning to lose the 3D structure associated with ENH's function via thermal denaturation. In contrast, there are no inflection points observed for UVF at any point, due to the melting temperature not being sampled in this study. RMSD values rapidly increase beyond the melting temperature, as the kinetic energy about thermal fluctuations is greater than the potential energy imparted from protein-protein interactions, giving ENH its secondary and tertiary structure. For the thermostabilized, NOMELT-translated protein sequence, a very similar trend is observed: inflection point centered about a melting point followed by a sharp increase towards complete, global denaturation. Consistent with the hypothesis of this study, the inflection point in RMSD for the NOMELT sequence occurs at a higher temperature than the wildtype ENH. Critically, this inflection point is approximately 20K higher than the ENH inflection point with statistical significance. This agrees with the overall hypothesis of NOMELT; thermophilicity of protein sequences can be learned and protein sequences can be translated from mesophilic to thermophilic, analogous to translation between human languages. Furthermore, a 20K increase in melting temperature is significant, and a similar increase to a catalytic enzyme's melting temperature could be enough to push a biologically relevant enzyme towards industrial applications.[11,34] Compared to the literature variant UVF, NOMELT is not quite as successful at raising the melting temperature as a bottom-up redesign of an entire sequence. This is not completely unexpected, nor does it invalidate the success of NOMELT. Understanding where this application of LLMs stands relative to other examples of protein design is important to engineers considering engineering proteins of their own and contributes to the progeny of LLM applications in biology and chemistry. The novelty of NOMELT's application of LLMs to target thermophilicity is demonstrated via MD simulations of

input and output sequences, with success compared to contemporary protein design techniques found in a comparison with UVF.

Chapter 3. ADVANCED APPLICATIONS OF NOMELT

The MD simulations performed on ENH, proposed NOMELT sequences throughout development of the model, and literature variants were instrumental to developing and validating NOMELT. One of the primary challenges faced during the conception of these MD studies was the choice of CV and how to appropriately define the melting point of a protein. RMSD is a coarse metric, resulting in many degenerate configurations with the same RMSD values yet vastly different structures. Depending on the context, "melting point" can be the temperature at which 50% of a protein sample is globally unfolded, maintaining no semblance of folded structure and non-crystalline. Similarly, yet very distinct in the case of fluorescent proteins, the melting point can also be defined as the temperature at which protein activity begins to rapidly decrease.[73,74]

ENH is a very simple protein with its *entire* secondary and tertiary structure imparting function. Small changes in the orientation of ENH's helices are expected and are likely critical to the protein family's broad specificity towards DNA's double helix.[62,63] FPs are much different; the fluorescence (activity) of FPs is largely dependent on the identity and internal environment of the chromophore, encased by the β-barrel scaffold conserved in all FPs.[39,40] In contrast to ENH, a vast majority of the protein-protein interactions of the β-barrel may remain intact while the internal environment of the chromophore is disturbed, limiting or eliminating fluorescence altogether. Assuming the same approach as ENH, this would be associated with a negligible increase in RMSD and very low signal-to-noise ratio. Additionally, using FPs as an input sequence offers an additional opportunity to understand if NOMELT can preserve complex protein function while promoting thermostability and contribute to a notoriously difficult and slow area of protein development in FPs.

3.1 Choice of Fluorescent Protein

To validate the success of NOMELT and understand its inputs/outputs for FPs, hypotheses stemming from the ENH study will be tested, and the sequence space of FPs will be thoroughly explored. The goal of this study will be to increase the melting temperature of an RFP without affecting fluorescent properties while learning more about NOMELT's ability to translate protein stability and maintain function. As illustrated above, FPs are vastly different from highly conserved proteins such as ENH, so it is important to choose an RFP that can be reasonably improved and understood by NOMELT. After these careful considerations, a monomeric variant of the protein named Eos, after the Greek goddess of dawn, was chosen as a target to modify (mEos).[84]

The primary consideration for using mEos as an input sequence for NOMELT is a lack of thermostability of the monomeric variant, specifically. With an excitation/emission wavelength of 569nm/581nm, respectively, this is considered a photoconvertible, near-red FP. An irreversible, UV-induced backbone cleavage of the chromophore gives a green and red form of mEos, with emission wavelengths as 516 and 581nm, respectively. This color change is useful for some applications, but I will strictly be focusing on the red structure of the protein. Originally isolated from *Lobophyllia hemprichii* and generally stable at 37°C, a monomeric variant was prepared and found to have poor fluorescence above 30°C.[84] This monomeric variant, intended to be used for single-molecule experiments where oligomerization is an issue, such as FRET, is mutated with V123T and T158H from the wildtype. Introducing these polar, charged side chains on the exterior of the β-barrel disrupts the quaternary structure found between subunits of the wildtype tetramer and appears to have little effect on the photophysical properties and physical properties other than melting temperature.[84]

Compared to ENH, the sequence space of FPs is much smaller with a few conserved structural motifs. Residues involved in chromophore formation are highly conserved and correlated with excitation and emission wavelength.[45,46] FPs with a tripeptide sequence similar to avGFP, Ser65-Tyr66-Gly67, are often green, while RFP chromophores owe their red-shifted emission wavelength to the extended the π-conjugation with His-Tyr-Gly tripeptide.[46] However, light-induced backbone cleavages, chromophore isomerization and protonation states, and other irreversible chemical changes are subject to changing photophysical properties of any FP. The ubiquitous β-barrel for FPs is the most variable part of the sequence, and residues associated with the β-barrel are largely inconsequential to photophysical properties, such as emission wavelength. β-barrel stability and OGT have been linked in literature, as this scaffold must stay intact at

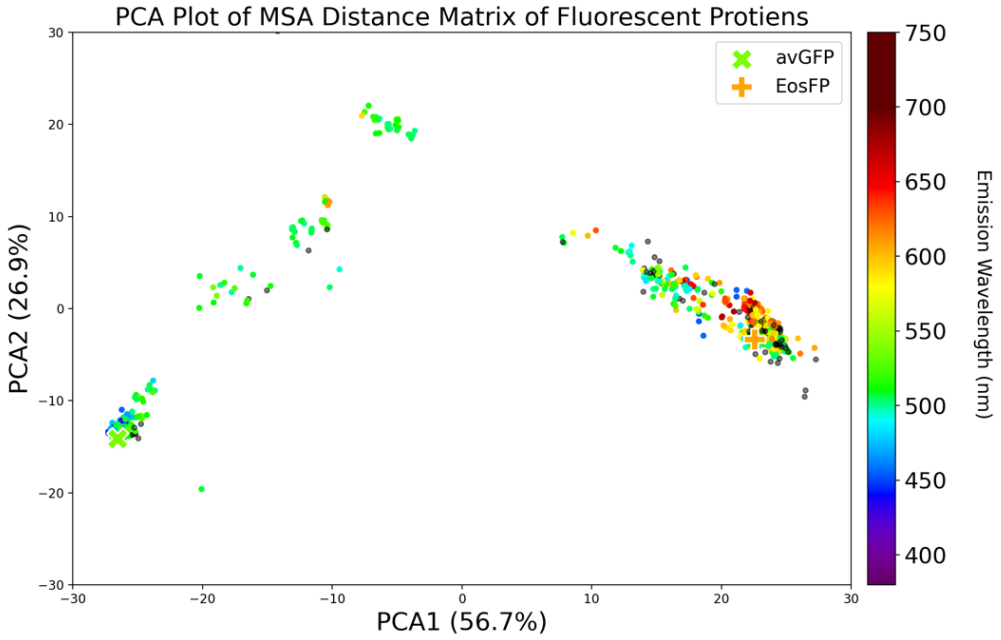


Figure 6: Principal Component Analysis of Multiple-Sequence Alignment.

Dimensionality reduction of Multiple Sequence Alignment of FP sequences from FPBase. Explained variance by each PCA is listed in axis labels. Points are colored by emission wavelength.

elevated temperatures to ensure fluorescence, assuming that the action of fluorescence is/once was an evolutionarily advantageous characteristic.[41]

The principal component analysis (PCA) plot in **Figure 6** shows a dimensionality reduction, colored by emission wavelength, of a Clustal-W[85] multiple sequence alignment (MSA) of 900 sequences from FPbase–a community-editable and open-source database of FPs founded in 2019[50]–where clusters of proteins and their familial mutants can be observed. The tight cluster in the lower left represents avGFP and all familial mutants, including the variants discussed in **Chapter 1**. These variants have few point mutations and highly similar sequences to the wildtype avGFP, as the proximity of the points in the PCA plot would suggest. Just out of view lies a sparse cluster in the upper left and consists of FPs that require cofactors and are not autocatalytic. These proteins are consistently larger, have a more complex fluorescence mechanism, and are beyond the focus of this study. The more diffuse, elliptical cluster on the lower right represents proteins like DsRed, including similar wildtypes from stony corals and their familial mutants. Qualitatively, these clusters appear to be correlated with emission wavelength; more diversity in red-shifted emission wavelengths are seen in the lower-right elliptical cluster than the lower-left cluster centered around avGFP. Because the residues associated with the β-barrel far outnumber those responsible for the chromophore, **Figure 6** supports the idea that β-barrel residue sequences are consistently different between RFPs and GFPs. mEos lies at the center of the cluster dominated by RFPs, suggesting that a couple of mutations to the sequence would still result in a red-shifted FP. This is critical to further test NOMELT's ability to maintain function, i.e., emission wavelength color, while increasing thermal stability. Furthermore, mEos mostly represents a wildtype protein, albeit with two point mutations to ensure limited monomerization. While testing NOMELT's ability to translate human-engineered sequences to a higher melting temperature, the model

performed poorly by suggesting few, if any, mutations to thermally stabilized mutants. Possible explanations for this behavior include the classification of meso- and thermophilic sequences being defined as <40°C and >40°C, respectively, in the training set.[31,36] Because the training set contains protein pairs, NOMELT can relate mesophilic sequences to thermophilic sequences reliably but may struggle to translate an already thermophilic sequence to an even more thermophilic sequence. Additionally, sequences of human constructs could be introducing variation in primary structure that does not occur naturally or appear in the training set, limiting NOMELT's ability to propose a library of possible sequences.

3.2 Computational Setup

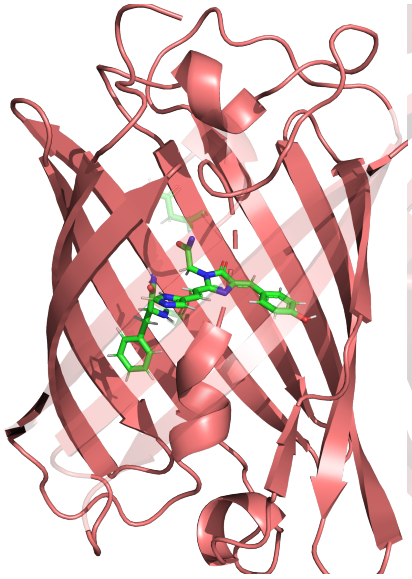
The computational tools and environment used in this study are nearly identical to those outlined in **Chapter 2**. Equilibrations were instead run in NVT then NPT ensembles for 500ps and 800ps, respectively, while production runs were run in the NPT ensemble for 800ns. Longer equilibration times were chosen to relax unfavorable or unrealistic starting configurations stemming from the computational setup of these proteins. The NPT production ensemble was chosen to correct for possible density artifacts, ensuring uniform density as water infiltrates the β-barrel cavity at elevated temperatures. For the computational model of mEos, one subunit of the red form PDB structure for the tetramer EosFP (PDB:2BTJ) was mutated via ChimeraX.[86,87] This red structure differs from the green form by an irreversible, UV light-induced backbone cleavage between residues Phe61 and His62 and side-chain cyclization of the chromophore's parent residues.[84,88] Both the cleaved phenylalanine residue and chromophore were parameterized independently. This results in subunits A and B and being held together with quaternary structure in both mEos and the NOMELT variant.

While the chromophores are modified amino acids, the interactions and atom types are different to reflect the irreversible chemical changes resulting from chromophore formation. The force field parameters–including atom types, bond force constants, simple bond angles, and proper and improper dihedrals–were created with CGenFF.[89] Redundant or equivalent interactions already characterized by the Charmm36m force field were used where possible.[75] For assigning partial charges to the nuclei of modified residues, the Gaussian software package[90] was used with the Hartree-Fock 6-31G(d) basis set assuming a net neutral charge for every chromophore. To most accurately replicate the environment inside the chromophore, peptide bonds were substituted with hydrogen atoms, and the difference in charge added by these capping hydrogens was distributed through the remainder of the atoms in the modified residue. The chromophore for mEos and the NOMELT variant are identical but differ slightly from some literature structures. The exact protonation state and structure of chromophores in FPs is heavily studied,[30,39,40] and there exist many equivalent resonance structures and ionic states in varying states of equilibrium.[91,92] This is especially true for RFPs with red-shifted, heavily conjugated π-systems, making the individual setup of every resonance form of RFP enormously time-consuming. To ensure adequate comparisons between simulations in literature and aide in CV definition, the exterior ring was treated as a hydroxyphenyl group as observed in **Figure 1** and **Figure 7**. The chromophore environment of mEos is depicted below in **Figure 7** and is identical to the NOMELT variant. To prepare the mEos sequence optimized by NOMELT, the AF2-predicted structure was grafted onto the modified side chains mentioned above and properly connected to replicate the environment after autocatalytic formation of the chromophore in the red form.

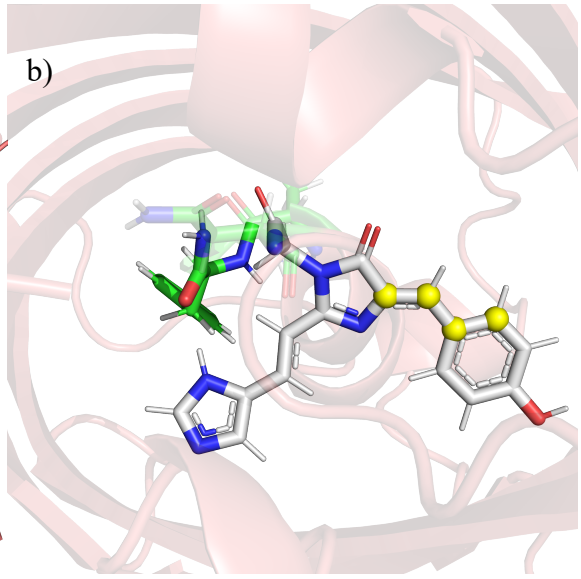
Tasking NOMELT with suggesting a library of possibly thermostabilizing mutations and searching this space for the most thermostable variant while maintaining function via a

complicated, autocatalytic pathway is a large leap in complexity from the ENH application. Furthermore, the fluorescence mechanism of FPs means global denaturation of the scaffold will happen well after the protein loses its function, meaning a new CV to quantify unfolding needs to be defined. While many examples of MD simulation campaigns of GFP and its closely related analogs exist in literature,[30,39,40,93–95] there are far fewer MD studies of RFPs. These GFP studies

a)



b)



c)

```
NOMELT_______ ---MKPDMKINLRMEGNVNGHHFVIDGDGTGKPFEGKQSMDLEVKEGGPLPFAFDILTTA
mEos_________  MSAIKPDMKINLRMEGNVNGHHFVIDGDGTGKPFEGKQSMDLEVKEGGPLPFAFDILTTA

NOMELT_______ FHYGNRVFAEYPDHIQDYFKQSFPKGYSWERSLTFEDGGICIARNDITMEGDTFYNKVRF
mEos_________  FHYGNRVFAEYPDHIQDYFKQSFPKGYSWERSLTFEDGGICIARNDITMEGDTFYNKVRF

NOMELT_______ HGTNFPANGPVMQKKTLKWEPSTEKMYVKDGVLTGDIHMALLLEGNAHYRCDFRTTYKAK
mEos_________  HGTNFPANGPVMQKKTLKWEPSTEKMYVRDGVLTGDIHMALLLEGNAHYRCDFRTTYKAK

NOMELT_______ EKGVKLPGYHFVDHCIEILSHDKDYNKVKLYEHAVAHSGLPDNARR
mEos          EKGVKLPGYHFVDHCIEILSHDKDYNKVKLYEHAVAHSGLPDNARR
```

Figure 7: Key Features of mEos and CV Definition.

(a) Cutaway of mEos (PDB: 2BTJ) showing internal chromophore in solid stick representation. (b) Top-down view of chromophore (gray) and atoms defined in CV definition (yellow). Semi-transparent stick representations in green show termini of subunits A and B. (c) Sequence alignment of NOMELT variant and mEos wildtype, mutations in bold.

often relate a proper dihedral angle within the π-conjugated system to planarity of the chromophore,[39,93,95] as coplanar configurations of the hydroxyphenyl and imidazoline rings are associated with bright, reliable fluorescence.[95,96] Twisting about the double bond between these two rings, sometimes called a hula twist, opens the door to more energetically favorable nonradiative pathways as the system relaxes to a vibronic ground state.[93,96] The atoms defined in this dihedral are shown in **Figure 7** in yellow and calculated every 10ps with Plumed2's TORSION function.[78]

3.3 NOMELT Output, Results, and Discussion

To understand if NOMELT can recreate a thermostable sequence while maintaining protein function and structure of the input sequence, the mutation spaces for ENH and mEos input sequences should be compared. As mentioned in **Chapter 2**, FPs are fundamentally different than ENH with structural motifs unique to FPs leading to a much more stable overall protein structure. Where NOMELT suggested 14 possible mutations to ENH, only 3 mutations were proposed to be impart thermal stability; a deletion of the first 3 residues, I4M, and R148K was projected to be the most thermostable permutation. Fewer suggested mutations support the hypothesis that NOMELT has an inherent understanding of protein structure and stability, despite only having prokaryotic protein sequence pairs as a training set. As thermophilic protein sequences utilize a different distribution of amino acids to maximize hydrogen bonding and minimize hydrophobic surface area, mEos appears to already have many of the features that impart stability to thermostable proteins. Alternatively, few mutations could simply represent a failure mode of NOMELT that is struggling to propose mutations that promote thermostability because the cutoff for a thermophilic sequence in the training set is significantly below the globally unfolded definition of melted FPs. While training data artifacts may be the cause of the deletion of three N-terminus residues and

methionine insertion, I believe the R148K mutation, swapping an arginine with a lysine, is evidence of NOMELT's capability to "read between the lines" of the primary structure of proteins when suggesting thermostable sequences.

Upon inspection of the simulations and ensemble-averaged dihedral values as a function of time for mEos and the NOMELT variant, it is unclear what effect this mutation had directly on the chromophore. This residue lies on the exterior of the β-barrel far from the chromophore, unlike the S65T mutation from **Chapter 1**, but not all that different from the thermostabilizing mutation S147P. Looking at just the dihedral data in **Figure 8** and considering the definition of folded for this study, it is not immediately obvious the effect the R148K mutation has on thermostability. Both variants appear to begin unfolding at 350K around 300ns, meaning the R148K mutation does

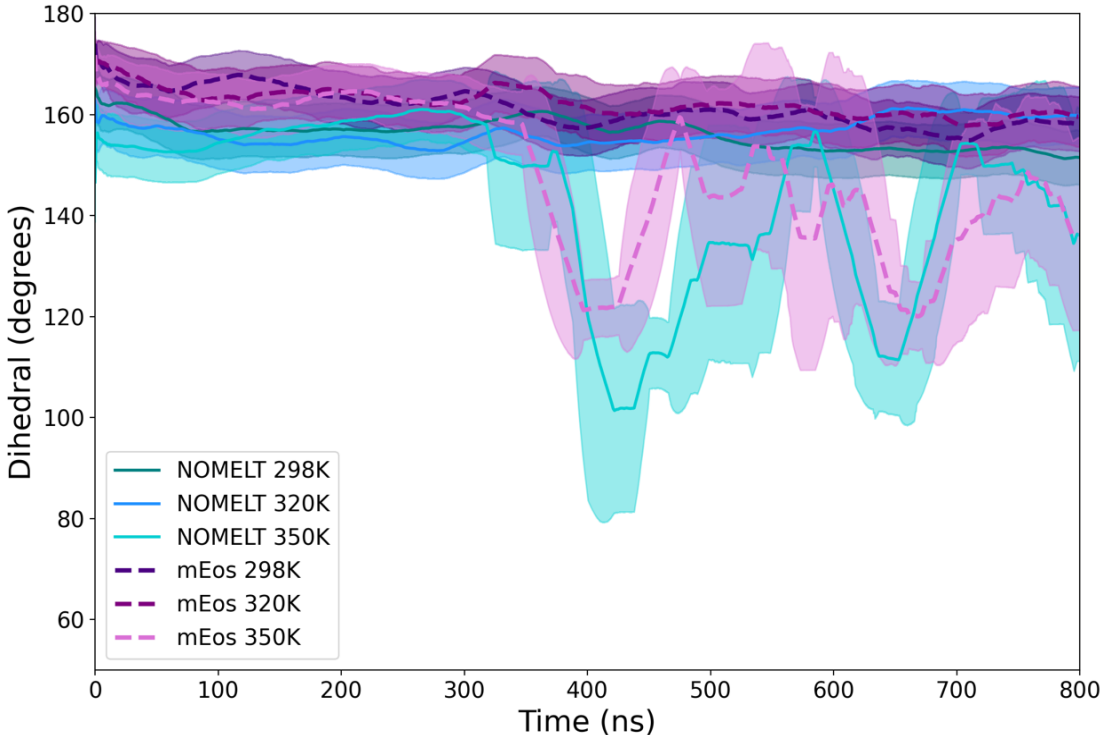


Figure 8: Ensemble-Averaged (n=3) Dihedral Angles of mEos and NOMELT.

Moving average of all replicas and temperatures as a function of time. Perfectly planar chromophore is defined as 180°.

not delay unfolding, nor does it impart stability to the chromophore at identical elevated temperatures. Furthermore, there isn't a single statistically significant finding that can be gathered here as all error bars at all temperatures for both systems are overlapping at some point. Given more sampling via replicate simulations, it is possible the NOMELT variant's chromophore is less planar than mEos, given the slightly lower dihedral angle at the beginning of the simulations. However, this is likely an artifact of computational setup and starting configurations, as the dihedral angles at 298K and 320K are nearly identical for both systems by the end of the simulations.

Understanding the β-barrel's role in the fluorescence of FPs is crucial to quantifying and rationalizing the effects of mutations distal to the chromophore on thermostability. First and foremost, the R148K mutation is not part of the parent residues, and it can be assumed that photophysical properties such as excitation and emission wavelength will remain unchanged from the wildtype. There are examples of thermostable FP variants exhibiting increased protein-protein interactions with or near the chromophore,[10,97] as well as mutations on the exterior of the β-barrel present in literature.[98] Considering lysine is a charged, polar amino acid, the most favorable orientation for the side chain is going to be facing outside the β-barrel and interacting with solvent. These charged surface residues can form salt bridges with other charged residues, a characteristic consistent with thermostable proteins in general, as well as promote entropically favorable repacking of adjacent residues.[7,99] It is unclear what other effects this mutation to mEos could have on other physical properties, such as maturation time, refolding kinetics, and fluorescence lifetime, and synthesizing the protein may be the only way to begin hypothesizing what other effects this R148K mutation has.

When root-mean square fluctuation (RMSF) is plotted as a function of residue index in **Figure 9**, the stabilizing effects of the mutation are obvious. RMSF is frequently used to assess the stability and activity of proteins and enzymes, as lower RMSF values are correlated with increased enzymatic activity and thermostability.[100] This stabilizing effect of this R148K is significant over residues 125-160 and nearly centered about the mutation. A significant decrease in alpha-carbon fluctuations is observed in the NOMELT variant, suggesting that this mutation may have a stabilizing effect that propagates to neighboring residues. This analysis is interesting and not very different from how designed proteins have been vetted for thermostability in literature.[67,70,100]

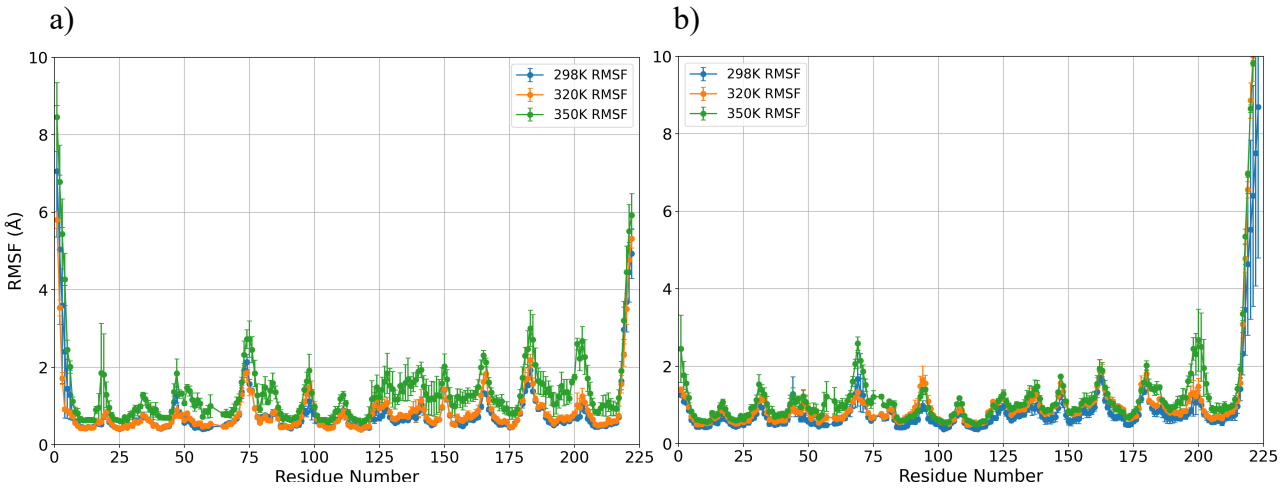


Figure 9: Ensemble-Averaged (n=3) RMSF Values as a Function of Residue Index. (a) mEos and (b) NOMELT variant. Gap in residue index 64 is artifact from computational setup of chromophore.

But what does this range of residues have to do with function, in mEos and FPs globally, other than contribute to the larger β-barrel scaffold? Some of these residues may have side chains interacting with the chromophore by pointing inward of the β-barrel, but there is another structural motif between certain strands of the β-barrel that is critical to fluorescence. Previously mentioned in **Chapter 1,** the proton transfer resulting in radiative emission of FPs occurs through a water

molecule coordinated with the chromophore. It has been proposed that the crystallographic waters present in PDB structures of FPs are the culprit here and can diffuse out of the β-barrel after fluorescence.[94] Highlighted in **Figure 10**, the dominant pathway of water exchange for this process is believed to be through a hole in the β-barrel, particularly between β-strands 7 and 10.[94] Upon heating, the distances between the atoms of these residues increase via increased thermal fluctuations, weakening the protein-protein interactions that facilitate this water exchange. As the protein unfolds, water is likely pouring into the β-barrel, affecting the planarity of the chromophore with extra water contacts. This is observed qualitatively in the simulations for both mEos and the NOMELT variant. Because the R148K mutation has a stabilizing effect on the residues associated with the β-barrel hole, I believe that the NOMELT variant can maintain fluorescence at higher temperatures with more favorable protein-protein interactions, therefore facilitating the water exchange at higher temperatures. This structural analysis of a thermostable protein proposed by an LLM trained only on prokaryotic protein sequence pairs suggests that the model can target structurally dynamic subsections of a larger structure for mutations while maintaining complex protein function.

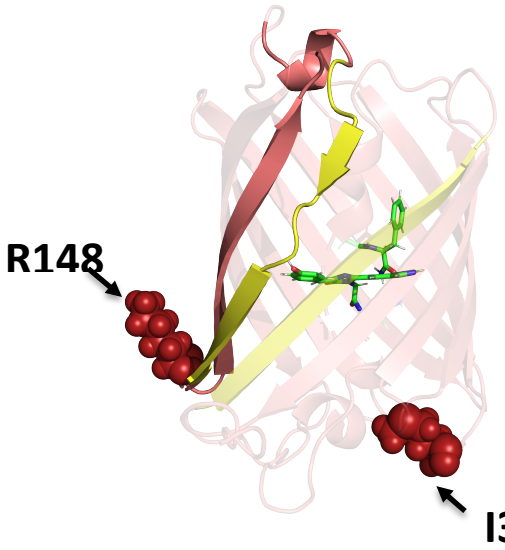


Figure 10: Cutaway of NOMELT Variant.

Mutations are highlighted with red spheres and labeled. Residues associated with water exchange are highlighted in yellow. Residues stabilized by R148K are shown in solid cartoon representation.

3.4 Final Thoughts and Future Work

Over the course of three chapters, I have used MD simulations to evaluate two very different protein designs for stability and function. With traditional protein design techniques, such as directed evolution, circular permutation, and other rational design techniques, just one of these protein designs could realistically take an entire PhD. Hundreds of student hours, tens of thousands of dollars for reagents and instruments, and luck sometimes aren't enough to yield designed protein. The current explosion of computational resources available for protein design is hard to ignore, and the NOMELT model is simply one of many tools used to understand the sequence-structure relationship. The MD campaigns in this study provided valuable insight into the performance of NOMELT and removed a lot of the guesswork from traditional protein design techniques. Since the stability of these protein designs has been demonstrated via MD simulations,

the next step for this study is the expression, purification, and evaluation of both NOMELT variants in the lab.

Thanks to connections within the Beck Research Lab, both NOMELT variants of ENH and mEos are slated to be expressed in living organisms and fully characterized by Professor Michelle McCully at Santa Clara University. Once synthesized, the physical and photophysical properties of the NOMELT variants can be assessed. The helix-turn-helix motif found in structural snapshots and AF2 predications of the NOMELT variant are most certainly the global fold that would be observed in the synthesized protein. However, evaluating NOMELT's ENH variant activity and affinity towards DNA would be extremely valuable. Secondary structure is tied to function, but how have the mutations NOMELT introduced affected protein activity towards DNA? If the NOMELT variant had both higher thermostability and greater activity, that would be a huge win for NOMELT and LLMs applied to protein sequences in general. Regarding the mEos variant, there are many more interesting physical properties that can be explored with a synthesized protein. If adding polar residues to the exterior of the β-barrel of tetrameric EosFP created mEos, does the R148K mutation reduce unwanted dimerization at higher concentrations?[84] What effects does this mutation have on maturation time, quantum yield, and excitation/emission wavelength? Given the relative ease of expressing FPs in living cells like *E. coli*, the best way to answer these questions is going to be synthesis and characterization in the lab. It is entirely possible that the mutations proposed have unintended consequences, or benefits, to the usability of these proteins. Part of this uncertainty is what makes applying LLMs to protein sequences so powerful.

Pattern recognition is one of the defining characteristics of being human and is something that we do better than any other organism. There is a long history of humans noticing patterns in nature; we can create more disease-resistant and productive crops, create entirely new species of

animals with desired characteristics, and engineer new proteins to serve a specific purpose. The limit of our ability to recognize patterns is reached when tasked with identifying trends over millions of protein sequences with potentially hundreds of homologs and permutation libraries with billions of sequences. Proteins are inherently high-dimensional, and applying LLMs, such as NOMELT, represents a large leap forward in our understanding of the sequence-function relationship. With MD simulation campaigns and a thorough structural analysis of two different applications of NOMELT, the abilities of NOMELT to translate protein sequences to higher melting temperatures while preserving complex protein function is demonstrated. It is my hope that upon synthesis and isolation of both proposed NOMELT variants, researchers will have open-source access to NOMELT and an entirely new RFP ready for imaging experiments of all kinds.

REFERENCES

(1) Sadowski, M. I.; Jones, D. T. The Sequence–Structure Relationship and Protein Function Prediction. *Curr. Opin. Struct. Biol.* **2009**, *19* (3), 357–362. https://doi.org/10.1016/j.sbi.2009.03.008.

(2) Anfinsen, C. B. Principles That Govern the Folding of Protein Chains. *Science* **1973**, *181* (4096), 223–230. https://doi.org/10.1126/science.181.4096.223.

(3) Zhang, J.; Yang, J.-R. Determinants of the Rate of Protein Sequence Evolution. *Nat. Rev. Genet.* **2015**, *16* (7), 409–420. https://doi.org/10.1038/nrg3950.

(4) Trivedi, R.; Nagarajaram, H. A. Intrinsically Disordered Proteins: An Overview. *Int. J. Mol. Sci.* **2022**, *23* (22), 14050. https://doi.org/10.3390/ijms232214050.

(5) Rabert, C.; Weinacker, D.; Pessoa Jr, A.; Farías, J. G. Recombinants Proteins for Industrial Uses: Utilization of Pichia Pastoris Expression System. *Braz. J. Microbiol.* **2013**, *44* (2), 351–356. https://doi.org/10.1590/S1517-83822013005000041.

(6) Barone, G. D.; Emmerstorfer-Augustin, A.; Biundo, A.; Pisano, I.; Coccetti, P.; Mapelli, V.; Camattari, A. Industrial Production of Proteins with Pichia Pastoris—Komagataella Phaffii. *Biomolecules* **2023**, *13* (3), 441. https://doi.org/10.3390/biom13030441.

(7) Close, D. W.; Paul, C. D.; Langan, P. S.; Wilce, M. C. J.; Traore, D. A. K.; Halfmann, R.; Rocha, R. C.; Waldo, G. S.; Payne, R. J.; Rucker, J. B.; Prescott, M.; Bradbury, A. R. M. Thermal Green Protein, an Extremely Stable, Nonaggregating Fluorescent Protein Created by Structure-guided Surface Engineering. *Proteins Struct. Funct. Bioinforma.* **2015**, *83* (7), 1225–1237. https://doi.org/10.1002/prot.24699.

(8) Cava, F.; De Pedro, M. A.; Blas-Galindo, E.; Waldo, G. S.; Westblade, L. F.; Berenguer, J. Expression and Use of Superfolder Green Fluorescent Protein at High Temperatures *in Vivo* : A Tool to Study Extreme Thermophile Biology. *Environ. Microbiol.* **2008**, *10* (3), 605–613. https://doi.org/10.1111/j.1462-2920.2007.01482.x.

(9) Balabanova, L.; Golotin, V.; Podvolotskaya, A.; Rasskazov, V. Genetically Modified Proteins: Functional Improvement and Chimeragenesis. *Bioengineered* **2015**, *6* (5), 262–274. https://doi.org/10.1080/21655979.2015.1075674.

(10) Kimata, Y.; Iwaki, M.; Lim, C. R.; Kohno, K. A Novel Mutation Which Enhances the Fluorescence of Green Fluorescent Protein at High Temperatures. *Biochem. Biophys. Res. Commun.* **1997**, *232* (1), 69–73. https://doi.org/10.1006/bbrc.1997.6235.

(11) Bharatiy, S. K.; Hazra, M.; Paul, M.; Mohapatra, S.; Samantaray, D.; Dubey, R. C.; Sanyal, S.; Datta, S.; Hazra, S. In Silico Designing of an Industrially Sustainable Carbonic Anhydrase Using Molecular Dynamics Simulation. *ACS Omega* **2016**, *1* (6), 1081–1103. https://doi.org/10.1021/acsomega.6b00041.

(12) Atkinson, J. T.; Jones, A. M.; Zhou, Q.; Silberg, J. J. Circular Permutation Profiling by Deep Sequencing Libraries Created Using Transposon Mutagenesis. *Nucleic Acids Res.* **2018**, *46* (13), e76–e76. https://doi.org/10.1093/nar/gky255.

(13) Pan, X.; Kortemme, T. Recent Advances in de Novo Protein Design: Principles, Methods, and Applications. *J. Biol. Chem.* **2021**, *296*, 100558. https://doi.org/10.1016/j.jbc.2021.100558.

(14) Huttanus, H. M.; Triola, E.-K. H.; Velasquez-Guzman, J. C.; Shin, S.-M.; Granja-Travez, R. S.; Singh, A.; Dale, T.; Jha, R. K. Targeted Mutagenesis and High-Throughput Screening

of Diversified Gene and Promoter Libraries for Isolating Gain-of-Function Mutations. *Front. Bioeng. Biotechnol.* **2023**, *11*, 1202388. https://doi.org/10.3389/fbioe.2023.1202388.
(15) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.
(16) The UniProt Consortium; Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bye-A-Jee, H.; Cukura, A.; Denny, P.; Dogan, T.; Ebenezer, T.; Fan, J.; Garmiri, P.; Da Costa Gonzales, L. J.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Joshi, V.; Jyothi, D.; Kandasaamy, S.; Lock, A.; Luciani, A.; Lugaric, M.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Pundir, S.; Qi, G.; Raj, S.; Raposo, P.; Rice, D. L.; Saidi, R.; Santos, R.; Speretta, E.; Stephenson, J.; Totoo, P.; Turner, E.; Tyagi, N.; Vasudev, P.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A. J.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A. H.; Axelsen, K. B.; Bansal, P.; Baratin, D.; Batista Neto, T. M.; Blatter, M.-C.; Bolleman, J. T.; Boutet, E.; Breuza, L.; Gil, B. C.; Casals-Casas, C.; Echioukh, K. C.; Coudert, E.; Cuche, B.; De Castro, E.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gaudet, P.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz, N.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Kerhornou, A.; Le Mercier, P.; Lieberherr, D.; Masson, P.; Morgat, A.; Muthukrishnan, V.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Poux, S.; Pozzato, M.; Pruess, M.; Redaschi, N.; Rivoire, C.; Sigrist, C. J. A.; Sonesson, K.; Sundaram, S.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Zhang, J. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D523–D531. https://doi.org/10.1093/nar/gkac1052.
(17) Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. GenBank. *Nucleic Acids Res.* **2016**, *44* (D1), D67–D72. https://doi.org/10.1093/nar/gkv1276.
(18) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.
(19) Mouvet, F.; Villard, J.; Bolnykh, V.; Rothlisberger, U. Recent Advances in First-Principles Based Molecular Dynamics. *Acc. Chem. Res.* **2022**, *55* (3), 221–230. https://doi.org/10.1021/acs.accounts.1c00503.
(20) Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* **2002**, *9* (9), 646–652. https://doi.org/10.1038/nsb0902-646.
(21) Bunker, A.; Róg, T. Mechanistic Understanding From Molecular Dynamics Simulation in Pharmaceutical Research 1: Drug Delivery. *Front. Mol. Biosci.* **2020**, *7*, 604770. https://doi.org/10.3389/fmolb.2020.604770.
(22) *Biomolecular Simulations: Methods and Protocols*; Bonomi, M., Camilloni, C., Eds.; Methods in molecular biology; Humana Press: New York, NY, 2019.
(23) Hénin, J.; Lelièvre, T.; Shirts, M. R.; Valsson, O.; Delemotte, L. Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0]. *Living J. Comput. Mol. Sci.* **2022**, *4* (1). https://doi.org/10.33011/livecoms.4.1.1583.

(24) Childers, M. C.; Daggett, V. Insights from Molecular Dynamics Simulations for Computational Protein Design. *Mol. Syst. Des. Eng.* **2017**, *2* (1), 9–33. https://doi.org/10.1039/C6ME00083E.
(25) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci.* **2002**, *99* (20), 12562–12566. https://doi.org/10.1073/pnas.202427399.
(26) Schäfer, T. M.; Settanni, G. Data Reweighting in Metadynamics Simulations. *J. Chem. Theory Comput.* **2020**, *16* (4), 2042–2052. https://doi.org/10.1021/acs.jctc.9b00867.
(27) Karmani, R. K.; Agha, G.; Squillante, M. S.; Seiferas, J.; Brezina, M.; Hu, J.; Tuminaro, R.; Sanders, P.; Träffe, J. L.; Geijn, R. A.; Träff, J. L.; Geijn, R. A.; Sander, M. B.; Gustafson, J. L.; Dror, R. O.; Young, C.; Shaw, D. E.; Lin, C.; Lee, J.-K.; Chang, R.-G.; Kuan, C.-B.; Kollias, G.; Grama, A. Y.; Li, Z.; Whaley, R. C.; Vuduc, R. W. Anton, A Special-Purpose Molecular Simulation Machine. In *Encyclopedia of Parallel Computing*; Padua, D., Ed.; Springer US: Boston, MA, 2011; pp 60–71. https://doi.org/10.1007/978-0-387-09766-4_199.
(28) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614. https://doi.org/10.1002/jcc.21287.
(29) Robustelli, Paul; Piana, Stefano; Shaw, David E. Developing a Molecular Dynamics Forcefield for Both Folded and Disordered Protein States. *PNAS 115* (21).
(30) Breyfogle, K. L.; Blood, D. L.; Rosnik, A. M.; Krueger, B. P. Molecular Dynamics Force Field Parameters for the EGFP Chromophore and Some of Its Analogues. *J. Phys. Chem. B* **2023**, *127* (26), 5772–5788. https://doi.org/10.1021/acs.jpcb.3c01486.
(31) Komp, E.; Phillips, C.; Alanzi, H. N.; Zorman, M.; Beck, D. A. C. *A Learnable Transition from Low Temperature to High Temperature Proteins with Neural Machine Translation*; preprint; Bioinformatics, 2024. https://doi.org/10.1101/2024.02.06.579188.
(32) Ahmed, Z.; Zulfiqar, H.; Tang, L.; Lin, H. A Statistical Analysis of the Sequence and Structure of Thermophilic and Non-Thermophilic Proteins. *Int. J. Mol. Sci.* **2022**, *23* (17), 10116. https://doi.org/10.3390/ijms231710116.
(33) Kumar, S.; Tsai, C.-J.; Nussinov, R. Factors Enhancing Protein Thermostability. *Protein Eng. Des. Sel.* **2000**, *13* (3), 179–191. https://doi.org/10.1093/protein/13.3.179.
(34) Valentini, G.; Malchiodi, D.; Gliozzo, J.; Mesiti, M.; Soto-Gomez, M.; Cabri, A.; Reese, J.; Casiraghi, E.; Robinson, P. N. The Promises of Large Language Models for Protein Design and Modeling. *Front. Bioinforma.* **2023**, *3*, 1304099. https://doi.org/10.3389/fbinf.2023.1304099.
(35) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; Rost, B. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. July 12, 2020. https://doi.org/10.1101/2020.07.12.199554.
(36) Komp, E.; Alanzi, H. H.; Francis, R.; Vuong, C.; Roberts, L.; Mossallanejad, A.; Beck, D. A. C. Homologous Pairs of Low and High Temperature Originating Proteins Spanning the Known Prokaryotic Universe. *Sci. Data* **2023**, *10* (1), 682. https://doi.org/10.1038/s41597-023-02553-w.

(37) Reid, B. G.; Flynn, G. C. Chromophore Formation in Green Fluorescent Protein. *Biochemistry* **1997**, *36* (22), 6786–6791. https://doi.org/10.1021/bi970281w.
(38) Zheng, J.; Guo, N.; Huang, Y.; Guo, X.; Wagner, A. High Temperature Delays and Low Temperature Accelerates Evolution of a New Protein Phenotype. *Nat. Commun.* **2024**, *15* (1), 2495. https://doi.org/10.1038/s41467-024-46332-6.
(39) Patnaik, S. S.; Trohalaki, S.; Pachter, R. Molecular Modeling of Green Fluorescent Protein: Structural Effects of Chromophore Deprotonation. *Biopolymers* **2004**, *75* (6), 441–452. https://doi.org/10.1002/bip.20156.
(40) Federico Coppola; Fulvio Perrella; Alessio Petrone; Greta Donati; Nadia Rega. A Not Obvious Correlation Between the Structure of Green Fluorescent Protein Chromophore Pocket and Hydrogen Bond Dynamics: A Choreography From Ab Initio Molecular Dynamics. *Front Mol Biosci* **2020**, *7*. https://doi.org/10.3389/fmolb.2020.569990.
(41) Day, R. N.; Davidson, M. W. The Fluorescent Protein Palette: Tools for Cellular Imaging. *Chem. Soc. Rev.* **2009**, *38* (10), 2887. https://doi.org/10.1039/b901966a.
(42) Shimomura, O.; Johnson, F. H.; Saiga, Y. Extraction, Purification and Properties of Aequorin, a Bioluminescent Protein from the Luminous Hydromedusan, *Aequorea*. *J. Cell. Comp. Physiol.* **1962**, *59* (3), 223–239. https://doi.org/10.1002/jcp.1030590302.
(43) Zimmer, M. GFP: From Jellyfish to the Nobel Prize and Beyond. *Chem. Soc. Rev.* **2009**, *38* (10), 2823. https://doi.org/10.1039/b904023d.
(44) Prasher, D. C.; Eckenrode, V. K.; Ward, W. W.; Prendergast, F. G.; Cormier, M. J. Primary Structure of the Aequorea Victoria Green-Fluorescent Protein. *Gene* **1992**, *111* (2), 229–233. https://doi.org/10.1016/0378-1119(92)90691-H.
(45) Shcherbakova, D. M.; Verkhusha, V. V. Chromophore Chemistry of Fluorescent Proteins Controlled by Light. *Curr. Opin. Chem. Biol.* **2014**, *20*, 60–68. https://doi.org/10.1016/j.cbpa.2014.04.010.
(46) Nienhaus, K.; Nienhaus, G. U. Chromophore Photophysics and Dynamics in Fluorescent Proteins of the GFP Family. *J. Phys. Condens. Matter* **2016**, *28* (44), 443001. https://doi.org/10.1088/0953-8984/28/44/443001.
(47) Heim, R.; Cubitt, A. B.; Tsien, R. Y. Improved Green Fluorescence. *Nature* **1995**, *373* (6516), 663–664. https://doi.org/10.1038/373663b0.
(48) Hung, L.; Terwilliger, T. C.; Waldo, G. S.; Nguyen, H. B. Engineering Highly Stable Variants of *Corynactis Californica* Green Fluorescent Proteins. *Protein Sci.* **2024**, *33* (2), e4886. https://doi.org/10.1002/pro.4886.
(49) Kiss, C.; Temirov, J.; Chasteen, L.; Waldo, G. S.; Bradbury, A. R. M. Directed Evolution of an Extremely Stable Fluorescent Protein. *Protein Eng. Des. Sel.* **2009**, *22* (5), 313–323. https://doi.org/10.1093/protein/gzp006.
(50) Lambert, T. J. FPbase: A Community-Editable Fluorescent Protein Database. *Nat. Methods* **2019**, *16* (4), 277–278. https://doi.org/10.1038/s41592-019-0352-8.
(51) Baird, G. S.; Zacharias, D. A.; Tsien, R. Y. Biochemistry, Mutagenesis, and Oligomerization of DsRed, a Red Fluorescent Protein from Coral. *Proc. Natl. Acad. Sci.* **2000**, *97* (22), 11984–11989. https://doi.org/10.1073/pnas.97.22.11984.
(52) Matz, M. V.; Fradkov, A. F.; Labas, Y. A.; Savitsky, A. P.; Zaraisky, A. G.; Markelov, M. L.; Lukyanov, S. A. Fluorescent Proteins from Nonbioluminescent Anthozoa Species. *Nat. Biotechnol.* **1999**, *17* (10), 969–973. https://doi.org/10.1038/13657.
(53) Shaner, N. C.; Campbell, R. E.; Steinbach, P. A.; Giepmans, B. N. G.; Palmer, A. E.; Tsien, R. Y. Improved Monomeric Red, Orange and Yellow Fluorescent Proteins Derived

from Discosoma Sp. Red Fluorescent Protein. *Nat. Biotechnol.* **2004**, *22* (12), 1567–1572. https://doi.org/10.1038/nbt1037.
(54) Wiedenmann, J.; Schenk, A.; Röcker, C.; Girod, A.; Spindler, K.-D.; Nienhaus, G. U. A Far-Red Fluorescent Protein with Fast Maturation and Reduced Oligomerization Tendency from *Entacmaea Quadricolor* (Anthozoa, Actinaria). *Proc. Natl. Acad. Sci.* **2002**, *99* (18), 11646–11651. https://doi.org/10.1073/pnas.182157199.
(55) Imamura, H.; Otsubo, S.; Nishida, M.; Takekawa, N.; Imada, K. Red Fluorescent Proteins Engineered from Green Fluorescent Proteins. *Proc. Natl. Acad. Sci.* **2023**, *120* (45), e2307687120. https://doi.org/10.1073/pnas.2307687120.
(56) Shakoori, A. R. Fluorescence In Situ Hybridization (FISH) and Its Applications. In *Chromosome Structure and Aberrations*; Bhat, T. A., Wani, A. A., Eds.; Springer India: New Delhi, 2017; pp 343–367. https://doi.org/10.1007/978-81-322-3673-3_16.
(57) Shcherbakova, D. M.; Subach, O. M.; Verkhusha, V. V. Red Fluorescent Proteins: Advanced Imaging Applications and Future Design. *Angew. Chem. Int. Ed.* **2012**, *51* (43), 10724–10738. https://doi.org/10.1002/anie.201200408.
(58) Cui, C.; Shu, W.; Li, P. Fluorescence In Situ Hybridization: Cell-Based Genetic Diagnostic and Research Applications. *Front. Cell Dev. Biol.* **2016**, *4*. https://doi.org/10.3389/fcell.2016.00089.
(59) Song, J.; Dean, Z. A Short Review of Deep Tissue Imaging Techniques and Applications. *J. Phys. Conf. Ser.* **2022**, *2287* (1), 012028. https://doi.org/10.1088/1742-6596/2287/1/012028.
(60) Structural Genomics Consortium; Architecture et Fonction des Macromolécules Biologiques; Berkeley Structural Genomics Center; China Structural Genomics Consortium; Integrated Center for Structure and Function Innovation; Israel Structural Proteomics Center; Joint Center for Structural Genomics; Midwest Center for Structural Genomics; New York Structural GenomiX Research Center for Structural Genomics; Northeast Structural Genomics Consortium; Oxford Protein Production Facility; Protein Sample Production Facility, Max Delbrück Center for Molecular Medicine; RIKEN Structural Genomics/Proteomics Initiative; SPINE2-Complexes. Protein Production and Purification. *Nat. Methods* **2008**, *5* (2), 135–146. https://doi.org/10.1038/nmeth.f.202.
(61) Kamerzell, T. J.; Middaugh, C. R. Prediction Machines: Applied Machine Learning for Therapeutic Protein Design and Development. *J. Pharm. Sci.* **2021**, *110* (2), 665–681. https://doi.org/10.1016/j.xphs.2020.11.034.
(62) Draganescu, A.; Tullius, T. D. The DNA Binding Specificity of Engrailed Homeodomain. *J. Mol. Biol.* **1998**, *276* (3), 529–536. https://doi.org/10.1006/jmbi.1997.1567.
(63) Clarke, N. D.; Kissinger, C. R.; Desjarlais, J.; Gilliland, G. L.; Pabo, C. O. Structural Studies of the Engrailed Homeodomain. *Protein Sci.* **1994**, *3* (10), 1779–1787. https://doi.org/10.1002/pro.5560031018.
(64) Fraenkel, E.; Rould, M. A.; Chambers, K. A.; Pabo, C. O. Engrailed Homeodomain-DNA Complex at 2.2 å Resolution: A Detailed View of the Interface and Comparison with Other Engrailed Structures 1 1Edited by T. Richmond. *J. Mol. Biol.* **1998**, *284* (2), 351–361. https://doi.org/10.1006/jmbi.1998.2147.
(65) Shah, Premal S.; Hom, Geoffery K.; Ross, Scott A.; Lassila, Jonathan Kyle; Crowhurst, Karin A.; Mayo, Stephen L. Full-Sequence Computational Design and Solution STructure of a Thermostable Protein Variant. *J Mol Biol 372*, 1–6.

(66) Tripp, Katerine W.; Sternke, Matt; Majumdar, Ananya; Barrick, Doug. Creating a Homeodomain with High STability and DNA Binding Affinity by Sequence Averaging. *J Am Chem Soc* **2017**, *139* (14), 5051–5060. https://doi.org/10.1021/jacs.6b11323.
(67) McCully, Michelle E.; Beck, David A.C.; Daggett, Valerie. Promiscuous Contacts and Heightened Dynamics Increase Thermostability in an Engineered Variant of the Engrailed Homeodomain. *Protein Eng. Des. Sel. 26* (1), 35–45.
(68) McCully, M. E.; Beck, D. A. C.; Daggett, V. Microscopic Reversibility of Protein Folding in Molecular Dynamics Simulations of the Engrailed Homeodomain. *Biochemistry* **2008**, *47* (27), 7079–7089. https://doi.org/10.1021/bi800118b.
(69) Nguyen, C.; Yearwood, L. M.; McCully, M. E. Thermostabilization Mechanisms in Thermophilic versus Mesophilic Three-helix Bundle Proteins. *J. Comput. Chem.* **2022**, *43* (3), 197–205. https://doi.org/10.1002/jcc.26782.
(70) Gonzalez, N. A.; Li, B. A.; McCully, M. E. The Stability and Dynamics of Computationally Designed Proteins. *Protein Eng. Des. Sel.* **2022**, *35*, gzac001. https://doi.org/10.1093/protein/gzac001.
(71) McCully, M. E.; Beck, D. A. C.; Fersht, A. R.; Daggett, V. Refolding the Engrailed Homeodomain: Structural Basis for the Accumulation of a Folding Intermediate. *Biophys. J.* **2010**, *99* (5), 1628–1636. https://doi.org/10.1016/j.bpj.2010.06.040.
(72) Mayor, U.; Guydosh, N. R.; Johnson, C. M.; Grossmann, J. G.; Sato, S.; Jas, G. S.; Freund, S. M. V.; Alonso, D. O. V.; Daggett, V.; Fersht, A. R. The Complete Folding Pathway of a Protein from Nanoseconds to Microseconds. *Nature* **2003**, *421* (6925), 863–867. https://doi.org/10.1038/nature01428.
(73) Seelig, J.; Seelig, A. Chemical Protein Unfolding – A Simple Cooperative Model. *J. Phys. Chem. B* **2023**, *127* (39), 8296–8304. https://doi.org/10.1021/acs.jpcb.3c03558.
(74) Seelig, J.; Seelig, A. Protein Unfolding—Thermodynamic Perspectives and Unfolding Models. *Int. J. Mol. Sci.* **2023**, *24* (6), 5457. https://doi.org/10.3390/ijms24065457.
(75) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31* (4), 671–690. https://doi.org/10.1002/jcc.21367.
(76) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.
(77) The PLUMED consortium. Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **2019**, *16* (8), 670–673. https://doi.org/10.1038/s41592-019-0506-8.
(78) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.* **2014**, *185* (2), 604–613. https://doi.org/10.1016/j.cpc.2013.09.018.
(79) Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8, 2015.
(80) Bernetti, M.; Bussi, G. Pressure Control Using Stochastic Cell Rescaling. *J. Chem. Phys.* **2020**, *153* (11), 114107. https://doi.org/10.1063/5.0020514.

(81) Grubmüller, H.; Heller, H.; Windemuth, A.; Schulten, K. Generalized Verlet Algorithm for Efficient Molecular Dynamics Simulations with Long-Range Interactions. *Mol. Simul.* **1991**, *6* (1–3), 121–142. https://doi.org/10.1080/08927029108022142.
(82) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An $N \cdot \log(N)$ Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092. https://doi.org/10.1063/1.464397.
(83) Gowers, R.; Linke, M.; Barnoud, J.; Reddy, T.; Melo, M.; Seyler, S.; Domański, J.; Dotson, D.; Buchoux, S.; Kenney, I.; Beckstein, O. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations; Austin, Texas, 2016; pp 98–105. https://doi.org/10.25080/Majora-629e541a-00e.
(84) Wiedenmann, J.; Ivanchenko, S.; Oswald, F.; Schmitt, F.; Röcker, C.; Salih, A.; Spindler, K.-D.; Nienhaus, G. U. EosFP, a Fluorescent Marker Protein with UV-Inducible Green-to-Red Fluorescence Conversion. *Proc. Natl. Acad. Sci.* **2004**, *101* (45), 15905–15910. https://doi.org/10.1073/pnas.0403668101.
(85) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Res.* **1994**, *22* (22), 4673–4680. https://doi.org/10.1093/nar/22.22.4673.
(86) Meng, E. C.; Goddard, T. D.; Pettersen, E. F.; Couch, G. S.; Pearson, Z. J.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX : Tools for Structure Building and Analysis. *Protein Sci.* **2023**, *32* (11), e4792. https://doi.org/10.1002/pro.4792.
(87) Shapovalov, M. V.; Dunbrack, R. L. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **2011**, *19* (6), 844–858. https://doi.org/10.1016/j.str.2011.03.019.
(88) Nienhaus, K.; Nienhaus, G. U.; Wiedenmann, J.; Nar, H. Structural Basis for Photo-Induced Protein Cleavage and Green-to-Red Conversion of Fluorescent Protein EosFP. *Proc. Natl. Acad. Sci.* **2005**, *102* (26), 9156–9159. https://doi.org/10.1073/pnas.0501874102.
(89) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52* (12), 3144–3154. https://doi.org/10.1021/ci300363c.
(90) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian˜16 Revision C.01, 2016.
(91) Scharnagl, C.; Raupp-Kossmann, R.; Fischer, S. F. Molecular Basis for pH Sensitivity and Proton Transfer in Green Fluorescent Protein: Protonation and Conformational Substates from Electrostatic Calculations. *Biophys. J.* **1999**, *77* (4), 1839–1857. https://doi.org/10.1016/S0006-3495(99)77028-1.

(92) Brejc, K.; Sixma, T. K.; Kitts, P. A.; Kain, S. R.; Tsien, R. Y.; Ormö, M.; Remington, S. J. Structural Basis for Dual Excitation and Photoisomerization of the *Aequorea Victoria* Green Fluorescent Protein. *Proc. Natl. Acad. Sci.* **1997**, *94* (6), 2306–2311. https://doi.org/10.1073/pnas.94.6.2306.
(93) Maddalo, S. L.; Zimmer, M. The Role of the Protein Matrix in Green Fluorescent Protein Fluorescence. *Photochem. Photobiol.* **2006**, *82* (2), 367–372. https://doi.org/10.1562/2005-04-11-RA-485.
(94) Shinobu, A.; Agmon, N. The Hole in the Barrel: Water Exchange at the GFP Chromophore. *J. Phys. Chem. B* **2015**, *119* (8), 3464–3478. https://doi.org/10.1021/jp5127255.
(95) Park, J. W.; Rhee, Y. M. Electric Field Keeps Chromophore Planar and Produces High Yield Fluorescence in Green Fluorescent Protein. *J. Am. Chem. Soc.* **2016**, *138* (41), 13619–13629. https://doi.org/10.1021/jacs.6b06833.
(96) Mukherjee, S.; Manna, P.; Hung, S.-T.; Vietmeyer, F.; Friis, P.; Palmer, A. E.; Jimenez, R. Directed Evolution of a Bright Variant of mCherry: Suppression of Nonradiative Decay by Fluorescence Lifetime Selections. *J. Phys. Chem. B* **2022**, *126* (25), 4659–4668. https://doi.org/10.1021/acs.jpcb.2c01956.
(97) Scott, D. J.; Gunn, N. J.; Yong, K. J.; Wimmer, V. C.; Veldhuis, N. A.; Challis, L. M.; Haidar, M.; Petrou, S.; Bathgate, R. A. D.; Griffin, M. D. W. A Novel Ultra-Stable, Monomeric Green Fluorescent Protein For Direct Volumetric Imaging of Whole Organs Using CLARITY. *Sci. Rep.* **2018**, *8* (1), 667. https://doi.org/10.1038/s41598-017-18045-y.
(98) Campbell, B. C.; Paez-Segala, M. G.; Looger, L. L.; Petsko, G. A.; Liu, C. F. Chemically Stable Fluorescent Proteins for Advanced Microscopy. *Nat. Methods* **2022**, *19* (12), 1612–1621. https://doi.org/10.1038/s41592-022-01660-7.
(99) Pédelacq, J.-D.; Cabantous, S.; Tran, T.; Terwilliger, T. C.; Waldo, G. S. Engineering and Characterization of a Superfolder Green Fluorescent Protein. *Nat. Biotechnol.* **2006**, *24* (1), 79–88. https://doi.org/10.1038/nbt1172.
(100) Burgin, T.; Pfaendtner, J.; Beck, D. A. C. Quick and Accurate Estimates of Mutation Effects on Transition-State Stabilization of Enzymes from Molecular Simulations with Restrained Transition States. *J. Phys. Chem. B* **2022**, *126* (48), 9964–9970. https://doi.org/10.1021/acs.jpcb.2c04802.