Engineering the kinetic stability of a β-trefoil protein by tuning its topological complexity

by

Delaney Anderson

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Chemistry

Waterloo, Ontario, Canada, 2022

Author’s Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Delaney Anderson was the sole author for Chapters 1 and 5, which were written under the supervision of Dr. Elizabeth Meiering and were not written for publication.

This thesis consists in part of one manuscript submitted for publication, which has not yet been accepted at the time of final submission of this thesis. Exception to sole authorship of material is as follows:

Research presented in Chapters 2, 3, and 4:

This research was conducted at the University of Waterloo and the Tata Institute for Fundamental Research in India under the supervision of Dr. Elizabeth Meiering and Dr. Shachi Gosavi, respectively. Dr. Elizabeth Meiering and Dr. Shachi Gosavi were listed as co-supervisors for the NSERC Michael Smith Foreign Study Supplement, which funded travel to India to facilitate learning computational methods used herein.

Delaney Anderson designed computation methods with consultation from Dr. Shachi Gosavi, Lakshmi Jayanthi, and Arkadeep Banerjee, as well as Dr. Elizabeth Meiering. Delaney Anderson performed all computational methods. Lakshmi Jayanthi and Arkadeep Banerjee contributed to analysis of simulation trajectories. Delaney Anderson drafted the manuscript, and each author provided intellectual input on manuscript drafts.

Citation:

Anderson, D.M., Jayanthi, L.P., Gosavi, S., and Meiering, E.M. (2022) Engineering the kinetic stability of a β-trefoil protein by tuning its topological complexity. *Front. Mol. Biosciences.* (Under review).

Research presented in Chapter 2.2.5, 3.9, and 3.10 and Appendix C

This research was conducted at the University of Waterloo the supervision of Dr. Elizabeth Meiering and Dr. Todd Holyoak. Delaney Anderson grew csHisH90G crystals and Delaney Anderson and Iain McDonald grew fa-csHisH90G crystals. Norman Tran collected and refined the data for csHisH90G and fa-csHisH90G. Structures were analyzed by Delaney Anderson. Norman Tran contributed to Table 3.5. Iain McDonald compiled Table C1 of Appendix C.

Abstract

Kinetic stability is crucially important for engineering stable proteins suitable for use in industrial, research, and medical applications. Specifically, kinetic stability is central to determining a protein's functional lifetime, where high kinetic stability generally correlates with high resistance against chemical and thermal denaturation, as well as proteolytic degradation. Despite its significance, few studies address the rational design of kinetic stability, and specific mechanisms of kinetic stability remain largely unknown. Here, we describe a method for designing protein kinetic stability by engineering long-range intramolecular interactions by taking advantage of conserved residue interactions in structurally homologous proteins. Specifically, we base our design on the extreme difference in kinetic stability observed between the β-trefoil proteins hisactophilin and ThreeFoil, which we partially attribute to a marked difference in the number of long-range interactions across the protein core. We report the design and characterization of a kinetically stabilized hisactophilin variant, core-swapped hisactophilin, mutated to contain ThreeFoil core residues in order to enhance long-range contacts in the hisactophilin core. Further, we show that kinetic stability predictions for core-swapped hisactophilin based on long-range order, absolute contact order, and simulated free energy barriers of unfolding are in good agreement with experimentally determined kinetic unfolding rates. In addition to highlighting the predictive power of simple measures of protein topology for changes in protein kinetic stability, these results emphasize core engineering as an attractive and tractable target for improving kinetic stability, particularly in proteins with symmetric cores. Thus, this work provides fundamental insight towards advancing our predictive understanding of kinetic stability and better enabling successful protein design.

Acknowledgments

Science does not take place in a vacuum, and research is nothing if not an extended exercise in collaboration. Many people contributed to this work, in ways big and small, and all deserve my heartfelt gratitude. To start, I want to thank Dr. Elizabeth Meiering for the opportunity to pursue this work, for her enthusiasm and insight toward the project, and for her aid in acquiring funding for travel to India to learn computational modeling essential for this project. Special thanks is given to Shachi Gosavi, Lakshmi Jayanthi, Arkadeep Banerjee, and the rest of the Gosavi group for volunteering their time and effort to teach me coarse-grained structure-based modeling and for graciously hosting me at the Tata Institute for Fundamental Research in India. Importantly, this trip was funded by the NSERC Michael Smith Foreign Study Supplement. I also want to thank my committee members, Dr. Andrew Doxey and Dr. Subha Kalyaanamoorthy, for their valuable advice on the bioinformatic and simulation methods used in this thesis, respectively. An enormous thank you is extended to Dr. Todd Holyoak and Norman Tran, whose advice and expertise were essential to obtaining the experimental structure for the core-swapped hisactophilin variant facsHisH90G described herein. Thanks is also given to Dr. Jean Duhamel and members of his group for the use of their fluorometer, which was crucially important for the experimental portion of this project. I also thank Cathy Van Esch, whose knowledge of administrative affairs was extremely helpful throughout my degree.

Acknowledgements would be woefully incomplete without taking time to address friends and family who have supported me throughout my time in academia. Firstly, I want to thank all members of the Meiering lab, past and present, who not only enriched what I learned throughout my degree, but also made the experience much more fun. I thank Dr. Susannah Gagnon, Dr. Max Legg, and Dr. Jonathan Thacker, who instilled in me my love of research and who started me on

this journey. As always, I thank my parents, Brian and Jennifer Anderson, and my sister, Paisley, for their unwavering love and support. Thanks is also given to Chelsea Walker, Alaina Trenholm, Naomi Smith, and Morgan King for their friendship and willingness to listen to my science rants (#sorrynotsorry). Finally, I want to thank my partner, Shane Ferko, for his advice and patience. Though your attempts to instill in me some semblance of a work-life balance have seen limited success so far, I appreciate that you keep trying and have hope for greater success in the future.

Table of contents

List of figures

List of tables

List of abbreviations

3Foil	ThreeFoil
wtHis	Wild type hisactophilin
MD	Molecular dynamics
TIM	Triosephosphate isomerase
ACO	Absolute contact order
LRO	Long-range order
C_α-SBM	Structure-based models coarse-grained to a C_α bead
T_f	Protein folding temperature
csHisH90G	Core-swapped hisactophilin with H90G
PROSS	Protein Repair One-Stop Shop
HisH90G	Pseudo-wild type hisactophilin
CM	Comparative modeling
SBM	Structure-based model
TSE	Transition state ensemble
Q	Fraction of native contacts
PSSM	Position-specific substitution matrix
MSA	Multiple sequence alignment
HMM	Hidden Markov model
fa-csHisH90G	Pfam fascin csHisH90G variant
ph-csHisH90G	HMMER phmmer csHisH90G variant
mu-csHisH90G	MUSTANG Mustguseal csHisH90G variant
GuHCl	Guanidine hydrocholoride

N	Native
U	Unfolded
$\boldsymbol{m_{eq}}$	Equilibrium m-value
β_T	β-Tanford
CMCF-ID	Canadian Macromolecular Crystallography Facility ID
fa	Fascin
ph	phmmer
mu	Mustguseal
GuSCN	Guanidine thiocyanate
$\boldsymbol{m_f}$	Denaturant dependence of folding
$\boldsymbol{m_u}$	Denaturant dependence of unfolding
RMSD	Root-mean-square deviation
csHis	Core-swapped hisactophilin (without H90G)
cs3Foil	Core-swapped ThreeFoil

1 Introduction

1.1 Proteins and protein engineering

Proteins are macromolecules that boast an incredibly versatile range of functions. Proteins serve as signaling molecules, form structural scaffolds, transport molecules into and between cells and organelles, and catalyze a myriad of chemical reactions with exquisite efficiency. Beyond biological systems, proteins are increasingly used in industrial, medical, and research applications. Enzymes offer faster reaction rates, require less energy, produce specific and consistent products, and are more environmentally friendly compared to traditional syntheses used in industrial processes (Bornscheuer et al., 2012; Singh et al., 2016; Sindhu et al., 2017). Protein therapeutics offer greater functional complexity and specificity against disease targets compared to small-molecule drugs (Gurevich and Gurevich, 2014). Protein-based biosensors enable specific detection of reporter molecules towards the advancement of medical and fundamental research worldwide (Mehrotra, 2016; Vigneshvar et al., 2016). Thus, protein applications have clear benefits, and novel protein innovation is desirable.

Until recently, the same characteristics that make proteins desirable for practical application, *i.e.* their specificity and selectivity, prevented them from being widely useful (Lutz and Iamurri, 2018). Industrial processes often require chemical reactions not observed in nature. Additionally, industrial processes often employ harsh reaction conditions such as high temperature, non-neutral pH, high concentrations of protein denaturants, or organic solvents (Liu et al., 2019). However, most proteins can only effectively catalyze their natural function and substrate at near-physiological conditions, significantly limiting their practical applications. The advent of protein engineering technologies lessened these limitations. Protein engineering techniques allow researchers to alter a protein's amino acid sequence and modify its functional properties (Engqvist and Rabe, 2019). Through protein engineering, researchers are able to artificially adapt proteins to recognize non-natural substrates, function in harsh environments, and catalyze novel chemical reactions (Lutz and Iamurri, 2018). Thus, protein engineering represents a vast potential to increase the distribution of industrially-useful proteins (Lutz and Iamurri, 2018).

Recent decades have seen considerable expansion in protein applications and engineering technologies. Initial protein designs changed minimal residues since techniques for creating site-directed mutations were novel and non-trivial. Despite these limitations, many groups reported engineering industrial-grade proteins. For example, Estelle *et al.* (1985) mutated Met222 to Ser in subtilisin, creating an oxidation-resistant protease for use in commercial laundry detergent. A new wave of protein engineering came with directed evolution, which allowed researchers to create random protein mutations on a larger scale. In a seminal paper, Chen and Arnold (1993) reported using sequential rounds of random mutagenesis to engineer a ten-point mutant subtilisin E that displayed 500-fold increased protease activity in organic solvent (Chen and Arnold, 1993). Today, proteins engineered for industrial purposes include: insect-specific bacterial toxins for pest control in the agriculture sector (Tian et al., 2013; Shao et al., 2016; Mao et al., 2017); various ketoreductase enzymes for manufacturing chiral intermediates in the pharmaceutical industry (Bornscheuer et al., 2012); and thermo-stable lipases in the food, pharmaceutical, and detergents industries (Bornscheuer et al., 2002; Kumar et al., 2019), among many others (Choi et al., 2015). Despite clear success, traditional protein engineering still typically requires iterative rounds of experimental design and testing, making it costly and laborious (Bornscheuer et al., 2012; Broom et al., 2017).

Modern protein engineering is an amalgamation of bioinformatic, computational, and experimental techniques (Jones et al., 2017; Lutz and Iamurri, 2018). Researchers routinely use homologous protein structures and sequences to identify functionally and structurally important residues and to inform specific amino acid substitutions. In particular, consensus-based sequence design shows considerable success for improving protein stability and function (Broom et al., 2012; Longo et al., 2012; Feng et al., 2016; Sternke et al., 2019). Protein modeling and design software enhances protein engineering by allowing researchers to create and screen vast libraries of mutant sequences *in silico*. Programs like Rosetta perform extensive searches in sequence and conformational space to identify non-intuitive and synergistic mutations (Rohl et al., 2004; Kaufmann et al., 2010; Goldenzweig et al., 2016; Kuhlman, 2019). Further, structure and sequence information can guide computational design by identifying residues to exclude from mutation or by restricting mutations to residues in homologous sequences. Sequences predicted to best accomplish design goals are validated using a battery of experimental techniques, most often including biophysical and functional assays to select top-performing mutants. Finally, designed sequences may undergo additional rounds of experimental sequence optimization (*e.g.* random mutagenesis or consensus design) to further enhance design attributes (Khoury et al., 2014). Thus, sequence- and structure-based design, computational design, and experimental methods are complementary techniques that better enable protein engineering.

This thesis considers several combinatorial strategies for engineering protein stability for biotechnical protein applications. Specifically, the β-trefoil proteins ThreeFoil (3Foil) and hisactophilin (wtHis) are redesigned to probe protein kinetic stability, which remains a largely unresolved challenge in fundamental protein science. Key challenges in protein engineering will be discussed, including the inverse protein folding problem and protein kinetic stability. Pertinent design strategies will be considered throughout Chapter 1 and described in detail in Chapter 2.

1.2 Protein folding: a mystery in fundamental protein science

Protein activity is determined by a protein's three-dimensional structure. Since proteins are synthesized as a linear chain of amino acids, proteins must fold to achieve their functional, native form. Research into protein folding began in earnest over 60 years ago. In 1961, Anfinsen and colleagues observed that completely reduced and denatured ribonuclease A slowly regains its catalytic activity over time (Anfinsen et al., 1961). Anfinsen reasoned that the protein must refold into its native state, allowing the ribonuclease's re-oxidized sulfhydryl groups to reform functionally important disulfide bonds. Anfinsen concluded that proteins fold spontaneously and that all the information needed for a protein to achieve its native conformation must be contained in its amino acid sequence. Anfinsen (1973) later proposed the thermodynamic hypothesis, which states that a proteins' native structure is the lowest Gibbs free energy state in a system. Working in parallel, Levinthal (1969) observed that it is impossible for a folding protein to randomly sample all possible configurations in a biologically relevant timescale. For example, it would take a small protein with 101 residues 10^{27} years to sample all possible configurations. However, proteins routinely fold in time-scales of microseconds to minutes (Martínez, 2014). This inconsistency sparked a series of hypotheses for protein folding mechanisms to account for experimental folding rates.

Several folding models arose in response to Levinthal's paradox. Each model posits an "intermediate" that forms early in the folding process and then guides the pathway towards the native conformation. In the framework model, folding occurs in a stepwise manner as the unfolded polypeptide forms hydrogen-bonded secondary structures, which interact to form the tertiary structure (Kim and Baldwin, 1982; Ptitsyn, 1987). The collision-diffusion model similarly postulates that local parts of the polypeptide chain form microdomains consisting of α-helices, β-strands, or hydrophobic clusters, which collide and coalesce into multi-microdomain intermediates (Ptitsyn and Rashin, 1975; Karplus and Weaver, 1994). The hydrophobic collapse model proposes that protein folding initially occurs as hydrophobic amino acids interact through the hydrophobic effect and form the protein's core (Kanehisa and Tsong, 1978; Lapidus et al., 2007). Finally, the nucleation model posits that a nucleation event serves as a scaffold onto which polypeptide segments can be rapidly added (Levinthal, 1969; Wetlaufer, 1973). These models are difficult to differentiate experimentally because high resolution structure-determining methods like NMR and X-ray crystallography do not report fast-decaying structures or low populated states (Englander and Mayne, 2014). While spectroscopic methods measure real-time folding, they report non-specific structural features that may be common to several models (Englander and Mayne, 2014). Thus, experimental efforts have failed to provide compelling evidence to support a specific folding model over others.

Today, we understand that protein folding is more complicated than any individual model discussed above. While burial of hydrophobic surface area is generally accepted to drive protein folding (Dill, 1990), hydrophobic burial is achievable through many possible folding pathways. Protein folding involves all manner of interatomic interactions, including hydrophobic interactions between non-polar residues, electrostatic interactions between charged residues, hydrogen bonding to produce secondary structures, and van der Waals interactions between tightly packed residues (Dill and Maccallum, 2012). Protein conformations are also constrained by sterically-allowed and favorable backbone angles in folding and native structures (Kim and Baldwin, 1982). Researchers now view protein folding in terms of a funnel-shaped energy landscape (Figure 1.1), where folding may occur through multiple pathways, protein conformations become sequentially lower in energy (Anfinsen, 1973), and lower-energy states are able to access fewer conformations (Dill and Maccallum, 2012; Englander and Mayne, 2014; Martínez, 2014). So, small perturbations in peptide conformation that favor the native state incrementally stabilize the protein and lead to the native structure (Dill, 1990; Martínez, 2014). Rough energy landscapes also account for concepts like folding frustration, where the peptide creates transiently-favorable non-native contacts and becomes temporarily trapped in a local energy well (Figure 1.1) (Martínez, 2014). Despite these insights, protein folding is still a largely unpredictable process, and specific folding mechanisms remain obscure for many proteins.

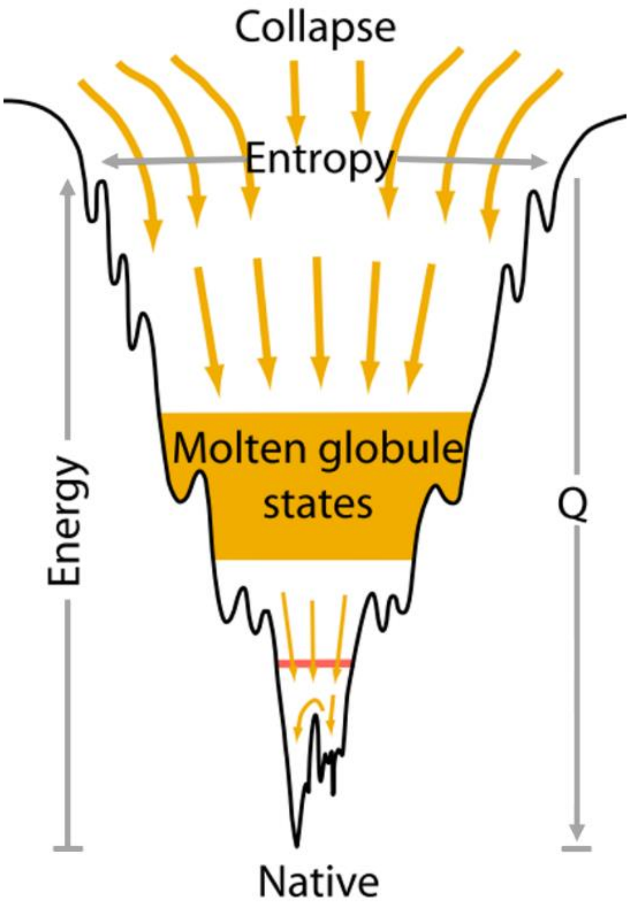


Figure 1.1. Protein folding is modeled as a funnel-shaped energy-landscape. Protein folding occurs through multiple pathways (indicated by yellow arrows). Unfolded peptide chains initially have high entropy (represented by the width of the funnel) and can access many conformations. As proteins fold, native contacts form, and fewer conformations are accessible. Native-like contacts lower the peptide's Gibbs free energy (represented by the depth of the funnel), and the protein's native state occupies the global minimum free energy. Q represents the fraction of native contacts formed. The red line indicates the transition state. Adapted from Englander and Mayne (2014).

Protein folding remains an enduring challenge in protein engineering. Specifically, protein design is hampered by the inverse protein folding problem: reliably predicting amino acid sequences that will fold into a target structure is difficult and often inaccurate (Yue and Dill, 1992; Dill and Maccallum, 2012).

Computational methods (*e.g.* Rosetta (Rohl et al., 2004; Kaufmann et al., 2010)) strive to mitigate the inverse protein folding problem by extensively sampling sequence- and conformational space (Khoury et al., 2014; Baek et al., 2021; Jumper et al., 2021; Mirdita et al., 2022). Computational methods substantially increase the number of putative sequences that can be explored, and thus increase the probability of finding an optimal sequence for a given fold. Design algorithms use template-guided search methods and implement physics- and knowledge-based force-fields to mimic molecular behavior in natural sequences (Raugei et al., 2006; Khoury et al., 2014). Importantly, template-guided searches bias design algorithms to choose amino acids that favor backbone conformations and residue interactions known to adopt the desired fold in nature. Additionally, predicted biophysical metrics (*e.g.* ΔΔG) can be used to score the fitness of designed sequences *in silico*, eliminating the need for large-scale experimental evaluation (Goldenzweig et al., 2016). Thus, computational protein design attempts to bypass the inverse protein folding problem by testing many sequences in many conformations with less need to restrict sample size due to limited experimental resources. However, these methods are not infallible. Proteins that lack homologous sequences or structures are especially difficult to model, and their design often results in failure (Khoury et al., 2014). Even design of well-characterized proteins is not trivial since many high-scoring constructs fail to express at appreciable levels or lack sufficient stability *in vitro* (Broom et al., 2017, 2020; Rocklin et al., 2017). Thus, protein design requires considerable attention to protein folding and stability.

1.3 Kinetic stability: an enduring challenge in protein engineering

Protein stability is a key determinant of a protein's capacity to remain folded. Most natural proteins display moderate stability and are poorly adapted to harsh conditions used in protein applications (Magliery, 2015). Tremendous efforts have been invested in improving stability in industrial proteins (Estell et al., 1985; Chen and Arnold, 1993; Bornscheuer et al., 2002, 2012). However, stability design remains a laborious and poorly understood undertaking. Beneficial mutations are difficult to predict, and the vast majority of mutations are neutral or destabilizing (Magliery, 2015; Nisthal et al., 2019; Broom et al., 2020). Stability design is further complicated by inconsistencies in the scientific literature. While protein stability ideally should be discussed in terms of thermodynamic and kinetic stabilities, most studies confound the two under the blanket term thermostability. Since thermodynamic and kinetic stabilities are functionally distinct and operate using different molecular mechanisms, this confuses stability design and obscures understanding of protein stability (Sanchez-Ruiz, 2010).

Protein stability is a complex interplay between thermodynamic and kinetic stabilities (Figure 1.2). Thermodynamic stability refers to the free energy difference between a protein's native and unfolded states (Figure 1.2A) (Brissos et al., 2014). Thermodynamically stable proteins have lower Gibbs free energy in their native states relative to their unfolded states, and their equilibrium favors the folded protein (Sanchez-Ruiz, 2010). Kinetic stability is defined by the free energy barrier between the native state and the transition state (Figure 1.2B) (Sanchez-Ruiz, 2010; Sun et al., 2019). Kinetically stable proteins have high free energy barriers and unfold slowly (Broom et al., 2015b). So, thermodynamic stability determines the population of native protein, while kinetic stability dictates the protein's functional lifetime. Proteins can exert their biological function with low thermodynamic stability as long as they have sufficient kinetic stability to maintain their tertiary structure over a sufficiently long physiological time scale (Sanchez-Ruiz, 2010). Kinetically stable proteins tend to be resistant to irreversible processes like proteolytic degradation and aggregation, resulting in drastically reduced protein turnover (Broom et al., 2015b; Colón et al., 2017). Further, increasing kinetic stability correlates with improved catalytic yield (McLendon G, 1978), longer shelf-life (Luo et al., 2002), and higher tolerance to denaturing conditions (Brissos et al., 2014). Thus, kinetic stability is critically important for designing proteins for practical applications.

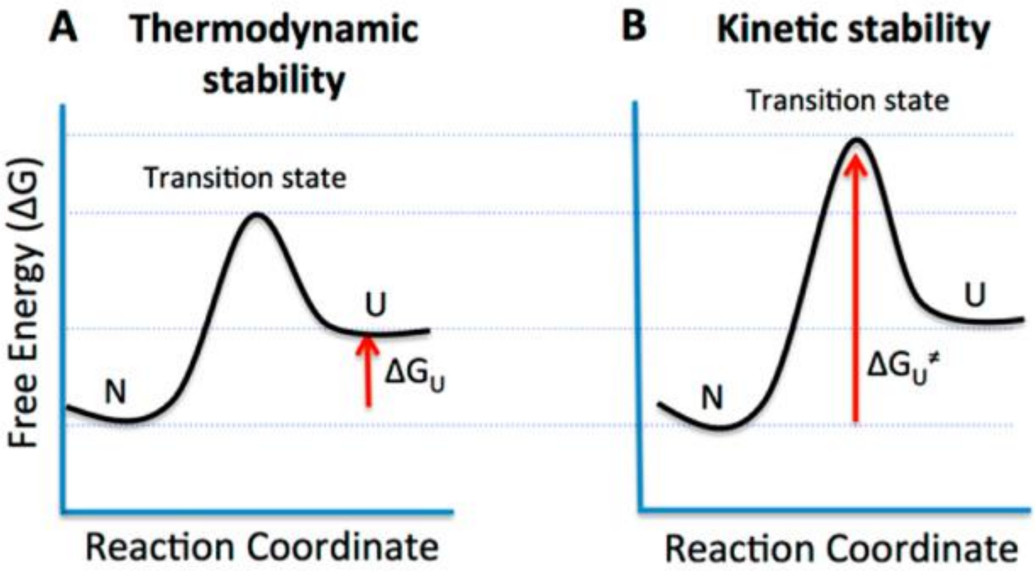


Figure 1.2. Protein thermodynamic and kinetic stabilities are energetically different. (A) Thermodynamic stability (ΔG_U) is given by the difference in Gibbs free energy between the protein's native (N) and unfolded (U) states. **(B)** Kinetic stability ($\Delta G^{\neq}$) is the difference in Gibbs free energy between the proteins native state and the transition state. Adapted from Sun *et al*. (2019).

Few studies address the rational design of kinetic stability. Previous strategies for improving kinetic stability focused on improving protein kinetic stability by reducing local unfolding or increasing protein rigidity. For example, introducing disulfide bonds and proline residues enhanced kinetic stability in barnase and thermolysin-like protease by decreasing local unfolding and peptide mobility (Mansfeld et al., 1997; Van den Berg et al., 2010). Kim *et al*. (2012) increased protein rigidity and half-life in xylase by engineering hydrophobic networks in the protein core. However, prioritizing protein rigidity is often detrimental to protein function because proteins must undergo conformational changes to achieve substrate binding and catalysis (Kim et al., 2012; Xie et al., 2014). Alternatively, researchers have targeted residue B-factors for designing kinetic stability (Sun et al., 2019). B-factors indicate the conformational flexibility for a given residue in a protein’s crystal structure, and residues with high B-factors are considered “hot spots” for instability. Mutating residues with high B-factors has increased kinetic stability in a variety of industrially-important proteins, including lipase b (Le et al., 2012), sucrose isomerase (Duan et al., 2016), and pullulnase (Chen et al., 2015). While promising, B-factor engineering significantly limits which proteins can be targeted for kinetic stability improvement as it requires a crystal structure of the protein. Circumventing this limitation, Quezada *et al.* (2018) used high-temperature molecular dynamics (MD) simulations to identify thermally flexible residues and enhance the kinetic stability of a triosephosphate isomerase (TIM) by mutating 35 thermoflexible residues to structurally equivalent thermostable residues from a homologous TIM protein. The success of this strategy decreased with fewer point mutations, suggesting that kinetic stability may result from concerted effects between many interacting residues. Recently, Liu *et al.* (2021) combined rational disulfide engineering or B-factor engineering with additional selection criteria including MD simulations, change in Gibbs free energy of unfolding calculations, and visual inspection to exclude unreasonable mutations to great effect, reporting substantial improvement in kinetic stability in 10 of 42 mutants without complete loss of activity. However, this method was extremely labor-intensive, and successful mutants showed no unified mechanism for stabilization. So, despite variable success, molecular determinants of kinetic stability remain poorly understood and no reliable approach for engineering kinetic stability is established (Sanchez-Ruiz, 2010; Sun et al., 2019; Musil et al., 2019).

1.4 Relating protein topology to protein folding and kinetic stability

Protein topology describes the geometric conformation of a protein’s backbone, which encompasses local α-helical and β-stranded secondary structures and their tertiary arrangement.

Simple empirical parameters for structural complexity have been proposed. In particular, Absolute Contact Order (ACO) and Long-Range Order (LRO) are known to correlate with protein folding rates (Figure 1.3A-D). ACO reflects the relative importance of local and non-local contacts in the native structure. ACO is given by:

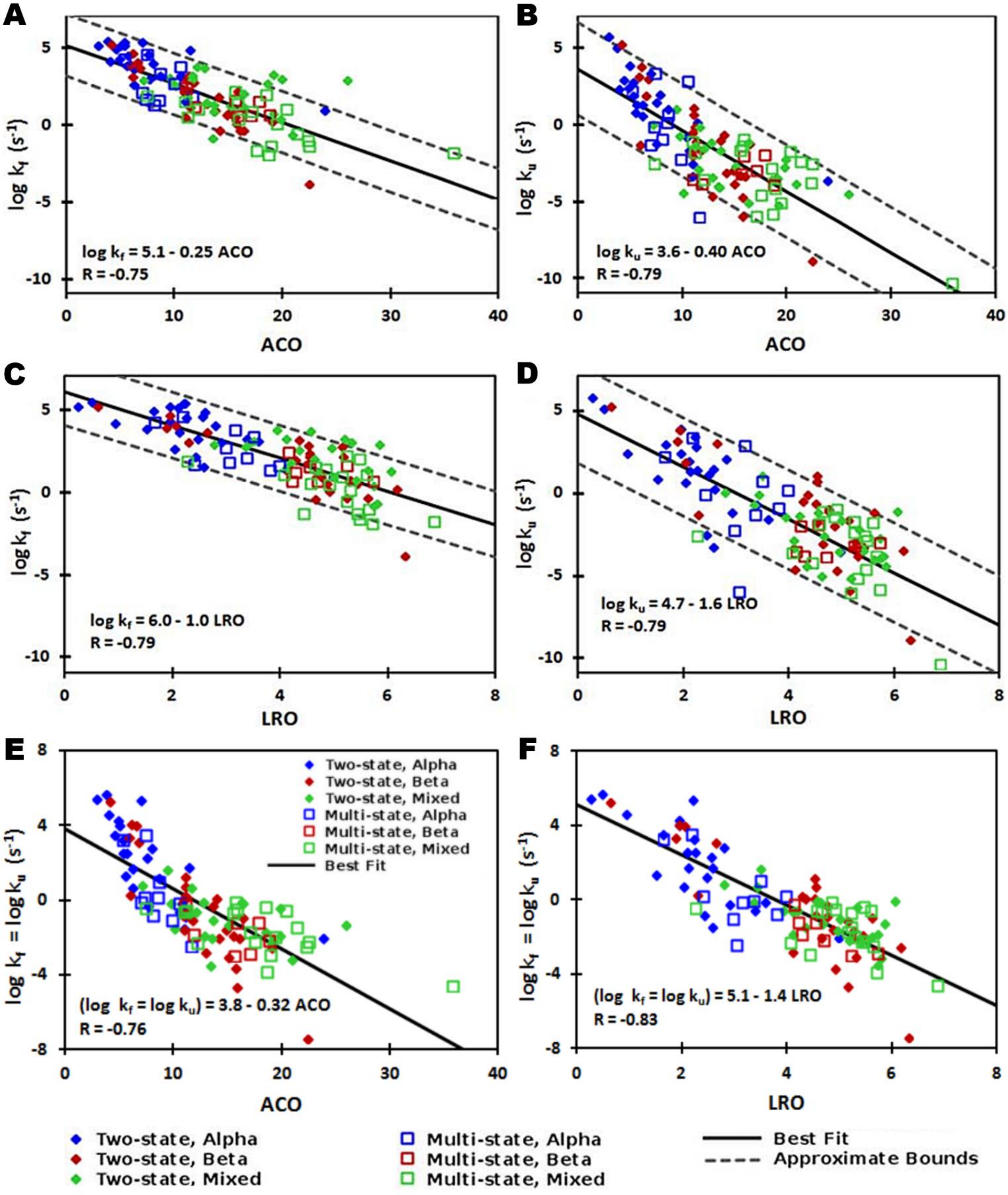
$$ACO = \frac{1}{N_c}\sum_{i,j}^{N_c} |i - j| \qquad (\mathbf{1})$$

where N_c is the total number of contacting atom pairs (where heavy atoms are considered to be contacting if they are less than 6 Å apart, and hydrogen atoms are ignored), and $|i\text{-}j|$ is the number of residues in the sequence separating residue pairs *i* and *j*. ACO is generally smaller for proteins mainly stabilized by short-range contacts and large for those with many long-range contacts (Ivankov et al., 2003). LRO reports structural complexity from the number of long-range contacts normalized to chain length, and is given by:

$$LRO = \frac{1}{L}\sum_{i,j}^{R_c} n_{i,j} \qquad (\mathbf{2})$$

where L is the chain length, R_c is the total number of contacting residue pairs, and $n_{i,j}$ is equal to one when $|i\text{-}j| \geq 12$ and 0 otherwise (Gromiha and Selvaraj, 2001). Broom *et al*. (2015a) recently showed that both ACO and LRO correlate equally well with folding and unfolding rates, where both rates decrease with increasing structural complexity (Figure 1.3A-D).

Measures of structural complexity may also enable predictions for protein stability. Broom *et al*. (2015a) showed that increasing LRO and ACO correlate strongly with decreasing protein unfolding rates at the transition midpoint (Figure 1.3E). This suggests that LRO may report on the structural complexity and relative energy of the transition state under conditions of thermodynamic equilibrium (Broom et al., 2015a). Critically, comparing proteins under conditions of equal thermodynamic stability at the transition midpoint, *i.e.* where the equilibrium free energy of unfolding (ΔG_U) is zero, allows identification of differences in kinetic unfolding barrier heights (Gosavi, 2013; Broom et al., 2015a, 2015b). In a separate study, Broom *et al*. (2015b) showed that unfolding free energy barriers simulated using structure-based models coarse-grained to a C_α bead (C_α-SBM) increased at higher LRO and that proteins with high LROs tend to have longer half-lives. Thus, LRO may be useful in broadly predicting kinetic stability for native and designed proteins.



A
log k_f (s^{-1})
5
0
-5
-10
0
10
20
30
40
ACO
log k_f = 5.1 - 0.25 ACO
R = -0.75
B
log k_u (s^{-1})
5
0
-5
-10
0
10
20
30
40
ACO
log k_u = 3.6 - 0.40 ACO
R = -0.79
C
log k_f (s^{-1})
5
0
-5
-10
0
2
4
6
8
LRO
log k_f = 6.0 - 1.0 LRO
R = -0.79
D
log k_u (s^{-1})
5
0
-5
-10
0
2
4
6
8
LRO
log k_u = 4.7 - 1.6 LRO
R = -0.79
E
log k_f = log k_u (s^{-1})
8
4
0
-4
-8
0
10
20
30
40
ACO
Two-state, Alpha
Two-state, Beta
Two-state, Mixed
Multi-state, Alpha
Multi-state, Beta
Multi-state, Mixed
Best Fit
(log k_f = log k_u) = 3.8 - 0.32 ACO
R = -0.76
F
log k_f = log k_u (s^{-1})
8
4
0
-4
-8
0
2
4
6
8
LRO
(log k_f = log k_u) = 5.1 - 1.4 LRO
R = -0.83
◆ Two-state, Alpha
◆ Two-state, Beta
◆ Two-state, Mixed
□ Multi-state, Alpha
□ Multi-state, Beta
□ Multi-state, Mixed
Best Fit
Approximate Bounds

Figure 1.3. Structural complexity correlates strongly with protein folding and unfolding rates. A dataset of 108 two-state and multi-state folders from across structural protein classes was analyzed to determine correlations between measures of structural complexity, ACO and LRO, and protein folding and unfolding rates. Correlations are shown for **(A)** ACO and folding rate, **(B)** ACO and unfolding rate, **(C)** LRO and folding rate, **(D)** LRO and unfolding rate, **(E)** ACO and folding and unfolding rates at the transition midpoint, and **(F)** LRO and folding and unfolding rates at the transition midpoint. Lines of best fit (solid black) and correlation values are given for the whole dataset. Values for two-state (filled diamonds), multi-state (open squares), alpha (blue), beta (red) and mixed (green) protein subsets are given in Table 1 of Broom *et al.* (2015a). Dashed lines in panels A-D represent $\pm$10-fold and $\pm$100-fold variation in folding and unfolding rate constants, respectively. Adopted from Broom *et al.* (2015a).

1.5 Coarse-grained structure-based models as predictive measures of kinetic stability

Beyond simple LRO calculations, C_α-SBM folding simulations represent a promising method for probing trends in unfolding free energy barriers to predict relative protein kinetic stability (Chavez et al., 2004; Broom et al., 2015b). In C_α-SBM, amino acids are represented as single beads located at the C_α atom connected along the protein backbone. Atom-atom interactions are extracted from the protein's native structure, and all interactions for a given residue are projected onto its C_α atom (Clementi et al., 2000). The potential energy of the system is given by the Hamiltonian

$$\mathcal{H}_{C\alpha}(\vec{x},\vec{x}^{o}) = \sum_{ij\in bonds} \frac{\varepsilon_b}{2}\left(r_{ij}-r_{ij}{}^{o}\right)^2 + \sum_{ijk\in angles} \frac{\varepsilon_\theta}{2}\left(\theta_{ijk}-\theta_{ijk}{}^{o}\right)^2 + \sum_{ijkl\in dihedrals} \epsilon_D F_D\left(\varphi_{ijkl}-\varphi_{ijkl}{}^{o}\right)$$

$$+ \textstyle\sum_{ij\in contacts} \epsilon_C \left[5\left(\frac{r_{ij}{}^{o}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{ij}{}^{o}}{r_{ij}}\right)^{10}\right] + \textstyle\sum_{ij\notin contacts} \epsilon_{NC} \left(\frac{\sigma_{NC}}{r_{ij}}\right)^{12} \qquad \textbf{(3)}$$

where $F_D(\delta\varphi) = [1-\cos(\delta\varphi)] + \frac{1}{2}[1-\cos(3\delta\varphi)]$ (Clementi et al., 2000; Noel and Onuchic, 2012; Noel et al., 2016). By treating native contacts as attractive interactions and non-native contacts as repulsive interactions, the Hamiltonian imposes a native bias on the system (Noel and Onuchic, 2012). Importantly, this bias ensures that the global energy minimum is represented by the native protein structure and directs the folding trajectory to achieve the correct fold. The free energy of unfolding can be modeled at the protein's folding temperature (T_f), which is functionally equivalent to the transition midpoint (Broom et al., 2015b). Significantly, Chavez *et al.* (2004) found that the heights of the free energy barrier for unfolding correlate extremely well with protein folding rates at the protein folding temperature for small globular proteins (Figure 1.4) (Chavez et al., 2004). Thus, C_α-SBM modeling represents a powerful tool for predicting protein kinetic stability. Accordingly, we will use C_α-SBM simulations to evaluate protein kinetic stability in our β-trefoil model.

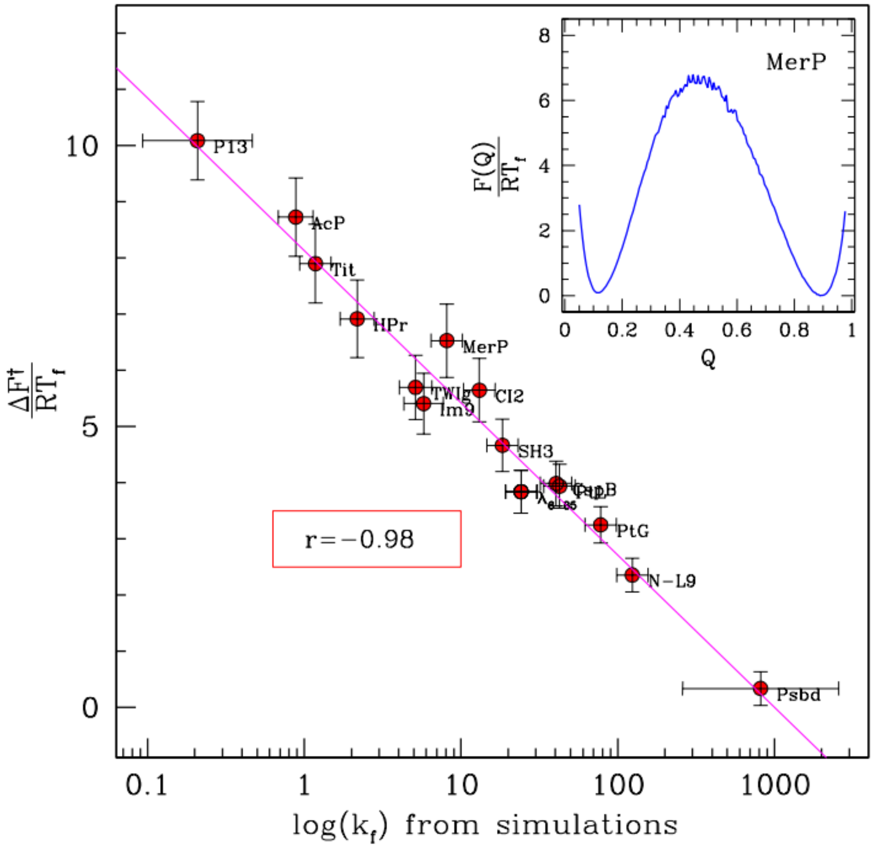


Figure 1.4. Coarse-grained unfolding free energy barriers correlate strongly with protein folding rates. Folding rates from molecular dynamic simulations [log(k_f)] for a set of energetically unfrustrated single domain proteins are plotted against coarse-grained barrier heights ($\Delta F^{\ddagger}/RT_f$) extracted from free energy profiles as a function of Q, the fraction of native contacts formed. Folding rates are measured at the protein folding temperature (T_f) and correlate well with experimentally defined folding rates. Folding rates correlate strongly with unfolding free energy barrier heights. A representative free energy barrier of unfolding is shown in the inset, where MerP is a mixed β-stranded and α-helical protein with a folding rate constant of ~1. A list of relevant dataset proteins is given in Table 1 of Chavez *et al.* (2004) Adapted from Chavez *et al.* (2004).

1.6 β-trefoils as a model system for kinetic stability.

The β-trefoil fold is a common motif in natural proteins. β-trefoil domains currently include 23 superfamilies in the Pfam database, and over 8000 sequences are predicted to adopt the fold (Terada et al., 2017; Mistry et al., 2021). β-trefoils have three structural repeats, or “trefoils”, and display internal symmetry (Figure 1.5). β-trefoils are comprised of 12 β-stands arranged into a six-stranded β-barrel and a six-stranded triangular cap, where one hairpin from the barrel and one from the cap form each foil (Figure 1.5) (Murzin et al., 1992; Broom et al., 2012). Residues along each β-strand alternate pointing into the protein interior or outwards towards the protein surface (Murzin et al., 1992). Inward-facing residues form the protein core, which consists of 18 highly conserved hydrophobic residues (Murzin et al., 1992). β-strands are connected by turns and loops of variable length, which account for up to 70% of the fold's exposed surface area (Murzin et al., 1992). Despite their common fold, β-trefoil proteins show low sequence identity, diverse ligand-binding functions, and variable binding-site localization (Ponting and Russell, 2000; Gosavi, 2013; Blaber, 2022). Due in part to their internal symmetry and functional diversity, β-trefoils have been used as model proteins to study protein folding (Gosavi, 2013), evolution (Broom et al., 2012; Longo et al., 2014), and design (Lee and Blaber, 2011; Broom et al., 2012; Terada et al., 2017).

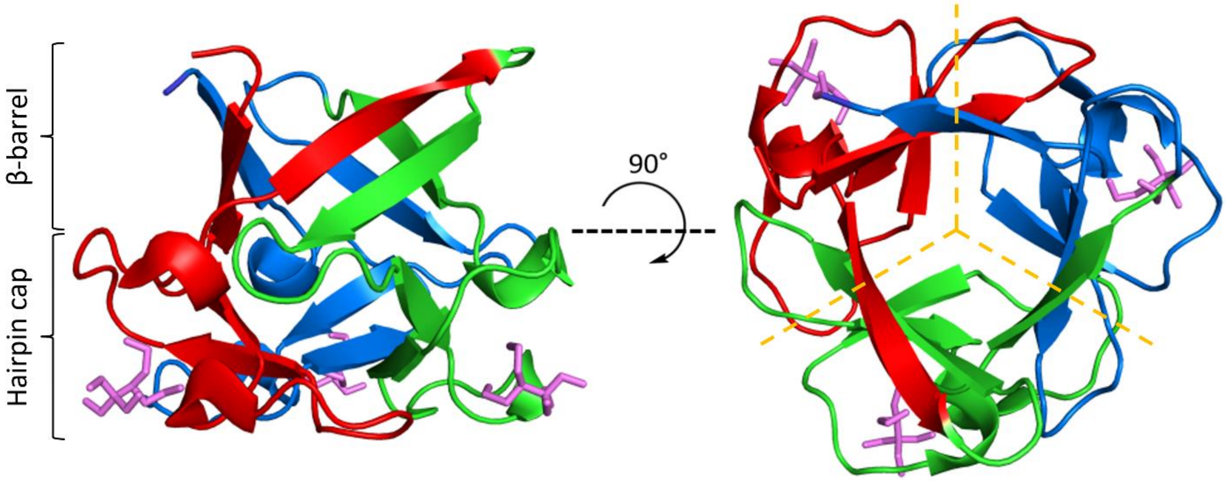


Figure 1.5. The β-trefoil fold. ThreeFoil (PDB ID: 3PG0) is a representative model of the β-trefoil fold. β-trefoils consist of 12 β-strands arranged in a β-barrel and a hair-pin cap (left). β-strands are connected by loops and turns of variable length. β-trefoils bind diverse ligands through their loops (ligands shown in purple). β-trefoils display internal pseudo three-fold symmetry (right). Symmetric foils are indicated by dashed yellow lines.

β-trefoils offer a powerful system for studying the molecular basis of kinetic stability. Despite their common fold, β-trefoils display great diversity in primary sequences and a wide range of stabilities (Murzin et al., 1992; Ponting and Russell, 2000; Broom et al., 2012; Gosavi, 2013). This diversity offers a large sequence space for sampling favorable amino acid mutations (Broom et al., 2012; Wu et al., 2019). Of particular interest are the β-trefoil proteins 3Foil (Figure 1.5) and wtHis. 3Foil is a completely symmetric designed protein that displays extreme kinetic stability (Longo et al., 2012; Broom et al., 2015b). Extraordinarily, 3Foil has an unfolding half-life of approximately 8 years (Broom et al., 2015b). In contrast, the natural β-trefoil hisactophilin has only moderate kinetic stability and a typical unfolding half-life in the range of minutes to hours (Broom et al., 2015b). Thus, 3Foil and wtHis together encompass a range of kinetic stabilities inherent in β-trefoil proteins, making them good candidate proteins for developing a kinetic stability model.

3Foil and wtHis provide a compelling model for using LRO, ACO, and C_α-SBM unfolding free energy barriers to guide kinetic stability design. Previous C_α-SBM simulations show that 3Foil's remarkable kinetic stability arises from numerous long-range contacts between loop residues and across the protein core, which result in an unusually large unfolding free energy barrier (Figure 1.6) (Broom et al., 2015b). C_α-SBM simulations predict that deleting long-range contacts significantly lowers the 3Foil free energy barrier of unfolding, thus decreasing its kinetic stability (Mut2 in Figure 1.6) (Broom et al., 2015b). In contrast, the natural β-trefoil protein wtHis has markedly fewer long-range contacts, a low unfolding barrier (Figure 1.6) (Gosavi, 2013; Broom et al., 2015b), and moderate kinetic stability with a typical unfolding half-life of minutes to hours (Smith et al., 2010). Notably, conserved core residues in wtHis form a large, functional cavity that spans the protein core (Smith et al., 2010; Shental-Bechor et al., 2012; Mackenzie et al., 2022). This cavity precludes the formation of long-range contacts between core residues and contributes to wtHis' low LRO and unfolding free energy barrier. Equivalent residues in 3Foil are tightly packed and form many stabilizing long-range intramolecular contacts that contribute significantly to 3Foil's high LRO. These core residues offer a clear starting point for modulating kinetic stability in wtHis and 3Foil.

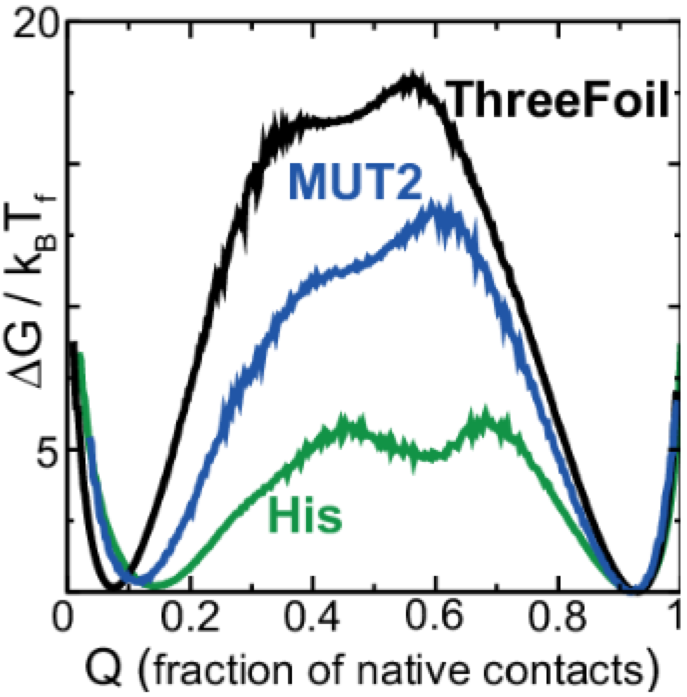


Figure 1.6. Long-range contacts modulate the free energy barrier of unfolding. Free energy barriers for unfolding for ThreeFoil (3Foil) (black) and hisactophilin (His) (green) were simulated using C_α-SBM (right). The free energy barrier for Mut2 (blue), a 3Foil mutant where long-range contacts have been deleted (left), was also simulated. Folding free energies are plotted at the transition midpoint (T_f) as a function of the fraction of native contacts (Q) and free energy ($\Delta G/k_BT_f$). The unfolding free energy barrier height decreases with fewer long-range contacts. Long-range contacts deleted in Mut2 are listed in the Appendix of Broom *et al*. (2015b). Adapted from Broom *et al.* (2015b).

Here, we describe the design and characterization of a hisactophilin variant, core-swapped hisactophilin (csHisH90G), mutated to contain 3Foil core residues. The goal of this core-swapped design is two-fold: 1) to assess using LRO, ACO, and C_α-SBM unfolding free energy barriers as predictive measures to rationally design kinetic stability; and 2) to test enhancing protein kinetic stability within the confines of the protein's existing chain length and fold (Thirumalai, 1995). Using LRO, ACO, and C_α-SBM, we predict a moderate increase in csHisH90G topological complexity and unfolding free energy barrier height, both indicating improved kinetic stability relative to wtHis. Addressing concern for csHisH90G thermodynamic stability, we use the protein stability prediction tool Protein Repair One-Stop Shop (PROSS) to design stabilized csHisH90G variants (Goldenzweig et al., 2016), which we also screen for kinetic stability using LRO, ACO, and C_α-SBM simulations. Based on these predictions, we experimentally expressed and purified a single csHisH90G design and three PROSS mutants, which we show to be well-behaved and well-folded *in vitro*. Experimental kinetic folding and unfolding measurements confirm that csHisH90G and fa-csHisH90G display greater kinetic stability compared to their parent proteins. Further, kinetic data and folding simulations show that csHisH90G displays folding behavior intermediate to wtHis and 3Foil. We discuss the advantages and limitations of using LRO, ACO, and C_α-SBM simulations for protein kinetic stability design compared to prevailing methods. Finally, we propose that engineering protein cores, either by swapping conserved hydrophobic core residues in homologous protein folds or using *de novo* core packing software, may offer a feasible and robust strategy for improving kinetic stability in designed proteins.

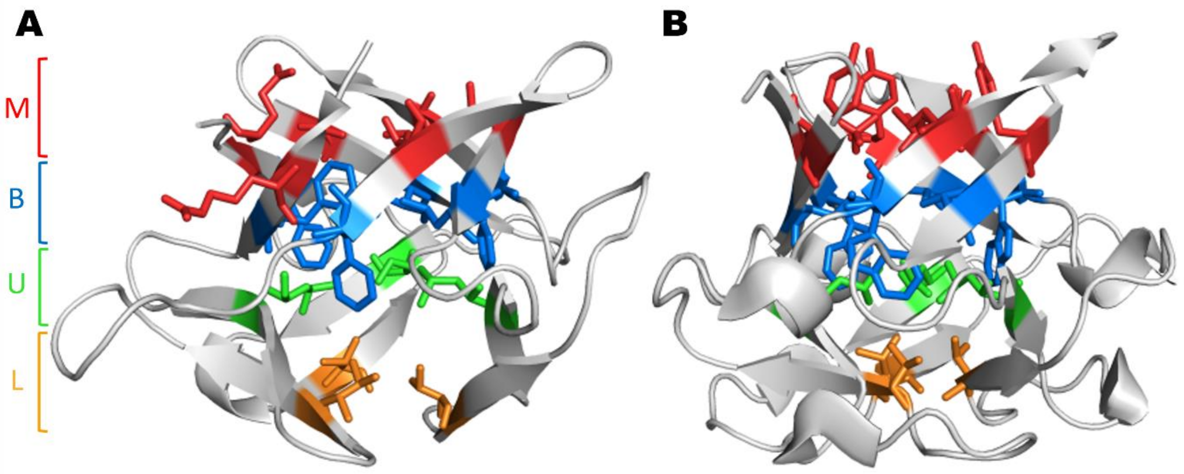
2 Materials and Methods

2.1 In silico

LRO, ACO, and C_α-SBM free energy barriers were used as predictive values to guide kinetic stability design in wtHis and 3Foil variants. *In silico* methods focused primarily on identifying structurally equivalent residues in wtHis and 3Foil that, when swapped into wtHis, sufficiently increase the LRO, ACO, and unfolding barrier heights in C_α-SBM simulations and, thus, predict increased kinetic stability of hisactophilin variants (Figure 3.1; Table 3.1).

2.1.1 Identifying target residues for kinetic stability design

Residues *i* and *j* in a given protein are said to be in contact if a pair of heavy atoms belonging to residues *i* and *j* are in close proximity in the protein's folded state. In C_α-SBM, used herein, a contact between a pair of atoms is projected onto the C_α atoms of the corresponding residues *i* and *j*. All contacting residue pairs are compiled in a list (see Appendix A). A contact map is a symmetric plot of this list with both x and y axes denoting residue numbers. Colored boxes are marked on this plot at (*i*, *j*) and (*j*, *i*) when a contact is present between residues *i* and *j* (Figures 3.1-3.4D). Contact maps were generated for energy-minimized structures of wtHis (PDB ID: 1HCD) and 3Foil (PDB ID: 3PG0) using the C_α Shadow algorithm available on the SMOG2 web server (Clementi et al., 2000; Noel et al., 2012, 2016). Shadow maps used default parameters of a 6 Å maximum contact cutoff and 1 Å atom "shadowing" radius (Noel et al., 2012, 2016). wtHis and 3Foil were aligned using a sequence-based structure alignment, and equivalent residues were identified (Figure 3.1A). Residue pairs in wtHis and 3Foil contact maps were compared to identify conserved networks of interacting, structurally equivalent residues in which wtHis residues make fewer contacts than those of 3Foil (Figure 3.1; see Chapter 3 Results for details). Identification of long-range interaction networks was prioritized over local interactions since alteration of long-range interactions is captured in ACO and LRO measures, while local interactions are only represented by ACO. Additionally, LRO provides a stronger, more linear correlation for proteins of variable size since it is normalized to chain length (Broom et al., 2015b). 3Foil and wtHis core residues were identified as promising targets for modulating LRO and ACO in a residue-swapped wtHis/3Foil hybrid since wtHis core residues make significantly fewer long-range contacts than 3Foil core residues (Figure 2.1).



C

	wtHis		3Foil		csHisH90G	
Layer	Residue	No. long-range contacts	Residue	No. long-range contacts	Residue	No. long-range contacts
M	R4	7.5	Y5	7.5	Y4	7.5
	V43	3	Y52	4	Y43	4
	V83	4.5	Y99	5	Y83	4.5
	V36	2	L44	4	L36	2.5
	L76	2	L91	4	L76	3.5
	E115	5	L138	5	L115	3.5
B	F6	5.5	L7	7	L6	5.5
	L45	4	L54	4.5	L45	5
	I85	3.5	L101	4.5	L85	3
	F34	11	W42	13	W34	12.5
	F74	9	W89	13	W74	10
	F113	10	W136	13	W113	9
U	L14	5	L16	7.5	L14	4.5
	L53	6	L63	8.5	L53	6.5
	I93	5.5	L110	8.5	L93	7
L	V21	2	V29	9	V21	4
	V61	3	V76	9	V61	3
	V101	3.5	V123	9	V101	3.5
Total LR contacts		92		136		99

Figure 2.1. ThreeFoil core residues make significantly more long-range contacts than hisactophilin core residues. (A) Hisactophilin (wtHis) and **(B)** ThreeFoil (3Foil) conserved core residues are colored by layer according to their position in the β-trefoil fold (indicated in panel C). **(C)** Conserved core residues are listed for wtHis, 3Foil, and core-swapped hisactophilin (csHisH90G), where 11 of 18 residues differ between wtHis and 3Foil. Long-range contacts made between two core residues are counted as 0.5 long-range contacts per residue. The number of core residue long-range contacts is markedly higher in 3Foil relative to wtHis. Note that the specific contacts made by a given residue may differ between proteins (*e.g.* wtHis R4, which points toward solvent, makes different contacts than 3Foil Y5, which points into the protein core).

The hisactophilin point mutant H90G (HisH90G) was identified in previous equilibrium denaturation experiments to be thermodynamically stabilized compared to wtHis (MacKenzie et al., 2022). Here, HisH90G was used as a pseudo-wild type parent protein for the core-swap design, and the H90G point mutation was included in core-swapped hisactophilin variants (see Results).

2.1.2 Generating the core-swapped model

Ten structural models of csHisH90G were generated using Robetta Comparative Modeling (CM) (Chivian et al., 2003; Song et al., 2013) using wtHis and 3Foil structures as templates. Template alignments were modified such that csHisH90G core residues and β-strands 1 and 12 (residues 1-7 and 112-115, respectively) were modeled after 3Foil residues only, and csHisH90G hairpin cap residues (residues 11-29, 48-73, and 89-100) were modeled after hisactophilin residues only (Figure 3.1A). Models with buried Y4/43/83 hydroxyl groups were discarded due to the high energy cost of burying the polar group in the protein's hydrophobic core. LRO and ACO scores were calculated for each model, and outliers were identified using the interquartile range method and discarded. Remaining csHisH90G models all displayed increased LRO and ACO scores compared to wtHis, as expected. Finally, models were assessed using PROCHECK (Laskowski et al., 1993; Laskowski et al., 1996) and MolProbity (Williams et al., 2018) Ramachandran scores, and the model with the most favorable Ramachandran score was chosen as the structural model for csHisH90G for all subsequent computational methods.

2.1.3 Predicting the unfolding rate constant of csHisH90G

The csHisH90G unfolding rate constant was predicted using the linear correlations for the unfolding rate constant and LRO or ACO reported by Broom *et al.* (2015a) for β proteins (Table 3.1). The linear correlation for LRO and the unfolding rate constant at the transition midpoint is given by:

$$k_{u,C_{mid}} = -1.70(LRO) + 6.6 \quad \textbf{(4)}$$

where $k_{u,C_{mid}}$ is the unfolding rate constant at the transition midpoint. The linear correlation for ACO and the unfolding rate constant at the transition midpoint is given by:

$$k_{u,C_{mid}} = -0.52(ACO) + 5.9 \quad \textbf{(5)}$$

Experimental structural information is unavailable for HisH90G. However, LRO, ACO, and C_α-SBMs are based on contact maps, and the H90G point mutation has little effect on the hisactophilin contact map. Specifically, H90 makes only two long-range intramolecular contacts (H90, K104 and H90, E105) and one short-range contact (K86, H90) in wtHis, and G90 makes identical contacts in the structural model for csHisH90G. Thus, the structural substitution of wtHis for HisH90G is reasonable for LRO, ACO, and C_α-SBM simulation predictions. It should, however, be noted that sequence specific effects (*e.g.* secondary-structural propensities) due to the H90G mutation that may affect hisactophilin folding are not captured in LRO, ACO, or C_α-SBM methods. Predicted LRO, ACO, and C_α-SBM unfolding free energy barrier heights for wtHis and csHisH90G are given in Table 3.1.

2.1.4 Predicting the free energy barrier of unfolding using protein folding simulations

All simulations were carried out using the GROMACS v.4.5.4 software package (Bekker, H., Berendsen, H. J. C., Dijkstra, E. J., Achterop, S., van Drunen, R. et al., 1993; Berendsen et al., 1995; Lindahl et al., 2001; Van Der Spoel et al., 2005; Hess et al., 2008). GROMACS geometry and topology files were generated for wtHis and csHisH90G using the AMBER99SB-ILDN force field and TIP3P water model (Jorgensen et al., 1983; MacKerell et al., 1998; Lindorff-Larsen et al., 2010). All protein hydrogens were ignored. Solvent molecules were replaced with Na^+ or Cl^- ions until the system reached net neutral charge. Energy minimization simulations were performed for 2000 steps using the method of steepest descent. Energy-minimized wtHis and csHisH90G structures were used in subsequent C_α-SBM simulations.

Protein folding for wtHis and csHisH90G was investigated using C_α-SBM simulations. Proteins fold on a biologically reasonable timescale because of a funnel-shaped energy landscape in which interactions (or contacts) present in the native state of the protein are more stabilizing than any non-native interactions that occur during protein folding (Bryngelson et al., 1995; Wolynes et al., 1995; Onuchic et al., 1997; Onuchic and Wolynes, 2004). Structure-based models (SBMs) encode this funnel in their potential energy functions by ignoring attractive non-native interactions and encoding attractive native interactions through inter-residue contacts calculated from the native structure. The coarse-grained C_α-SBM used here to simulate wtHis, csHisH90G, and 3Foil has previously been used successfully to simulate the folding of several proteins (Clementi et al., 2000; Chavez et al., 2004; Gosavi et al., 2006, 2008; Hills and Brooks, 2009; Hyeon and Thirumalai, 2011; Gosavi, 2013; Broom et al., 2015b; Giri Rao and Gosavi, 2018; Lalwani Prakash and Gosavi, 2021). The exact form of the potential energy function of this C_α-SBM is in equation 3 of Chapter 1 (Clementi et al., 2000). Geometry, topology, table, and parameter files required for C_α-SBM simulations were obtained from the SMOG2 webserver (Appendix B, Scheme 1) (Clementi et al., 2000; Noel et al., 2012, 2016). Contact maps were generated using the same criteria given above.

C_α-SBM simulations were performed using a stochastic dynamics integrator with a 0.0005 ps time step. All simulations were performed using the NVT ensemble. Proteins were simulated at their respective folding temperatures (T_f), which is defined as the temperature at which the folded and unfolded states are equally populated and folding transitions occur from both the unfolded and folded states to ensure reasonable sampling of the transition state ensemble (TSE). Unfolded protein geometry files for wtHis and csHisH90G were obtained by running short, high temperature (T = 230 K) simulations. Note that coarse-grained GROMACS simulations use reduced temperature units that do not directly correspond to experimental temperatures. Preliminary simulations were initiated using the native protein geometry and the unfolded protein geometry for each temperature and performed for $1\text{x}10^8$ time steps (50 ns). Folding temperatures for wtHis and csHisH90G were determined by performing preliminary simulations over iteratively smaller temperature ranges until folding transitions (*i.e.* the folded state transitioned to the unfolded state or *vice versa*) occurred from folded and unfolded structures and the populations of both states were approximately equal. Production runs were performed at the T_f for a total of $2\text{x}10^{10}$ time steps (10 μs). Folding simulations for 3Foil, which required enhanced sampling due its unusually large free energy barrier, are described in Appendix B.

Since through space attractive interactions are primarily encoded in the C_α-SBM through native contacts, the fraction of native contacts (Q) is often used as a progress coordinate (Clementi et al., 2000; Chavez et al., 2004). Here, we plot the unfolding free energy barrier as a function of Q (Figure 3.1-3.4E). The number of formed contacts is calculated for every simulation snapshot.

A contact is said to be formed if the distance between the contacting residues is less than 1.2 times their distance in the folded structure. The Q of a given snapshot is the number of contacts formed in that snapshot divided by the total number of native contacts. Snapshots are then pooled and binned based on their Q into a histogram P(Q). The free energy F(Q) is then equal to -ln(P(Q)) and is plotted as a function of Q. This plot has at least two minima: one at low Q that represents the unfolded minimum, and one at high Q that represents the folded minimum. The free energy barrier separating these minima is the unfolding free energy barrier. To compare free energy barriers of different proteins, simulations of each protein are reweighted such that the folded and unfolded minima have the same free energy, which is set to 0 (Figure 3.1-3.4E). We assume free energy barriers to be experimentally distinguishable if their heights differ by ~2 k_BT_f (Onuchic and Wolynes, 2004; Gosavi, 2013).

In order to understand any changes in the folding pathway, we also plotted average contact maps of wtHis, csHisH90G, and 3Foil near the transition state ensemble (Q = 0.40) (Figure 3.7; Figure B2). To plot these contact maps, all simulation snapshots at the required Q are pooled. In each snapshot, the value of a formed contact is set to 1 and the value of an unformed contact is set to 0. The value of a contact in an average contact map calculated at Q is the value of that contact averaged overall all snapshots at the Q. In the average contact map, the boxes marking the contacts are colored according to their value. Consequently, the average contact map is a visual representation of the average partially folded structure of the protein at Q.

2.1.5 Increasing kinetic and thermodynamic stability in csHisH90G

PROSS was used to identify additional mutations to enhance csHisH90G thermodynamic stability and solubility. PROSS implements a position-specific substitution matrix (PSSM) derived from a multiple sequence alignment (MSA), Rosetta mutation scanning, and Rosetta combinatorial design to identify stabilizing mutations (Goldenzweig et al., 2016). Three MSAs were generated to apply PROSS. Distinct sequence selection criteria were used when curating each alignment to enrich the pool of possible residue substitutions chosen by PROSS.

The first MSA was curated to probe stability enhancing mutations from evolutionarily related sequences. wtHis has no close sequence homologues, but is distantly related to the fascin family of the β-trefoil fold (Ponting and Russell, 2000). Therefore, we downloaded all sequences in the curated Pfam fascin family (PF06268) (Mistry et al., 2021), which totaled 3064 sequences and included wtHis. Curated sequences were filtered with redundancy removed at 99%, leaving 807 sequences including wtHis. Finally, BLASTp was used to align csHisH90G to the remaining fascin sequences with an E-value of 0.05. The BLASTp alignment resulted in 134 sequences with > 30% identity to csHisH90G, which is the minimum sequence identity suggested by PROSS. The final sequence alignment was verified visually using Jalview (Waterhouse et al., 2009) and AliView (Larsson, 2014).

A second, sequence-based alignment was generated using the HMMER phmmer algorithm to probe homologous protein sequences not restricted to the β-trefoil fascins (Potter et al., 2018). The HMMER phmmer algorithm uses profile hidden Markov models (HMMs) to detect and align remote sequence homologues based on the probability that two sequences are related (Eddy, 2004). The phmmer algorithm was used to search the csHisH90G query sequence against the non-redundant UniProtKB database with an E-value of 0.0001 and all other parameters set to default values (Bateman et al., 2021). The query resulted in 354 significant matches, all of which were labeled as fascins or as uncharacterized proteins. Sequence redundancy was removed at 95%. The final phmmer MSA consisted of 101 sequences.

Finally, a structure-based alignment was also generated to capture sequence information for non-homologous sequences that share a homologous fold with wtHis. Specifically, 3Foil, wtHis, csHisH90G, and the representative fascin protein human fascin 1 (PDB ID: 3LLP; split into four β-trefoil domains) were structurally aligned using MUSTANG (Konagurthu et al., 2006). Additional β-trefoil sequences from the UniProtKB and Swiss-Prot databases were aligned to each of the previous structures using the Mustguseal Web server (Suplatov et al., 2018). Using Mustguseal Mode 2 and Scenario 1 input options, a maximum of 350 sequences were aligned to each query structure and redundancy for all sequences was filtered at 90%. The dissimilarity filter threshold was set at 0.25 bit score per column. No sequence length filter was included. The resulting MSA contained 683 sequences (350 ricin and 333 fascin sequences). The alignment quality was visually inspected using Jalview (Waterhouse et al., 2009) and AliView (Larsson, 2014) using 3Foil and wtHis core residues as reference markers, and sequences were manually realigned as needed. Any sequences missing one or more trefoils relative to the representative structures were removed. Following manual curation, 663 sequences remained in the Mustguseal MSA.

csHisH90G was submitted to PROSS with one of the Pfam fascin, Mustguseal, or phmmer MSAs. Core residues and G90 were either fixed or allowed to vary. The talaris2014 energy function was used in all PROSS submissions. 162 single or multi mutant csHisH90G variants were generated in total. PROSS variants were examined using several additional *in silico* predictors to confirm that protein thermodynamic stability and solubility were improved compared to csHisH90G. The thermodynamic stability of each PROSS mutant compared to csHisH90G was measured using MAESTRO (Laimer et al., 2015) and FoldX (Guerois et al., 2002) stability predictors (Figures 3.2-3.4C). All destabilizing PROSS designs were discarded. Additionally, designs predicted to be stabilized by less than 0.8 kcal/mol relative to csHisH90G by MAESTRO were also discarded. Solubility for remaining PROSS variants was confirmed using the sequence-based predictors AGGRESCAN (Conchillo-Solé et al., 2007) and CamSol (Sormanni et al., 2015) (Figures 3.2-3.4C). Finally, LRO and ACO values were generated for all designs as described above, and those with LRO or ACO values lower than csHisH90G were discarded (Table 3.2). One variant generated from each MSA was selected for experimental validation. The variants were named Pfam fascin csHisH90G (fa-csHisH90G) (Figure 3.2), phmmer csHisH90G (ph-csHisH90G) (Figure 3.3), and Mustguseal csHisH90G (mu-csHisH90G) (Figure 3.4), respectively. C_α-SBM simulations were performed for fa-csHisH90G, ph-csHisH90G, and mu-csHisH90G as described above. Predicted LRO, ACO, and C_α-SBM unfolding free energy barrier heights for fa-csHisH90G, ph-csHisH90G, and mu-csHisH90G are given in Table 3.2.

2.2 Experimental

2.2.1 Protein expression

wtHis and pseudo-wild type HisH90G were expressed using the pHW plasmid in Escherichia coli BL21 cells as previously described (Wong et al., 2004). csHisH90G and fa-csHisH90G were expressed using the pET28a+ expression vector in *E. coli* BL21 (DE3) cells with pLysS. All cell strains were inoculated into 2TY media and grown at 37°C with shaking for approximately three hours. Upon reaching OD_{600} 0.7, cells were induced with 0.5 mM IPTG. Post induction, cells were grown at 25°C with shaking for 20-24 hours. Protein expression of the novel constructs csHisH90G and fa-csHisH90G was confirmed using whole-cell SDS-PAGE analysis (Figure 3.6A). Cells were harvested at 5000g, and cell pellets were stored at -80°C until cell lysis.

2.2.2 Cell lysis and protein purification

Cells were resuspended in 50 mM Tris buffer pH 8.0 with 0.1 M NaCl, 1 mM MgCl2, and 0.1 mM PMSF. Once homogenous, DNase I was added to the resuspension, and cells were lysed at >17 000 psi for 5 minutes using the Emulsiflex®-C5 High Pressure Homogenizer (AVESTIN, Inc, ON, Canada). Lysate was centrifuged twice at 20 000 rpm for 22 minutes, and the supernatant was filtered using a 0.45 µm syringe filter.

wtHis and hisactophilin variants bind NTA-Ni resin without the use of a His tag due to their high histidine content (31 of 117 residues in wtHis). As such, wtHis and hisactophilin variants were purified via nickel immobilized metal affinity chromatography using Profinity IMAC resin (Profinity IMAC, BioRad Laboratories Inc, CA, USA) using a BioRad low-pressure chromatography system (BioRAD BioLogic LP, BioRad Laboratories Inc, CA, USA). The nickel affinity column was equilibrated with 50 mM Tris buffer pH 8.0 with 0.1 M NaCl, 1 mM MgCl2, and 0.1 mM PMSF. Following loading of filtered lysate at a flow rate of 2 mL/min, the column was washed with 50 mM sodium phosphate buffer pH 6.3 with 0.1 M NaCl, 20 mM imidazole, and 0.1 mM PMSF at 3 mL/min. Then, wtHis or the hisactophilin variant was eluted using 50 mM sodium phosphate buffer pH 6.3 with 0.1 M NaCl, 0.25 to 0.5 M imidazole, and 0.1 mM PMSF. Purified protein was dialyzed three times against 25 mM ammonium carbonate pH 8.88 using 10 kDa molecular cutoff dialysis tubing (Repligen Spectra/Por 6 molecularporous membrane tubing, Spectrum Laboratories Inc, CA, USA). Protein was concentrated to 5-10 mg/mL using an Amicon® Stirred Cell (EMD Millipore Corporation, MA, USA) and a 10 kDa molecular weight cut-off membrane (Ultra Cel® 10 kDa Ultrafiltration Discs, EMD Millipore Corporation, MA, USA). Following concentration, protein was lyophilized and stored at -80°C.

2.2.3 GuHCl equilibrium denaturation

Lyophilized wtHis, HisH90G, csHisH90G, or fa-csHisH90G was dissolved in 50 mM potassium phosphate buffer pH 7.7 with 1 mM DTT to a final concentration of 2 mg/mL (~150 μM). wtHis and HisH90G protein stocks were diluted to 10 μM and csHisH90G and fa-csHisH90G were diluted to 6 μM in various concentrations of guanidine hydrocholoride (GuHCl) ranging from 0 to 7.5 M in 50 mM potassium phosphate buffer. All samples were equilibrated at 27°C for at least ten half-lives of unfolding. Equilibrium fluorescence scans were collected for each sample using a PTI QuantaMaster™ Series fluorometer (QM-0875-21 Modular Research Fluorometer, Horiba Scientific, ON, Canada). wtHis and HisH90G unfolding equilibria were measured using tyrosine fluorescence at 306 nm with excitation at 277 nm (Figure 3.5A,C) (Wong et al., 2004). csHisH90G and fa-csHisH90G unfolding equilibria were measured using tryptophan fluorescence at 314 nm with excitation at 280 nm (Figure 3.5B, C; Figure 3.6 B, C) (Broom et al., 2012). All scans were done with 1 nm excitation and 5 nm emission slit widths. The resulting curves were fit to a linear extrapolation model:

$$Y = (Y_N + S_N[GuHCl]) - \frac{\left((Y_N+S_N[GuHCl])-(Y_U+S_U[GuHCl])\right)\left(e^{\frac{\Delta G_{U-F}-m_{eq}[GuHCl]}{RT}}\right)}{1+ e^{\frac{\Delta G_{U-F}-m_{eq}[GuHCl]}{RT}}} \qquad \textbf{(6)}$$

where *Y* is the optical signal of the native (*N*) or unfolded (*U*) state, *S* in the [GuHCl]-dependence of the optical signal, $\Delta G_{U\text{-}F}$ is the free energy of unfolding in water, m_{eq} is the [GuHCl]-dependence of $\Delta G_{U\text{-}F}$, *R* is the gas constant 1.987 cal K^{-1} mol^{-1}, and *T* is the temperature in Kelvin (Figure 3.5C; Figure 3.6C). The data are well fit by the 2-state unfolding model. The fitted experimental *m* values decrease with increasing C_{mid} for the variants studied here, which is consistent with the known nonlinear denaturant-dependence of stability (Liu et al., 2001; Wong et al., 2004). Values for equilibrium fits are given in Table 3.3 for wtHis, HisH90G, and csHisH90G and in Table 3.4 for fa-csHisH90G.

2.2.4 GuHCl refolding and unfolding kinetics

Kinetic unfolding and refolding experiments were carried out using manual mixing on a PTI QuantaMaster™ Series fluorometer. For refolding experiments, lyophilized protein was dissolved in concentrated buffered GuHCl (~8 M) to 2 mg/mL (~150 μM). For unfolding experiments lyophilized protein was dissolved in phosphate buffer to 2 mg/mL (~150 μM). Protein stocks were diluted to 10 μM (wtHis or HisH90G) or 6 μM (csHisH90G or fa-csHisH90G) at various GuHCl concentrations ranging approximately 1 M GuHCl on either side of the kinetic midpoint. Mixing dead times were ~2-5s. Sample fluorescence was measured for at least 10 half-lives, using the same fluorometer settings as for equilibrium experiments (above). The kinetic rate constants were obtained by fitting the data to either a single exponential model:

$$Y = A\left(e^{-t/t_1}\right) + Y_0 \qquad \textbf{(7)}$$

or a single exponential model with a linear drift:

$$Y = A\left(e^{-t/t_1}\right) + Y_0 + dt \qquad \textbf{(8)}$$

where *A* is the amplitude of the change in fluorescence, *t* is time in seconds, t_1 is the inverse of the rate constant, *k*, Y_0 is the intensity of the fluorescence at *t* = 0 seconds, and *d* is the drift. Rate constants were then fit to a 2-state model as previously described (Liu et al., 2001) given by:

$$\ln(k_{obs}) = \ln\left(k_f^{H_2O} e^{\left(\frac{m_f[GuHCl]}{RT}\right)} + k_u^{H_2O} e^{\left(\frac{m_u[GuHCl]}{RT}\right)}\right) \qquad \textbf{(9)}$$

where k_{obs} is the measured rate constant, m_f/RT and m_u/RT are the linear [GuHCl]-dependences of the folding and unfolding rate constants, respectively (Figure 3.5D; Figure 3.6D), and $k_f^{H_2O}$ and $k_u^{H_2O}$ are the folding and unfolding rate constants in water, respectively. The equilibrium *m*-value (m_{eq}) was calculated by:

$$m_{eq} = m_u - m_f \qquad \textbf{(10)}$$

and reflects the total increase in solvent accessible surface area between the protein's folded and unfolded states. The β-Tanford value (β_T) for folding reflects the change in solvent accessible surface area of the transition state relative to the unfolded state, and is given by:

$$\beta_T = m_f/m_{eq} \qquad \textbf{(11)}$$

where a value of 1 indicates a native-like transition state and a value of 0 indicates an unfolded-like transition state. The equilibrium Gibbs free energy of unfolding was calculated by:

$$\Delta G_{U-F} = -RT \ln\left(\frac{k_u^{H_2O}}{k_f^{H_2O}}\right) \qquad \textbf{(12)}$$

Measured and calculated kinetic parameters are given in Table 3.3 for wtHis, HisH90G, and csHisH90G and in Table 3.4 for fa-csHisH90G.

2.2.5 X-ray crystallography

Lyophilized csHisH90G and fa-csHisH90G were dissolved in 50 mM TRIS pH 7.5 to a final concentration of 10 mg/mL. High throughput screening of crystallization conditions was carried out using an Art Robbins Instruments Gryphon robot with Gryphon system control software and 96-well Hampton INTELLI-PLATEs™ (Hampton Research, Aliso Vieji, CA, USA).

csHisH90G and fa-csHisH90G were plated with JCSG-*plus* (Molecular Dimensions Ltd., Maumee, OH, USA), BCS (Molecular Dimensions Ltd., Maumee, OH, USA), MCSG1 (Molecular Dimensions Ltd., Maumee, OH, USA), MCSG4 (Molecular Dimensions Ltd., Maumee, OH, USA), and PACT (Molecular Dimensions Ltd., Maumee, OH, USA) screening conditions at 1:1 ratios of 10 mg/mL csHisH90G : mother liquor. Crystals grew in several conditions in screens for csHisH90G and fa-csHisH90G. csHisH90G screens produced small, tear-drop-shaped crystals with rounded edges. fa-csHisH90G screens produced cube-shaped crystals and clusters of rectangular rods. Crystals were confirmed to be protein using a combination of izit dye, crush, and diffraction tests.

csHisH90G hanging drop optimization plates were set up for MCSG1 condition A1 (0.1 M HEPES pH 7.5, 20 % PEG 8000). PEG concentrations were varied by 10% of the initial mother liquor concentration, and pH was varied by 0.2 pH units. Tear drop-shaped crystals with sharp edges were observed in 0.1 M HEPES pH 7.9, and 22 % PEG 8000. Further optimization of this condition (0.1 M HEPES pH 8.1, 17.6 % PEG 8000) produced diamond-shaped crystals. Finally, the additive praseodymium (III) acetate hydrate (Additive Screen, Hampton Research, Aliso Viejo, CA, USA) was added to optimized hanging drops to improve crystal diffraction. csHisH90G crystals were soaked in 20% PEG400 and flash frozen in liquid nitrogen before being shot on the Canadian Macromolecular Crystallography Facility ID (CMCF-ID) (O8ID) beamline (0.9686 Å wavelength) at the Canadian Light Source (Saskatoon, SK). The detector was set 400 mm from the mounted crystal, and diffraction data were collected at 0.5° oscillations with 0.15 second exposure time. Despite achieving diffraction to a resolution of 2.85 Å (Appendix C, Figure C1), the crystal was twinned, and diffraction data could not be solved. s fa-csHisH90G hanging drop optimization plates were set up for JCSG-*plus* condition A6 (0.2 M $LiSO_4$ M, 0.1 M phosphate/citrate buffer pH 4.2, and 20 % PEG 1000). PEG concentrations and pH were varied as above. Large rectangular rod clusters were attained in 0.2 M $LiSO_4$ M, 0.1 M phosphate/citrate buffer pH 3.8, and 20 % PEG 1000 (Appendix C, Figure C2). Clusters were broken apart, and individual rods were soaked in 20% PEG400 and flash frozen in liquid nitrogen. Crystals were shot on the University of Waterloo home source, equipped with a Rigaku rotating copper anode X-ray generator (Cu K_α = 1.54 Å) and an R-axis IV++ detector (Rigaku Americas Corporation, USA). The detector was set 150 mm from the mounted crystal and diffraction data were collected over 1° intervals with 10° 2θ offset with 120 seconds exposure time to a complete 360° dataset. Crystals diffracted to 1.70 Å (Appendix C, Figure C2). Diffraction data were indexed, integrated, and scaled using HKL2000 (HKL Research Inc., Charlottesville, VA, USA). Data were then imported into the CCP4i suite with Scalepack2Mtz and solved using molecular replacement with Molrep (Vagin and Teplyakov, 1997; Potterton et al., 2003). Significantly, Molrep could not solve the data using the structural model used to simulated fa-csHisH90G. Instead, the data were solved using a new model generated by ColabFold (Mirdita et al., 2022) from the fa-csHisH90G primary sequence. Data were refined using Phenix.refine (Afonine et al., 2012; Headd et al., 2012) in conjunction with manual model building in COOT (Emsley et al., 2010). Model geometry was analyzed and optimized based on suggestions by MolProbity (Williams et al., 2018). Data collection and refinement were carried out by Norman Tran from the Holyoak group. Refinement statistics are given in Table 3.5.

3 Results

3.1 Ensuring sufficient thermodynamic stability in the parent protein

While improving wtHis kinetic stability is the primary objective of our design, thermodynamic stability must also be considered. The vast majority of mutations in proteins are neutral or destabilizing (Magliery, 2015; Doyle et al., 2016; Goldenzweig et al., 2016; Broom et al., 2017, 2020; Rocklin et al., 2017; Nisthal et al., 2019). 3Foil has only moderate thermodynamic stability, and it is reasonable to expect that placing 3Foil core residues into wtHis may decrease hisactophilin thermodynamic stability. To promote making a foldable designed protein, we introduce a thermodynamically stabilizing point mutation in addition to swapping core residues. Previous equilibrium denaturation studies on hisactophilin show that the point mutation H90G is thermodynamically stabilizing (MacKenzie et al., 2022). Glycine is highly conserved in this position in all three trefoils of this symmetric fold in other β-trefoil proteins (Ponting and Russell, 2000). Glycine is also present at the equivalent positions in wtHis' other two trefoils, and 3Foil has glycine in all three structurally equivalent positions (Figure 3.1A) (Ponting and Russell, 2000; Broom et al., 2012). To improve the thermodynamic stability of our parent protein and to increase the probability of expressing a well-folded core-swapped protein, we use the H90G point mutant as a stabilized pseudo-wild type (HisH90G) from which to engineer our core-swapped design.

3.2 3Foil core residues promote long-range contact formation and increase topological complexity in hisactophilin by decreasing core cavity volume

To increase kinetic stability in wtHis, we sought to engineer additional long-range intramolecular contacts to increase the topological complexity of wtHis, as measured by LRO, ACO, and unfolding free energy barrier heights from C_{α}-SBM simulations. Toward this end, we compared intramolecular contacts in wtHis to those of 3Foil, which displays extreme kinetic stability and shares a common fold with wtHis (Broom et al., 2015b) (Figure 3.1A, B). Here, we focus on differences in contacts made by 18 residues that are conserved as core residues in b-trefoils (Murzin et al., 1992; Ponting and Russell, 2000). 3Foil core residues contribute 136 long-range contacts to its LRO, while those of wtHis contribute only 92 (Figure 2.1). 3Foil and wtHis differ at 11 of the 18 core residues (Figure 3.1A) and display markedly different core packing (Figure 3.1B, C). Notably, R4 and E115 in wtHis twist away from the hydrophobic core to point toward solvent (Figure 2.1A, B; Figure 3.1B, C). In comparison to 3Foil, the wtHis core is largely composed of relatively small residues like valine and phenylalanine, which are less densely packed compared to other β-trefoil proteins (Lee and Blaber, 2011; Broom et al., 2012; Terada et al., 2017; Blaber, 2021). Together, wtHis' unusual R4 and E115 backbone conformations and diminished core packing create a cavity through the protein core with a cavity volume of 65.0 Å^3, as calculated using CASTp (Tian et al., 2018; Figure D1). In contrast, the 3Foil core is closely packed with no detectable cavity and contains larger tyrosine and tryptophan residues that all point inwards to make long-range interactions throughout the core (Figure 3.1B). Further, 3Foil core residues have a combined volume of 3.2x10^3 Å^3, whereas equivalent residues in wtHis have ~10% decreased volume of 2.9x10^3 Å^3 (Perkins, 1986). While not originally intended in our design strategy, introducing 3Foil's larger core residues into wtHis may reduce the cavity in the wtHis core, increase core packing, and achieve the formation of additional long-range contacts across the protein, as intended, by bringing core residue side chains into closer proximity. As such, replacing wtHis core residues with those of 3Foil is expected to increase hisactophilin's LRO, ACO, and kinetic stability.

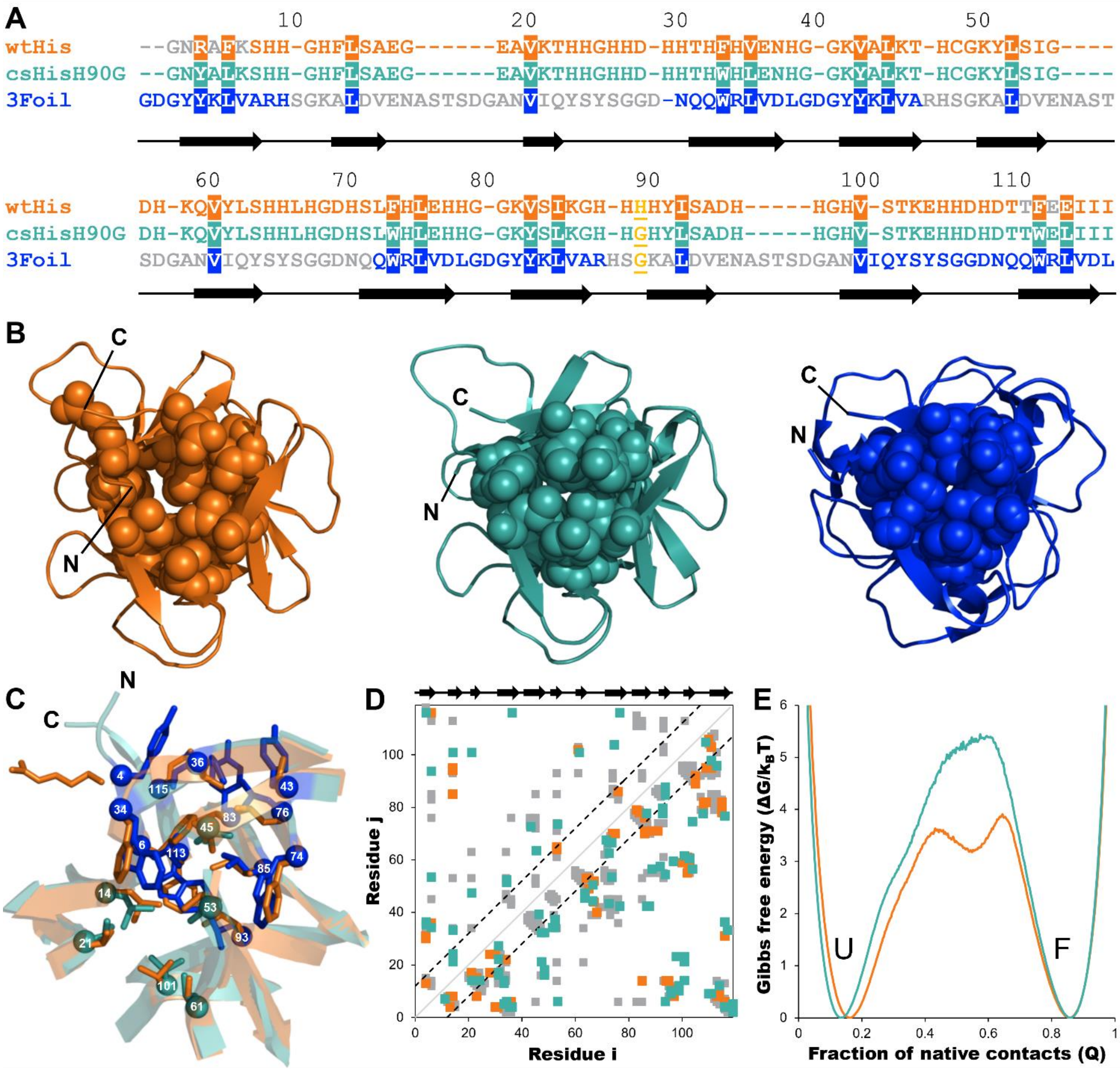

A
10
20
30
40
50
wtHis --GNRAFKSHH-GHFLSAEG------EAVKTHHGHHD-HHTHFHVENHG-GKVALKT-HCGKYLSIG----
csHisH90G --GNYALKSHH-GHFLSAEG------EAVKTHHGHHD-HHTHWHLENHG-GKYALKT-HCGKYLSIG----
3Foil GDGYYKLVARHSGKALDVENASTSDGANVIQYSYSGGD-NQQWRLVDLGDGYYKLVARHSGKALDVENAST
60
70
80
90
100
110
wtHis DH-KQVYLSHHLHGDHSLFHLEHHG-GKVSIKGH-HHHYISADH-----HGHV-STKEHHDHDTTFEEIII
csHisH90G DH-KQVYLSHHLHGDHSLWHLEHHG-GKYSLKGH-HGHYLSADH-----HGHV-STKEHHDHDTTWELIII
3Foil SDGANVIQYSYSGGDNQQWRLVDLGDGYYKLVARHSGKALDVENASTSDGANVIQYSYSGGDNQQWRLVDL
B
C
N
C
N
C
N
C
N
C
4
36
115
43
34
76
83
45
6
113
74
85
14
53
21
93
101
61
D
Residue j
Residue i
0
20
40
60
80
100
E
Gibbs free energy (ΔG/k$_B$T)
Fraction of native contacts (Q)
0
0.2
0.4
0.6
0.8
1
1
2
3
4
5
6
U
F

Figure 3.1. Engineering long-range intramolecular contacts between core residues enhances the hisactophilin unfolding free energy barrier. Hisactophilin (wtHis; orange) core residues were replaced with those of ThreeFoil (3Foil; blue) to give the core-swapped hisactophilin variant csHisH90G (cyan). **(A)** Sequences for wtHis, csHisH90G, and 3Foil are given as a structure-based sequence alignment with the 18 conserved core residues targeted for engineering highlighted. The thermodynamically stabilizing point mutation H90G is underlined and given in yellow. Secondary structure for wtHis and csHisH90G is indicated below the alignment, with β-strands represented as arrows. Residues are numbered relative to wtHis. wtHis and 3Foil residues that were excluded from structural templates used in Rosetta Comparative Modeling to generate the csHisH90G model are given in grey. **(B)** Native structures for wtHis (orange, left), csHisH90G (cyan, middle), and 3Foil (blue, right) are given looking down the β-barrel (*i.e.* the N- and C-termini facing out of the page) with the 18 conserved core residues shown in space-filled representation. Improved core packing density is evident from wtHis to csHisH90G and from csHisH90G to 3Foil. **(C)** wtHis and csHisH90G are overlaid to illustrate mutated core residues. wtHis core residues are given in orange. csHisH90G core residues derived from 3Foil are shown in blue, and csHisH90G core residues that are unchanged from wtHis (*i.e.* equivalent 3Foil residues already had the same amino acid identity as wtHis) are given in cyan. C_α atoms are shown as spheres and are numbered according to the alignment given in (A). Loop residues are removed for simplicity. **(D)** Difference contact map for wtHis and csHisH90G. Contacts common to both proteins are shown in grey. The top left portion shows contact pairs made by core residues to any other residue. The bottom right portion shows all residue pairs for each protein. Long-range contacts, in which residues *i* and *j* are more than 11 residues apart in the primary sequence, are all contacts outside of the back dashed lines. Secondary structure is indicated above. C_α contact maps were generated using the Shadow map algorithm available through SMOG2 using default parameters (*i.e.* 6 Å maximum contact cutoff and 1 Å atom occlusion). All simulations for wtHis and csHisH90G were completed using SMOG2 Shadow maps. **(E)** wtHis and csHisH90G unfolding free energy barriers were simulated using C_α-SBMs. Simulations were run at each protein's folding temperature, and unfolding free energy barriers were solved using the Boltzmann reweighting method described by Gosavi *et al.* (2006). Unfolding free energy barriers are given along the reaction coordinate Q, the fraction of native contacts. The unfolded (U) and folded (F) states are indicated. The unfolding free energy barrier predicted for csHisH90G is 1.5 k_BT_f larger than that predicted for wtHis. Unfolding free energy barrier heights are given in Table 3.1.

Incorporating 3Foil core residues into wtHis markedly increases core packing density, as illustrated by the decrease in core cavity size in space-filled models from wtHis to csHisH90G (Figure 1.3B). In contrast to R4 and E115 in wtHis, the corresponding residues Y4 and L115 in csHisH90G point into the protein core, eliminating the twisted backbone conformations of wtHis and reducing the size of the cavity. Indeed, CASTp predicts only 2.57 $Å^3$ of space in the csHisH90G core (Tian et al., 2018) (Figure D1), and ProteinVolume calculates an increase in protein volume from $15.6x10^3$ $Å^3$ in wtHis to $16.3x10^3$ $Å^3$ in csHisH90G (Chen and Makhatadze, 2015). Introducing 3Foil core residues into wtHis increases the number of long-range contacts made in csHisH90G relative to wtHis (Figure 3.1D; Appendix A). While several new contacts are formed between core residues, many of the new long-range interactions are made between core residue backbone groups and loop or mini-core residues (Figure 3.1D) (Dubey et al., 2005). Of the new contacts made between core residue side chains in csHisH90G, most are less than 12 residues apart and do not qualify as long-range contacts according to the definition used to calculate LRO (see section 1.4, equation 2). In wtHis and csHisH90G, no core residues in adjacent β-strands in the hairpin cap or in neighboring trefoils in the beta-barrel, with the exception of β-strands 1 and 12, are more than 11 residues apart in the primary sequence. So, bringing core residue side chains of neighboring trefoils into closer proximity by introducing larger side chains does not increase LRO in csHisH90G, and long-range contacts are gained primarily between β-barrel core residues from the same trefoil. This is owing to hisactophilin's relatively short β2-β3 loops and tight hairpin turns, which are longer in other β-trefoil proteins (Murzin et al., 1992; Broom et al., 2012; Gosavi, 2013; Terada et al., 2017; Kimura et al., 2020). In 3Foil, adjacent β-strands in the hairpin cap region are preceded by longer β2-β3 loops such that core residues in the hairpins are 13 residues apart in the primary sequence. Additionally, the 3Foil β-barrel includes longer turns between sequential β-strands such that core residues in the B layer are 12 residues apart in the primary sequence for neighboring trefoils. So, due to longer loops, hairpin cap and B layer core residues in 3Foil may form long-range contacts to all adjacent β-strands, both within and between trefoils.

Despite hisactophilin's more limited capacity to form long-range contacts between core residues, contact maps for wtHis and csHisH90G show that csHisH90G gains eight long range core-core contacts and 17 additional long-range contacts between core residues and loop or mini-core residues that are not present in wtHis (see Appendix A). Thus, our model for csHisH90G suggests that 3Foil core residues successfully reduce core cavity volume and increase long-range contacts in the hisactophilin core.

Parallel to the observed enrichment of long-range contacts in csHisH90G, LRO increases from 4.1 in wtHis to 4.5 in csHisH90G (Table 3.1). Using the linear correlation between LRO and unfolding rate constants for two-state β proteins reported by Broom *et al*. (2015a), this difference in LRO predicts a 4.1-fold decrease in csHisH90G's unfolding rate constant compared to wtHis at the transition midpoint and a corresponding 4.1-fold increase in unfolding half-life. ACO calculations also indicate that csHisH90G is kinetically stabilized compared to wtHis. csHisH90G's ACO increases to 13.1 from 12.2 in wtHis (Table 3.1). The linear correlation for ACO and unfolding rate constants in two-state β proteins at the transition midpoint predicts that csHisH90G's unfolding rate constant is 2.8-fold slower than wtHis (Broom et al., 2015a). Since both LRO and ACO predict higher topological complexity and slower unfolding rates for csHisH90G, we continued to investigate whether this core-swapped design increases hisactophilin kinetic stability. To obtain a higher resolution model of the change in kinetic stability, we performed C_α-SBM simulations to model free energy barriers of unfolding for wtHis and csHisH90G.

Table 3.1. Predicted and experimental unfolding kinetics for wtHis, csHisH90G, and 3Foil.

	wtHis	**csHisH90G**	**3Foil**
LRO	4.1	4.5	6.2
$\mathbf{k}_{\mathbf{u},\,\mathbf{C}_{\mathbf{mid}}}$ $(\mathbf{LRO})$ [1]	$4.3\text{x}10^{-1}$	$1.0\text{x}10^{-1}$	$9.8\text{x}10^{-5}$
ACO	12.2	13.1	22.6
$\mathbf{k}_{\mathbf{u},\,\mathbf{C}_{\mathbf{mid}}}$ $(\mathbf{ACO})$ [2]	$3.6\text{x}10^{-1}$	$1.3\text{x}10^{-1}$	$1.5\text{x}10^{-6}$
Free energy barrier height ($\mathbf{\Delta G_u/k_B T}$) [3]	3.9	5.4	17.0†
Experimental $\mathbf{k}_{\mathbf{u},\,\mathbf{C}_{\mathbf{mid}}}$ ($\mathbf{x10^{-3}\,s^{-1}}$)	**6.7±3.1**⁎ 3.4 ±1.1	3.4±1.8	$1.9\pm0.8\text{x}10^{-5}$

1. $\mathbf{k}_{\mathbf{u},\,\mathbf{C}_{\mathbf{mid}}}$ $(\mathbf{LRO})$ is calculated using the linear relation $\mathbf{k}_{\mathbf{u},\,\mathbf{C}_{\mathbf{mid}}} = -1.7(LRO) + 6.6$ for two-state β-proteins at the transition midpoint (Broom et al., 2015a).
2. $\mathbf{k}_{\mathbf{u},\,\mathbf{C}_{\mathbf{mid}}}$ $(\mathbf{ACO})$ is calculated using the linear relation $\mathbf{k}_{\mathbf{f},\,\mathbf{C}_{\mathbf{mid}}} = -0.52(ACO) + 5.9$ for two-state β-proteins at the transition midpoint (Broom et al., 2015a).
3. Free energy barrier heights were generated from C_α-SBM folding simulations at the protein folding temperature, the simulation temperature equivalent to the transition midpoint (see section 2.1.4).

†C_α-SBM folding simulations for 3Foil are given in Appendix B.

⁎Experimental $\mathrm{k}_{\mathrm{u},\,\mathrm{C}_{\mathrm{mid}}}$ for HisH90G, the pseudo-wild type parent for csHisH90G.

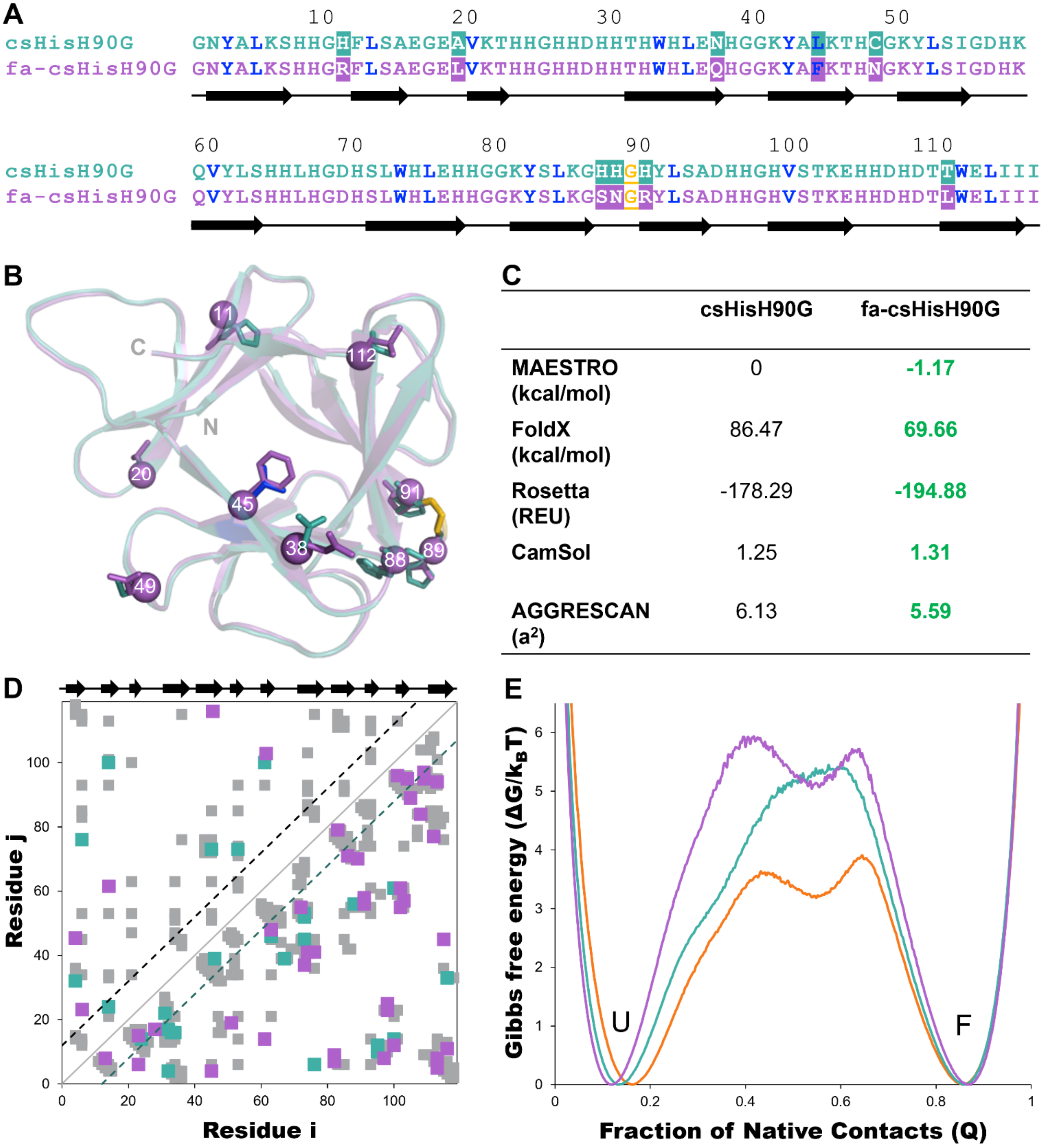
SBMs encode a protein's native contacts in their energy function and are useful in probing the relationship between protein folding and protein topology (Nymeyer et al., 1998; Chavez et al., 2004; Hyeon and Thirumalai, 2011). Here, we applied C_α-SBM simulations to gain insight into wtHis and csHisH90G unfolding free energy barriers, which we use as a predictive measure for relative kinetic stability. Specifically, we used C_α-SBM simulations to model each protein's free energy barrier for unfolding at the transition midpoint, where larger barrier heights are correlated with higher kinetic stability *in vitro* (Kramers, 1940; Chavez et al., 2004; Gosavi, 2013; Broom et al., 2015b). Using a Shadow contact map with a 6 Å maximum contact distance and a 1 Å atom "shadowing" radius, wtHis is predicted to have a free energy barrier of unfolding of 3.9 k_BT_f (Table 3.1), which is in good agreement with previously reported barrier heights for wtHis using CSU maps, an alternate form of contact map that uses the same potential energy function (Gosavi, 2013; Broom et al., 2015b). C_α folding simulations for csHisH90G predict a larger maximum unfolding free energy barrier height of 5.4 k_BT_f for csHisH90G (Table 3.1), with an average increase in barrier height of 1.8 k_BT_f over wtHis for the transition region ($Q = 0.45$ to 0.65) and a maximum barrier height difference of 2.2 k_BT_f at $Q = 0.54$ (Figure 3.1E). Thus, LRO, ACO, and C_α-SBM unfolding free energy barrier heights all predict a modest but measurable increase in protein kinetic stability for csHisH90G. We therefore proceeded to validate the design experimentally.

3.3 Thermodynamic stabilization of csHisH90G by PROSS increases structure complexity in core-swapped hisactophilin

As stated previously, due to 3Foil's moderate thermodynamic stability relative to wtHis and the prevalence of destabilizing point mutations in the literature (Magliery, 2015), we reasoned that core-swapped hisactophilin may be thermodynamically destabilized compared to wtHis. MAESTRO predicts that 3Foil core residues destabilize wtHis by ~1.0 kcal/mol (Laimer et al., 2015). So, in addition to including H90G in the parent scaffold, we implemented PROSS to identify additional thermostabilizing point mutations (Goldenzweig et al., 2016). We posited that since PROSS works to design multiple mutations within a given scaffold (Goldenzweig et al., 2016), PROSS provides another opportunity to increase long-range contacts and improve kinetic stability. Thus, in addition to screening PROSS mutants for increased stability and sufficient solubility, we also measured each protein's LRO and ACO to select for increased topological complexity. One csHisH90G variant was selected from each of the three pools of PROSS mutants from the fascin (fa), phmmer (ph), or mustguseal (mu) MSAs, and C_α-SBM simulations were performed for each of fa-csHisH90G, ph-csHisH90G, and mu-csHisH90G.

All three csHisH90G variants have multiple mutations suggested by PROSS (Figures 9-11A), and all differ by at least five residues. fa-csHisH90G, ph-csHisH90G, and mu-csHisH90G are all predicted to be thermodynamically stabilized relative to csHisH90G by MAESTRO and FoldX (Figures 9-11C) (Guerois et al., 2002; Laimer et al., 2015). Additionally, selected csHisH90G variants display improved solubility scores in CamSol and diminished aggregation propensity in AGGRESCAN (Figures 9-11C) (Conchillo-Solé et al., 2007; Sormanni et al., 2015). Finally, fa-csHisH90G, ph-csHisH90G, and mu-csHisH90G have increased LRO and ACO values relative to csHisH90G (Table 3.1; Table 3.2), and C_α-SBM simulations predict larger free energy barriers of unfolding than wtHis (Figures 9-11E). Thus, the selected PROSS variants for csHisH90G are all predicted to have increased thermodynamic relative to csHisH90G and increased kinetic stability relative to wtHis.

fa-csHisH90G has nine additional point mutations compared to csHisH90G (Figure 3.2A, B). Notably, the core residue at position 45 is mutated to phenylalanine in fa-csHisH90G (Figure 3.2A, B), whereas the csHisH90G core residue is L45. wtHis displays a core phenylalanine in four of six core residues in the B layer (Figure 2.1), and the Pfam fascin (PF06268) family HMM logo shows strong preference for phenylalanine in several conserved core residue positions (Mistry et al., 2021). So, PROSS likely selected F45 based on the prevalence of phenylalanine in the Pfam fascin MSA, which suggests that phenylalanine may be strongly favorable in fascin and fascin-like protein cores (Goldenzweig et al., 2016). Mutation of core residues was unexpected since the csHisH90G model shows tight packing of 3Foil core residues (Figure 3.2B). The addition of F45 slightly increases the combined volume of core residues by $2.2x10^2$ $Å^3$ in fa-csHisH90G compared to csHisH90G (Perkins, 1986). Accordingly, the hisactophilin core cavity is further reduced to 0.23 Å^3, as calculated using CASTp (Tian et al., 2018; Figure D2). Despite its larger size compared to L45 in csHisH90G, F45 in fa-csHisH90G does not significantly alter the number or identity of contacts made by core residues in fa-csHisH90G and csHisH90G (Figure 3.2D). All other mutations in fa-csHisH90G are to surface residues (Figure 3.2B).



	csHisH90G	fa-csHisH90G
MAESTRO (kcal/mol)	0	-1.17
FoldX (kcal/mol)	86.47	69.66
Rosetta (REU)	-178.29	-194.88
CamSol	1.25	1.31
AGGRESCAN (a^2)	6.13	5.59

Figure 3.2. PROSS mutations based on the β-trefoil fascin family increase long-range contacts and enhance the hisactophilin unfolding free energy barrier. The Protein Repair One-Stop Shop (PROSS) was used to identify thermodynamically stabilizing point mutations in core-swapped hisactophilin (csHisH90G; cyan) from a multiple sequence alignment of curated sequences from the PFam fascin family (PF06268) to give the variant fascin-csHisH90G (fa-csHisH90G; purple) (Goldenzweig et al., 2016; Mistry et al., 2021). **(A)** Sequences for csHisH90G and fa-csHisH90G are given as a structure-based sequence alignment with the 18 core-swapped residues from 3Foil in blue and PROSS mutations highlighted. The thermodynamically stabilizing point mutation H90G is in yellow and underlined. Secondary structure for csHisH90G and fa-csHisH90G is indicated below the alignment, with β-strands represented as arrows. Residues are numbered relative to csHisH90G. **(B)** csHisH90G and fa-csHisH90G are overlaid to illustrate PROSS mutations. csHisH90G residues are given in cyan and the core residue L45 is given in blue. fa-csHisH90G residues derived from PROSS are shown in purple. G90 is given in yellow in stick representation. C_α atoms are shown as spheres and are numbered according to the alignment given in (A). **(C)** fa-csHisH90G stability relative to csHisH90G was measured using the stability predictors MAESTRO and FoldX (Guerois et al., 2002; Laimer et al., 2015), and PROSS Rosetta scores (given in “Rosetta energy units”). fa-csHisH90G solubility and arrogation propensity relative to csHisH90G were predicted using CamSol (given as a solubility score) and AGGRESCAN (Conchillo-Solé et al., 2007; Sormanni et al., 2015), respectively. fa-csHisH90G values predicted to be improved compared to csHisH90G are given in green. **(D)** Difference contact map for csHisH90G and fa-csHisH90G. Contacts common to both proteins are shown in grey. The top left portion shows contact pairs made by core residues to any other residue. The bottom right portion shows all residue pairs for each protein. Long-range contacts, in which residues *i* and *j* are more than 11 residues apart in the primary sequence, are all contacts outside of the back dashed lines. Secondary structure is indicated above. C_α contact maps were generated using the Shadow map algorithm available through SMOG2 using default parameters (*i.e.* 6 Å maximum contact cutoff and 1 Å atom occlusion). All simulations for csHisH90G and fa-csHisH90G were completed using SMOG2 Shadow maps. **(E)** wtHis (orange), csHisH90G (cyan), and fa-csHisH90G (purple) unfolding free energy barriers were simulated using C_α-SBMs. Simulations were run at each protein’s folding temperature, and unfolding free energy barriers were solved using the Boltzmann reweighting method described by Gosavi *et al.* (2006). Unfolding free energy barriers are given along the reaction coordinate Q, the fraction of native contacts. The unfolded (U) and folded (F) states are indicated. The unfolding free energy barrier predicted for fa-csHisH90G is 2.7 k_BT_f larger than that predicted for wtHis. Unfolding free energy barrier heights are given in Table 3.2.

LRO and ACO values indicate that fa-csHisH90G has greater structural complexity than csHisH90G and wtHis. fa-csHisH90G has a LRO value of 4.8 and an ACO of 14.1 (Table 3.2).

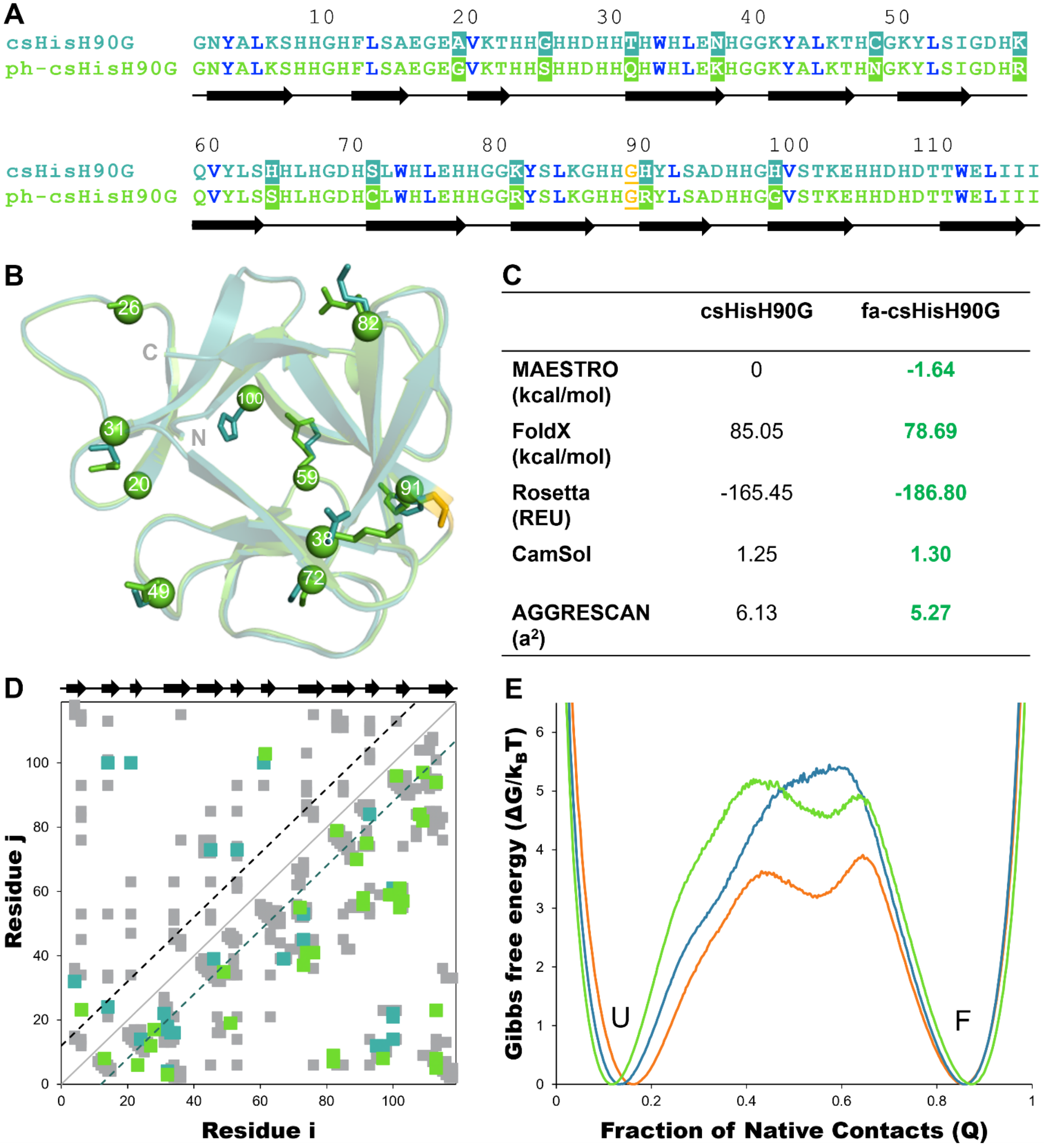
Using the linear correlation between LRO and unfolding rates for two-state β proteins reported by Broom *et al.* (2015a), this difference in LRO predicts a 3.9-fold decrease in the fa-csHisH90G unfolding rate constant compared to csHisH90G, which corresponds to a 3.9-fold increase in unfolding half-life. Similarly, the difference in ACO predicts a 3.5-fold decrease in the fa-csHisH90G unfolding rate constant. Thus, both LRO and ACO predict enhanced kinetic stability for fa-csHisH90G compared to csHisH90G. C_α-SBM simulations also predicts greater kinetic stability for fa-csHisH90G relative to wtHis. Using C_α-SBM simulations, fa-csHisH90G is predicted to have an unfolding free energy barrier height of 5.9 k_BT_f (Table 3.2), with an average increase in barrier height of 0.8 k_BT_f over csHisH90G for the transition region (Q = 0.30 to 0.65) and a maximum barrier height difference of 2.1 k_BT_f at Q = 0.36 (Figure 3.2E). Similarly, the fa-csHisH90G unfolding free energy barrier has an average increase of 2.0 k_BT_f over wtHis for the transition region (Q = 0.30 to 0.70) and a maximum barrier height difference of 2.7 k_BT_f at Q = 0.36 (Figure 3.2E). Since we consider unfolding free energy barriers to be experimentally distinguishable if their heights differ by ~2 k_BT_f, the change in kinetic stability predicted from fa-csHisH90G and csHisH90G is ambiguous as only the maximum difference in barrier heights exceeds 2 k_BT_f. However, both the average and maximum difference in unfolding free energy barrier heights for fa-csHisH90G and wtHis exceed 2 k_BT_f. Thus, LRO, ACO, and C_α-SBM unfolding free energy barrier heights predict that fa-caHisH90G will have greater kinetic stability than wtHis, and LRO and ACO suggest that fa-csHisH90G will unfold more slowly than csHisH90G. We therefore proceeded to experimentally validate fa-csHisH90G.

Table 3.2. Predicted and experimental unfolding kinetics for csHiSH90G variants.

	fa-csHisH90G	ph-csHisH90G [4]	mu-csHisH90G [4]
LRO	4.8	4.7	4.6
$\mathbf{k_{u,\,C_{mid}}\,(LRO)}$ [1]	$2.6 \mathrm{x} 10^{-2}$	$4.1 \mathrm{x} 10^{-2}$	$6.0 \mathrm{x} 10^{-2}$
ACO	14.1	13.5	13.4
$\mathbf{k_{u,\,C_{mid}}\,(ACO)}$ [2]	$3.6 \mathrm{x} 10^{-2}$	$7.5 \mathrm{x} 10^{-2}$	$8.2 \mathrm{x} 10^{-2}$
Free energy barrier height ($\mathbf{\Delta G_u / k_B T}$) [3]	5.9	5.2	6.0
Experimental $\mathbf{k_{u,\,C_{mid}}}$ (x10^{-3})	2.6±1.6		

1. $\mathbf{k_{u,\,C_{mid}}\,(LRO)}$ is given by the linear relation $\mathbf{k_{u,\,C_{mid}}} = -1.7(LRO) + 6.6$ for two-state β-proteins at the transition midpoint (Broom et al., 2015a).
2. $\mathbf{k_{u,\,C_{mid}}\,(ACO)}$ is given by the linear relation $\mathbf{k_{f,\,C_{mid}}} = -0.52(ACO) + 5.9$ for two-state β-proteins at the transition midpoint (Broom et al., 2015a).
3. Free energy barrier heights were generated from C_{α}-SBM folding simulations at the protein folding temperature, the simulation temperature equivalent to the transition midpoint (see section 2.1.4).
4. Experimental validation of ph-csHisH90G and mu-csHisH90G is still in progress.

ph-csHisH90G has 11 additional mutations compared to csHisH90G (Figure 3.3A, B). Unlike fa-csHisH90G, core residues in ph-csHisH90G were not mutated from 3Foil core residues. As such, contacts made by core residues in ph-csHisH90G are not significantly altered compared to csHisH90G (Figure 3.3D), and changes to core contacts are attributed to side chain repacking and side chain and backbone minimization carried out by PROSS (Goldenzweig et al., 2016). Instead, all mutations were to solvent-facing residues (Figure 3.3B). PROSS mutations in ph-csHisH90G contribute relatively few new long-range contacts to csHisH90G, with 11 contacts gained and five contacts lost by residues mutated by PROSS compared to csHisH90G. Of the PROSS mutations, only R82 and R91 contribute more than one additional contact, with R82 making three contacts to charged and polar residues in the β1-β2 turn and R91 making two contacts to residues in the β6-β7 turn. R82 and β1-β2 turn residues interact across the interface between the N- and C-terminal β-sheets, and thus may further stabilize the β-barrel. In similar manner, R91 and β6-β7 turn residues interact along the interface between hairpin cap hairpins in the central and N-terminal trefoils. So, the R91 point mutation may function to strengthen the hairpin interface. Notably, both R82 and R91 have longer side chains than the native residues K82 and H91, which likely enables the formation of these novel contacts. Thus, PROSS-derived mutations in ph-csHisH90G may stabilize csHisH90G by strengthening interactions between secondary structures whose residues are distant in the primary sequence.



	csHisH90G	fa-csHisH90G
MAESTRO (kcal/mol)	0	-1.64
FoldX (kcal/mol)	85.05	78.69
Rosetta (REU)	-165.45	-186.80
CamSol	1.25	1.30
AGGRESCAN (a^2)	6.13	5.27

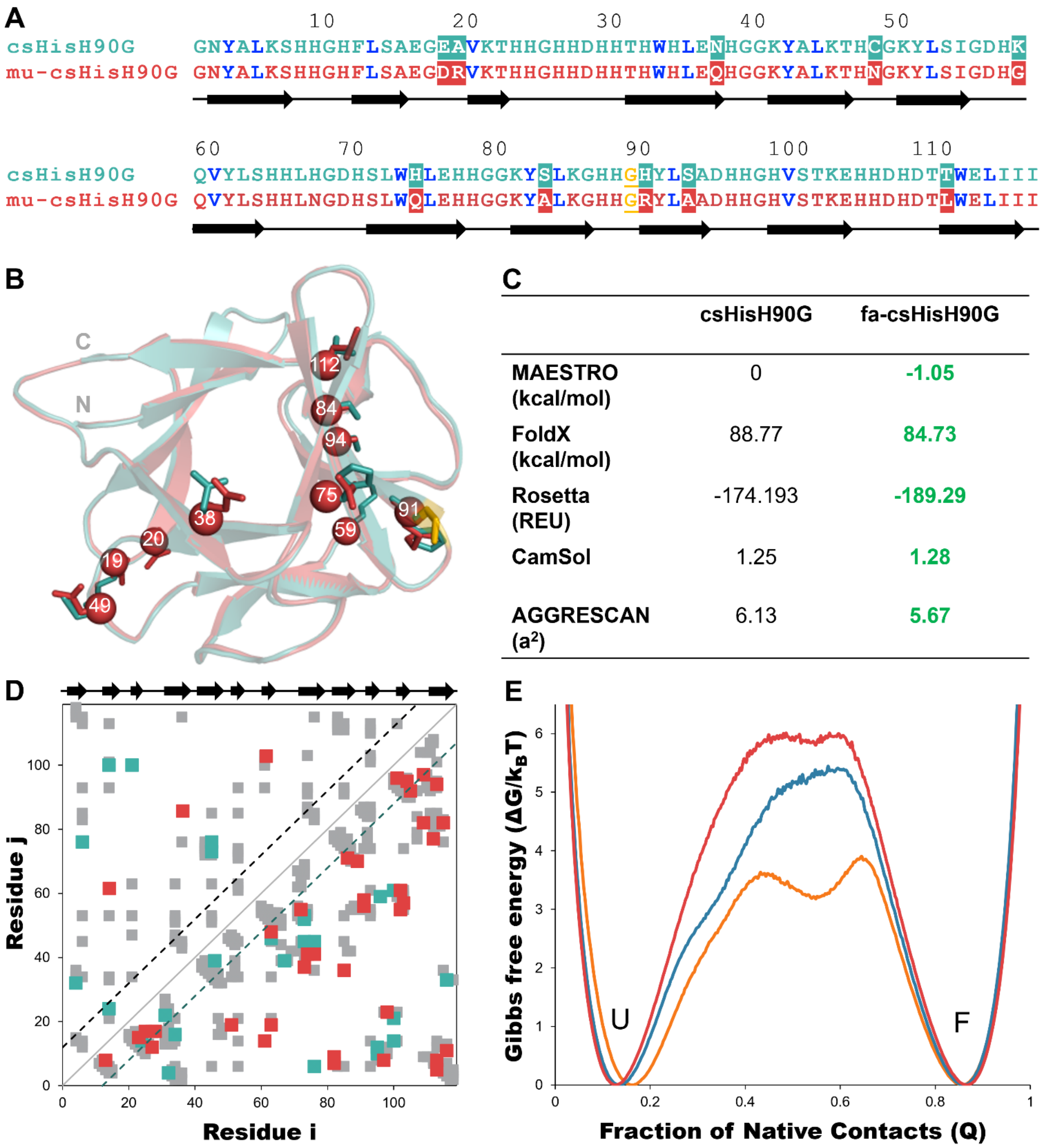
Figure 3.3. PROSS mutations based on sequence homology increase long-range contacts enhance the hisactophilin unfolding free energy barrier. The Protein Repair One-Stop Shop (PROSS) was used to identify thermodynamically stabilizing point mutations in core-swapped hisactophilin (csHisH90G; cyan) from a multiple sequence alignment generated using the HMMER phmmer algorithm (E-value of 0.0005) to give the variant phmmer-csHisH90G (ph-csHisH90G; green) (Goldenzweig et al., 2016; Potter et al., 2018). **(A)** Sequences for csHisH90G and ph-csHisH90G are given as a structure-based sequence alignment with the 18 core-swapped residues from 3Foil in blue and PROSS mutations highlighted. The thermodynamically stabilizing point mutation H90G is in yellow and underlined. Secondary structure for csHisH90G and ph-csHisH90G is indicated below the alignment, with β-strands represented as arrows. Residues are numbered relative to csHisH90G. **(B)** csHisH90G and ph-csHisH90G are overlaid to illustrate PROSS mutations. csHisH90G residues are given in cyan. ph-csHisH90G residues derived from PROSS are shown in green. G90 is given in yellow in stick representation. C_α atoms are shown as spheres and are numbered according to the alignment given in (A). **(C)** ph-csHisH90G stability relative to csHisH90G was measured using the stability predictors MAESTRO and FoldX (Guerois et al., 2002; Laimer et al., 2015), and PROSS Rosetta scores (given in "Rosetta energy units"). ph-csHisH90G solubility and arrogation propensity relative to csHisH90G were predicted using CamSol (given as a solubility score) and AGGRESCAN (Conchillo-Solé et al., 2007; Sormanni et al., 2015), respectively. ph-csHisH90G values predicted to be improved compared to csHisH90G are given in green. **(D)** Difference contact map for csHisH90G and ph-csHisH90G. Contacts common to both proteins are shown in grey. The top left portion shows contact pairs made by core residues to any other residue. The bottom right portion shows all residue pairs for each protein. Long-range contacts, in which residues *i* and *j* are more than 11 residues apart in the primary sequence, are all contacts outside of the back dashed lines. Secondary structure is indicated above. C_α contact maps were generated using the Shadow map algorithm available through SMOG2 using default parameters (*i.e.* 6 Å maximum contact cutoff and 1 Å atom occlusion). All simulations for csHisH90G and fa-csHisH90G were completed using SMOG2 Shadow maps. **(E)** wtHis (orange), csHisH90G (cyan), and ph-csHisH90G (green) unfolding free energy barriers were simulated using C_α-SBMs. Simulations were run at each protein's folding temperature, and unfolding free energy barriers were solved using the Boltzmann reweighting method described by Gosavi *et al.* (2006). Unfolding free energy barriers are given along the reaction coordinate Q, the fraction of native contacts. The unfolded (U) and folded (F) states are indicated. The unfolding free energy barrier predicted for ph-csHisH90G is 1.6 k_BT_f larger than that predicted for wtHis. Unfolding free energy barrier heights are given in Table 3.2.

Kinetic stability predictions based on protein topology and C_α-SBM simulations were more ambiguous for ph-csHisH90G than for fa-csHisH90G and csHisH90G. fa-csHisH90G LRO and

ACO are both increased relative to csHisH90G, with values of 4.7 and 13.5 (Table 3.2), respectively. As before, linear correlation of LRO and ACO with unfolding rate constants for β-proteins predicts 2.6-fold and 1.4-fold slower unfolding rate constants for ph-csHisH90G compared to csHisH90G (Broom et al., 2015a), indicating greater kinetic stability for ph-csHisH90. However, C_α-SBM simulations predict an unfolding free energy barrier height of 5.2 k_BT_f (Table 3.2), indicating slightly lower kinetic stability in ph-csHisH90G relative to csHisH90G. Thus, kinetic stability predictions from ph-csHisH90G protein topology and C_α-SBM simulations are contradictory, and it is unclear whether ph-csH90G will display greater kinetic stability than csHisH90G *in vitro*. Conversely, ph-csHisH90G LRO, ACO, and C_α-SBM unfolding free energy barriers all suggest improved kinetic stability over wtHis (Table 3.1; Table 3.2; Figure 3.3E). Compared to wtHis, the ph-csHisH90G unfolding free energy barrier height is increased by 1.3 k_BT_f on average over the transition region ($Q = 0.45$ to 0.65) and has a maximum barrier height difference of 1.6 k_BT_f at $Q = 0.46$ (Figure 3.3E). However, neither the average unfolding free energy barrier height difference nor the maximum barrier height difference exceeds 2 k_BT_f, so ph-csHisH90G does not meet our theoretical threshold for experimentally distinguishable unfolding kinetics. Despite this shortcoming, we decided to experimentally validate ph-csHisH90G because its LRO value falls between those of fa-csHisH90G and mu-csHisH90G (below), and we were interested in characterizing hisactophilin variants with a range of structural complexity.

mu-csHisH90G has 10 additional mutations compared to csHisH90G (Figure 3.4A, B). As with the ph-csHisH90G variant, mu-csHisH90G maintains all 3Foil core residues, and contacts made by core contacts do not differ significantly from csHisH90G (Figure 3.4D). Instead, all mutations were to solvent-facing residues (Figure 3.4B). Again, PROSS mutations in mu-csHisH90G add few additional contacts compared to csHisH90G, with only six long-range contacts gained and one contact lost. As in ph-csHisH90G, R91 makes two novel contacts to residues in the β6-β7 turn along the hairpin cap interface for the central and C-terminal trefoils. Notably, PROSS suggested several mutations to β-strand residues in mu-csHisH90G, whereas most mutations suggested for fa-csHisH90G and ph-csHisH90G occur in loops and turns. While some mutations to β-sheets appear to increase β structure propensity (*e.g.* N38Q, H75Q), others do not (*e.g.* S84A, S94A) (Creighton, 1992). Overall, beneficial effects of specific mu-csHisH90G mutations for csHisH90G are ambiguous from the primary sequence and structural model.

As with fa-csHisH90G and ph-csHisH90G, LRO and ACO values predict decreased unfolding rates, and therefore greater kinetic stability, for mu-csHisH90G relative to csHisH90G. mu-csHisH90G has a LRO of 4.6 and an ACO of 13.4 (Table 3.2), corresponding to 1.7-fold and 1.3-fold slower predicted unfolding rates compared to csHisH90G. C_α-SBM simulations show similar results to fa-csHisH90G. Specifically, the mu-csHisH90G free energy barrier is not sufficiently increased over that of csHisH90G to reliably predict as experimentally distinguishable kinetics by C_α-SBM simulations. The mu-csHisH90G unfolding free energy barrier is the largest predicted unfolding barrier of the three csHisH90G variants simulated, with a barrier height of 6.0 k_BT_f (Table 3.2). However, mu-csHisH90G has an average increase in barrier height of only 0.7 k_BT_f over csHisH90G for the transition region (Q = 0.40 to 0.70) and a maximum barrier height increase of only 1.5 k_BT_f at Q = 0.41 (Figure 3.4E). Again, mu-csHisH90G instead has a sufficiently large increase in barrier height over the wtHis barrier to reasonably predict improved kinetic stability, with an average barrier increase of 2.2 k_BT_f for the transition region (Q = 0.40 to 0.70) and a maximum barrier increase of 2.7 k_BT_f at Q = 0.56 (Figure 3.4E). Notably, mu-csHisH90G gives the lowest LRO and ACO scores of the three csHisH90G variants but is predicted to have the largest increase in its unfolding free energy barrier by C_α-SBM simulations.



	csHisH90G	fa-csHisH90G
MAESTRO (kcal/mol)	0	-1.05
FoldX (kcal/mol)	88.77	84.73
Rosetta (REU)	-174.193	-189.29
CamSol	1.25	1.28
AGGRESCAN (a^2)	6.13	5.67

Figure 3.4. PROSS mutations based on the β-trefoil architecture increase long-range contacts and enhances the core-swapped hisactophilin unfolding free energy barrier. The Protein Repair One-Stop Shop (PROSS) was used to identify thermodynamically stabilizing point mutations in core-swapped hisactophilin (csHisH90G; cyan) from a structure-based multiple sequence alignment generated using MUSTANG and Mustguseal to give the variant mustgueal-csHisH90G (mu-csHisH90G; red) (Konagurthu et al., 2006; Goldenzweig et al., 2016; Suplatov et al., 2018). The MUSTANG structural alignment is comprised of the wtHis, csHisH90G, ricin structure 3Foil, and the four fascin domains of human fascin 1. **(A)** Sequences for csHisH90G and mu-csHisH90G are given as a structure-based sequence alignment with the 18 core-swapped residues from 3Foil in blue and PROSS mutations highlighted. The thermodynamically stabilizing point mutation H90G is in yellow and underlined. Secondary structure for csHisH90G and mu-csHisH90G is indicated below the alignment, with β-strands represented as arrows. Residues are numbered relative to csHisH90G. **(B)** csHisH90G and mu-csHisH90G are overlaid to illustrate PROSS mutations. csHisH90G residues are given in cyan. mu-csHisH90G residues derived from PROSS are shown in red. G90 is given in yellow in stick representation. C_α atoms are shown as spheres and are numbered according to the alignment given in (A). **(C)** mu-csHisH90G stability relative to csHisH90G was measured using the stability predictors MAESTRO and FoldX (Guerois et al., 2002; Laimer et al., 2015), and PROSS Rosetta scores (given in "Rosetta energy units"). mu-csHisH90G solubility and arrogation propensity relative to csHisH90G were predicted using CamSol (given as a solubility score) and AGGRESCAN (Conchillo-Solé et al., 2007; Sormanni et al., 2015), respectively. mu-csHisH90G values predicted to be improved compared to csHisH90G are given in green. **(D)** Difference contact map for csHisH90G and mu-csHisH90G. Contacts common to both proteins are shown in grey. The top left portion shows contact pairs made by core residues to any other residue. The bottom right portion shows all residue pairs for each protein. Long-range contacts, in which residues *i* and *j* are more than 11 residues apart in the primary sequence, are all contacts outside of the back dashed lines. Secondary structure is indicated above. C_α contact maps were generated using the Shadow map algorithm available through SMOG2 using default parameters (*i.e.* 6 Å maximum contact cutoff and 1 Å atom occlusion). All simulations for csHisH90G and mu-csHisH90G were completed using SMOG2 Shadow maps. **(E)** wtHis (orange), csHisH90G (cyan), and mu-csHisH90G (red) unfolding free energy barriers were simulated using C_α-SBMs. Simulations were run at each protein's folding temperature, and unfolding free energy barriers were solved using the Boltzmann reweighting method described by Gosavi *et al.* (2006). Unfolding free energy barriers are given along the reaction coordinate Q, the fraction of native contacts. The unfolded (U) and folded (F) states are indicated. The unfolding free energy barrier predicted for mu-csHisH90G is 2.7 k_BT_f larger than that predicted for wtHis. Unfolding free energy barrier heights are given in Table 3.2.

3.4 csHisH90G fluorescence suggests a 3Foil-like core

First, we assessed whether csHisH90G successfully adopts a well-folded tertiary structure with 3Foil-like core packing by comparing csHisH90G fluorescence to that of wtHis and 3Foil. csHisH90G fails to unfold in 7 M urea, so we used a stronger denaturant, guanidine hydrochloride (GuHCl). Native wtHis and HisH90G exhibit a fluorescence emission maximum at 306 nm in GuHCl (Figure 3.5A), in good agreement with previous equilibrium experiments performed in urea (Liu et al., 2001; MacKenzie et al., 2022). Since wtHis and HisH90G lack tryptophan residues and fluorescence arises predominately from tyrosines, no shift in maximum emission wavelength is observed upon wtHis or HisH90G chemical denaturation, and protein denaturation is instead manifested by an increase in fluorescence intensity. In contrast, csHisH90G displays maximum emission at ~325 nm in the native state and ~350 nm in the denatured state (Figure 3.5B), consistent with tryptophan fluorophores going from a buried hydrophobic environment to a solvent-exposed environment upon GuHCl denaturation (Vivian and Callis, 2001). csHisH90G denaturation shows striking similarity to 3Foil, which upon unfolding also undergoes a pronounced red shift from 323 nm in the native state to ~360 nm in the unfolded state in guanidine thiocyanate (GuSCN) (Broom et al., 2012). The unusually strong blue shift observed in native 3Foil is attributed to its densely packed core, which renders 3Foil tryptophan residues completely inaccessible to solvent (Vivian and Callis, 2001; Broom et al., 2012). The similar blue shift for csHisH90G supports the core of csHisH90G also being well-packed and resembling that of 3Foil. However, since csHisH90G is slightly less blue shifted compared to 3Foil and exhibits a less pronounced red shift upon unfolding, the csHisH90G core may be more accessible to solvent than the 3Foil core. This interpretation agrees with our structural model of csHisH90G, which shows core packing similar to, but not as close packed as, 3Foil (Figure 3.5B). Alternatively, these small differences in emission profiles may be related to using different denaturants (*i.e.* GuHCl for csHisH90G and GuSCN for 3Foil), as seen previously in 3Foil (unpublished data). Nevertheless, these data show that csHisH90G is well-folded *in vitro*, indicating successful engineering of 3Foil core residues into wtHis.

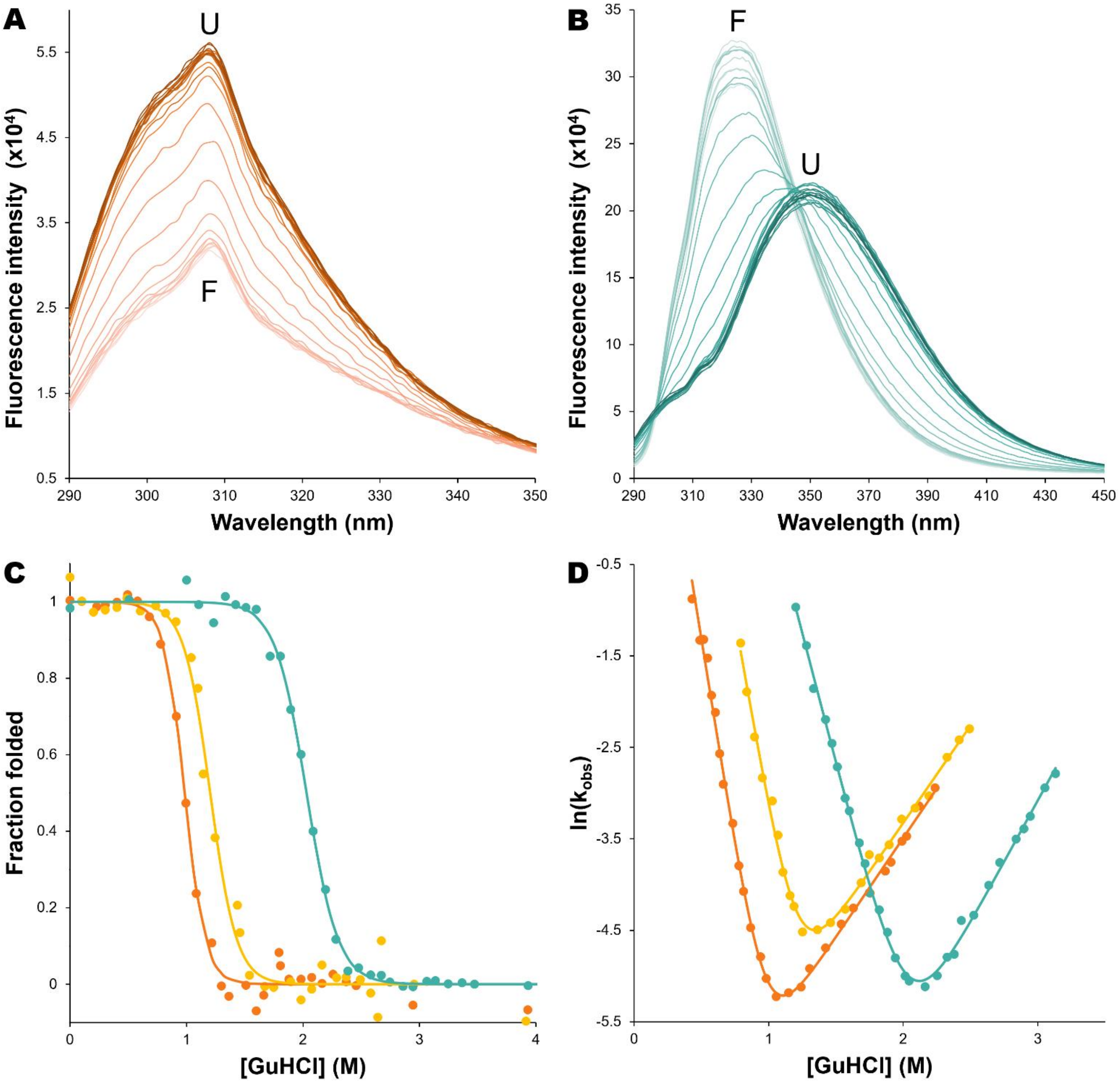

A
U
F
Fluorescence intensity (x10⁴)
5.5
4.5
3.5
2.5
1.5
0.5
290
300
310
320
330
340
350
Wavelength (nm)
B
F
U
Fluorescence intensity (x10⁴)
35
30
25
20
15
10
5
0
290
310
330
350
370
390
410
430
450
Wavelength (nm)
C
Fraction folded
1
0.8
0.6
0.4
0.2
0
0
1
2
3
4
[GuHCl] (M)
D
ln(k_obs)
-0.5
-1.5
-2.5
-3.5
-4.5
-5.5
0
1
2
3
[GuHCl] (M)

Figure 3.5. csHisH90G displays successful kinetic stabilization over HisH90G. Fluorescence emission spectra for unfolding equilibria for **(A)** wtHis and **(B)** csHisH90G in 0 to 4 M GuHCl (darker color indicates higher denaturant concentration). csHisH90G shows a pronounced red shift from the folded state (F) at ~325 nm to the unfolded state (U) at ~350 nm. **(C)** Fluorescence-monitored GuHCl denaturation curves displayed as the fraction of folded protein for wtHis (orange), HisH90G (yellow), and csHisH90G (cyan). Solutions contained 50 mM potassium phosphate pH 7.7, 0 to 4 M GuHCl, 1 mM DTT, and 10 uM protein for wtHis and HisH90G or 6 uM protein for csHisH90G. csHisH90G is significantly thermodynamically stabilized compared to wtHis and HisH90G. All samples were equilibrated at 27°C for at least 10 half-lives. **(D)** Chevron plots for observed folding and unfolding rate constants for wtHis, HisH90G, and csHisH90G at 27°C. wtHis and HisH90G kinetics were monitored at 306 nm with excitation at 277 nm. csHisH90G kinetics were monitored at 314 nm with excitation at 280 nm. csHisH90G is kinetically stabilized compared to its parent protein, HisH90G. Values for equilibrium and kinetic fits are given in Table 3.3. Models used to fit equilibrium and kinetic data are given in sections 2.2.3 and 2.2.4.

3.5 3Foil core residues enhance hisactophilin thermodynamic stability

To further validate that the csHisH90G design results in a well-behaved, cooperatively folded protein, we next measured its folding equilibrium and thermodynamic stability by chemical denaturation. The fraction of folded protein with increasing GuHCl concentration is shown in Figure 3.5C, and fitted parameters for two-state equilibrium denaturation curves are given in Table 3.3. Relative stabilities are assessed by C_{mid}, which is the most accurate measure of relative stability (Pace, 1986; Fersht, 1999). Folding is fully reversible for wtHis, HisH90G, and csHisH90G. The denaturation curve for HisH90G is shifted to higher GuHCl concentration compared to wtHis, with C_{mid}values of 1.20 M GuHCl and 0.98 M GuHCl, respectively. Therefore, HisH90G has increased thermodynamic stability compared to wtHis, as expected (MacKenzie et al., 2022). Notably, csHisH90G shows significant thermodynamic stabilization compared to wtHis and HisH90G, with a C_{mid} of 2.03 M GuHCl. Kinetic midpoints for wtHis, HisH90G, and csHisH90G show excellent agreement with equilibrium midpoints, consistent with 2-state unfolding transitions (Table 3.3).

Table 3.3. Equilibrium and kinetic parameters for wtHis, HisH90G, csHisH90G, and 3Foil.

		wtHis	**HisH90G**	**csHisH90G**	**3Foil***
$\mathbf{C_{mid}}$ *(M)*	*Eq*	0.98±0.01	1.20±0.02	2.03±0.01	
	Kin	0.98±0.05	1.21±0.09	2.03±0.08	0.79±0.04
$\mathbf{m_{eq}}$ *(kcal·mol^{-1}·M^{-1})*	*Eq*	6.66±1.09	5.25±1.10	4.39±0.47	
	Kin	6.74±0.17	6.29±0.23	4.99±0.10	9.42±0.23
$\mathbf{\Delta G_{U-F}}$ *(kcal·mol^{-1})*	*Eq*	6.55±1.10	6.32±1.33	8.95±0.97	
	Kin	6.61±0.16	7.63±0.27	10.13±0.19	7.41±0.16
$\mathbf{k_u^{H_2O}}$ *(x10^{-4} s^{-1})*		4.0±0.5	5.1 ±0.8	0.2 ±0.03	2.8±0.1x10^{-6}
$\mathbf{m_u}$ *(s^{-1}·M^{-1})*		1.29±0.05	1.27±0.05	1.60±0.04	3.20±0.03
$\mathbf{k_f^{H_2O}}$ *(s^{-1})*		26.0 ±3.5	182.3 ±54.1	346.4 ±50.5	7.0±0.4x10^{-5}
$\mathbf{m_f}$ *(s^{-1}·M^{-1})*		-5.46 ±0.12	-5.02±0.19	-3.39±0.06	-6.22±0.20
$\mathbf{k_{u,C_{mid}}}$ *(x10^{-3} s^{-1})*		**3.4 ±1.1**	**6.7±3.1**	**3.4±1.8**	**1.9±0.8x10^{-5}**
$\mathbf{\beta_T}$ *(mol·kcal^{-1}·s^{-1})*		**0.81±0.04**	**0.80±0.06**	**0.68±0.02**	**0.66±0.04**

* 3Foil kinetics were obtained using GuSCN activity (Broom et al., 2015b). Equilibrium data could not be measured for 3Foil due to its extreme kinetic stability.

csHisH90G's considerable thermodynamic stabilization was unexpected, given that 3Foil has moderate thermodynamic stability compared to wtHis. 3Foil core residues increasing hisactophilin thermodynamic stability suggests that the 3Foil core is thermodynamically stable and 3Foil is thermodynamically destabilized by other features. A trade-off between stability and function is a well-known phenomenon and is also evident in b-trefoil proteins (Fersht, 1999; Liu et al., 2001; Broom et al., 2015b). As such, 3Foil's functional loops are a likely source of diminished thermodynamic stability in 3Foil (Broom et al., 2015b). Similarly, Symfoil contains core residues similar to those of 3Foil and gained hyperthermostability with the concurrent loss of heparin-binding functionality over several iterative rounds of design (Lee and Blaber, 2011; Lee *et al.*, 2011). wtHis core residues also contribute to function by forming a deep pocket that accommodates the hydrophobic acyl chain of a covalently attached N-terminal myristoyl group (Smith et al., 2010; Shental-Bechor et al., 2012; MacKenzie et al., 2022). When the myristoyl group is buried in the wtHis core, the acyl chain makes additional stabilizing interactions with hydrophobic residues, increasing thermodynamic stability (Smith et al., 2010; MacKenzie et al., 2022). Since 3Foil core residues are larger and more closely packed than those of wtHis, core residues in csHisH90G may accomplish a similar thermodynamically stabilizing effect as the myristoyl group in wtHis. So, wtHis core residues may limit non-myristoylated wtHis thermodynamic stability, and 3Foil core residues in csHisH90G ameliorate this limitation by eliminating the core pocket functionality and facilitating augmented core packing.

3.6 3Foil core residues enhance kinetic stability in csHisH90G

To assess the outcome of our kinetic stability design, we measured the folding kinetics of wtHis, HisH90G, and csHisH90G using chemical denaturation (Figure 3.5D). wtHis displays moderate kinetic stability in GuHCl, with an unfolding rate constant of $3.4 \text{x} 10^{-3}$ s^{-1} and a half-life of ~3.5 minutes at the transition midpoint (Table 3.3). The H90G point mutation decreases kinetic stability relative to wtHis, increasing the HisH90G unfolding rate constant to $6.7 \text{x} 10^{-3}$ s^{-1} and reducing its half-life to ~1.7 minutes at the transition midpoint (Table 3.3). We hypothesize that the molecular basis for the accelerated unfolding and folding kinetics may be related to favoring folding and formation of a distinctive turn-like conformation in β-trefoils, where this glycine is strongly conserved (Ponting and Russell, 2000). Molecular details of this sequence-specific effect cannot be captured by LRO, ACO, or C_{α}-SBM simulations. Notably, HisH90G retains similar denaturant dependence of folding (m_f) and unfolding kinetics (m_u) to wtHis (Table 3.3), indicating that the changes in solvent accessible surface area from the folded or unfolded state to the transition state are maintained. Accordingly, the HisH90G β-Tanford (β_T) value, which reports on the structure of the transition state, is nearly unchanged compared to wtHis (Table 3.3). Thus, we can conclude that despite lower kinetic stability in HisH90G, the HisH90G folding pathway and transition state are similar to those of wtHis.

csHisH90G has enhanced kinetic stability compared to HisH90G. This is evident in Figure 3.5D, as the csHisH90G chevron is shifted downward to slower unfolding kinetics relative to HisH90G. Indeed, csHisH90G has an unfolding rate constant of 3.4×10^{-3} s^{-1} at the kinetic midpoint in GuHCl, which is 2.0-fold slower than that of HisH90G (Table 3.3). So, substituting 3Foil core residues into HisH90G doubles the unfolding half-life, in excellent agreement with the predicted modest effect of these mutations based on LRO, ACO, and C_{α}-SBM simulations. Thus, the strategy of using 3Foil core residues to engineer kinetic stability in hisactophilin by increasing long-range intramolecular contacts improved hisactophilin kinetic stability. Significantly, this method of designing protein kinetic stability through the consideration of LRO, ACO, and simulated unfolding free energy barriers provides a rational and predictable route to engineering targeted protein kinetic stability.

3.7 Thermostability design using PROSS can increase kinetic stability

We initially investigated the solubility of fa-csHisH90G, ph-csHisH90G, and mu-csHisH90G using SDS-PAGE to determine their suitability for fluorescence measurements, which are highly sensitive to precipitating protein that could disrupt the light beam. All three show good expression and express in the soluble fraction (Figure 3.6A). Though a moderately populated band is observed in the insoluble fraction for ph-csHisH90G (Figure 3.6A), this is likely due to incomplete cell lysis rather than insoluble protein since the majority of ph-csHisH90G is present in the soluble fraction. Both fa-csHisH90G and mu-csHisH90G are almost completely present in the soluble fraction (Figure 3.6A). Thus, all csHisH90G PROSS variants are suitable for characterization by fluorescence.

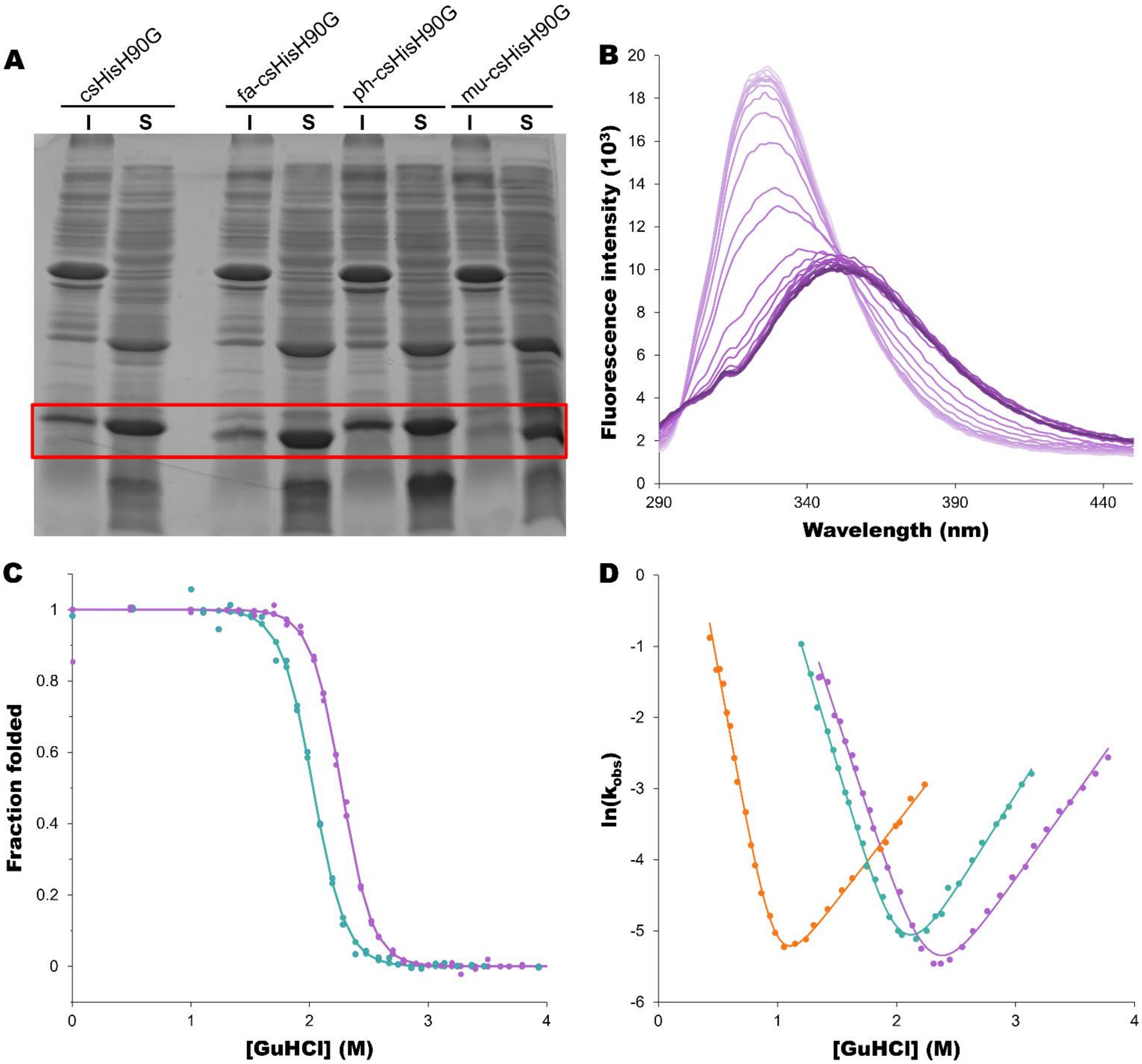

A
csHisH90G
fa-csHisH90G
ph-csHisH90G
mu-csHisH90G
I
S
I
S
I
S
I
S
B
Fluorescence intensity (10^3)
0
2
4
6
8
10
12
14
16
18
20
290
340
390
440
Wavelength (nm)
C
Fraction folded
0
0.2
0.4
0.6
0.8
1
0
1
2
3
4
[GuHCl] (M)
D
$\ln(k_{obs})$
0
-1
-2
-3
-4
-5
-6
0
1
2
3
4
[GuHCl] (M)

Figure 3.6. fa-csHisH90G displays successful kinetic stabilization over csHisH90G. (A) SDS-PAGE shows that fascin-csHisH90G (fa-csHisH90G), phmmer-csHisH90G (ph-csHisH90G), and Mustguseal-csHisH90G (mu-csHisH90G) express in the soluble fraction. csHisH90G, fa-csHisH90G, ph-csHisH90G, and mu-csHisH90G were induced with 0.5 mM IPTG at OD_{600} 0.7 and grown for 6 hours in LB media. The insoluble (I) and soluble fractions (S) were harvested by centrifugation. csHisH90G variants are indicated by a red box. csHisH90G was included as a positive control. **(B)** Fluorescence emission spectra for unfolding equilibria for fa-csHisH90G in 0 to 4 M GuHCl (darker color indicates higher denaturant concentration). fa-csHisH90G shows a pronounced red shift from the folded state (F) at ~325 nm to the unfolded state (U) at ~350 nm. **(C)** Fluorescence-monitored GuHCl denaturation curves displayed as the fraction of folded protein for csHisH90G (cyan) and fa-csHisH90G (purple). Solutions contained 50 mM potassium phosphate pH 7.7, 0 to 4 M GuHCl, 1 mM DTT, and 6 uM protein for csHisH90G and fa-csHisH90G. fa-csHisH90G is more thermodynamically stable than csHisH90G. All samples were equilibrated at 27°C for at least 10 half-lives. **(D)** Chevron plots for observed folding and unfolding rate constants for wtHis, csHisH90G, and fa-csHisH90G at 27°C. wtHis kinetics were monitored at 306 nm with excitation at 277 nm. csHisH90G and fa-csHisH90G kinetics were monitored at 314 nm with excitation at 280 nm. fa-csHisH90G is kinetically stabilized compared to wtHis and csHisH90G. Values for equilibrium and kinetic fits are given in Table 3.4. Models used to fit equilibrium and kinetic data are given in the sections 2.2.3 and 2.2.4. Experimental validation of ph-csHisH90G and mu-csHisH90G is still in progress.

fa-csHisH90G exhibits a fluorescence profile similar to those of csHisH90G and 3Foil. As with csHisH90G, fa-csHisH90G displays maximum emission at ~325 nm for the folded protein and ~350 nm in the unfolded protein (Figure 3.6B). Notably, the fa-csHisH90G core has a phenylalanine residue at position 45 (Figure 3.2A, B), whereas the equivalent residue in csHisH90G is a leucine (Figure 3.2A). Despite the addition of several hydrocarbons from the F45 aromatic ring, fa-csHisH90G tryptophan emission maxima are largely unchanged compared to csHisH90G. This is likely because F45 retains fa-csHisH90G core hydrophobicity relative to the csHisH90G core, so fa-csHisH90G tryptophans remain solvent inaccessible in a completely hydrophobic local environment. Thus, fluorescence measurements support that fa-csHisH90G is well-folded with a densely packed core that completely buries the tryptophan residues.

fa-csHisH90G shows moderate thermodynamic stabilization relative to csHisH90G (Figure 3.6C). Specifically, the denaturation curve for fa-csHisH90G has a C_{mid} of 2.27 M GuHCl, which is 0.24 M GuHCl higher than that of csHisH90G (Table 3.3). Thus, fascin-based PROSS mutations were successful in increasing thermodynamic stability in core-swapped hisactophilin. fa-csHisH90G is also kinetically stabilized compared to csHisH90G (Figure 3.6D). fa-csHisH90G displays an unfolding rate of 2.56x10^{-3} s^{-1} and a half-life of ~4.52 minutes at the transition midpoint (Table 3.4), which is 1.3-fold slower than both csHisH90G and wtHis. Again, kinetic data for fa-csHisH90G shows good agreement with the predicted values from LRO, ACO, and C$_{\alpha}$-SBM unfolding free energy barriers, which all predicted a moderate decrease in unfolding kinetic rates compared to csHisH90G. So, in addition to being an effective strategy for predicting change in kinetic stability due to several clustered mutations, such as in the csHisH90G core relative to wtHis, this method is also applicable for predicting the change in kinetic stability for multiple, non-interacting point mutations, such as in fa-csHisH90G relative to csHisH90G.

Table 3.4. Equilibrium and kinetic parameters for fa-csHisH90G

	Equilibrium	Kinetic
$\mathbf{C_{mid}}$ *(M)*	2.27 ±0.01	2.28±0.11
$\mathbf{m_{eq}}$ *(kcal·mol^{-1}·M^{-1})*	4.60±0.63	4.45±0.11
$\mathbf{\Delta G_{U-F}}$ *(kcal·mol^{-1})*	10.46±1.43	10.13±0.24
$\mathbf{k_u^{H_2O}}$ *(x10^{-5} s^{-1})*		1.2±0.3
$\mathbf{m_u}$ *(s^{-1}·M^{-1})*		1.41±0.04
$\mathbf{k_f^{H_2O}}$ *(s^{-1})*		282.1±53.8
$\mathbf{m_f}$ *(s^{-1}·M^{-1})*		-3.04±0.07
$\mathbf{k_{u,C_{mid}}}$ *(x10^{-3} s^{-1})*		2.6±0.0
$\mathbf{\beta_T}$ *(mol·kcal^{-1}·s^{-1})*		0.68±0.03

3.8 csHisH90G folding kinetics and C$_\alpha$-SBM simulations suggest a 3Foil-like folding pathway

In contrast with HisH90G, the β_T value for csHisH90G resembles that of 3Foil rather than wtHis (Broom et al., 2015b) (Table 3.3). csHisH90G and 3Foil have β_T values of 0.68 and 0.66, respectively, whereas wtHis exhibits a $\beta_T$value of 0.81, indicating that csHisH90G and 3Foil have less native-like transition states than wtHis. Strikingly, C$_\alpha$-SBM folding simulations also suggest distinct folding pathways for csHisH90G and wtHis (Figure 3.7A, B). Folding in csHisH90G initiates from the central trefoil, while folding in wtHis is initiated in its C-terminal trefoil. Folding simulations for 3Foil show advanced folding in its central foil similar to csHisH90G at the same Q, suggesting that hisactophilin gains 3Foil-like folding character upon swapping core residues (Figure B2). However, 3Foil also shows concurrent wtHis-like folding in its C-terminal trefoil

(Figure B2), suggesting that observed folding pathways for 3Foil and csHisH90G do not arise solely from 3Foil core residues. Contact analysis of each protein's C-terminal trefoil reveals that several stabilizing interactions present in wtHis are lost in csHisH90G (Figure 3.1D). Specifically, in wtHis, the backbone conformation of E115 is twisted such that the E115 sidechain points toward solvent rather than the protein core. Through this unusual conformation, E115 makes long-range, stabilizing charge-charge and charge-polar contacts to residues in the β8-β9 turn. Additionally, the twisted β12 backbone conformation in wtHis also facilitates numerous interactions between residues in β1 and β12. In contrast, L115 in csHisH90G points towards the protein core to form hydrophobic contacts with residues Y4, L36, and L76. This ameliorates the twisted β12 backbone conformation observed in wtHis, but also results in the loss of stabilizing interactions to the β8-β9 turn and between several β1/β12 residues. So, while 3Foil core residues in hisactophilin relieve steric strain in β-strands 1 and 12, concurrent loss of stabilizing interactions from β12 may hinder folding in the C-terminal trefoil.

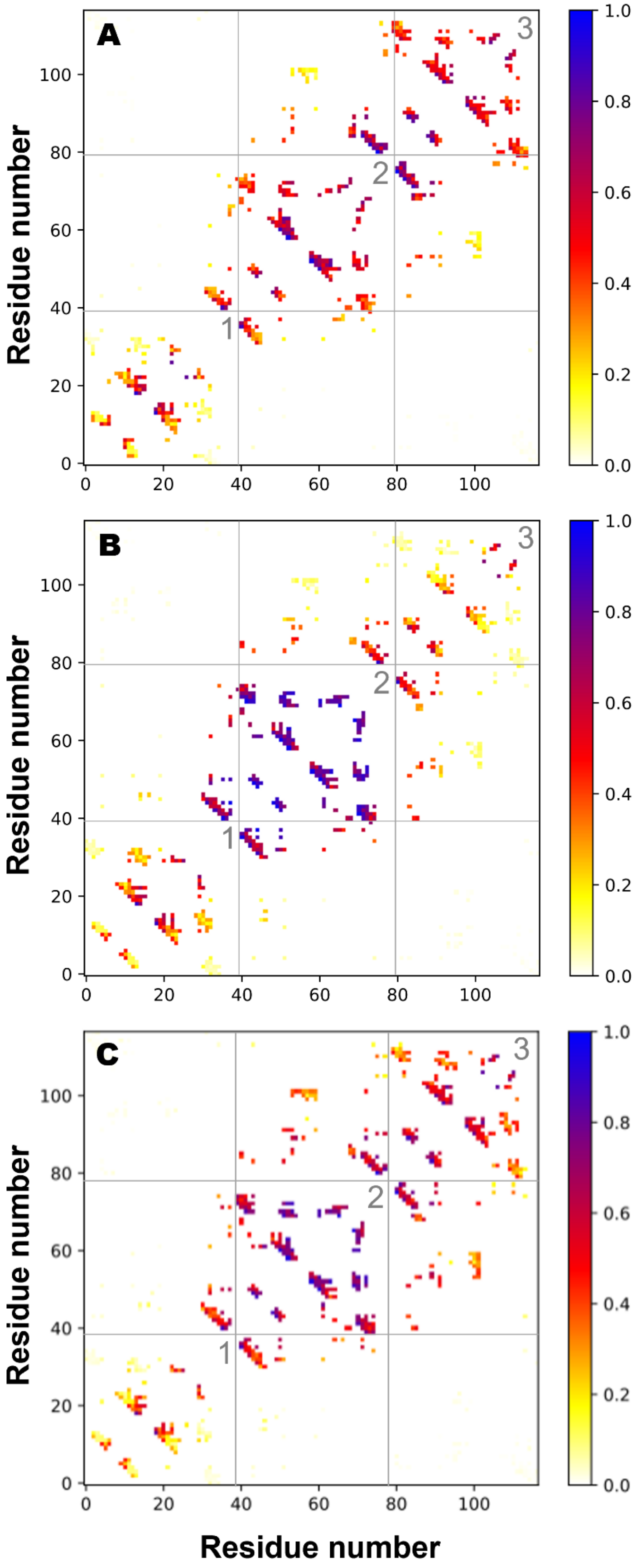

A
B
C
1
2
3
Residue number
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0

Figure 3.7. C$_\alpha$ structure-based simulations reveal distinct folding pathways for wtHis and csHisH90G. Average contact maps for the transition state ensemble for **(A)** wtHis **(B)** csHisH90G, and **(C)** fa-csHisH90G at Q = 0.4. Contacts are colored based on degree of structure, with 1 indicating fully formed and 0 indicates unformed. The N-terminal, central, and C-terminal trefoils are labeled 1, 2, and 3, respectively. wtHis initiates folding from its C-terminal trefoil, while csHisH90G folding occurs from its central trefoil. The transition state ensemble for 3Foil is given in Figure B2 of Appendix B.

Interestingly, C$_\alpha$-SBM folding simulations for fa-csHisH90G show concurrent folding of the central and C-terminal trefoils, indicating rescue of 3Foil and wtHis-like folding in the C-terminal trefoil (Figure 3.7C). fa-csHisH90G has a β_T value of 0.68, indicating a transition state with a similar degree of native-like structure to csHisH90G and 3Foil (Table 3.3; Table 3.4). Contact maps comparing csHisH90G and fa-csHisH90G show that fa-csHisH90G gains contacts in its C-terminal trefoil. Significantly, these contacts are in the same regions where contacts were lost in csHisH90G relative to wtHis. While residues F45, N89, and L112 collectively gain four long-range contacts ((F45, L115), (F45, Y4), (N89, E105), (L112, E77)), the majority of residues contributing to new contacts in the fa-csHigH90G C-terminal trefoil were not mutated by PROSS. However, all residues underwent side chain packing and side chain and backbone minimization, which may be a source of alternate contacts from non-mutated residues.

Several studies show that mutating core residues can alter protein folding pathways (Ventura and Serrano, 2004; Dalessio et al., 2005; Wensley et al., 2010; Longo et al., 2014; Blaber, 2022). Given that burial of hydrophobic surface area is widely accepted to drive protein folding (Dill, 1990), it is unsurprising that 3Foil core residues change the hisactophilin folding pathway. In replacing wtHis core residues with those of 3Foil, we estimate that the hydrophobic surface area of core residues increases from 1750 Å^2 in wtHis to 2240 Å^2 in csHisH90G. 3Foil core residues also introduce several additional long-range interactions, which are expected to increase the entropic cost of chain ordering during folding (Baker, 2019). Completely symmetric designed proteins (*i.e.* Symfoil and Phifoil) have been shown to have multiple redundant folding nuclei (Longo et al., 2014; Blaber, 2022; Tenorio et al., 2022). Thus, 3Foil core residues may change the hisactophilin folding pathway by providing an alternate folding nucleus (or nuclei) from which to initiate protein folding.

3.9 3Foil-like core packing enhances hisactophilin three-fold symmetry

To validate the computational models used to determine LRO, ACO, and C_α-SBM unfolding free energy barriers for csHisH90G and variants, we next sought to experimentally characterize the structure of csHisH90G and fa-csHisH90G. While csHisH90G successfully grew large, diamond-shaped crystals that diffracted to 2.85 Å, data refinement revealed that the crystals were twinned (*i.e.* the crystals contained two or more unit cells of different dimensions and orientations such that reflections from one unit cell could not be distinguished from other unit cells present (Rhodes, 1993)) (Appendix C, Figure C1), and the data could not be solved. In contrast, fa-csHisH90G produced large, rod-shaped crystal clusters of high diffraction quality, enabling the solution of a 1.70 Å structure (Appendix C, Figure C2; Figure 3.8A, B). Interestingly, diffraction data for fa-csHisH90G could not be solved using the Rosetta-derived model for fa-csHisH90G that we previously used for LRO, ACO, and C_α-SBM predictions. Instead, molecular replacement was achieved using a new fa-csHisH90G model generated using the publicly available AlphaFold2 server, ColabFold (Mirdita et al., 2022) (Appendix C, Figure C3A). Data refinement resulted in a well-defined electron density map (Figure 3.8B), and reasonable electron density is present for all residues in the final structure. Data collection and refinement statistics are given in Table 3.5.

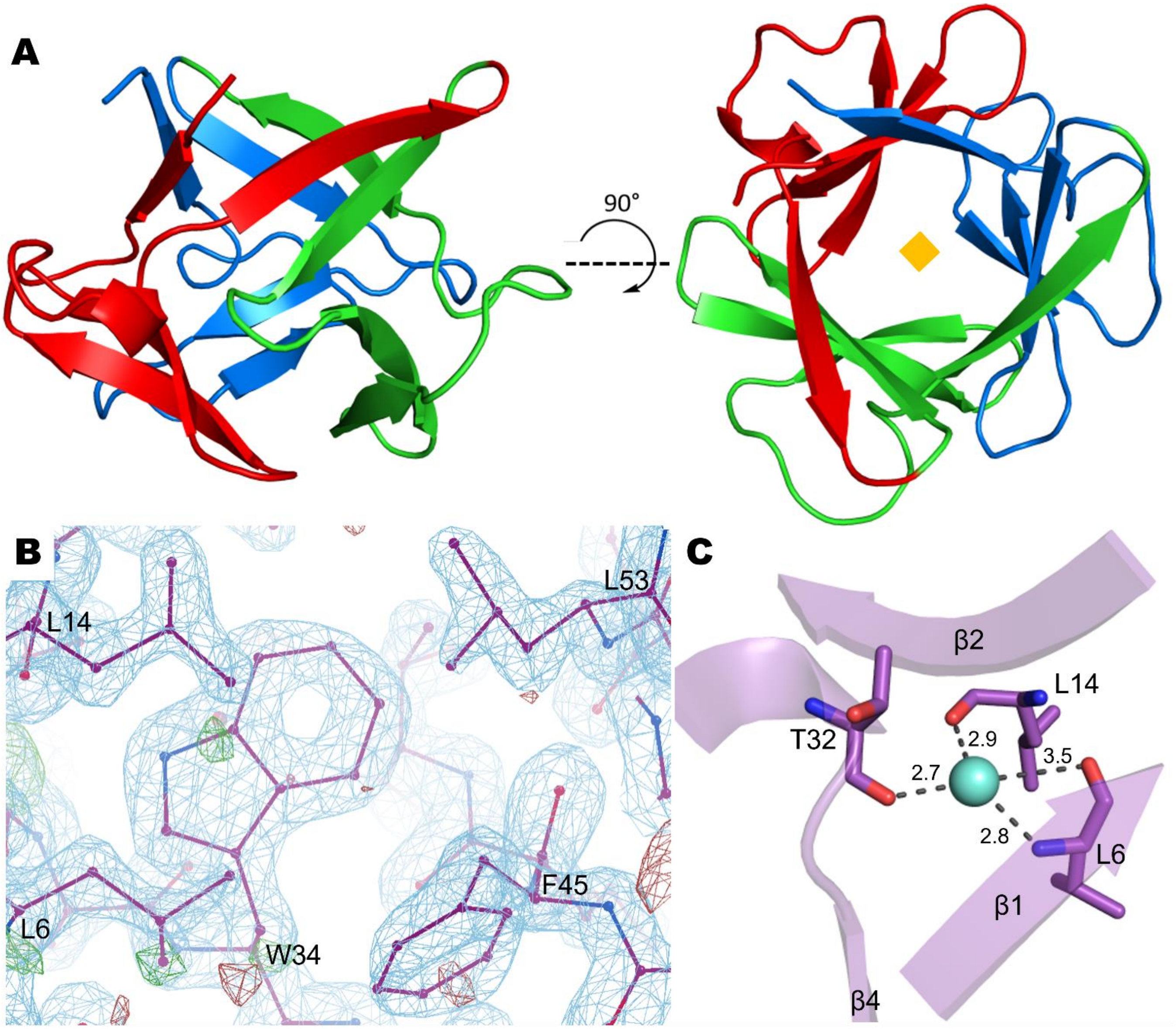

A
90°
B
L14
L53
L6
W34
F45
C
β2
L14
T32
2.9
2.7
3.5
2.8
L6
β1
β4

Figure 3.8. The fa-csHisH90G crystal structure shows enhanced symmetry characteristic of the β-trefoil architecture. (A) fa-csHiSH90G crystals were grown in 0.2 M $LiSO_4$ M, 0.1 M phosphate/citrate buffer pH 3.8, and 20 % PEG 1000, soaked in 20% PEG400, flash frozen in liquid nitrogen, and shot on the University of Waterloo home source (Cu K_α = 1.54 Å) (Figure C2). Data were solved by CCP4i Molrep (Vagin and Teplyakov, 1997; Potterton et al., 2003) using a fa-csHisH90G structural model generated by ColabFold (Mirdita et al., 2022). Data were refined to 1.7 Å using Phenix.refine and COOT (Emsley et al., 2010; Afonine et al., 2012; Headd et al., 2012). fa-csHisH90G displays internal pseudo three-fold symmetry. The N-terminal, central, and C-terminal trefoils are colored red, green, and blue, respectively, for ease of comparison with 3Foil (Figure 1.5). The fa-csHisH90G central axis is represented as a yellow diamond. **(B)** $2F_o$-F_c electron density map is contoured at 2σ. Core residues are shown as an example of the excellent agreement between electron density and the molecular structure. Data collection and refinement statistics are given in Table 3.5. Data were collected and refined by Norman Tran of the Holyoak group. **(C)** The structurally conserved buried water molecule (aqua sphere) is shown coordinated by L6, L14, and T32 in the N-terminal trefoil. Symmetric buried water molecules are coordinated by F45, L53, and S72 in the central trefoil and by L85, L03 and T111 in the C-terminal trefoil. Buried water molecules participate in bridging hydrogen bonds (dashed lines) between β1, β2, and β4 of each trefoil.

The fa-csHisH90G crystal structure reveals a well-folded protein with characteristic β-trefoil structure (Figure 3.8A). The fa-csHisH90G crystal structure also reveals three symmetric buried water molecules common to the β-trefoil architecture, which are thought to be important in promoting structural organization of local hydrophobic residues during folding (Figure 3.8C) (Murzin et al., 1992; Broom et al., 2012; Blaber, 2020). Each water molecule is hydrogen bonded to three residues within a given trefoil. In fa-csHisH90G, the buried water molecule is coordinated by L6, L14, and T32 in the N-terminal trefoil (Figure 3.8C). Equivalent residues coordinate water molecules in the central (F45, L53, S72) and C-terminal trefoils (L85, L93, T111). These residues are structurally equivalent to the 3Foil residues L7, L16, and Q40, which coordinate the buried water molecule in 3Foil (Broom et al., 2012). Thus, our designed core-swapped protein captures folding mechanisms specific to the β-trefoil fold in similar fashion to 3Foil.

Table 3.5. fa-csHisH90G data collection and refinement statistics

Data Collection	
Resolution (Å)	50.50-1.70
Data collection wavelength (Å)	1.54
Space group	P1
a (Å)	33.40
b (Å)	33.46
c (Å)	54.89
α (°)	75.22
β (°)	75.42
γ (°)	89.02
R_{merge} (%)	0.072 (0.115)
R_{meas} (%)	0.084 (0.191)
R_{pim} (%)	0.044 (0.135)
$CC_{1/2}$	0.982 (0.956)
Completeness (%)	89.12
Unique reflections	21637
Redundancy	2.9
Mean I/σ(I)	13.69 (3.45)
Wilson B (Å^2)	17.86
Refinement	
Reflections used	21622
R_{free} reflections	1050
R_{work} (%)	18.25
R_{free} (%)	22.62
No. water	145
No. non-H atoms protein	1975
No. non-H atoms ligands	73
r.m.s. bond (Å)	0.005
r.m.s angle (°)	0.734
Ramachandran outliers (%)	0.00
Clash score	5.38
Mean B-factor (Å^2)	
Water	28.71
Protein	25.71
Ligands	39.55

*Statistics for the highest resolution shell are given in parentheses. Data collection and refinement of fa-csHisH90G was performed by Norman Tran of the Holyoak group.

Compared to the wtHis NMR structure, the fa-csHisH90G crystal structure shows greater 3-fold symmetry similar to that of other β-trefoil proteins (Figure 3.8A; Figure 3.9A). This enhanced symmetry appears to be due to repacking of the hairpin cap of the N-terminal trefoil. The N-terminal and central hairpins in the wtHis hairpin cap are not as tightly associated as in other β-trefoil proteins, which disrupts the β-trefoil pseudo three-fold symmetry (Figure 3.9A). In fa-csHisH90G, β3 of the N-terminal hairpin is shifted toward the protein's central axis by at 4.0 Å (Figure 3.9A,B), restoring the fold's structural symmetry. Puzzlingly, V21 in β3 is not mutated from wtHis to fa-csHisH90G, and thus cannot account for the change in β3 orientation. While the PROSS mutation A20L is present in β3 (Figure 3.9B), A20 in wtHis and L20 in fa-csHisH90G share similar hydrogen bonding between β-sheets 2 and 3 and make no other intramolecular contacts, making this mutation unlikely to be solely responsible for the observed shift in β3. Alignment of fa-csHisH90G and wtHis core residues shows repacking throughout the protein core that may account the change in position observed for β3 (Figure 3.9C). In addition to V21 in β3, C_α carbons for residues 4 and 61 show significantly different spatial coordinates between wtHis and fa-csHisH90G, indicating movement in the local backbone between structures. While C_α carbons of remaining core residues generally show good alignment between proteins, side chain chi angles differ substantially for several residues. For example, U layer residues L14 and L53 (Figure 2.1), which are not mutated from wtHis to fa-csHisH90G, have chi_1 values of -58.62° and -93.98°, respectively, in wtHis and chi_1 values of -169.71° and -178.59° in fa-csHisH90G (Appendix C, Table C1). Altered packing in the U layer likely results from accommodation of larger tryptophan residues in the B layer. Further, U layer packing in fa-csHisH90G appears to reorder packing in the L layer of the hairpin cap such that V21 and V61 are shifted towards the protein's central axis. Thus, global repacking of 3Foil core residues in the hisactophilin scaffold results in local backbone remodeling in the hairpin cap that increases symmetry in hisactophilin trefoils.

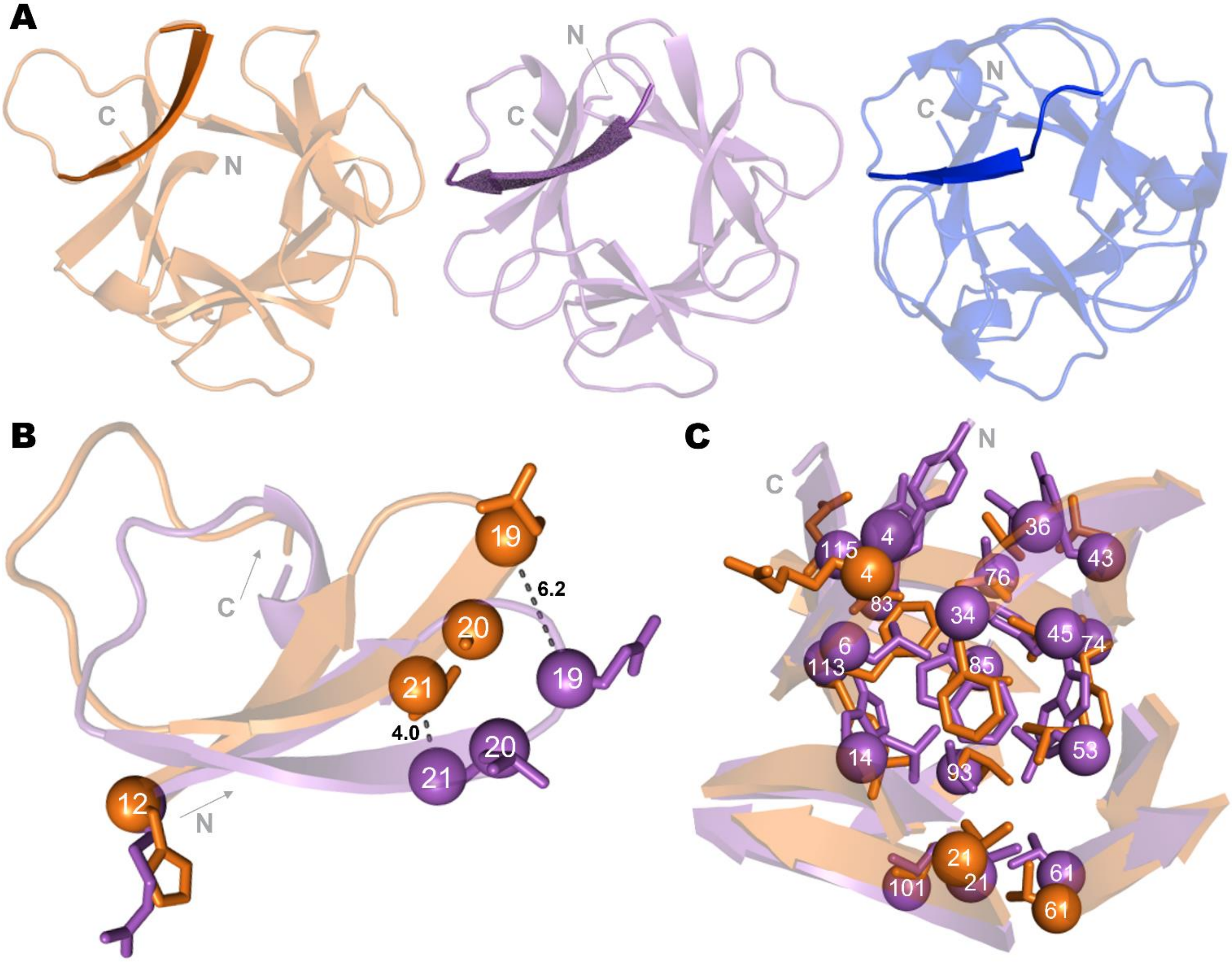


Figure 3.9. Repacking of core residues increases symmetry in the hairpin cap. (A) Native structures for wtHis (orange, left), fa-csHisH90G (purple, middle), and 3Foil (blue, right) are given looking up the β-barrel (*i.e.* the N- and C-termini facing into the page) with β3 highlighted to show the shift in the β-sheet position from wtHis to fa-csHisH90G. In fa-csHisH90G, β3 is brought closer to the protein's central axis (Figure 3.8A) in close resemblance to other β-trefoil proteins, such as 3Foil. **(B)** β3 is moved 4.0 Å closer to the center (measured between wtHis and fa-csHisH90G V21 C_α atoms), and the hairpin turn shifts 6.2 Å (measured between E19 C_α atoms). The PROSS mutations H12R and A20L are shown. **(C)** fa-csHisH90G (purple) is overlaid with wtHis (orange) to illustrate core residue repacking in fa-csHisH90G. Core residues in the wtHis NMR structure show significantly different side chain orientations in the N- and C-termini and hairpin cap compared to the fa-csHisH90G crystal structure. Loops are removed for simplicity.

Comparing core residues in the fa-csHisH90G and 3Foil crystal structures reveals remarkably similar side chain orientations for nearly all core residues (Figure 3.10A). In fact, fa-csHisH90G and 3Foil core residues have a root-mean-square deviation (RMSD) of only 0.40 Å. This is in excellent agreement with fluorometry data, which show that fa-csHisH90G core tryptophans share a similar environment with those of 3Foil, as evident in their near identical maximum emission wavelengths and characteristic far red shift going from folded to unfolded protein (Figure 3.6B; Figure S3 of Broom et al., 2012). While the fa-csHisH90G contains the additional mutation F45 from PROSS, csHisH90G and fa-csHisH90G display almost identical fluorescence (Figure 3.5B; Figure 3.6B), suggesting similar core-packing around tryptophan residues. Thus, the fa-csHisH90G crystal structure supports that csHisH90G adopts 3Foil-like core packing. fa-csHisH90G core residues do not align as well with core residues in the computational model for csHisH90G, with particular attention drawn to V21 and V61 as in the wtHis alignment (Figure 3.10B). However, fa-csHisH90G and csHisH90G core residues have a small overall RMSD of 0.67 Å. Similarly, comparing core residues in the fa-csHisH90G crystal structure and computational model generated using Rosetta gives an only slightly increased RMSD of 0.74 Å. Further, CASTp predicts a core cavity with a volume of 0.34 Å^3 for the fa-csHisH90G, which closely resembles that predicted for the Rosetta model of fa-csHisH90G (Appendix D, Figure D2). So, with the exception of V21 and V61, core residues are well-captured in our computational models for csHisH90G and fa-csHisH90G.

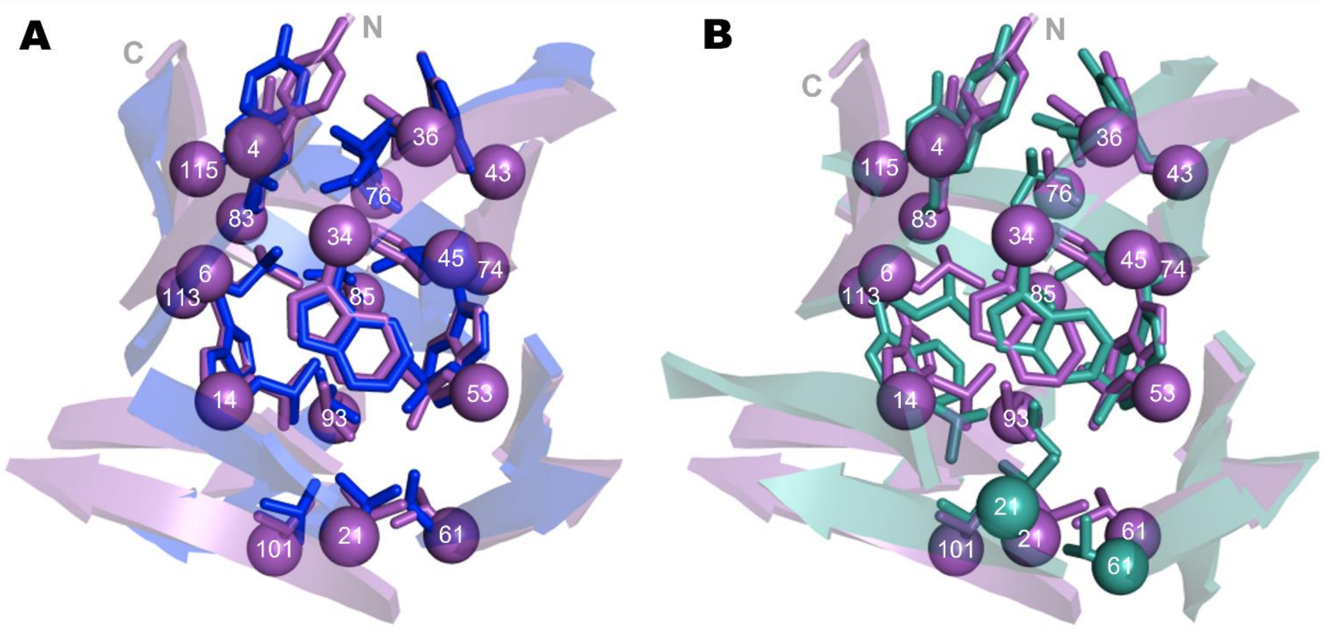


Figure 3.10. Core residues in the fa-csHisH90G crystal structure align well with those of 3Foil. fa-csHisH90G (purple) is overlaid with **(A)** 3Foil (blue) and **(B)** csHisH90G (cyan) to compare packing of core residues in the protein cores. Core residues in the csHisH90G computational model are largely consistent with those of fa-csHisH90G, with the exception of V21 and V61. fa-csHisH90G and 3Foil core residues align extremely well in almost all cases, with an RMSD of 0.40 Å. Loops are removed for simplicity.

3.10 Structural differences in wtHis, csHisH90G, and fa-csHisH90G are captured by computational models

In addition to increased symmetry between trefoils, the fa-csHisH90G crystal structure shows key deviations from the wtHis NMR structure in the N- and C-terminal β-strands (Figure 3.11A). β1 and β12 are largely remodeled in fa-csHisH90G compared to wtHis owing to the substitution of the hydrophobic residues Y4 and L115 for the charged residues R4 and E115, respectively. In wtHis, the side chains of both R4 and E115 face toward solvent despite these residues occupying conserved core-facing positions, forcing the β1 and β12 backbones to adopt a twisted conformation (Figure 3.11A; Figure 3.9C; Figure 2.1A). In contrast, both residues point toward the protein core in fa-csHisH90G, eliminating the twisted backbone conformation (Figure

3.11A, B). The N- and C-terminal β-strands in our model for csHisH90G closely resembles those of the fa-csHisH90G crystal structure, with residues Y4, L6, and L115 displaying similar rotamer conformations in both structures. While W113 has different rotamers in the fa-csHisH90G and csHisH90G structures (Figure 3.11B), the csHisH90G W113 orientation captures 13 of 16 contacts made by the W113 rotamer in fa-csHisH90G. Thus, the csHisH90G model is in good agreement with the experimentally derived conformation of the N- and C-terminal β-strands of fa-csHisH90G. Notably, this structural data supports our hypothesis that 3Foil core residues change intramolecular contacts in the C-terminal trefoil during protein folding since the core-facing L115 in fa-csHisH90G cannot make long-range stabilizing contacts to residues in the β8-β9 turn as E115 does in wtHis (see section 3.8). Therefore, the experimental structure for fa-csHisH90G lends credibility to C_α-SBM simulation data and, consequently, to the computational models for csHisH90G and fa-csHisH90G.

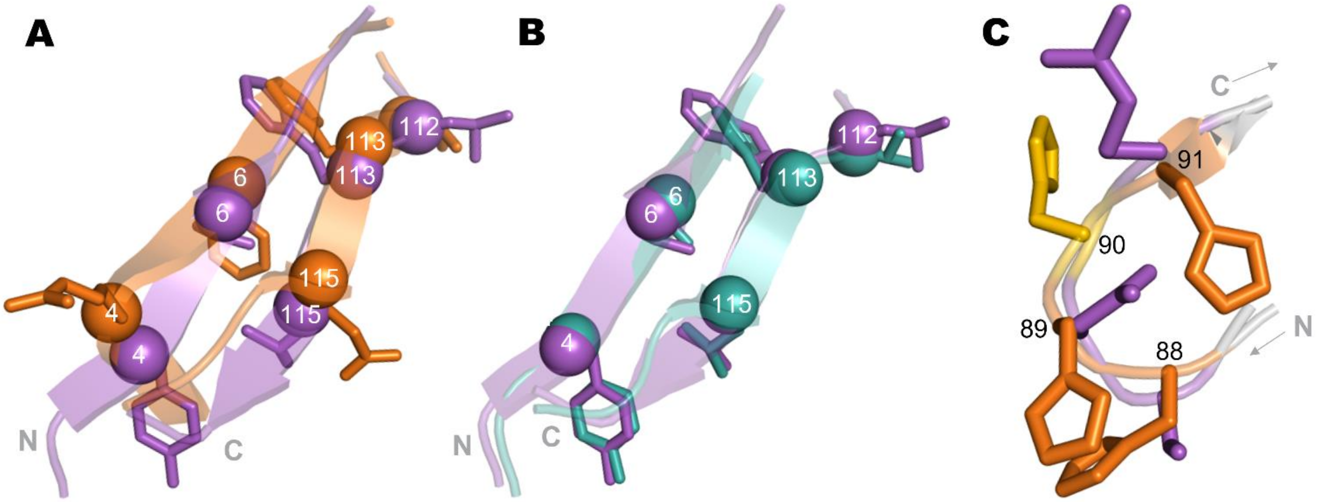


Figure 3.11. The fa-csHisH90G structure reveals key deviations from wtHis. (A) fa-csHisH90G N- and C-termini are remodeled such that core residues Y4 and L115 face inward while R4 and E115 in wtHis twist the backbone to face toward solvent. The PROSS mutation T112L is shown. **(B)** The N- and C-termini show good agreement between the fa-csHisH90G crystal structure and the Rosetta-generated csHisH90G model. **(C)** The β9-β10 turn is completely remodeled in fa-csHisH90G compared to wtHis. Despite the sequential mutation of four residues, the wtHis and fa-csHisH90G backbones are well-aligned. C_{α}-SBM contacts in the β9-β10 turn are largely similar between wtHis and fa-csHisH90G (see Table C3 in Appendix C).

Finally, the β9-β10 turn is completely remodeled in fa-csHisH90G compared to wtHis owing to the addition of the H90G point mutant and several local PROSS mutations (Figure 3.11C). Specifically, PROSS suggested the point mutations H88S, H89N, and H91R. These mutations are notable over other PROSS mutants for several reasons. H88S, H89N, and H91R are the only localized substitutions suggested by PROSS in all three csHisH90G variants, with the exception of E19D and A20R in mu-csHisH90G (Figures 9-11A, B). Additionally, H91R is present in all csHisH90G variants, suggesting that R91 is particularly favorable in csHisH90G as modeled using the PROSS energy function. Lastly, these mutations surround the H90G point mutant, which is kinetically destabilizing (Figure 3.5; Table 3.3). A precise molecular mechanism for kinetic destabilization by G90 has not been experimentally investigated. However, local chain rigidity is known to modulate kinetic stability in other designed proteins (Clarke and Fersht, 1993; Mansfeld et al., 1997; den Burg et al., 1999; Pikkemaat et al., 2002; Liu et al., 2021), and glycine is an exceptionally flexible residue. Therefore, G90 may reduce kinetic stability in hisactophilin by increasing mobility in the β9-β10 turn. PROSS previously stabilized human acetylcholinesterase by increasing backbone rigidity in loop residues (Goldenzweig et al., 2016), and so may be capable of enhancing csHisH90G thermostability in similar manner. Notably, the remodeled β9-β10 turn displays lower B-factors than its structurally equivalent turns, suggesting greater rigidity in the β9-β10 turn (Appendix C, Figure C3B). Unfortunately, B-factors are unavailable for wtHis and HisH90G, so the rigidity of residues in the β9-β10 turn cannot be directly compared for wtHis, HisH90G, and fa-csHisH90G. Despite mutation to four sequential residues, the β9-β10 turn backbone is strikingly similar in wtHis and fa-csHisH90G (Figure 3.11C). Additionally, C_α-SBM contacts in the β9-β10 turn do not significantly change except for residue 89, which makes two long-range contacts in wtHis and five long-range contacts in fa-csHisH90G (Appendix C, Table C3). Therefore, the effect of these localized PROSS mutations in the β9-β10 turn is ambiguous in both C_α-SBM simulations and in structural data for wtHis and fa-csHisH90G, and further investigation is needed.

4 Discussion

4.1 Using simple descriptors of protein topology to engineer kinetic stability

Interactions present in the native state dominate non-native interactions during protein folding (Bryngelson et al., 1995; Wolynes et al., 1995; Onuchic et al., 1997; Onuchic and Wolynes, 2004). Thus, protein topology or structure as encoded by the network of interactions or contacts present in the native state, rather than detailed protein energetics, determine how proteins fold. Simple functions of these contact networks, such as LRO and ACO, seem to be able to capture the principal features of protein topology and correlate quantitatively with unfolding free energy barrier heights (Chavez et al., 2004; Broom et al., 2015a). MD simulations of SBMs, which encode these contact networks and ignore attractive non-native interactions, are not only able to model these barrier heights (Chavez et al., 2004; Gosavi et al., 2006; Gosavi, 2013), but can also be used to understand barrier shapes, the population of intermediates, and the folding path (Hills and Brooks, 2009; Noel and Onuchic, 2012; Kmiecik et al., 2016). We used these simple descriptors of protein topology to model and then modulate unfolding barrier heights to directly engineer protein kinetic stability. Specifically, in designing csHisH90G, we aimed to establish engineering long-range intramolecular interactions as a credible strategy for modulating protein kinetic stability and to show that LRO, ACO, and C_α-SBM unfolding free energy barriers may serve as valuable predictive measures of changes in kinetic stability.

Engineering long-range interaction networks is attractive for several reasons. Previous C_α-SBM studies of 3Foil show that deleting long-range contacts made by loop residues lowers the 3Foil free energy barrier of unfolding (Broom et al., 2015b), while experimental kinetics from the 3Foil mutant Q71I show that eliminating long-range contacts in the mini-core decreases kinetic stability (Dubey et al., 2005; Broom et al., 2017). Long-range interactions are encompassed by both LRO and ACO, while short-range contacts are considered directly only by ACO (Gromiha and Selvaraj, 2001; Ivankov et al., 2003). Further, compared to ACO, LRO provides a stronger, more linear correlation with protein unfolding rates for proteins of larger size and variable structure (Broom et al., 2015a). Here, LRO and ACO calculations predicted 4.1-fold and 2.8-fold slower unfolding rate constants at the transition midpoint, respectively, for csHisH90G compared to wtHis (Table 3.1), which we use as a structural proxy for the pseudo-wild type parent protein HisH90G. Experimental folding kinetics show that csHisH90G has 2.0-fold slower unfolding kinetics relative to HisH90G (Table 3.3; Figure 3.1D). Similarly, LRO and ACO values for fa-csHisH90G predict a 3.9-fold and 3.5-fold decrease in unfolding kinetics compared to csHisH90G (Table 3.2), while experimental kinetics for fa-csHisH90G are 1.3-fold slower than those of csHisH90G (Table 3.4). So, in the case of both csHisH90G and fa-csHisH90G, ACO may provide a more accurate prediction of the unfolding rate constant than LRO. Additional kinetic stability designs using proteins of variable size and structure must be pursued to investigate whether LRO or ACO is more accurate in predicting protein kinetic stability.

4.2 Alternate methods of engineering kinetic stability

Our method for engineering protein kinetic stability is founded on the definition of kinetic stability, *i.e.* tuning the height of the Gibbs free energy barrier between the native protein and the transition state. In contrast, previous studies aimed to modulate kinetic stability indirectly by engineering protein characteristics that show experimental correlation with protein kinetic stability. For example, engineering strategies that target disulfide bonds (Mansfeld et al., 1997; den Burg et al., 1999; Liu et al., 2021), residues with high B-factors (Le et al., 2012; Chen et al., 2015; Duan et al., 2016; Liu et al., 2021), or residues with high thermal flexibility (Pikkemaat et al., 2002; Xie et al., 2014; Quezada et al., 2018; Liu et al., 2021) all aim to reduce protein mobility and, consequently, protein unfolding. One way to reduce the flexibility of an amino acid is through mutations that increase its intra-protein contacts, and such mutations are likely to be of a similar nature across all methods. However, local rigidity could also be increased by mutations that tune secondary structural propensities (Geiger-Schuller et al., 2018), and such mutations are likely to be complementary to those seen in our method. This method for engineering kinetic stability relies on global structural measures such as LRO and ACO and is unlikely to be able to capture mutations that tune local sequence energetics and promote the formation of specific secondary structural elements or loops, the effect of the H90G mutation on wtHis being one potential example. Other examples studied here may include our PROSS-based csHisH90G variants, as preliminary experimental kinetics indicate that ph-csHisH90G and mu-csHisH90G may have reduced kinetic stability compared to csHisH90G despite increased LRO and ACO measures predicting slower unfolding kinetics (see section 5.1). These PROSS variants consist of multiple point mutations not related through a network of interacting residues and therefore may be subject to additional sequence effects not suitably modeled by the current method. It should be noted that increasing thermodynamic stability in many proteins will slow protein unfolding and increase kinetic stability in native-like conditions (Abkevich et al., 1995). However, such increases in kinetic stability may not hold when comparisons are made at the mid-point of the transition.

4.3 Calculating relative barrier heights: minutiae of the protein structure may not matter

Several forms and flavors of structure-based models exist that encode the protein structure at different levels of coarse-graining and in slightly different ways (Clementi et al., 2000; Hyeon and Thirumalai, 2011; Noel and Onuchic, 2012; Yadahalli et al., 2014; Noel et al., 2016). Similarly, although a canonical method for determining contacts for LRO and ACO calculations exists (Gromiha and Selvaraj, 2001; Ivankov et al., 2003), these measures could be calculated using other contact definitions (Broom et al., 2015b). These differences in the potential energy function or contact calculation are likely to not have a significant effect on relative barrier heights except in proteins where specific functional features affect folding and need to be encoded accurately (Azia and Levy, 2009; Yadahalli and Gosavi, 2016). This is true because overall protein topology, rather than the details of energetics, determines how a protein folds (Bryngelson et al., 1995; Wolynes et al., 1995; Onuchic et al., 1997; Onuchic and Wolynes, 2004). Conversely, a detailed structure of the protein, such as a high-resolution crystal structure, may not be required to predict relative barrier heights from C_α-SBM simulations, ACO, and LRO. In fact, higher kinetic stability was achieved here in csHisH90G and fa-csHisH90G despite the absence of a hisactophilin crystal structure.

Lack of a high-resolution crystal structure is not an uncommon hurdle in protein engineering. Crystal structures currently comprise only ~100 000 unique proteins of billions of known protein sequences, and known protein sequences outnumber solved protein structures 736 times over (Muhammed and Aki-Yalcin, 2019; Jumper et al., 2021). Recent advances in structure prediction, *e.g.* Alphafold2 (Jumper et al., 2021), ColabFold (Mirdita et al., 2022), and RoseTTAFold (Baek et al., 2021), now enable the generation of structural models with significant confidence from just the primary sequence that are suitable for C_α-SBM. Consequently, our method may be applied to the large proportion of proteins that lack high-resolution crystal structures. That being said, a “reasonable” and complete structure is required for C_α-SBM simulations and LRO and ACO calculations.

csHisH90G and fa-csHisH90G provide an excellent example of “reasonable” yet imperfect structural models being implemented effectively to predict and improve kinetic stability. The experimental structure of fa-csHisH90G shows key deviations from the structural models used to determine LRO, ACO, and C_α-SBM unfolding barriers (Appendix C, Figure C3A). For example, repacking of β-strands 3 and 7 in the hairpin cap is poorly captured in computational models, as is evident in the distinct spatial coordinates for V21 and V61 C_α carbons in structural alignments (Figure 3.10B). Thus, contacts in the hairpin cap may not be accurate to experimental structure in csHisH90G and fa-csHisH90G C_α-SBM simulations. Significantly, these erroneous contacts have the potential to adversely affect the unfolding free energy barrier height predicted by simulations. Unfolding rates predicted by LRO and ACO from faulty structural models may suffer similar inaccuracies. However, remaining core residues and the N- and C-termini are generally well-modeled for csHisH90G and fa-csHisH90G (Figure 3.10B; Figure 3.11A), with good agreement between contact maps based on the crystal structure and computational models. Ultimately, csHisH90G and fa-csHisH90G unfolding free energy barriers capture a moderate increase in barrier height for both designed proteins compared to their parent protein (Table 3.1; Table 3.2), which agrees with observed experimental improvements in kinetic stability (Figure 3.5D; Table 3.3; Figure 3.6D; Table 3.4). Thus, while not all structural features match the experimental structure, our models sufficiently capture csHisH90G and fa-csHisH90G to enable prediction and design of kinetic stability.

4.4 Core engineering can be performed without close sequence homologues

Many contemporary protein stability design strategies require the use of close sequence homologues and an MSA. For example, MSAs are used in consensus design (Broom et al., 2012; Feng et al., 2016; Sternke et al., 2019) and coevolution analysis (Reynolds et al., 2013; Ovchinnikov et al., 2014; Swint-Kruse, 2016). As with high-resolution crystal structures, many proteins lack a sufficient number of homologous sequences to benefit from these strategies. In fact, of the sequences in UniProtKB, 23% match no Pfam entry (Mistry et al., 2021). Despite belonging to the β-trefoil lineage, hisactophilin lacks closely related sequence homologues, precluding application of consensus or coevolution design. However, hisactophilin is an ideal test protein for our method for designing kinetic stability because it has close structural homologues with distinctly different kinetic stabilities. Further, strong conservation of core residue hydrophobicity but not identity across β-trefoil sequences, including wtHis and 3Foil, suggested that the β-trefoil core may tolerate the re-engineering of its interaction network, the basis of our strategy to design kinetic stability.

The results presented here recommend engineering of protein cores as an attractive and accessible target for increasing protein kinetic stability. Namely, swapping entire networks of core interactions between homologous proteins may be widely applicable. Improving protein core packing is generally associated with augmented van der Waals interactions, more favorable core residue side chain steric interactions, and increased burial of hydrophobic surface area, all of which are known to increase protein stability (Ventura and Serrano, 2004; Borgo and Havranek, 2012; Kim et al., 2012). Further, core engineering is among the most well-developed, predictable, and feasible strategies in protein design. Mutagenesis studies show that protein core sequences can be highly amenable to mutation to other hydrophobic residues, including complete core redesign, while retaining the protein fold (Kuhlman and Baker, 2000; Ng et al., 2007; Murphy et al., 2012; Ben-David et al., 2019; Koga et al., 2020). Protein design software, including SCWRL (Krivov et al., 2009), OSCAR (Liang et al., 2011), RASP (Miao et al., 2011), Rosetta (Kuhlman and Baker, 2000), SCCOMP (Eyal et al., 2004)), and FoldX (Guerois et al., 2002), achieve higher accuracy for predicting favorable core residue conformations compared to surface residues, offering higher rates of success for *de novo* core mutant designs (Peterson et al., 2014; Gaines et al., 2017; Broom et al., 2020). Core residues are often highly conserved as hydrophobic (Kuhlman and Baker, 2000;

Ben-David et al., 2019), making protein core engineering amenable to bioinformatic design strategies such as consensus and covarying residue design. Significant to our method for designing kinetic stability, consideration of alternate core packing is well suited to optimizing both van der Waals and steric interactions to increase a protein's LRO and ACO. Further, since core residue positions tend to be conserved among homologous proteins while specific residue identities can vary, core residues lend themselves as promising targets for designs that aim to swap entire networks of protein interactions, such as that reported here.

Beyond the demonstrated success of engineering core residues, this approach may allow protein stabilization while generally maintaining function. For example, in many β-trefoil proteins engineering core residues may have minimal effect on function because β-trefoils achieve ligand-binding through surface loop residues (Figure 1.5) (Brych et al., 2004; Olsen et al., 2004; Gosavi et al., 2008; Broom et al., 2012; Terada et al., 2017; Blaber, 2022). While engineering 3Foil core residues into hisactophilin appears to substantially restructure hairpin cap residues, including loops, the hisactophilin hairpin cap is more structurally distinct than most β-trefoils, for which less pronounced repacking is expected. Despite having residues conserved as hydrophobic at key positions contributing to the core, β-trefoil proteins display considerable plasticity in their core packing arrangements (Murzin et al., 1992; Ponting and Russell, 2000; Longo and Blaber, 2013; Blaber, 2022). So, β-trefoil proteins with desirable functionality but only moderate stability may gain kinetic and thermodynamic stability from the replacement of core residues with those of another β-trefoil or from *de novo* core repacking. In addition, developing a scaffold with high kinetic stability may provide a particularly useful starting point for subsequent engineering of function, which often impairs stability (Liu et al., 2001; Gosavi, 2013; Tenorio et al., 2022). Core residue engineering may be even more accessible for proteins containing repetitive structures and structurally symmetric proteins, as noted for β-trefoils, TIM barrels, or repeat proteins (Meiering et al., 1991; Sancho et al., 1991; Broom et al., 2015b, 2016; Vrancken et al., 2020). Such proteins are of interest for multivalent binding of identical or distinct ligands at binding sites. Finally, as observed for fa-csHisH90G, stabilizing proteins by engineering core residues may better promote protein crystallization and aid structural characterization of surface binding and catalytic sites.

5 Future directions

5.1 Experimental characterization of ph-csHisH90G and mu-csHisH90G

Of immediate interest is to complete experimental characterization of ph-csHisH90G and mu-csHisH90G. Specifically, equilibrium denaturation and kinetic folding and unfolding experiments must be completed to further validate our method of using protein topology measures and C_{α}-SBM simulations to predict and modulate kinetic stability. Preliminary measurements indicate that both ph-csHisH90G and mu-csHisH90G show improved thermodynamic stability compared to csHisH90G but decreased kinetic stability (data not shown). However, LRO and ACO values predict slower unfolding rates for ph-csHisH90G and mu-csHisH90G than for csHisH90G (Table 3.2). Thus, if preliminary results for ph-csHisH90G and mu-csHisH90G hold, our method fails to predict diminished kinetic stability in these csHisH90G variants. The experimental structure for fa-csHisH90G suggests that the computational models used for LRO, ACO, and C_{α}-SBM unfolding free energy barrier predictions for csHisH90G, ph-csHisH90G, and mu-csHisH90G may have several structural inaccuracies (Figure 3.9; Figure18; Appendix C, Figure C3). Further, it remains unclear whether our predictions are inaccurate due to the current computational model or due to sequences effects from PROSS mutations that our method cannot capture. Accordingly, crystal screens for ph-csHisH90G and mu-csHisH90G were plated toward obtaining experimental structures for both variants. If crystal structures are obtained, new LRO, ACO, and C_{α}-SBM unfolding barrier predictions may be calculated and compared to experimental kinetics. If our predictive measures still overestimate kinetic stability in ph-csHisH90G and mu-csHisH90G, experimental structures may be used to gain insight into structural differences between the kinetically destabilized variants and csHisH90G and fa-csHisH90G. So, acquiring experimental folding kinetics and crystal structures for ph-csHisH90G and mu-csHisH90G will not only further define the nuances of our method for engineering kinetic stability, but may also enhance our understanding of the molecular determinants of protein kinetic stability.

Similarly, experimental characterization of PROSS mutations as single point mutants or as double or triple mutants may enable us to determine which mutations are responsible for improved kinetic stability in fa-csHisH90G and apparent diminished kinetic stability in ph-csHisH90G and mu-csHisH90G. Here, we implemented multiple mutations in our designs because multiple mutations may rescue a protein should a destabilizing point mutation be present in the design (Magliery, 2015; Goldenzweig et al., 2016; Khersonsky et al., 2018). Unfortunately, the presence of multiple mutations, though beneficial in avoiding non-productive protein designs, confound the effect of the individual mutations. Thus, to understand the effect of specific mutations on kinetic stability, mutations suggested by PROSS in fa-csHisH90G, ph-csHisH90G, and mu-csHisH90G should be expressed and characterized as single point mutants in the csHisH90G scaffold. Moreover, mutations that appear to act in concert, *e.g.* H88S, H89N, H90G, H91R in the β9-β10 loop in fa-csHisH90G, should also be investigated as combinatorial mutants. Single point mutants are listed in Table 5.1.

Table 5.1. PROSS mutations expressed as single point mutants in csHisH90G

csHisH90G residue	fa-csHisH90G mutant	ph-csHisH90G mutant	mu-csHisH90G mutant
H12	R		
E19			D
A20	L	G	R
G26		S	
T32		Q	
N38	Q	K	Q
L45	F		
C49	N	N	N
K59		R	G
H65		S	
S72		C	
H75			Q
K82		R	
S84			A
H88	S		
H89	N		
H91	R	R	R
S94			A
H100		G	
T112	L		L

5.2 csHis as a scaffold for engineering loop function into hisactophilin

In addition to serving as a proof of concept for our kinetic stability method, the core-swapped design provides an excellent starting scaffold for engineering loop functionality. csHisH90G is significantly thermodynamically stabilized, and in the absence of the H90G point mutation, it is also expected to be approximately twice as kinetically stable as wtHis. Since the introduction of functional residues often results in a thermodynamic stability trade-off (Meiering et al., 1991; Sancho et al., 1991; Fersht, 1999; Liu et al., 2001; Broom et al., 2015b), and since any novel function will gain longevity from greater kinetic stability, core-swapped hisactophilin without the H90G point mutation (csHis) is the best scaffold from which to engineer functionality into hisactophilin. Additionally, engineering csHis loops provides the opportunity to further increase hisactophilin kinetic stability by engineering additional long-range contacts and increasing peptide length (Ivankov et al., 2003; Broom et al., 2015a, 2015b). In keeping with the present design, a logical loop engineering strategy to introduce both function and additional kinetic stability is to swap csHis β2-β3 loop residues for those of 3Foil. 3Foil displays lactose-binding β2-β3 loops that contribute significantly to 3Foil's remarkable kinetic stability through numerous long-range surface contacts (Broom et al., 2015b). In fact, C_α-SBM simulations show that deleting these loop contacts significantly lowers the 3Foil unfolding free energy barrier (Mut 1 in Figure 5.1). Additionally, 3Foil loops are longer than those of csHis, and the additional residues will increase the primary sequence separation of core residues such that core interactions engineered in the present thesis will gain long-range character and further contribute to kinetic stability in a loop-swapped protein. Thus, designing a loop-swapped hisactophilin protein will not only build on the work presented here, but will also lay the groundwork for engineering stable proteins with useful functions.

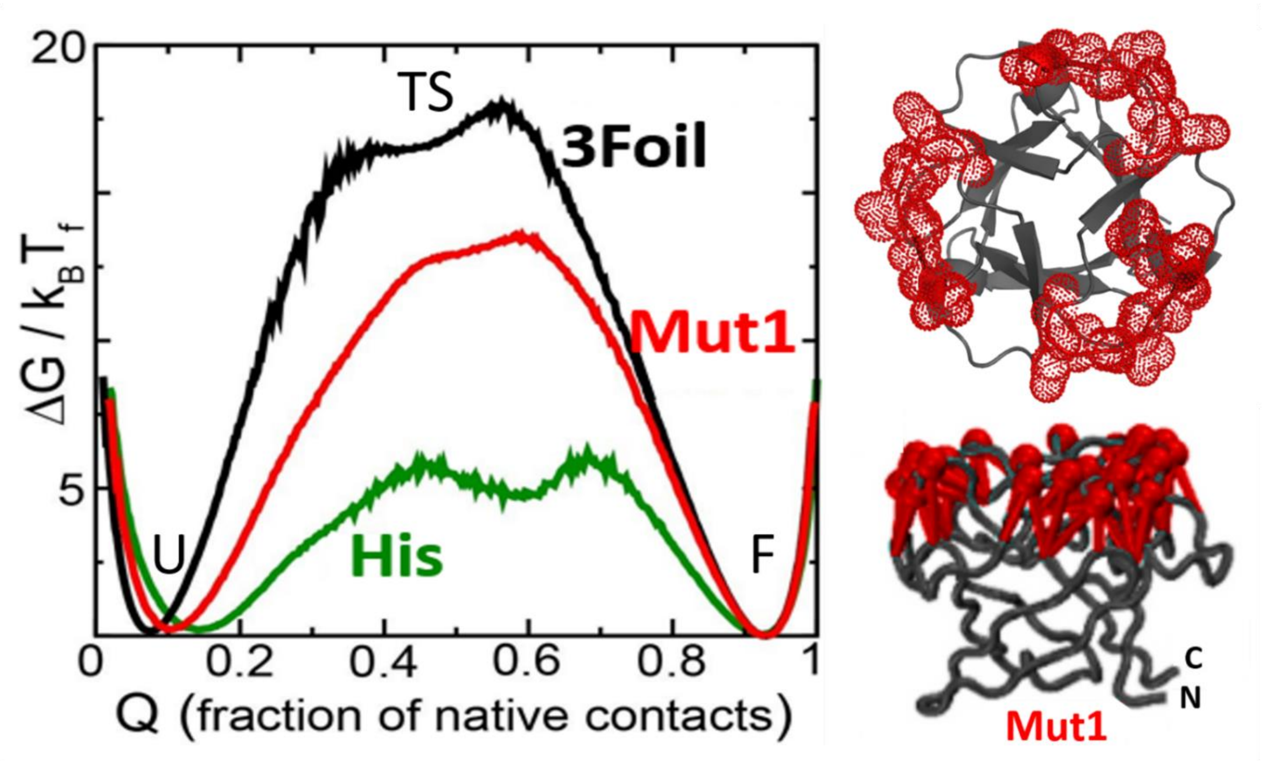


Figure 5.1. Long-range loop contacts modulate the free energy barrier of unfolding in 3Foil. Free energy barriers for unfolding for ThreeFoil (3Foil) (black) and hisactophilin (His) (green) were simulated using C_α-SBM simulations (right). The free energy barrier for Mut1 (red), a 3Foil mutant where long-range contacts made by loop residues have been deleted (left), was also simulated. Folding free energies are plotted at the transition midpoint (T_f) as a function of the fraction of native contacts (Q) and free energy ($\Delta G/k_B T_f$). The barrier height for the folding free energy decreases with fewer long-range loop contacts. A list of deleted long-range contacts in Mut1 is in the Appendix in Broom *et al.* (2015b). Adapted from Broom *et al.* (2015b).

5.3 Investigating key residues implicated in β-trefoil kinetic stability

Kinetic stability may also be modulated by mutating key residues that contribute many or very few long-range contacts to LRO and ACO. For example, Q78 in 3Foil contributes 12 long-range contacts, including three stabilizing hydrogen bonds (Figure 5.2A) (Broom et al., 2017), while the equivalent residue L63 in wtHis makes only one long-range contact. This trend is seen at symmetrically equivalent positions for both proteins. Thus, these residues represent a difference of approximately 33 long-range contacts between 3Foil and wtHis. The 3Foil mutant Q78I was shown to significantly decrease kinetic stability *in vitro* owing to the inability of the non-polar side chain to form long-range hydrogen bonds (Figure 5.2B). Conversely, substituting a polar residue at position 63 and symmetric positions 23 and 103 in hisactophilin may introduce additional long-range contacts and lead to increased LRO and kinetic stability. L63, T23, and T103 are of particular interest in wtHis and csHisH90G because these positions are cited as conserved core residues by Blaber (2019), so incorporation of 3Foil residues at these positions may further improve core packing in csHisH90G. Additionally, the polar nature of 3Foil residues Q31, Q78, and Q125 is paralleled in the hisactophilin residues T23 and T103, but not in L63. The interface between the hairpins of the N-terminal and central trefoils in the wtHis hairpin cap region show increased separation compared to other β-trefoils (Figure 3.9A), possibly in part because L63 cannot enable the formation of long-range hydrogen bonds as in Q78I. Thus, an additional core-swapped hisactophilin variant with T23Q/L63Q/T103Q should be investigated.

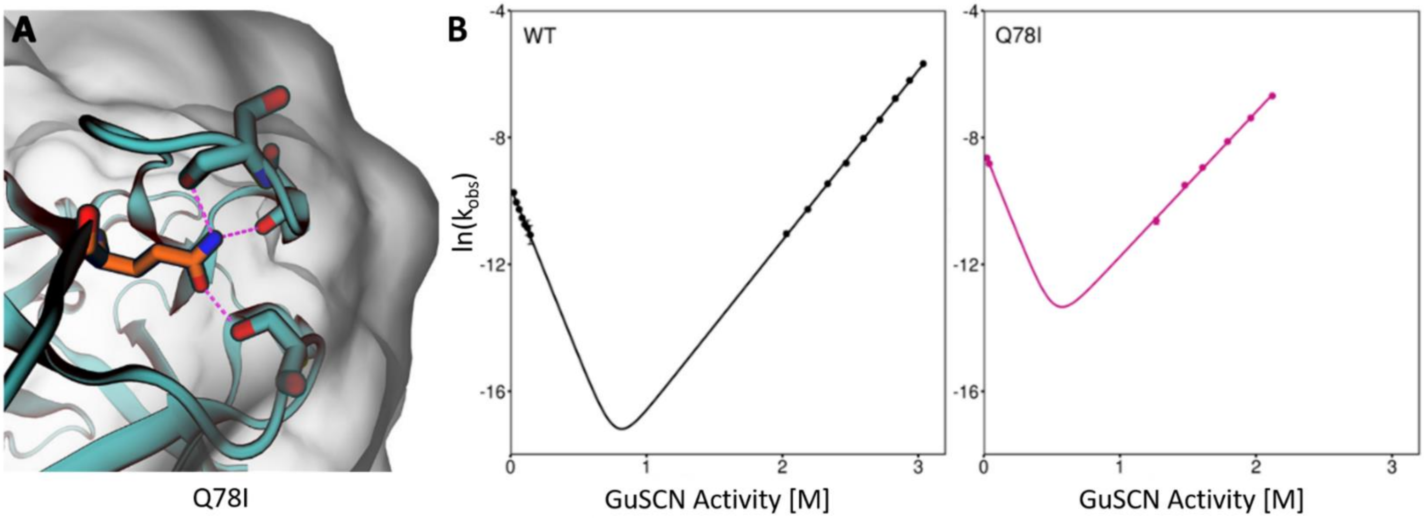


Figure 5.2. Eliminating long-range hydrogen bonds in ThreeFoil decreases kinetic stability. (A) Q78 in 3Foil forms three loop-stabilizing long-range hydrogen bonds that are lost upon mutation to the non-polar residue isoleucine. (B) Folding kinetics show that the rate of unfolding is significantly increased for the mutant Q78I compared to wild type (WT) 3Foil. An increased unfolding rate corresponds to a lower free energy barrier for unfolding and decreased kinetic stability. Adapted from Broom *et al.* (2017).

5.4 Designing diminished kinetic stability in core-swapped 3Foil

While improving kinetic stability is of obvious benefit for proteins used in practical applications, the ability to rationally decrease kinetic stability in select circumstances may also prove useful. For example, tunable protein kinetic and thermodynamic stability may be particularly useful in therapeutic applications, where high kinetic stability is beneficial for long-term storage of the protein therapeutic but can lead to *in vivo* toxicity due to slow protein degradation (Pey et al., 2008). At the least, understanding molecular mechanisms that increase and decrease protein kinetic stability may allow researchers to fine tune a protein's stability and longevity toward optimizing its functionality for a given application.

We propose to expand on the current work and implement our strategy of swapping core residues in wtHis and 3Foil to decrease kinetic stability in 3Foil. Long-range contacts across the protein core are central to 3Foil's high kinetic stability (Broom et al., 2015b). As reported in this thesis, wtHis core residues make far fewer long-range contacts than those of 3Foil. Since swapping 3Foil core residues into wtHis increases long-range contacts across the hisactophilin core and improves kinetic stability in csHisH90G, we reason that swapping wtHis core residues into 3Foil (cs3Foil) will decrease long-range contacts and kinetic stability in 3Foil. Importantly, designing and experimentally characterizing cs3Foil will allow us to pursue several questions that remain ambiguous in the current work, including whether ACO or LRO is the better measure for predicting unfolding rates from protein topology, and whether the method used to generate the structural model (*e.g.* Rosetta Comparative Modeling (Chivian et al., 2003; Song et al., 2013) vs. ColabFold (Mirdita et al., 2022), etc) significantly influences unfolding rate predictions made from LRO, ACO, and C_α-SBM unfolding free energy barriers heights.

6 Conclusions

Here, we increased kinetic stability in the β-trefoil protein hisactophilin by engineering its conserved core residues to promote the formation of additional long-range contacts. In doing so, we show that modulating long-range intramolecular interaction networks is a valid and promising approach to designing protein kinetic stability, which is an important ongoing question in the field of protein engineering. Significantly, we demonstrate that prediction of protein kinetic stability using unfolding free energy barriers simulated using C_α-SBMs and protein topology as measured by LRO and ACO are useful for engineering protein kinetic stability. Further, in modeling and experimentally validating a modest increase in kinetic stability in our designed proteins, csHisH90G and fa-csHisH90G, we show that our method may predict even small changes in kinetic stability for multi mutant proteins. In addition to predicting *in vitro* kinetic stabilization in csHisH90G and fa-csHisH90G, this method has several advantages over other prevalent strategies for rationally engineering protein kinetic stability. These advantages include *in silico* quantitative prediction of change in kinetic stability for mutant proteins, accommodation of structural models *in lieu* of experimentally determined protein structures, and relatively simple and inexpensive computational methods. Further, experimental structural characterization of fa-csHisH90G suggests that computational models used in LRO, ACO, and C_α-SBM unfolding barrier height predictions need only be reasonable approximations of the protein structure to enable kinetic stability design.

Our strategy for predicting and modulating protein kinetic stability is readily applicable to other protein families. Networks of long-range contacts, particularly those of conserved, hydrophobic core residues, can be compared between homologous proteins of differing kinetic stabilities to identify residues contributing to kinetic stability within a protein fold. Due to the nature of using structurally homologous proteins, parent proteins need not have closely related primary sequences so long as they share a common fold resulting in conserved interaction networks between structurally equivalent residues. Self-contained long-range interaction networks, such as those between loops or core residues, may then be swapped between homologous proteins to achieve the targeted kinetic stability. Importantly, using LRO, ACO, and C_α-SBM unfolding free energy barrier predictions, hybrid proteins that fail to reach the desired kinetic stability are identifiable before experimental validation. Fundamentally, these predictive measures can be used to evaluate the change in kinetic stability for any multi mutant protein relative to its parent protein. We envision this method may be extended to also incorporate novel long-range interactions, which may also be designed *de novo*. Thus, our method of engineering long-range intramolecular interactions and using protein topology measures and coarse-grained free energy barriers to modulate and predict *in vitro* kinetic stability is widely applicable in the design of hybrid and multi mutant proteins for fundamental and practical purposes

References

Abkevich, V. I., Gutin, A. M., and Shakhnovich, E. I. (1995). Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.* 252, 460–471. doi: 10.1006/jmbi.1995.0511.

Afonine, P. ., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. ., Mustyakimov, M., et al. (2012). Towards automated crystallographic structure refinement with phenix.refine. *Acta Cryst.* D68, 352–367.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science (80-.).* 181, 223–230. doi: 10.1126/science.181.4096.223.

Anfinsen, C. B., Haber, E., Sela, E., and White, F. H. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.* 47, 1309–1314. doi: 10.1073/pnas.47.9.1309.

Azia, A., and Levy, Y. (2009). Nonnative Electrostatic Interactions Can Modulate Protein Folding: Molecular Dynamics with a Grain of Salt. *J. Mol. Biol.* 393, 527–542. doi: 10.1016/j.jmb.2009.08.010.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science (80-.).* 373, 871–876. doi: 10.1126/science.abj8754.

Baker, D. (2019). What has de novo protein design taught us about protein folding and biophysics? *Protein Sci.* 28, 678–683. doi: 10.1002/pro.3588.

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., et al. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. doi: 10.1093/nar/gkaa1100.

Bekker, H., Berendsen, H. J. C., Dijkstra, E. J., Achterop, S., van Drunen, R., van der, Spoel, D., Sijbers, A., Keegstra, H., Reitsma, B., Renardus, M. K. R., Bekker, H., Berendsen, H., Dijkstra, E., Achterop, S., et al. (1993). Gromacs: A parallel computer for molecular dynamics simulations. *Phys. Comput. 92* 92, 252–256.

Ben-David, M., Huang, H., Sun, M. G. F., Corbi-Verge, C., Petsalaki, E., Liu, K., et al. (2019). Allosteric Modulation of Binding Specificity by Alternative Packing of Protein Cores. *J. Mol. Biol.* 431, 336–350. doi: 10.1016/j.jmb.2018.11.018.

Berendsen, H. J. C., van der Spoel, D., and van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* 91, 43–56. doi: 10.1016/0010-4655(95)00042-E.

Blaber, M. (2020). Conserved buried water molecules enable the β-trefoil architecture. *Protein Sci.* 29, 1794–1802. doi: 10.1002/pro.3899.

Blaber, M. (2021). Cooperative hydrophobic interactions in the β-trefoil architecture. *Protein Sci.* 30, 956--965. doi: 10.1002/pro.4059.

Blaber, M. (2022). Variable and Conserved Regions of Secondary Structure in the β-Trefoil

Fold: Structure Versus Function. *Front. Mol. Biosci.* 9, 1–11. doi: 10.3389/fmolb.2022.889943.

Bonomi, M., Branduardi, D., Bussi, G., Camilloni, C., Provasi, D., Raiteri, P., et al. (2009). PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* 180, 1961–1972. doi: https://doi.org/10.1016/j.cpc.2009.05.011.

Borgo, B., and Havranek, J. J. (2012). Automated selection of stabilizing mutations in designed and natural proteins. *Proc. Natl. Acad. Sci. U. S. A.* 109, 1494–1499. doi: 10.1073/pnas.1115172109.

Bornscheuer, U. T., Bessler, C., Srinivas, R., and Hari Krishna, S. (2002). Optimizing lipases and related enzymes for efficient application. *Trends Biotechnol.* 20, 433–437. doi: 10.1016/S0167-7799(02)02046-2.

Bornscheuer, U. T., Huisman, G. W., Kazlauskas, R. J., Lutz, S., Moore, J. C., and Robins, K. (2012). Engineering the third wave of biocatalysis. *Nature* 485, 185–194. doi: 10.1038/nature11117.

Brissos, V., Gonçalves, N., Melo, E. P., and Martins, L. O. (2014). Improving kinetic or thermodynamic stability of an azoreductase by directed evolution. *PLoS One* 9. doi: 10.1371/journal.pone.0087209.

Broom, A., Doxey, A. C., Lobsanov, Y. D., Berthin, L. G., Rose, D. R., Howell, P. L., et al. (2012). Modular evolution and the origins of symmetry: Reconstruction of a three-fold symmetric globular protein. *Structure* 20, 161–171. doi: 10.1016/j.str.2011.10.021.

Broom, A., Gosavi, S., and Meiering, E. M. (2015a). Protein unfolding rates correlate as strongly as folding rates with native structure. *Proc. Natl. Acad. Sci. U.S.A.* 24, 580–587. doi: 10.1002/pro.2606.

Broom, A., Jacobi, Z., Trainor, K., and Meiering, E. M. (2017). Computational tools help improve protein stability but with a solubility tradeoff. *J. Biol. Chem.* 292, 14349–14361. doi: 10.1074/jbc.M117.784165.

Broom, A., Ma, S. M., Xia, K., Rafalia, H., Trainor, K., Colón, W., et al. (2015b). Designed protein reveals structural determinants of extreme kinetic stability. *Proc. Natl. Acad. Sci. U. S. A.* 112, 14605–14610. doi: 10.1073/pnas.1510748112.

Broom, A., Trainor, K., Jacobi, Z., and Meiering, E. M. (2020). Computational modelling of protein stability : quantitative analysis reveals solutions to pervasive problems. *Structure* 28, 1–13. doi: 10.1016/j.str.2020.04.003.

Broom, A., Trainor, K., Mackenzie, D. W. S., and Meiering, E. M. (2016). ScienceDirect Using natural sequences and modularity to design common and novel protein topologies. *Curr. Opin. Struct. Biol.* 38, 26–36. doi: 10.1016/j.sbi.2016.05.007.

Brych, S. R., Dubey, V. K., Bienkiewicz, E., Lee, J., Logan, T. M., and Blaber, M. (2004). Symmetric primary and tertiary structure mutations within a symmetric superfold: A solution, not a constraint, to achieve a foldable polypeptide. *J. Mol. Biol.* 344, 769–780. doi: 10.1016/j.jmb.2004.09.060.

Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins Struct. Funct. Bioinforma.* doi: 10.1002/prot.340210302.

Chavez, L. L., Onuchic, J. N., and Clementi, C. (2004). Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates. *J. Am. Chem. Soc.* 126, 8426–8432. doi: 10.1021/ja049510+.

Chen, A., Li, Y., Nie, J., McNeil, B., Jeffrey, L., Yang, Y., et al. (2015). Protein engineering of Bacillus acidopullulyticus pullulanase for enhanced thermostability using in silico data driven rational design methods. *Enzyme Microb. Technol.* 78, 74–83. doi: 10.1016/j.enzmictec.2015.06.013.

Chen, C. R., and Makhatadze, G. I. (2015). ProteinVolume: Calculating molecular van der Waals and void volumes in proteins. *BMC Bioinformatics* 16, 1–6. doi: 10.1186/s12859-015-0531-2.

Chen, K., and Arnold, F. H. (1993). Tuning the activity of an enzyme for unusual environments: Sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl. Acad. Sci. U. S. A.* 90, 5618–5622. doi: 10.1073/pnas.90.12.5618.

Chivian, D., Kim, D. E., Malstrom, L., Bradley, P., Robertson, T., Murphy, P., et al. (2003). Automated Prediction of CASP-5 Structures Using the Robetta Server. *Proteins Struct. Funct. Genet.* 53, 524–533.

Choi, J. M., Han, S. S., and Kim, H. S. (2015). Industrial applications of enzyme biocatalysis: Current status and future aspects. *Biotechnol. Adv.* 33, 1443–1454. doi: 10.1016/j.biotechadv.2015.02.014.

Clarke, J., and Fersht, A. R. (1993). Engineered disulfide bonds as probes of the folding pathway of barnase: Increasing the stability of proteins against the rate of denaturation. *Biochemistry* 32, 4322–4329. doi: 10.1021/bi00067a022.

Clementi, C., Nymeyer, H., and Onuchic, J. N. (2000). Topological and energetic factors: what determines the structural details of the transition state ensemble and “en-route” intermediates for protein folding? an investigation for small globular proteins. *J. Mol. Biol.* 298, 937–953. doi: 10.1006/JMBI.2000.3693.

Colón, W., Church, J., Sen, J., Thibeault, J., Trasatti, H., and Xia, K. (2017). Biological Roles of Protein Kinetic Stability. *Biochemistry* 56, 6179–6186. doi: 10.1021/acs.biochem.7b00942.

Conchillo-Solé, O., de Groot, N. S., Avilés, F. X., Vendrell, J., Daura, X., and Ventura, S. (2007). AGGRESCAN: A server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. *BMC Bioinformatics* 8. doi: 10.1186/1471-2105-8-65.

Creighton, T. E. (1992). *Proteins: Structures and Molecular Properties*. 2nd ed. W.H. Freeman and Company.

Dalessio, P. M., Boyer, J. A., McGettigan, J. L., and Ropson, I. J. (2005). Swapping core residues in homologous proteins swaps folding mechanism. *Biochemistry* 44, 3082–3090. doi: 10.1021/bi048125u.

den Burg, B., De Kreij, A., der Veek, P., Mansfeld, J., Venema, G., Van den Burg, B., et al. (1999). Characterization of a novel stable biocatalyst obtained by protein engineering. *Biotechnol. Appl. Biochem.* 30, 35–40. doi: 10.1111/j.1470-8744.1999.tb01156.x.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry* 29, 7133–7155. doi: 10.1021/bi00483a001.

Dill, K. A., and Maccallum, J. L. (2012). The Protein-Folding Problem , 50 Years On. *Science (80-.).* 338, 1042–1047.

Doyle, C. M., Rumfeldt, J. A., Broom, H. R., Sekhar, A., Kay, L. E., and Meiering, E. M. (2016). Concurrent Increases and Decreases in Local Stability and Conformational Heterogeneity in Cu, Zn Superoxide Dismutase Variants Revealed by Temperature-Dependence of Amide Chemical Shifts. *Biochemistry* 55, 1346–1361. doi: 10.1021/acs.biochem.5b01133.

Duan, X., Cheng, S., Ai, Y., and Wu, J. (2016). Enhancing the thermostability of Serratia plymuthica sucrose isomerase using B-factor-directed mutagenesis. *PLoS One* 11, 1–16. doi: 10.1371/journal.pone.0149208.

Dubey, V. K., Lee, J., and Blaber, M. (2005). Redesigning symmetry-related “mini-core” regions of FGF-1 to increase primary structure symmetry: Thermodynamic and functional consequences of structural symmetry. *Protein Sci.* 14, 2315–2323. doi: 10.1110/ps.051494405.

Eddy, S. R. (2004). What is a hidden Markov model? *Nat. Biotechnol.* 22, 1315–1316.

Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010). Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* 66, 486–501. doi: 10.1107/S0907444910007493.

Englander, S. W., and Mayne, L. (2014). The nature of protein folding pathways. *Proc. Natl. Acad. Sci. U. S. A.* 111, 15873–15880. doi: 10.1073/pnas.1411798111.

Engqvist, M. K. M., and Rabe, K. S. (2019). Applications of protein engineering and directed evolution in plant research. *Plant Physiol.* 179, 907–917. doi: 10.1104/pp.18.01534.

Estell, D. A., Graycar, T. P., and Wells, J. A. (1985). Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. *J. Biol. Chem.* 260, 6518–6521.

Eyal, E., Najmanovich, R., Mcconkey, B. J., Edelman, M., and Sobolev, V. (2004). Importance of Solvent Accessibility and Contact Surfaces in Modeling Side-Chain Conformations in Proteins. *J. Comput. Chem.* 25, 712–724. doi: 10.1002/jcc.10420.

Feng, X., Tang, H., Han, B., Lv, B., and Li, C. (2016). Enhancing the Thermostability of β-Glucuronidase by Rationally Redesigning the Catalytic Domain Based on Sequence Alignment Strategy. *Ind. Eng. Chem. Res.* 55, 5474–5483. doi: 10.1021/acs.iecr.6b00535.

Fersht, A. (1999). *Structure and Mechanism in Protein Science: A guide to Enzyme Catalysis and Protein Folding*. 2nd ed. W.H. Freeman.

Gaines, J. C., Virrueta, A., Buch, D. A., Fleishman, S. J., O’Hern, C. S., and Regan, L. (2017). Collective repacking reveals that the structures of protein cores are uniquely specified by steric repulsive interactions. *Protein Eng. Des. Sel.* 30, 387–394. doi: 10.1093/protein/gzx011.

Gallicchio, E., Andrec, M., Felts, A. K., and Levy, R. M. (2005). Temperature weighted histogram analysis method, replica exchange, and transition paths. *J. Phys. Chem. B* 109, 6722–6731. doi: 10.1021/jp045294f.

Geiger-Schuller, K., Sforza, K., Yuhas, M., Parmeggiani, F., Baker, D., and Barrick, D. (2018). Extreme stability in de novo-designed repeat arrays is determined by unusually stable short-range interactions. *Proc. Natl. Acad. Sci. U. S. A.* 115, 7539–7544. doi: 10.1073/pnas.1800283115.

Giri Rao, V. V. H., and Gosavi, S. (2018). On the folding of a structurally complex protein to its metastable active state. *Proc. Natl. Acad. Sci. U. S. A.* 115, 1998–2003. doi: 10.1073/pnas.1708173115.

Goldenzweig, A., Goldsmith, M., Hill, S. E., Gertman, O., Laurino, P., Ashani, Y., et al. (2016). Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* 63, 337–346. doi: 10.1016/j.molcel.2016.06.012.

Gosavi, S. (2013). Understanding the Folding-Function Tradeoff in Proteins. *PLoS One* 8. doi: 10.1371/journal.pone.0061222.

Gosavi, S., Chavez, L. L., and Jennings, P. A. (2006). Topological Frustration and the Folding of Interleukin-1 b. *J. Mol. Biol* 357, 986–996. doi: 10.1016/j.jmb.2005.11.074.

Gosavi, S., Whitford, P. C., and Jennings, P. A. (2008). Extracting function from a β-trefoil folding motif. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10384–10389.

Gromiha, M. M., and Selvaraj, S. (2001). Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: Application of long-range order to folding rate prediction. *J. Mol. Biol.* 310, 27–32. doi: 10.1006/jmbi.2001.4775.

Guerois, R., Nielsen, J. E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J. Mol. Biol.* 320, 369–387. doi: 10.1016/S0022-2836(02)00442-4.

Gurevich, V. V., and Gurevich, E. V. (2014). Therapeutic Potential of Small Molecules and Engineered Proteins. *Handb. Exp. Pharmacol.* 219, 1–11. doi: 10.1007/978-3-642-41199-1.

Headd, J. J., Echols, N., Afonine, P. ., Grosse-Kunstleve, R. W., Chen, V. B., Moriarty, N. W., et al. (2012). Use of knowledge-based restraints in phenix.refine to improve macromolecular refinement at low resolution. *Acta Cryst.* D68, 381–390.

Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* 4, 435–447. doi: 10.1021/ct700301q.

Hills, R. D., and Brooks, C. L. (2009). Insights from coarse-grained go models for protein folding and dynamics. *Int. J. Mol. Sci.* 10, 889–905. doi: 10.3390/ijms10030889.

Hyeon, C., and Thirumalai, D. (2011). Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nat. Commun.* 2, 1–11. doi: 10.1038/ncomms1481.

Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D., and Finkelstein, A. V. (2003). Contact order revisited: Influence of protein size on the folding rate. *Protein Sci.* 12, 2057–2062. doi: 10.1110/ps.0302503.

Jones, B. J., Lim, H. Y., Huang, J., and Kazlauskas, R. J. (2017). Comparison of Five Protein Engineering Strategies for Stabilizing an α/β-Hydrolase. *Biochemistry* 56, 6521–6532. doi: 10.1021/acs.biochem.7b00571.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79, 926–935. doi: 10.1063/1.445869.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi: 10.1038/s41586-O21-O3819-2.

Kanehisa, M. I., and Tsong, T. Y. (1978). Mechanisms of the multiphasic kinetics in the folding and unfolding of globular proteins. *J. Mol. Biol.* 124, 177–194. doi: https://doi.org/10.1016/0022-2836(78)90155-9.

Karplus, M., and Weaver, D. L. (1994). Protein folding dynamics: The diffusion-collision model and experimental data. *Protein Sci.* 3, 650–668. doi: 10.1002/pro.5560030413.

Kästner, J. (2011). Umbrella sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 1, 932–942. doi: 10.1002/wcms.66.

Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H., and Meiler, J. (2010). Practically Useful : What the ROSETTA Protein Modeling Suite Can Do for You. *Biochemistry* 49, 2987–2998. doi: 10.1021/bi902153g.

Khersonsky, O., Lipsh, R., Avizemer, Z., Ashani, Y., Goldsmith, M., Leader, H., et al. (2018). Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Mol. Cell* 72, 178-186.e5. doi: 10.1016/j.molcel.2018.08.033.

Khoury, G. A., Smadbeck, J., Kieslich, C. A., and Floudas, C. A. (2014). Protein folding and de novo protein design for biotechnological applications. *Trends Biotechnol.* 32, 99–109. doi: 10.1016/j.tibtech.2013.10.008.

Kim, P. S., and Baldwin, R. L. (1982). Specific Intermediates in the Folding Reactions of Small Proteins and the Mechanism of Protein Folding. *Annu. Rev. Biochem.* 51, 459–489. doi: 10.1146/annurev.bi.51.070182.002331.

Kim, T., Joo, J. C., and Yoo, Y. J. (2012). Hydrophobic interaction network analysis for thermostabilization of a mesophilic xylanase. *J. Biotechnol.* 161, 49–59. doi: 10.1016/j.jbiotec.2012.04.015.

Kimura, R., Aumpuchin, P., Hamaue, S., Shimomura, T., and Kikuchi, T. (2020). Analyses of the folding sites of irregular β-trefoil fold proteins through sequence-based techniques and

Gō-model simulations. *BMC Mol. Cell Biol.* 21, 1–17. doi: 10.1186/s12860-020-00271-4.

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., and Kolinski, A. (2016). Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* 116, 7898–7936. doi: 10.1021/acs.chemrev.6b00163.

Koga, R., Yamamoto, M., Kosugi, T., Kobayashi, N., Sugiki, T., Fujiwara, T., et al. (2020). Robust folding of a de novo designed ideal protein even with most of the core mutated to valine. *Proc. Natl. Acad. Sci. U. S. A.* 117, 31149–31156. doi: 10.1073/pnas.2002120117.

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., and Lesk, A. . (2006). MUSTANG: A Multiple Structural Alignment Algorithm. *Proteins Struct. Funct. Genet.* 64, 599–574. doi: 10.1002/prot.20921.

Kramers, H. A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* 7, 284–304. doi: 10.1016/S0031-8914(40)90098-2.

Krivov, G. ., Shapovalov, M. ., and Dunbrack, L. R. J. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Mol. Cell. Biochem.* 77, 778–795. doi: 10.1002/prot.22488.

Kuhlman, B. (2019). Designing protein structures and complexes with the molecular modeling program Rosetta. *J. Biol. Chem.* 294, 19436–19443. doi: 10.1074/jbc.AW119.008144.

Kuhlman, B., and Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10383–10388. doi: 10.1073/pnas.97.19.10383.

Kumar, R., Goomber, S., and Kaur, J. (2019). Engineering lipases for temperature adaptation: Structure function correlation. *Biochim. Biophys. Acta - Proteins Proteomics* 1867, 140261. doi: https://doi.org/10.1016/j.bbapap.2019.08.001.

Laimer, J., Hofer, H., Fritz, M., Wegenkittl, S., and Lackner, P. (2015). MAESTRO - multi agent stability prediction upon point mutations. *BMC Bioinformatics* 16, 1–13. doi: 10.1186/s12859-015-0548-6.

Lalwani Prakash, D., and Gosavi, S. (2021). Understanding the Folding Mediated Assembly of the Bacteriophage MS2 Coat Protein Dimers. *J. Phys. Chem. B* 125, 8722–8732. doi: 10.1021/acs.jpcb.1c03928.

Lapidus, L. J., Yao, S., McGarrity, K. S., Hertzog, D. E., Tubman, E., and Bakajiny, O. (2007). Protein hydrophobic collapse and early folding steps observed in a microfluidic mixer. *Biophys. J.* 93, 218–224. doi: 10.1529/biophysj.106.103077.

Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30, 3276–3278. doi: 10.1093/bioinformatics/btu531.

Le, Q. A. T., Joo, J. C., Yoo, Y. J., and Kim, Y. H. (2012). Development of thermostable Candida antarctica lipase B through novel in silico design of disulfide bridge. *Biotechnol. Bioeng.* 109, 867–876. doi: 10.1002/bit.24371.

Lee, J., and Blaber, M. (2011). Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. *Proc. Natl. Acad. Sci. U. S. A.* 108, 126–130. doi:

10.1073/pnas.1015032108.

Lee, J., Blaber, S. I., Dubey, V. K., and Blaber, M. (2011). A polypeptide “building block” for the β-trefoil fold identified by “top-down symmetric deconstruction.” *J. Mol. Biol.* 407, 744–763. doi: 10.1016/j.jmb.2011.02.002.

Levinthal, C. (1969). How to fold graciously. *Mössbauer Spectrosc. Biol. Syst. Proc.* 24, 22–24. doi: citeulike-article-id:380320.

Liang, S., Zheng, D., Zhang, C., and Standley, D. M. (2011). Fast and accurate prediction of protein side-chain conformations. *Bioinformatics* 27, 2913–2914. doi: 10.1093/bioinformatics/btr482.

Lindahl, E., Hess, B., and van der Spoel, D. (2001). GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Mol. Model. Annu.* 7, 306–317. doi: 10.1007/s008940100045.

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., et al. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinforma.* 78, 1950–1958. doi: 10.1002/prot.22711.

Liu, C., Chu, D., Wideman, R. D., Houliston, R. S., Wong, H. J., and Meiering, E. M. (2001). Thermodynamics of denaturation of hisactophilin, a β-trefoil protein. *Biochemistry* 40, 3817–3827. doi: 10.1021/bi002609i.

Liu, Q., Xun, G., and Feng, Y. (2019). The state-of-the-art strategies of protein engineering for enzyme stabilization. *Biotechnol. Adv.* 37, 530–537. doi: 10.1016/j.biotechadv.2018.10.011.

Liu, Z., Liang, Q., Wang, P., Kong, Q., Fu, X., and Mou, H. (2021). Improving the kinetic stability of a hyperthermostable β-mannanase by a rationally combined strategy. *Int. J. Biol. Macromol.* 167, 405–414. doi: 10.1016/j.ijbiomac.2020.11.202.

Longo, L., Lee, J., and Blaber, M. (2012). Experimental support for the foldability-function tradeoff hypothesis: Segregation of the folding nucleus and functional regions in fibroblast growth factor-1. *Protein Sci.* 21, 1911–1920. doi: 10.1002/pro.2175.

Longo, L. M., and Blaber, M. (2013). Prebiotic protein design supports a halophile origin of foldable proteins. *Front. Microbiol.* 4, 2013–2015. doi: 10.3389/fmicb.2013.00418.

Longo, L. M., Kumru, O. S., Middaugh, C. R., and Blaber, M. (2014). Evolution and design of protein structure by folding nucleus symmetric expansion. *Structure* 22, 1377–1384. doi: 10.1016/j.str.2014.08.008.

Luo, P., Hayes, R. J., Chan, C., Stark, D. M., Hwang, M. Y., Jacinto, J. M., et al. (2002). Development of a cytokine analog with enhanced stability using computational ultrahigh throughput screening. *Protein Sci.* 11, 1218–1226. doi: 10.1110/ps.4580102.

Lutz, S., and Iamurri, S. M. (2018). “Protein Engineering: Past, Present, and Future,” in *Protein Engineering: Methods and Protocols,* eds. U. T. Bornscheuer and M. Höhne (New York, NY: Springer New York), 1–12. doi: 10.1007/978-1-4939-7366-8_1.

MacKenzie, D. W. S., Schaefer, A., Steckner, J., Leo, C. A., Naser, D., Artikis, E., et al. (2022).

A fine balance of hydrophobic-electrostatic communication pathways in a pH-switching protein. *Proc. Natl. Acad. Sci.* 119, e2119686119. doi: 10.1073/pnas.2119686119/-/DCSupplemental.60.

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., et al. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* 102, 3586–3616. doi: 10.1021/jp973084f.

Magliery, T. J. (2015). Protein stability: Computation, sequence statistics, and new experimental methods. *Curr. Opin. Struct. Biol.* 33, 161–168. doi: 10.1016/j.sbi.2015.09.002.

Mansfeld, J., Vriend, G., Dijkstra, B. W., Veltman, O. R., Van Den Burg, B., Venema, G., et al. (1997). Extreme stabilization of a thermolysin-like protease by an engineered disulfide bond. *J. Biol. Chem.* 272, 11152–11156. doi: 10.1074/jbc.272.17.11152.

Mao, C., Xie, H., Chen, S., Valverde, B. E., and Qiang, S. (2017). Error-prone PCR mutation of Ls-EPSPS gene from Liriope spicata conferring to its enhanced glyphosate-resistance. *Pestic. Biochem. Physiol.* 141, 90–95. doi: https://doi.org/10.1016/j.pestbp.2016.12.004.

Martínez, L. (2014). Introducing the Levinthal's protein folding paradox and its solution. *J. Chem. Educ.* 91, 1918–1923. doi: 10.1021/ed300302h.

McLendon G, R. E. (1978). Is protein turnover thermodynaimcally controlled? *J. Chem. Inf. Model.* 253, 6335–6337. doi: 10.1017/CBO9781107415324.004.

Mehrotra, P. (2016). Biosensors and their applications - A review. *J. Oral Biol. Craniofacial Res.* 6, 153–159. doi: 10.1016/j.jobcr.2015.12.002.

Meiering, E. M., Bycroft, M., and Fersht, A. R. (1991). Characterization of Phosphate Binding in the Active Site of Barnase by Site-Directed Mutagenesis and NMR. *Biochemistry* 30, 11348–11356. doi: 10.1021/bi00111a022.

Miao, Z., Cao, Y., and Jiang, T. (2011). RASP: Rapid modeling of protein side chain conformations. *Bioinformatics* 27, 3117–3122. doi: 10.1093/bioinformatics/btr538.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nat. Methods* 19, 679–682. doi: 10.1038/s41592-022-01488-1.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913.

Muhammed, M. T., and Aki-Yalcin, E. (2019). Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chem. Biol. Drug Des.* 93, 12–20. doi: 10.1111/cbdd.13388.

Murphy, G. S., Mills, J. L., Miley, M. J., Machius, M., Szyperski, T., and Kuhlman, B. (2012). Increasing sequence diversity with flexible backbone protein design: The complete redesign of a protein hydrophobic core. *Structure* 20, 1086–1096. doi: 10.1016/j.str.2012.03.026.

Murzin, A. G., Lesk, M., and Chothiap, C. (1992). β-Trefoil Fold: Patterns of Structure and

Sequence in the Kunitz Inhibitors Interleukin-1β and 1α and Fibroblast Growth Factors. *J. Mol. Biol* 223, 531–543.

Ng, S. P., Billings, K. S., Ohashi, T., Allen, M. D., Best, R. B., Randles, L. G., et al. (2007). Designing an extracellular matrix protein with enhanced mechanical stability. *Proc. Natl. Acad. Sci. U. S. A.* 104, 9633–9637. doi: 10.1073/pnas.0609901104.

Nisthal, A., Wang, C. Y., Ary, M. L., and Mayo, S. L. (2019). Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* 116, 16367–16377. doi: 10.1073/pnas.1903888116.

Noel, J. K., Levi, M., Raghunathan, M., Lammert, H., Hayes, R. L., Onuchic, J. N., et al. (2016). SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLoS Comput. Biol.* 12, 1–14. doi: 10.1371/journal.pcbi.1004794.

Noel, J. K., Whitford, P. C., and Onuchic, J. N. (2012). The shadow map: A general contact definition for capturing the dynamics of biomolecular folding and function. *J. Phys. Chem. B* 116, 8692–8702. doi: 10.1021/jp300852d.

Noel, J., and Onuchic, J. (2012). *The Many Faces of Structure-Based Potentials: From Protein Folding Landscapes to Structural Characterization of Complex Biomolecules*. Springer, Boston, MA doi: 10.1007/978-1-4614-2146-7.

Nymeyer, H., García, A. E., and Onuchic, J. N. (1998). Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5921–5928. doi: 10.1073/pnas.95.11.5921.

Olsen, S. K., Ibrahimi, O. A., Raucci, A., Zhang, F., Eliseenkova, A. V., Yayon, A., et al. (2004). Insights into the molecular basis for fibroblast growth factor receptor autoinhibition and ligand-binding promiscuity. *Proc. Natl. Acad. Sci. U. S. A.* 101, 935–940. doi: 10.1073/pnas.0307287101.

Onuchic, J. N., Luthey-Schulten, Z., and Wolynes, P. G. (1997). THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* doi: 10.1146/annurev.physchem.48.1.545.

Onuchic, J. N., and Wolynes, P. G. (2004). Theory of protein folding. *Curr. Opin. Struct. Biol.* 14, 70–75. doi: 10.1016/j.sbi.2004.01.009.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* 2014, 1–21. doi: 10.7554/eLife.02030.

Pace, C. N. (1986). Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol.* 131, 266–280. doi: 10.1016/0076-6879(86)31045-0.

Perkins, S. J. (1986). Protein volumes and hydration effects. *Eur. J. Biochem.* 157, 169–180.

Peterson, L. X., Kang, X., and Kihara, D. (2014). Assessment of Protein Side-Chain Conformation Prediction Methods in Different Residue Environments. *Proteins* 82, 1971–1984. doi: 10.1002/prot.24552.

Pey, A. L., Rodriguez-Larrea, D., Bomke, S., Dammers, S., Godoy-Ruiz, R., Garcia-Mira, M. M., et al. (2008). Engineering proteins with tunable thermodynamic and kinetic stabilities. *Proteins Struct. Funct. Genet.* 71, 165–174. doi: 10.1002/prot.21670.

Pikkemaat, M. G., Linssen, A. B. M., Berendsen, H. J. C., and Janssen, D. B. (2002). Molecular dynamics simulations as a tool for improving protein stability. *Protein Eng.* 15, 185–192.

Ponting, C. P., and Russell, R. B. (2000). Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all β-trefoil proteins. *J. Mol. Biol.* 302, 1041–1047. doi: 10.1006/jmbi.2000.4087.

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res.* 46, W200–W204. doi: 10.1093/nar/gky448.

Potterton, E., Briggs, P., Turkenburg, M., and Dodson, E. (2003). A graphical user interface to the CCP4 program suite. *Acta Cryst.* D59, 1131–1137.

Ptitsyn, O. B. (1987). Protein folding: Hypotheses and experiments. *J. Protein Chem.* 6, 273–293. doi: 10.1007/BF00248050.

Ptitsyn, O. B., and Rashin, A. A. (1975). A model of myoglobin self-organization. *Biophys. Chem.* 3, 1–20. doi: https://doi.org/10.1016/0301-4622(75)80033-0.

Quezada, A. G., Cabrera, N., Piñeiro, Á., Díaz-Salazar, A. J., Díaz-Mazariegos, S., Romero-Romero, S., et al. (2018). A strategy based on thermal flexibility to design triosephosphate isomerase proteins with increased or decreased kinetic stability. *Biochem. Biophys. Res. Commun.* 503, 3017–3022. doi: 10.1016/j.bbrc.2018.08.087.

Raugei, S., Gervasio, F. L., and Carloni, P. (2006). Oct 2006. *Phys. Status Solidi BBasic Solid State Phys.* 243, 2500–2515. doi: 10.1007/978-1-4614-2146-7.

Reynolds, K. A., Russ, W. P., Socolich, M., and Ranganathan, R. (2013). “Chapter Ten - Evolution-Based Design of Proteins,” in *Methods in Protein Design*, ed. A. E. B. T.-M. in E. Keating (Academic Press), 213–235. doi: https://doi.org/10.1016/B978-0-12-394292-0.00010-2.

Rhodes, G. (1993). *Crystallography made crystal clear: a guide for users of macromolecular models, 2nd edition*. 2nd ed. California, USA: Academic Press.

Rocklin, G. J., Chidyausiku, T. M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., et al. (2017). Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science (80-.).* 357, 168–175. doi: 10.1126/science.aan0693.

Rohl, C. A., Strauss, C. E. M., Chivian, D., and Baker, D. (2004). Modeling Structurally Variable Regions in Homologous Proteins with Rosetta. *Proteins Struct. Funct. Genet.* 55, 656–677. doi: 10.1002/prot.10629.

Sabri Dashti, D., and Roitberg, A. E. (2013). Optimization of umbrella sampling replica exchange molecular dynamics by replica positioning. *J. Chem. Theory Comput.* 9, 4692–4699. doi: 10.1021/ct400366h.

Sanchez-Ruiz, J. M. (2010). Protein kinetic stability. *Biophys. Chem.* 148, 1–15. doi:

10.1016/j.bpc.2010.02.004.

Sancho, J., Meiering, E. M., and Fersht, A. R. (1991). Mapping transition states of protein unfolding by protein engineering of ligand-binding sites. *J. Mol. Biol.* 221, 1007–1014. doi: 10.1016/0022-2836(91)80188-Z.

Shao, E., Lin, L., Chen, C., Chen, H., Zhuang, H., Wu, S., et al. (2016). Loop replacements with gut-binding peptides in Cry1Ab domain II enhanced toxicity against the brown planthopper, Nilaparvata lugens (Stål). *Sci. Rep.* 6, 1–9. doi: 10.1038/srep20106.

Shental-Bechor, D., Smith, M. T. J., MacKenzie, D., Broom, A., Marcovitz, A., Ghashut, F., et al. (2012). Nonnative interactions regulate folding and switching of myristoylated protein. *Proc. Natl. Acad. Sci. U. S. A.* 109, 17839–17844. doi: 10.1073/pnas.1201803109.

Sindhu, R., Binod, P., Madhavan, A., Beevi, U. S., Mathew, A. K., Abraham, A., et al. (2017). Molecular improvements in microbial A-amylases for enhanced stability and catalytic efficiency. *Bioresour. Technol.* 245, 1740–1748. doi: 10.1016/j.biortech.2017.04.098.

Singh, R., Kumar, M., Mittal, A., and Mehta, P. K. (2016). Microbial enzymes: industrial progress in 21st century. *3 Biotech* 6. doi: 10.1007/s13205-016-0485-8.

Smith, M. T. J., Meissner, J., Esmonde, S., Wong, H. J., and Meiering, E. M. (2010). Energetics and mechanisms of folding and flipping the myristoyl switch in the β-trefoil protein, hisactophilin. *Proc. Natl. Acad. Sci. U. S. A.* 107, 20952–20957. doi: 10.1073/pnas.1008026107.

Song, Y., Dimaio, F., Wang, R. Y. R., Kim, D., Miles, C., Brunette, T., et al. (2013). High-resolution comparative modeling with RosettaCM. *Structure* 21, 1735–1742. doi: 10.1016/j.str.2013.08.005.

Sormanni, P., Aprile, F. A., and Vendruscolo, M. (2015). The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* 427, 478–490. doi: 10.1016/j.jmb.2014.09.026.

Sternke, M., Tripp, K. W., and Barrick, D. (2019). Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proc. Natl. Acad. Sci. U. S. A.* 166, 11275–11284. doi: 10.1073/pnas.1816707116.

Sun, Z., Liu, Q., Qu, G., Feng, Y., and Reetz, M. T. (2019). Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chem. Rev.* doi: 10.1021/acs.chemrev.8b00290.

Suplatov, D. A., Kopylov, K. E., Popova, N. N., Voevodin, V. V., and Švedas, V. K. (2018). Mustguseal: A server for multiple structure-guided sequence alignment of protein families. *Bioinformatics* 34, 1583–1585. doi: 10.1093/bioinformatics/btx831.

Swint-Kruse, L. (2016). Using Evolution to Guide Protein Engineering: The Devil IS in the Details. *Biophys. J.* 111, 10–18. doi: 10.1016/j.bpj.2016.05.030.

Tenorio, C. A., Parker, J. B., and Blaber, M. (2022). Functionalization of a symmetric protein scaffold: Redundant folding nuclei and alternative oligomeric folding pathways. *Protein Sci.* 31, 1–14. doi: 10.1002/pro.4301.

Terada, D., Voet, A. R. D., Noguchi, H., Kamata, K., Ohki, M., Addy, C., et al. (2017). Computational design of a symmetrical β-trefoil lectin with cancer cell binding activity. *Sci. Rep.* 7, 5943. doi: 10.1038/s41598-017-06332-7.

Thirumalai, D. (1995). From Minimal Models to Real Porteins: Time Scales for Protein Folding Kinetics. *J. Phys. I Fr.* 5, 1457–1467.

Tian, W., Chen, C., Lei, X., Zhao, J., and Liang, J. (2018). CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Res.* 46, W363–W367. doi: 10.1093/nar/gky473.

Tian, Y. S., Xu, J., Peng, R. H., Xiong, A. S., Xu, H., Zhao, W., et al. (2013). Mutation by DNA shuffling of 5-enolpyruvylshikimate-3-phosphate synthase from Malus domestica for improved glyphosate resistance. *Plant Biotechnol. J.* 11, 829–838. doi: 10.1111/pbi.12074.

Turner, P. J. (2005). XMGRACE, Version 5.1.19.

Vagin, A., and Teplyakov, A. (1997). MOLREP: An Automated Program for Molecular Replacement. *J. Appl. Crystallogr.* 30, 1022–1025. doi: 10.1107/S0021889897006766.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005). GROMACS: Fast, flexible, and free. *J. Comput. Chem.* 26, 1701–1718. doi: 10.1002/jcc.20291.

Ventura, S., and Serrano, L. (2004). Designing proteins from the inside out. *Proteins Struct. Funct. Genet.* 56, 1–10. doi: 10.1002/prot.20142.

Vigneshvar, S., Sudhakumari, C. C., Senthilkumaran, B., and Prakash, H. (2016). Recent advances in biosensor technology for potential applications - an overview. *Front. Bioeng. Biotechnol.* 4, 1–9. doi: 10.3389/fbioe.2016.00011.

Vivian, J. T., and Callis, P. R. (2001). Mechanisms of tryptophan fluorescence shifts in proteins. *Biophys. J.* 80, 2093–2109. doi: 10.1016/S0006-3495(01)76183-8.

Vrancken, J. P. M., Tame, J. R. H., and Voet, A. R. D. (2020). Development and applications of artificial symmetrical proteins. *Comput. Struct. Biotechnol. J.* 18, 3959–3968. doi: 10.1016/j.csbj.2020.10.040.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033.

Wensley, B. G., Batey, S., Bone, F. A. C., Chan, Z. M., Tumelty, N. R., Steward, A., et al. (2010). Experimental evidence for a frustrated energy landscape in a three-helix-bundle protein family. *Nature* 463, 685–688. doi: 10.1038/nature08743.

Wetlaufer, D. B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. U. S. A.* 70, 697–701. doi: 10.1073/pnas.70.3.697.

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., et al. (2018). MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* 27, 293–315. doi: 10.1002/pro.3330.

Wolynes, P. G., Onuchic, J. N., and Thirumalai, D. (1995). Navigating the folding routes. *Science (80-.).* 267, 1619 LP – 1620. doi: 10.1126/science.7886447.

Wong, H. J., Stathopulos, P. B., Bonner, J. M., Sawyer, M., and Meiering, E. M. (2004). Non-linear effects of temperature and urea on the thermodynamics and kinetics of folding and unfolding of hisactophilin. *J. Mol. Biol.* 344, 1089–1107. doi: 10.1016/j.jmb.2004.09.091.

Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J., and Arnold, F. H. (2019). Machine learning-assisted directed protein evolution with combinatorial libraries. 116. doi: 10.1073/pnas.1901979116.

Xie, Y., An, J., Yang, G., Wu, G., Zhang, Y., Cui, L., et al. (2014). Enhanced enzyme kinetic stability by increasing rigidity within the active site. *J. Biol. Chem.* 289, 7994–8006. doi: 10.1074/jbc.M113.536045.

Yadahalli, S., and Gosavi, S. (2016). Functionally Relevant Specific Packing Can Determine Protein Folding Routes. *J. Mol. Biol.* 428, 509–521. doi: 10.1016/j.jmb.2015.12.014.

Yadahalli, S., Hemanth Giri Rao, V. V., and Gosavi, S. (2014). Modeling Non-Native Interactions in Designed Proteins. *Isr. J. Chem.* 54, 1230–1240. doi: 10.1002/ijch.201400035.

Yue, K., and Dill, K. A. (1992). Inverse protein folding problem: designing polymer sequences. *Proc. Natl. Acad. Sci. U. S. A.* 89, 4163–4167. doi: 10.1073/pnas.89.9.4163.

Appendix A: Contact lists

wtHis

(2, 34), (2, 35), (2, 36), (2, 117), (3, 33), (3, 34), (3, 35), (3, 116), (3, 117), (3, 118), (4, 13), (4, 14), (4, 15), (4, 30), (4, 31), (4, 32), (4, 33), (4, 34), (4, 116), (4, 117), (4, 118), (5, 13), (5, 14), (5, 15), (5, 34), (5, 114), (5, 115), (5, 116), (5, 117), (6, 12), (6, 13), (6, 14), (6, 36), (6, 45), (6, 76), (6, 83), (6, 85), (6, 113), (6, 114), (6, 115), (6, 116), (7, 12), (7, 13), (7, 112), (7, 113), (7, 114), (7, 116), (8, 12), (8, 13), (8, 14), (8, 23), (8, 95), (8, 97), (8, 112), (8, 113), (9, 97), (9, 112), (9, 113), (9, 114), (10, 25), (10, 114), (11, 24), (11, 25), (11, 97), (12, 23), (12, 24), (12, 25), (12, 95), (12, 97), (13, 22), (13, 23), (13, 24), (13, 25), (13, 26), (13, 31), (13, 116), (14, 22), (14, 23), (14, 34), (14, 53), (14, 85), (14, 93), (14, 94), (14, 95), (14, 101), (14, 113), (15, 20), (15, 21), (15, 22), (15, 23), (15, 24), (15, 31), (15, 32), (15, 33), (15, 34), (16, 20), (16, 21), (16, 22), (16, 24), (16, 32), (16, 33), (16, 34), (16, 47), (17, 21), (17, 22), (17, 24), (17, 28), (17, 30), (17, 32), (18, 32), (21, 34), (21, 53), (21, 63), (22, 34), (23, 101), (24, 28), (24, 30), (24, 31), (24, 32), (25, 31), (26, 31), (31, 116), (33, 46), (33, 47), (33, 118), (34, 45), (34, 46), (34, 47), (34, 53), (34, 63), (35, 44), (35, 45), (35, 46), (35, 47), (36, 43), (36, 44), (36, 45), (36, 76), (37, 42), (37, 43), (37, 44), (37, 46), (38, 42), (38, 43), (39, 43), (39, 44), (39, 52), (39, 66), (39, 67), (39, 68), (39, 73), (40, 68), (41, 75), (42, 69), (42, 70), (42, 73), (42, 74), (42, 75), (43, 73), (43, 74), (43, 75), (43, 76), (44, 52), (44, 67), (44, 72), (44, 73), (44, 74), (45, 51), (45, 52), (45, 53), (45, 72), (45, 73), (45, 74), (45, 76), (45, 85), (46, 50), (46, 51), (46, 52), (47, 51), (49, 65), (50, 64), (50, 65), (51, 63), (51, 64), (52, 62), (52, 63), (52, 64), (52, 65), (52, 66), (52, 67), (52, 72), (52, 73), (53, 61), (53, 62), (53, 63), (53, 64), (53, 65), (53, 71), (53, 72), (53, 74), (53, 85), (53, 93), (54, 60), (54, 61), (54, 62), (54, 63), (54, 64), (54, 65), (54, 71), (54, 72), (54, 74), (55, 60), (55, 61), (55, 62), (55, 71), (55, 74), (55, 87), (55, 88), (55, 91), (55, 93), (55, 102), (55, 103), (56, 60), (56, 71), (57, 103), (58, 102), (58, 103), (59, 100), (59, 101), (59, 102), (59, 103), (60, 103), (61, 93), (61, 101), (61, 102), (61, 103), (62, 71), (64, 72), (65, 72), (67, 72), (67, 73), (68, 73), (69, 73), (70, 74), (70, 75), (70, 86), (70, 87), (70, 88), (71, 86), (71, 87), (71, 88), (71, 89), (71, 91), (73, 86), (73, 87), (74, 85), (74, 86), (74, 87), (74, 91), (74, 93), (75, 84), (75, 85), (75, 86), (76, 83), (76, 84), (76, 85), (76, 86), (77, 82), (77, 83), (77, 84), (78, 82), (78, 83), (78, 84), (78, 115), (79, 83), (79, 84), (79, 107), (79, 108), (79, 112), (81, 113), (81, 114), (81, 115), (82, 108), (82, 109), (82, 112), (82, 113), (82, 114), (82, 115), (83, 111), (83, 112), (83, 113), (83, 114), (83, 115), (84, 92), (84, 107), (84, 111), (84, 112), (84, 113), (85, 91), (85, 92), (85, 93), (85, 111), (85, 113), (86, 90), (86, 91), (86, 92), (87, 91), (89, 105), (90, 104), (90, 105), (91, 103), (91, 104), (91, 105), (92, 102), (92, 103), (92, 104), (92, 105), (92, 106), (92, 107), (92, 111), (93, 101), (93, 102), (93, 103), (93, 110), (93, 111), (93, 113), (94, 100), (94, 101), (94, 102), (94, 103), (94, 104), (94, 110), (94, 111), (95, 100), (95, 101), (95, 109), (95, 110), (95, 111), (95, 113), (96, 100), (96, 102), (96, 104), (96, 110), (101, 113), (103, 111), (104, 110), (104, 111), (106, 110), (106, 111), (107, 111), (107, 112), (108, 112)

Total contacts: 352

csHisH90G

(2, 34), (2, 35), (2, 36), (2, 117), (2, 118), (3, 33), (3, 34), (3, 35), (3, 116), (3, 117), (3, 118), (4, 14), (4, 15), (4, 32), (4, 33), (4, 34), (4, 35), (4, 36), (4, 115), (4, 116), (4, 117), (4, 118), (5, 13), (5, 14), (5, 15), (5, 33), (5, 34), (5, 114), (5, 115), (5, 116), (5, 118), (6, 12), (6, 13), (6, 14), (6, 34), (6, 45), (6, 53), (6, 76), (6, 83), (6, 85), (6, 93), (6, 113), (6, 114), (6, 115), (7, 11), (7, 12), (7, 13), (7, 112), (7, 113), (7, 114), (7, 116), (8, 12), (8, 95), (8, 112), (9, 95), (9, 97), (9, 109), (9, 110), (9, 112), (10, 25), (10, 95), (10, 97), (10, 98), (11, 24), (11, 25), (12, 23), (12, 24), (12, 25), (12, 95), (12, 97), (12, 98), (12, 113), (13, 22), (13, 23), (13, 24), (13, 25), (13, 26), (13, 31), (13, 33), (13, 116), (14, 21), (14, 22), (14, 23), (14, 24), (14, 33), (14, 34), (14, 53), (14, 93), (14, 100), (14, 101), (14, 113), (15, 20), (15, 21), (15, 22), (15, 24), (15, 31), (15, 32), (15, 33), (15, 34), (16, 20), (16, 21), (16, 22), (16, 24), (16, 31), (16, 32), (16, 33), (16, 34), (16, 47), (16, 48), (17, 22), (17, 24), (17, 30), (17, 31), (17, 32), (18, 32), (18, 48), (21, 34), (21, 47), (21, 53), (21, 63), (21, 100), (22, 31), (22, 100), (23, 97), (23, 100), (23, 101), (24, 31), (24, 32), (25, 31), (26, 31), (27, 31), (32, 47), (32, 48), (33, 46), (33, 47), (33, 116), (33, 118), (34, 45), (34, 46), (34, 47), (34, 51), (34, 52), (34, 53), (34, 63), (35, 44), (35, 45), (35, 46), (35, 47), (36, 43), (36, 44), (36, 45), (36, 76), (36, 115), (37, 42), (37, 43), (37, 44), (37, 46), (37, 52), (38, 42), (38, 43), (39, 43), (39, 44), (39, 46), (39, 52), (39, 66), (39, 67), (39, 68), (39, 73), (41, 75), (42, 69), (42, 70), (42, 73), (42, 74), (42, 75), (42, 76), (42, 86), (42, 87), (43, 72), (43, 73), (43, 74), (43, 75), (43, 76), (44, 52), (44, 66), (44, 72), (44, 73), (44, 74), (45, 51), (45, 52), (45, 53), (45, 63), (45, 72), (45, 73), (45, 74), (45, 76), (45, 85), (46, 50), (46, 51), (46, 52), (46, 63), (47, 51), (47, 63), (50, 64), (50, 65), (50, 66), (51, 63), (51, 64), (52, 62), (52, 63), (52, 64), (52, 65), (52, 66), (52, 72), (52, 73), (53, 61), (53, 62), (53, 63), (53, 71), (53, 72), (53, 73), (53, 74), (53, 85), (53, 93), (54, 60), (54, 61), (54, 62), (54, 63), (54, 64), (54, 65), (54, 71), (54, 72), (54, 74), (54, 93), (55, 60), (55, 61), (55, 62), (55, 71), (55, 74), (55, 87), (55, 88), (55, 91), (55, 93), (55, 103), (56, 60), (56, 71), (56, 88), (56, 103), (57, 91), (58, 102), (58, 103), (59, 96), (59, 101), (59, 102), (59, 103), (60, 102), (60, 103), (61, 93), (61, 100), (61, 101), (61, 103), (62, 71), (62, 72), (64, 72), (65, 72), (66, 72), (66, 73), (67, 71), (67, 72), (67, 73), (68, 73), (69, 73), (70, 87), (70, 88), (71, 87), (71, 88), (73, 86), (73, 87), (74, 85), (74, 86), (74, 87), (74, 91), (74, 93), (74, 103), (75, 84), (75, 85), (75, 86), (76, 83), (76, 84), (76, 85), (76, 115), (77, 82), (77, 83), (77, 84), (77, 86), (77, 92), (77, 107), (78, 82), (78, 83), (79, 84), (79, 92), (79, 107), (79, 112), (81, 113), (81, 114), (82, 112), (82, 113), (82, 114), (83, 111), (83, 112), (83, 113), (83, 114), (83, 115), (84, 92), (84, 93), (84, 107), (84, 111), (84, 112), (84, 113), (85, 91), (85, 92), (85, 93), (85, 111), (85, 113), (86, 90), (86, 91), (86, 92), (87, 91), (87, 93), (90, 104), (90, 105), (91, 103), (91, 104), (91, 105), (92, 102), (92, 103), (92, 104), (92, 106), (92, 107), (92, 111), (93, 101), (93, 102), (93, 103), (93, 110), (93, 111), (93, 113), (94, 100), (94, 101), (94, 102), (94, 103), (94, 104), (94, 110), (94, 111), (95, 100), (95, 101), (95, 110), (95, 112), (95, 113), (96, 100), (96, 102), (96, 104), (96, 110), (97, 110), (101, 113), (104, 108), (104, 110), (104, 111), (106, 111), (107, 111), (107, 112), (108, 112)

Total contacts: 370

3Foil

(2, 141), (2, 142), (3, 42), (3, 58), (3, 141), (3, 142), (4, 43), (4, 44), (4, 45), (4, 141), (4, 142), (5, 42), (5, 43), (5, 44), (5, 45), (5, 56), (5, 58), (5, 140), (5, 141), (5, 142), (6, 41), (6, 42), (6, 43), (6, 44), (6, 45), (6, 139), (6, 140), (6, 141), (6, 142), (7, 14), (7, 16), (7, 35), (7, 41), (7, 42), (7, 43), (7, 138), (7, 139), (7, 140), (7, 142), (8, 15), (8, 16), (8, 17), (8, 41), (8, 43), (8, 45), (8, 55), (8, 64), (8, 102), (8, 111), (8, 137), (8, 138), (8, 139), (8, 140), (9, 13), (9, 14), (9, 15), (9, 16), (9, 35), (9, 99), (9, 136), (9, 137), (9, 138), (9, 140), (10, 14), (10, 15), (10, 17), (10, 32), (10, 113), (10, 133), (10, 134), (10, 136), (10, 137), (10, 138), (11, 97), (11, 99), (11, 133), (11, 134), (11, 136), (11, 138), (12, 32), (12, 113), (12, 114), (12, 116), (12, 117), (12, 118), (12, 133), (12, 134), (12, 136), (13, 32), (13, 117), (13, 118), (14, 34), (14, 35), (14, 138), (15, 32), (15, 33), (15, 34), (15, 35), (15, 118), (15, 120), (15, 137), (16, 31), (16, 32), (16, 33), (16, 35), (16, 41), (17, 30), (17, 31), (17, 40), (17, 41), (17, 43), (17, 55), (17, 64), (17, 77), (17, 102), (17, 111), (17, 124), (17, 137), (18, 22), (18, 29), (18, 30), (18, 31), (18, 32), (18, 33), (18, 40), (18, 41), (19, 23), (19, 28), (19, 29), (19, 30), (19, 40), (19, 43), (19, 57), (19, 59), (19, 79), (20, 28), (20, 29), (20, 30), (20, 31), (20, 121), (21, 31), (21, 40), (22, 40), (22, 59), (23, 28), (23, 59), (23, 60), (23, 79), (24, 59), (24, 60), (24, 62), (24, 79), (25, 62), (25, 79), (26, 62), (26, 79), (26, 80), (26, 81), (27, 67), (27, 78), (27, 79), (27, 80), (28, 77), (28, 78), (28, 79), (29, 76), (29, 77), (29, 78), (29, 123), (29, 124), (30, 43), (30, 64), (30, 76), (30, 77), (30, 78), (30, 79), (30, 111), (30, 122), (30, 123), (30, 124), (31, 121), (31, 122), (31, 123), (31, 124), (31, 137), (32, 113), (32, 117), (32, 118), (32, 119), (32, 120), (32, 121), (32, 122), (32, 123), (32, 124), (32, 137), (33, 41), (33, 120), (33, 121), (34, 41), (34, 120), (35, 41), (35, 42), (35, 142), (36, 40), (36, 41), (37, 41), (37, 42), (38, 42), (39, 57), (39, 58), (39, 59), (40, 57), (40, 58), (40, 59), (41, 142), (42, 56), (42, 57), (42, 58), (42, 59), (42, 142), (43, 55), (43, 56), (43, 57), (43, 62), (43, 64), (43, 78), (43, 79), (43, 142), (44, 54), (44, 55), (44, 56), (44, 58), (45, 53), (45, 54), (45, 55), (45, 92), (45, 139), (46, 52), (46, 53), (46, 54), (46, 55), (46, 56), (47, 51), (47, 52), (47, 53), (48, 52), (48, 53), (48, 54), (48, 82), (48, 88), (48, 89), (48, 90), (49, 105), (50, 91), (50, 105), (51, 90), (51, 91), (51, 92), (52, 89), (52, 90), (52, 91), (52, 92), (52, 103), (52, 104), (52, 105), (53, 88), (53, 89), (53, 90), (53, 91), (53, 92), (54, 63), (54, 82), (54, 88), (54, 89), (54, 90), (55, 62), (55, 63), (55, 64), (55, 88), (55, 90), (55, 92), (55, 102), (55, 111), (56, 61), (56, 62), (56, 63), (57, 61), (57, 62), (57, 64), (57, 79), (59, 79), (60, 79), (61, 80), (61, 81), (61, 82), (62, 79), (62, 80), (62, 81), (62, 82), (63, 78), (63, 79), (63, 80), (63, 82), (63, 88), (64, 77), (64, 78), (64, 79), (64, 87), (64, 88), (64, 90), (64, 102), (64, 111), (64, 124), (65, 69), (65, 76), (65, 77), (65, 78), (65, 79), (65, 80), (65, 87), (65, 88), (66, 70), (66, 75), (66, 76), (66, 77), (66, 87), (66, 88), (66, 90), (66, 104), (66, 106), (66, 126), (67, 75), (67, 76), (67, 77), (67, 78), (68, 87), (69, 87), (69, 106), (70, 75), (70, 106), (70, 107), (70, 126), (71, 106), (71, 107), (71, 109), (71, 126), (72, 109), (72, 126), (73, 109), (73, 126), (73, 127), (73, 128), (74, 125), (74, 126), (74, 127), (75, 124), (75, 125), (75, 126), (76, 123), (76, 124), (76, 125), (77, 90), (77, 111), (77, 123), (77, 124), (77, 125), (77, 126), (80, 88), (81, 88), (82, 88), (82, 89), (83, 88), (84, 88), (84, 89), (85, 89), (86, 104), (86, 105), (86, 106), (87, 104), (87, 105), (87, 106), (89, 103), (89, 104), (89, 105), (89, 106), (90, 102), (90, 103), (90, 104), (90, 109), (90, 111), (90, 125), (90, 126), (91, 101), (91, 102), (91, 103), (91, 104), (91, 105), (91, 108), (92, 100), (92, 101), (92, 102), (92, 139), (93, 99), (93, 100), (93, 101), (93, 103), (94, 98), (94, 99), (94, 100), (95, 99), (95, 100), (95, 101), (95, 129), (95, 135), (95, 136), (95, 137), (97, 136), (98, 137), (98, 138), (98, 139), (99, 136), (99, 137), (99, 138), (100, 129), (100, 135), (100, 136), (100, 137), (100, 138), (100, 139), (101, 110), (101, 129), (101, 135), (101, 136), (101, 137), (102, 109), (102, 110), (102, 111), (102, 135), (102, 137), (102, 139), (103, 108), (103, 109), (103, 110), (103, 129), (104, 108), (104, 109), (104, 126), (106, 126), (107, 126), (108, 127), (108, 128), (108, 129), (109, 126), (109, 127), (109, 128), (109, 129), (110, 125), (110, 126), (110, 127), (110, 129), (110, 135), (111, 124), (111, 125), (111, 126), (111, 134), (111, 135), (111, 137), (112, 116), (112, 123), (112, 124), (112, 125), (112, 126), (112, 127), (112, 134), (112, 135), (113, 117), (113, 122), (113, 123), (113, 124), (113, 134), (113, 135), (113, 137), (114, 122), (114, 123), (114, 124), (114, 125), (115, 134), (116, 134), (117, 122), (124, 137), (127, 135), (128, 135), (129, 135), (129, 136), (130, 135), (131, 135), (131, 136), (132, 136)

Total contacts: 498

fa-csHisH90G

(2, 34), (2, 35), (2, 36), (2, 117), (2, 118), (3, 33), (3, 34), (3, 35), (3, 116), (3, 117), (3, 118), (4, 14), (4, 15), (4, 33), (4, 34), (4, 35), (4, 36), (4, 45), (4, 115), (4, 116), (4, 117), (4, 118), (5, 13), (5, 14), (5, 15), (5, 33), (5, 34), (5, 113), (5, 114), (5, 115), (5, 116), (5, 118), (6, 12), (6, 13), (6, 14), (6, 23), (6, 34), (6, 45), (6, 53), (6, 83), (6, 85), (6, 93), (6, 113), (6, 114), (6, 115), (7, 11), (7, 12), (7, 13), (7, 82), (7, 112), (7, 113), (7, 114), (7, 116), (8, 12), (8, 13), (8, 95), (8, 97), (8, 112), (8, 113), (9, 82), (9, 95), (9, 97), (9, 109), (9, 110), (9, 112), (10, 25), (10, 97), (10, 98), (11, 24), (11, 25), (11, 116), (12, 23), (12, 24), (12, 25), (12, 97), (12, 98), (12, 100), (12, 113), (13, 22), (13, 23), (13, 24), (13, 25), (13, 26), (13, 31), (13, 33), (13, 116), (14, 21), (14, 22), (14, 23), (14, 33), (14, 34), (14, 53), (14, 61), (14, 93), (14, 101), (14, 113), (15, 20), (15, 21), (15, 22), (15, 23), (15, 24), (15, 31), (15, 32), (15, 33), (15, 34), (16, 20), (16, 21), (16, 22), (16, 24), (16, 31), (16, 32), (16, 47), (16, 48), (17, 22), (17, 24), (17, 28), (17, 31), (18, 32), (18, 48), (19, 51), (21, 34), (21, 47), (21, 53), (21, 63), (21, 100), (22, 100), (23, 97), (23, 98), (23, 100), (23, 101), (24, 31), (24, 32), (25, 31), (25, 98), (26, 31), (27, 31), (32, 47), (32, 48), (33, 46), (33, 47), (33, 118), (34, 45), (34, 46), (34, 47), (34, 51), (34, 52), (34, 53), (34, 63), (35, 44), (35, 45), (35, 46), (35, 47), (36, 43), (36, 44), (36, 45), (36, 76), (36, 115), (37, 42), (37, 43), (37, 44), (37, 46), (37, 52), (37, 73), (38, 42), (38, 43), (39, 43), (39, 44), (39, 52), (39, 66), (39, 68), (39, 73), (41, 74), (41, 75), (41, 76), (42, 69), (42, 70), (42, 73), (42, 74), (42, 75), (42, 76), (42, 86), (42, 87), (43, 72), (43, 73), (43, 74), (43, 75), (43, 76), (44, 52), (44, 66), (44, 72), (44, 73), (44, 74), (45, 51), (45, 52), (45, 53), (45, 63), (45, 72), (45, 74), (45, 76), (45, 85), (45, 115), (46, 50), (46, 51), (46, 52), (47, 51), (47, 63), (48, 63), (50, 64), (50, 65), (50, 66), (51, 63), (51, 64), (52, 62), (52, 63), (52, 64), (52, 65), (52, 66), (52, 72), (53, 61), (53, 62), (53, 63), (53, 71), (53, 72), (53, 74), (53, 85), (53, 93), (54, 60), (54, 61), (54, 62), (54, 63), (54, 64), (54, 65), (54, 71), (54, 72), (54, 74), (54, 93), (55, 60), (55, 61), (55, 62), (55, 71), (55, 72), (55, 74), (55, 87), (55, 88), (55, 91), (55, 93), (55, 102), (55, 103), (56, 60), (56, 71), (56, 91), (56, 103), (57, 103), (58, 91), (58, 102), (58, 103), (59, 96), (59, 101), (59, 102), (59, 103), (60, 102), (60, 103), (61, 93), (61, 101), (61, 102), (61, 103), (62, 71), (62, 72), (64, 72), (65, 72), (66, 72), (66, 73), (67, 71), (67, 72), (67, 73), (68, 73), (69, 73), (70, 87), (70, 88), (70, 89), (71, 86), (71, 87), (71, 88), (73, 86), (73, 87), (74, 85), (74, 86), (74, 87), (74, 91), (74, 93), (74, 103), (75, 84), (75, 85), (75, 86), (76, 83), (76, 84), (76, 85), (76, 115), (77, 82), (77, 83), (77, 84), (77, 86), (77, 92), (77, 107), (77, 112), (78, 82), (78, 83), (79, 83), (79, 84), (79, 92), (79, 107), (79, 112), (81, 113), (81, 114), (82, 112), (82, 113), (82, 114), (83, 111), (83, 112), (83, 113), (83, 114), (83, 115), (84, 92), (84, 93), (84, 107), (84, 108), (84, 111), (84, 112), (84, 113), (85, 91), (85, 92), (85, 93), (85, 111), (85, 113), (86, 90), (86, 91), (86, 92), (87, 91), (87, 93), (89, 105), (90, 104), (90, 105), (91, 103), (91, 104), (91, 105), (92, 102), (92, 103), (92, 104), (92, 106), (92, 107), (92, 111), (93, 101), (93, 102), (93, 103), (93, 110), (93, 111), (93, 113), (94, 100), (94, 101), (94, 102), (94, 103), (94, 104), (94, 110), (94, 111), (94, 113), (95, 100), (95, 101), (95, 104), (95, 109), (95, 110), (95, 112), (95, 113), (96, 100), (96, 101), (96, 102), (96, 104), (96, 110), (97, 109), (97, 110), (101, 113), (104, 108), (104, 110), (104, 111), (106, 111), (107, 111), (107, 112), (108, 112)

Total contacts: 387

ph-csHisH90G

(2, 34), (2, 35), (2, 36), (2, 117), (2, 118), (3, 32), (3, 33), (3, 34), (3, 35), (3, 116), (3, 117), (3, 118), (4, 14), (4, 15), (4, 33), (4, 34), (4, 35), (4, 36), (4, 115), (4, 116), (4, 117), (4, 118), (5, 13), (5, 14), (5, 15), (5, 33), (5, 34), (5, 113), (5, 114), (5, 115), (5, 116), (5, 118), (6, 12), (6, 13), (6, 14), (6, 23), (6, 34), (6, 45), (6, 53), (6, 76), (6, 83), (6, 85), (6, 93), (6, 113), (6, 114), (6, 115), (7, 11), (7, 12), (7, 13), (7, 82), (7, 112), (7, 113), (7, 114), (7, 116), (8, 12), (8, 13), (8, 82), (8, 95), (8, 97), (8, 112), (8, 113), (9, 82), (9, 95), (9, 97), (9, 109), (9, 110), (9, 112), (10, 25), (10, 95), (10, 97), (10, 98), (11, 24), (11, 25), (12, 23), (12, 24), (12, 25), (12, 27), (12, 98), (12, 113), (13, 22), (13, 23), (13, 24), (13, 25), (13, 26), (13, 31), (13, 33), (13, 116), (14, 21), (14, 22), (14, 23), (14, 33), (14, 34), (14, 53), (14, 93), (14, 101), (14, 113), (15, 20), (15, 21), (15, 22), (15, 24), (15, 31), (15, 32), (15, 33), (15, 34), (16, 20), (16, 21), (16, 22), (16, 24), (16, 31), (16, 32), (16, 47), (16, 48), (17, 22), (17, 24), (17, 28), (17, 30), (17, 31), (18, 48), (19, 51), (21, 34), (21, 47), (21, 53), (21, 63), (23, 97), (23, 100), (23, 101), (23, 113), (24, 31), (24, 32), (25, 31), (26, 31), (27, 31), (32, 47), (32, 48), (33, 46), (33, 47), (33, 116), (33, 118), (34, 45), (34, 46), (34, 47), (34, 51), (34, 52), (34, 53), (34, 63), (35, 44), (35, 45), (35, 46), (35, 47), (35, 49), (36, 43), (36, 44), (36, 45), (36, 76), (36, 115), (37, 42), (37, 43), (37, 44), (37, 46), (37, 52), (37, 73), (38, 42), (38, 43), (39, 43), (39, 44), (39, 52), (39, 66), (39, 68), (39, 73), (41, 74), (41, 75), (41, 76), (42, 69), (42, 70), (42, 73), (42, 74), (42, 75), (42, 76), (42, 86), (42, 87), (43, 72), (43, 73), (43, 74), (43, 75), (43, 76), (44, 52), (44, 66), (44, 72), (44, 73), (44, 74), (45, 51), (45, 52), (45, 53), (45, 63), (45, 72), (45, 74), (45, 76), (45, 85), (46, 50), (46, 51), (46, 52), (46, 63), (47, 51), (47, 63), (50, 64), (50, 65), (50, 66), (51, 63), (51, 64), (52, 62), (52, 63), (52, 64), (52, 65), (52, 66), (52, 72), (52, 73), (53, 61), (53, 62), (53, 63), (53, 71), (53, 72), (53, 74), (53, 85), (53, 93), (54, 60), (54, 61), (54, 62), (54, 63), (54, 64), (54, 65), (54, 71), (54, 72), (54, 74), (54, 93), (55, 60), (55, 61), (55, 62), (55, 71), (55, 72), (55, 74), (55, 87), (55, 88), (55, 91), (55, 93), (55, 102), (55, 103), (56, 60), (56, 71), (56, 88), (56, 91), (56, 103), (57, 91), (57, 103), (58, 91), (58, 102), (58, 103), (59, 96), (59, 99), (59, 100), (59, 101), (59, 102), (59, 103), (60, 102), (60, 103), (61, 93), (61, 101), (61, 102), (61, 103), (62, 71), (62, 72), (64, 72), (65, 72), (66, 72), (66, 73), (67, 71), (67, 72), (67, 73), (68, 73), (69, 73), (70, 87), (70, 88), (70, 89), (71, 87), (71, 88), (73, 86), (73, 87), (74, 85), (74, 86), (74, 87), (74, 91), (74, 93), (74, 103), (75, 84), (75, 85), (75, 86), (75, 92), (76, 83), (76, 84), (76, 85), (76, 115), (77, 82), (77, 83), (77, 84), (77, 86), (77, 92), (77, 107), (78, 82), (78, 83), (79, 83), (79, 84), (79, 92), (79, 107), (79, 112), (81, 113), (81, 114), (82, 109), (82, 112), (82, 113), (82, 114), (83, 111), (83, 112), (83, 113), (83, 114), (83, 115), (84, 92), (84, 107), (84, 108), (84, 111), (84, 112), (84, 113), (85, 91), (85, 92), (85, 93), (85, 111), (85, 113), (86, 90), (86, 91), (86, 92), (87, 91), (87, 93), (90, 104), (90, 105), (91, 103), (91, 104), (91, 105), (92, 102), (92, 103), (92, 104), (92, 106), (92, 107), (92, 111), (93, 101), (93, 102), (93, 103), (93, 110), (93, 111), (93, 113), (94, 100), (94, 101), (94, 102), (94, 103), (94, 104), (94, 110), (94, 111), (94, 113), (95, 100), (95, 101), (95, 110), (95, 112), (95, 113), (96, 100), (96, 101), (96, 102), (96, 104), (96, 110), (97, 109), (97, 110), (101, 113), (104, 108), (104, 110), (104, 111), (106, 111), (107, 111), (107, 112), (108, 112)

Total contacts: 385

mu-csHisH90G

(2, 34), (2, 35), (2, 36), (2, 117), (2, 118), (3, 33), (3, 34), (3, 35), (3, 116), (3, 117), (3, 118), (4, 14), (4, 15), (4, 33), (4, 34), (4, 35), (4, 36), (4, 115), (4, 116), (4, 117), (4, 118), (5, 13), (5, 14), (5, 15), (5, 33), (5, 34), (5, 113), (5, 114), (5, 115), (5, 116), (5, 118), (6, 12), (6, 13), (6, 14), (6, 34), (6, 45), (6, 53), (6, 83), (6, 85), (6, 93), (6, 113), (6, 114), (6, 115), (7, 11), (7, 12), (7, 13), (7, 82), (7, 112), (7, 113), (7, 114), (7, 116), (8, 12), (8, 13), (8, 95), (8, 97), (8, 112), (8, 113), (9, 82), (9, 95), (9, 97), (9, 109), (9, 110), (9, 112), (10, 25), (10, 97), (10, 98), (11, 24), (11, 25), (11, 116), (12, 23), (12, 24), (12, 25), (12, 27), (12, 97), (12, 98), (12, 113), (13, 22), (13, 23), (13, 24), (13, 25), (13, 26), (13, 31), (13, 33), (13, 116), (14, 21), (14, 22), (14, 23), (14, 33), (14, 34), (14, 53), (14, 61), (14, 93), (14, 101), (14, 113), (15, 20), (15, 21), (15, 22), (15, 23), (15, 24), (15, 31), (15, 32), (15, 33), (15, 34), (16, 20), (16, 21), (16, 22), (16, 24), (16, 31), (16, 32), (16, 33), (16, 47), (16, 48), (17, 22), (17, 24), (17, 25), (17, 27), (17, 28), (17, 30), (17, 31), (17, 32), (18, 32), (18, 48), (19, 51), (19, 63), (21, 34), (21, 47), (21, 53), (21, 63), (22, 100), (23, 97), (23, 98), (23, 100), (23, 101), (24, 31), (24, 32), (25, 31), (26, 31), (27, 31), (32, 47), (32, 48), (33, 46), (33, 47), (33, 118), (34, 45), (34, 46), (34, 47), (34, 51), (34, 52), (34, 53), (34, 63), (35, 44), (35, 45), (35, 46), (35, 47), (36, 43), (36, 44), (36, 45), (36, 76), (36, 85), (36, 115), (37, 42), (37, 43), (37, 44), (37, 46), (37, 52), (37, 73), (38, 42), (38, 43), (39, 43), (39, 44), (39, 52), (39, 66), (39, 68), (39, 73), (41, 74), (41, 75), (41, 76), (42, 69), (42, 70), (42, 73), (42, 74), (42, 75), (42, 76), (42, 86), (42, 87), (43, 72), (43, 73), (43, 74), (43, 75), (43, 76), (44, 52), (44, 66), (44, 72), (44, 73), (44, 74), (45, 51), (45, 52), (45, 53), (45, 63), (45, 72), (45, 74), (45, 85), (46, 50), (46, 51), (46, 52), (47, 51), (47, 63), (48, 63), (50, 64), (50, 65), (50, 66), (51, 63), (51, 64), (52, 62), (52, 63), (52, 64), (52, 65), (52, 66), (52, 72), (53, 61), (53, 62), (53, 63), (53, 71), (53, 72), (53, 74), (53, 85), (53, 93), (54, 60), (54, 61), (54, 62), (54, 63), (54, 64), (54, 65), (54, 71), (54, 72), (54, 74), (54, 93), (55, 60), (55, 61), (55, 62), (55, 71), (55, 72), (55, 74), (55, 87), (55, 88), (55, 91), (55, 93), (55, 102), (55, 103), (56, 60), (56, 71), (56, 88), (56, 91), (56, 103), (57, 91), (57, 103), (58, 91), (58, 102), (58, 103), (59, 101), (59, 102), (59, 103), (60, 102), (60, 103), (61, 93), (61, 101), (61, 102), (61, 103), (62, 71), (62, 72), (64, 72), (65, 72), (66, 72), (66, 73), (67, 71), (67, 72), (67, 73), (68, 73), (69, 73), (70, 87), (70, 88), (70, 89), (71, 86), (71, 87), (71, 88), (73, 86), (73, 87), (74, 85), (74, 86), (74, 87), (74, 91), (74, 93), (74, 103), (75, 84), (75, 85), (75, 86), (76, 83), (76, 84), (76, 85), (76, 115), (77, 82), (77, 83), (77, 84), (77, 86), (77, 92), (77, 107), (77, 112), (78, 82), (78, 83), (79, 84), (79, 92), (79, 107), (79, 112), (81, 113), (81, 114), (82, 109), (82, 112), (82, 113), (82, 114), (82, 115), (83, 111), (83, 112), (83, 113), (83, 114), (83, 115), (84, 92), (84, 93), (84, 107), (84, 111), (84, 112), (84, 113), (85, 91), (85, 92), (85, 93), (85, 111), (85, 113), (86, 90), (86, 91), (86, 92), (87, 91), (87, 93), (90, 104), (90, 105), (91, 103), (91, 104), (91, 105), (92, 102), (92, 103), (92, 104), (92, 105), (92, 106), (92, 107), (92, 111), (93, 101), (93, 102), (93, 103), (93, 110), (93, 111), (93, 113), (94, 100), (94, 101), (94, 102), (94, 103), (94, 104), (94, 110), (94, 111), (94, 113), (95, 100), (95, 101), (95, 104), (95, 110), (95, 112), (95, 113), (96, 100), (96, 101), (96, 102), (96, 104), (96, 110), (97, 109), (97, 110), (101, 113), (104, 108), (104, 110), (104, 111), (106, 111), (107, 111), (107, 112), (108, 112)

Total contacts: 388

Appendix B: 3Foil C_{α}-SBM using replica exchange umbrella sampling

Modeling the 3Foil Gibbs free energy barrier of unfolding

Direct sampling using C_{α} coarse-grained folding simulations was unsuccessful in modeling the 3Foil folding trajectory. 3Foil folding transitions were exceedingly rare during direct sampling (*i.e.* unbiased) simulations owing to 3Foil's unusually high free energy barrier of unfolding (Broom et al., 2015b). Increasing the simulation length did not sufficiently increase the number of observed folding transitions, and 3Foil's T_f (*i.e.* the transition midpoint) could not be determined. Since high energy conformations in configuration space could not be sufficiently sampled, a reliable estimation of the 3Foil unfolding free energy barrier could not be obtained using direct sampling. Thus, we implemented the enhanced sampling method replica exchange umbrella sampling (REUS) at an estimated T_f to model 3Foil's folding trajectory and unfolding free energy barrier (Kästner, 2011; Giri Rao and Gosavi, 2018).

Umbrella sampling (US) enhances sampling of high-energy events by splitting the reaction coordinate, Q, into a series of windows with each window *i* centered on a unique value Q. Since Q is a measure of how native-like the structure is, each window represents a partially folded structure of Qi along the reaction coordinate. US applies an additional energy term, the bias potential, to each window *i* such that

$$E^{b}(r) = E^{u}(r) + \omega_{i}(\xi) \qquad \textbf{(B1)}$$

where $E^{b}(r)$ and $E^{u}(r)$ are the biased and unbiased potential energies for configuration *r*, respectively, and $\omega_{i}(\xi)$ is the bias potential, which depends only on the reaction coordinate (Kästner, 2011). This bias restrains the system to sample configuration space around a specific Qi for each window, allowing uniform sampling of the entire configuration space along the folding pathway. We enhanced US quality using Hamiltonian replica exchange (RE) (Kästner, 2011), in which configurations in neighboring windows attempt to exchange at specified intervals (Sabri Dashti and Roitberg, 2013). During exchange the bias $\omega_i(\xi)$ from window *i* is used to compute the total biased energy of the neighboring window *j* and the bias $\omega_j(\xi)$ from window *j* is used to calculate the total biased energy for window *i.* If the sum of the biased energies for *i* and *j* is less than the sum of their original energies, the coordinates from window *i* and *j* are exchanged, and the simulation continues (Kästner, 2011). Therefore, through replica exchange a configuration from window *i* may be subjected to all other bias potentials along the reaction coordinate, thus enhancing sampling of configuration space.

REUS simulations for 3Foil were carried out using GROMACS v.4.5.4 patched with PLUMED v.1.3, which allows enhanced-sampling methods and MPI processing (Bekker, H., Berendsen, H. J. C., Dijkstra, E. J., Achterop, S., van Drunen, R. et al., 1993; Berendsen et al., 1995; Lindahl et al., 2001; Van Der Spoel et al., 2005; Hess et al., 2008; Bonomi et al., 2009). Input files required for REUS GROMACS-PLUMED simulations are depicted in Scheme 2. GROMACS topology (.top), table (.xvg), and parameter (.mdp) files were obtained from the SMOG2 web server as described for C_α-SBM simulations (Clementi et al., 2000; Noel et al., 2012, 2016). In REUS protein folding simulations, geometry (.gro) files must span the entire reaction coordinate of interest, and each geometry file acts as the central coordinate for a given replica window. We chose to use the reaction coordinate Q, or the fraction of native contacts, to enhance sampling of partially folded structures along the folding pathway. Toward this end, a unique GROMACS geometry file was generated for each replica window using an in-house python script. Notably, equilibrium Q values, or the Q value upon which each replica window is centered, change based on the number of windows used in a given REUS simulation and on the placement of replica windows along the reaction coordinate. A tpr file, containing the unique starting structure, molecular topology, and simulation parameters for a given replica window, was generated for each window using the GROMACS **grompp** command. PLUMED data files (.dat), specifying the identity, force constant (κ), and equilibrium Q value for a given window, and the protein contact index (.ndx) file were also generated using in-house python scripts.

Trial REUS simulations for 3Foil were initially set up with 32 uniformly spaced replica windows with a force constant of 0.05. The trial simulations were run for 1x10^8 steps (50 ns). Replica exchanges were attempted every 5000 steps (2.5 ps), with 19 999 exchanges attempted in total. 3Foil simulations were completed at 160 K. Following REUS, the trial window set was evaluated to determine if it satisfied the criteria to undergo the weighted histogram analysis method (WHAM), which was used to unbiased the system's free energy (Gallicchio et al., 2005; Kästner, 2011). Specifically, WHAM criteria require that: 1) a given replica must exchange between all windows for all replicas in a simulation; 2) the probability of a given replica successfully exchanging with an adjacent window must approximate 0.2-0.4 for all adjacent windows; 3) the potential energy distribution of a given window must overlap with the potential energy distribution of the adjacent window for all windows. WHAM criteria were not met using 32 uniformly-spaced windows, so two additional windows were added. A second trial simulation for 34 non-uniformly-spaced windows was submitted, and WHAM criteria were achieved (Figure B1A-C). A production simulation for each of the 34 windows was run for 2x10^8 steps (100 ns) with a force constant of 0.05 and replica exchange attempted every 10 000 steps (5 ps; 19 999 exchange attempts total).

Following REUS production runs, WHAM was implemented to solve for the unbiased free energy of the 3Foil folding trajectory. WHAM iteratively solved the unknown values of $p_l{}^o$, the unbiased probability that the system had a given Q value and potential energy (equation B2), and f_i, the normalization constant (equation B3) (Gallicchio et al., 2005).

$$p_l{}^o = \sum_{i=1}^{S} n_{i,l} \,/\, \sum_{i=1}^{S} N_i\, f_i\, c_{i,l} \qquad \textbf{(B2)}$$

$$f_i{}^{-1} = \sum_{l=1}^{m} c_{i,l}\, p_l{}^o \qquad \textbf{(B3)}$$

where S denotes the total number of simulations, $n_{i,l}$ is the count in bin *l* of histogram *i*, N_i is the total number of counts in histogram *i*, and $c_{i,l}$ is the biasing factor given by

$$c_{i,l} = \exp[-(\beta_i - \beta_0)E_l]\,\exp\,[-\beta_i \omega_i(\xi)] \qquad \textbf{(B4)}$$

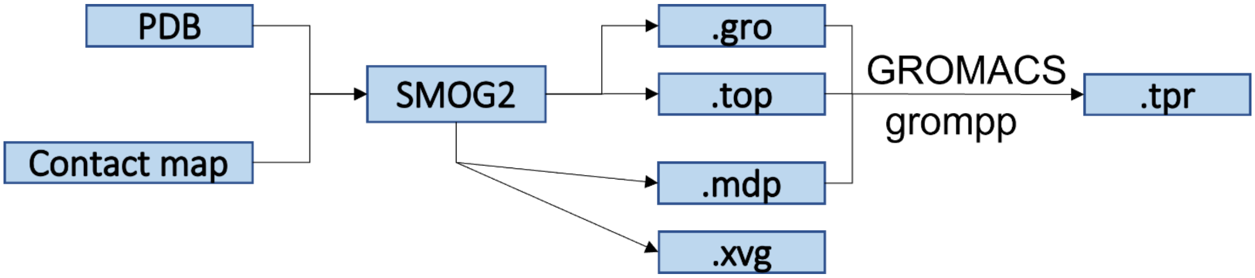
where β_i and β_0 are the inverse Boltzmann's temperature of window i and the reference inverse temperature, respectively, and E_l is the unbiased potential energy of bin *l* (Gallicchio et al., 2005). After solving for $p_l{}^o$ and f_i, the unbiased free energy of the system $A(\mathrm{Q})$ was calculated by

$$A(\mathrm{Q}) = -\frac{1}{\beta}\ln\, P(\mathrm{Q}) \qquad \textbf{(B5)}$$

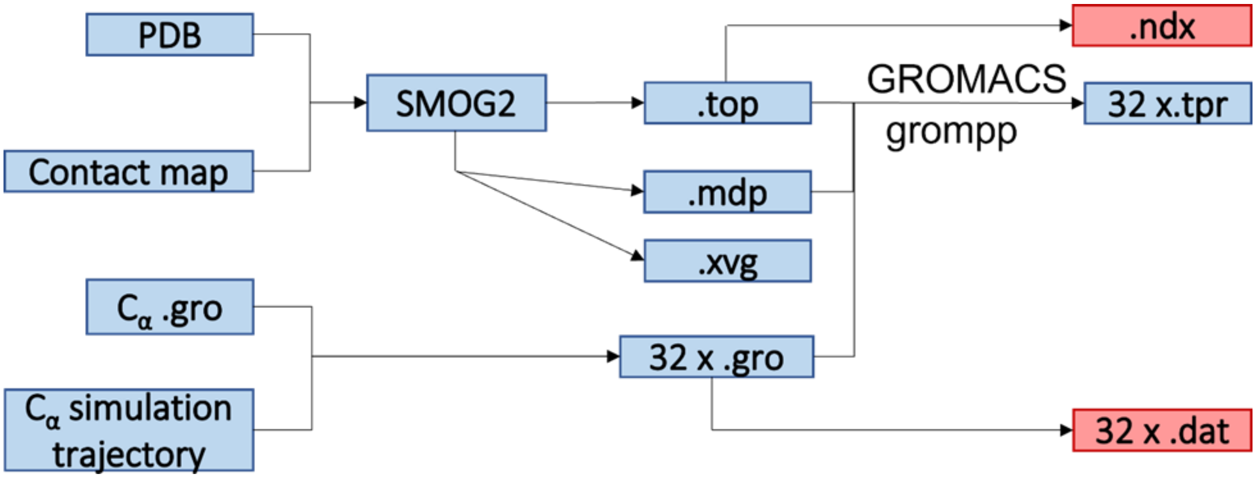
where $\beta = 1/k_BT$, k_B is the Boltzmann's constant, and $P(\mathrm{Q})$ is the distribution of the system along the reaction coordinate, Q, given by the unbiased probability distribution across all windows (Kästner, 2011). Finally, we modeled the 3Foil unbiased free energy barrier of unfolding at the T_f (Figure B1D), which was in good agreement with the previous unfolding free energy barrier modeled using a CSU contact map (Broom et al., 2015b).

As with wtHis and csHisH90G, the 3Foil folding pathway was assessed by examining the average contact maps along the progress coordinate Q. The average contact map represents the average probability of the formation of all native contacts at a given Q value along the folding pathway. As in the free energy profile of 3Foil, WHAM analysis was first performed to obtain unbiased probabilities of native contacts. The average contact maps are then calculated by averaging the contact probability matrices around a given Q value (± 0.025). Figure B2 shows the average contact map of 3Foil at the the transition state ensemble at Q ≈ 0.4 .

Schemes and Figures



Scheme 1. Generating input files for GROMACS C_α-SBM simulations using SMOG2. A protein's PDD file and a contact map listing contacting residues in the native structure may be uploaded to the SMOG2 web server (Noel and Onuchic, 2012). SMOG2 generates a geometry (.gro) and topology (.top) file from the PDB and contact map. An example parameter file (.mdp) for C_α modeling can be found on the SMOG2 website. The geometry, topology, and parameter file are compiled into a portable binary run input (.tpr) file using the GROMACS **grompp** command. The tpr file is used to start and run GROMACS simulations. A perl script used to generate the table (.xvg) file, which specifies the 10-12 Lennard-Jones potentials for all residue pairs used in C_α-SBM simulations, can also be found on the SMOG2 website (Noel and Onuchic, 2012).



Scheme 2. Generating input files for a GROMACS-PLUMED REUS simulation with 32 replica windows. The topology (.top), parameter (.mdp), and table (.xvg) files may be generated using SMOG2 as described in Scheme 1 (Noel and Onuchic, 2012). A geometry file for each replica window, each representing a partially folded protein structure with a unique Q value, is generated from a previous C_α-SBM simulation trajectory using an in-house python script. Importantly, these Q values are evenly spaced throughout the entire reaction coordinate. A tpr file for each replica window is compiled from the window's geometry, topology, and parameter files using the GROMACS **grompp** command. PLUMED data files (.dat) are generated for each window using another in-house python script. Each data file contains the replica window's Q value and force constant, κ. Finally, the PLUMED protein contacts index file (.ndx) is generated from the [pairs] section of the topology file. Files used by GROMACS and PLUMED are highlighted in blue and red, respectively.

A

Replica window	0		1		2		3		4		5		6		7		8		9		10		11
Q (10^{-2})	2.5		5.5		8.6		11.6		14.7		17.7		20.8		23.8		26.9		29.9		32.6		35.2
Replica exchange probability (10^{-2})		41		32		28		26		25		24		23		22		22		26		24	

Replica window	11		12		13		14		15		16		17		18		19		20		21		22
Q (10^{-2})	35.2		37.8		40.5		43.1		45.8		48.4		51.1		53.7		56.3		59.0		61.6		64.3
Replica exchange probability (10^{-2})		24		26		28		30		30		28		25		26		25		25		25	

Replica window	22		23		24		25		26		27		28		29		30		31		32		33
Q (10^{-2})	64.3		66.9		69.5		72.6		75.6		78.7		81.7		84.8		87.8		90.9		93.9		96.9
Replica exchange probability (10^{-2})		26		28		23		24		26		28		30		32		34		37		41	

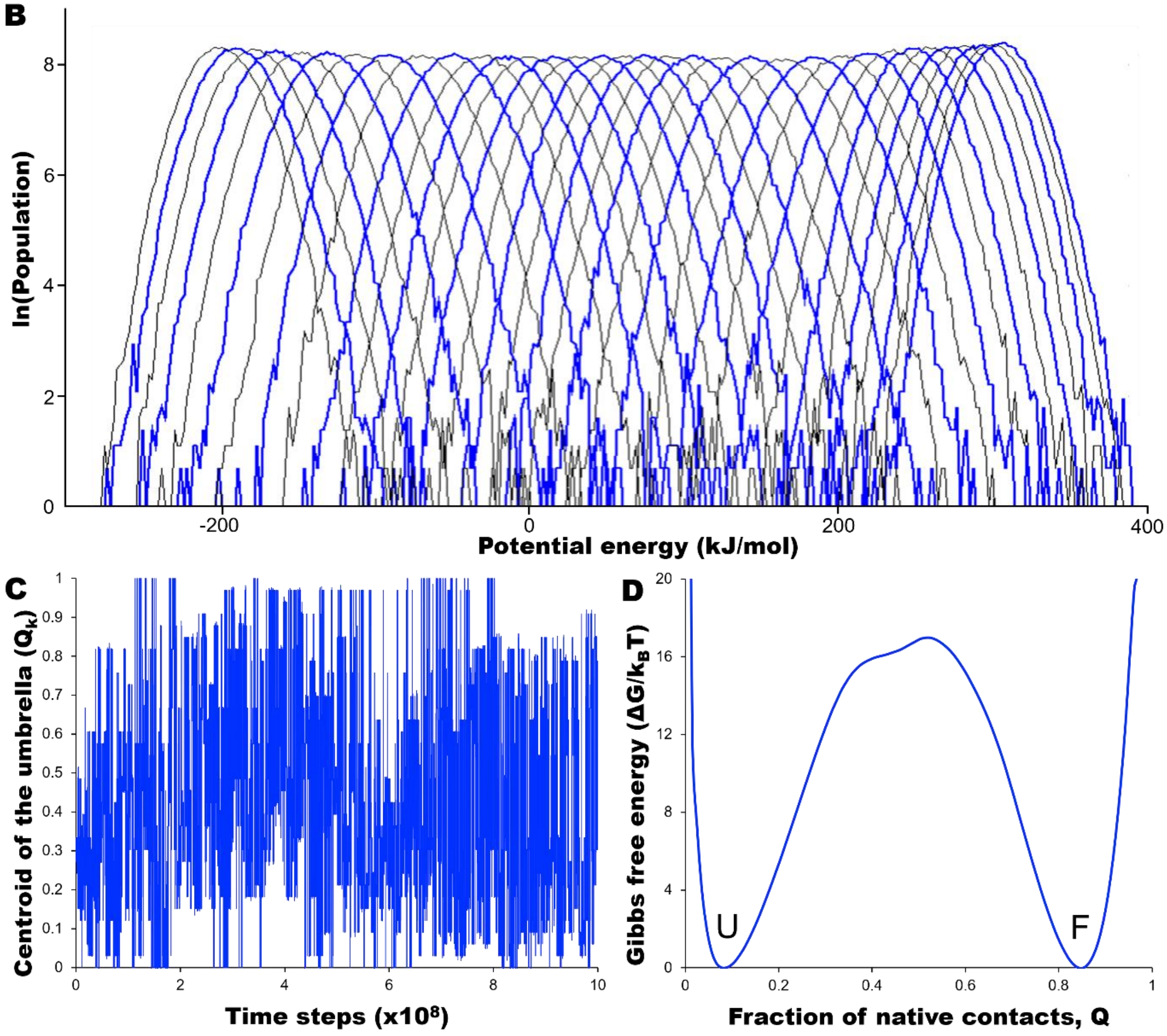


Figure B1. 3Foil REUS simulations meet WHAM criteria using 34 windows. 3Foil REUS simulations show that 34 non-uniformly spaced replica windows and a force constant of 0.05 result in good replica exchange statistics and window overlap for WHAM. **(A)** Following REUS, the probability of successful replica exchange between adjacent windows ranges from 0.22 to 0.41 for all windows, in good agreement with the recommended value of 0.2 to 0.4. Replica exchange was attempted every 10 000 steps (5 ps), and 19 999 replica exchanges were attempted in total. **(B)** The potential energy distribution of a given replica window overlaps with the potential energy distribution of adjacent replica windows for all windows. Odd replica windows are colored black and even replica windows are colored blue to better display the overlap of neighboring windows. Potential energy trajectories were generated using the GROMACS **g_energy** command. Replica exchange traces and potential energy distributions were visualized in XMGRACE (Turner, 2005). **(C)** Tracing exchange events for a given replica (*e.g.* replica 10 (above)) shows that each replica exchanged into all other replica windows. **(D)** The 3Foil unfolding free energy barrier modeled from unbiased REUS simulation data solved at T = 160.1 K using WHAM (Gallicchio et al., 2005). The unfolding free energy barrier is along the reaction coordinate Q, the fraction of native contacts. The unfolded (U) and folded (F) states are indicated. Unfolding free energy barrier heights are given in Table 3.1.

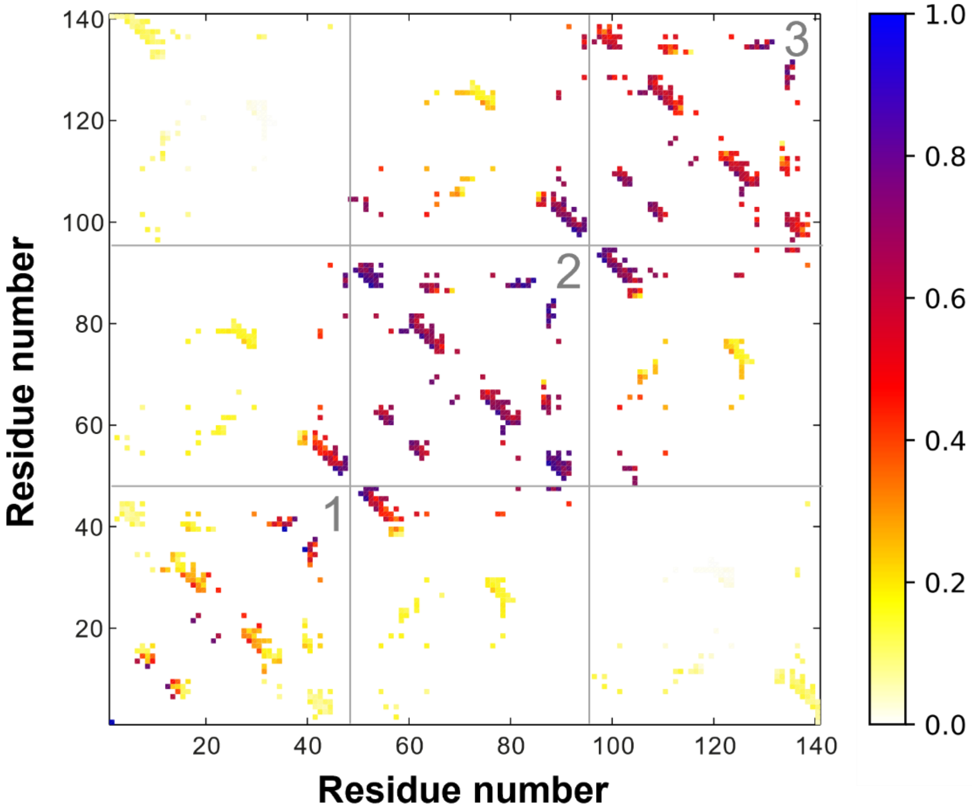


Figure B2. 3Foil simulations show simultaneous folding of the central and C-terminal trefoils. Average contact map for the transition state ensemble for 3Foil at Q = 0.4. Contacts are colored based on degree of structure, with 1 indicating native levels of structure and 0 indicating random coil. The N-terminal, central, and C-terminal trefoil are labeled 1, 2, and 3, respectively. 3Foil initiates folding from the C-terminal and folds the C-terminal and central foils together. Transition state ensembles for wtHis and csHisH90G are given in Figure 3.7.

Appendix C: X-ray crystallography for csHisH90G and fa-caHisH90G

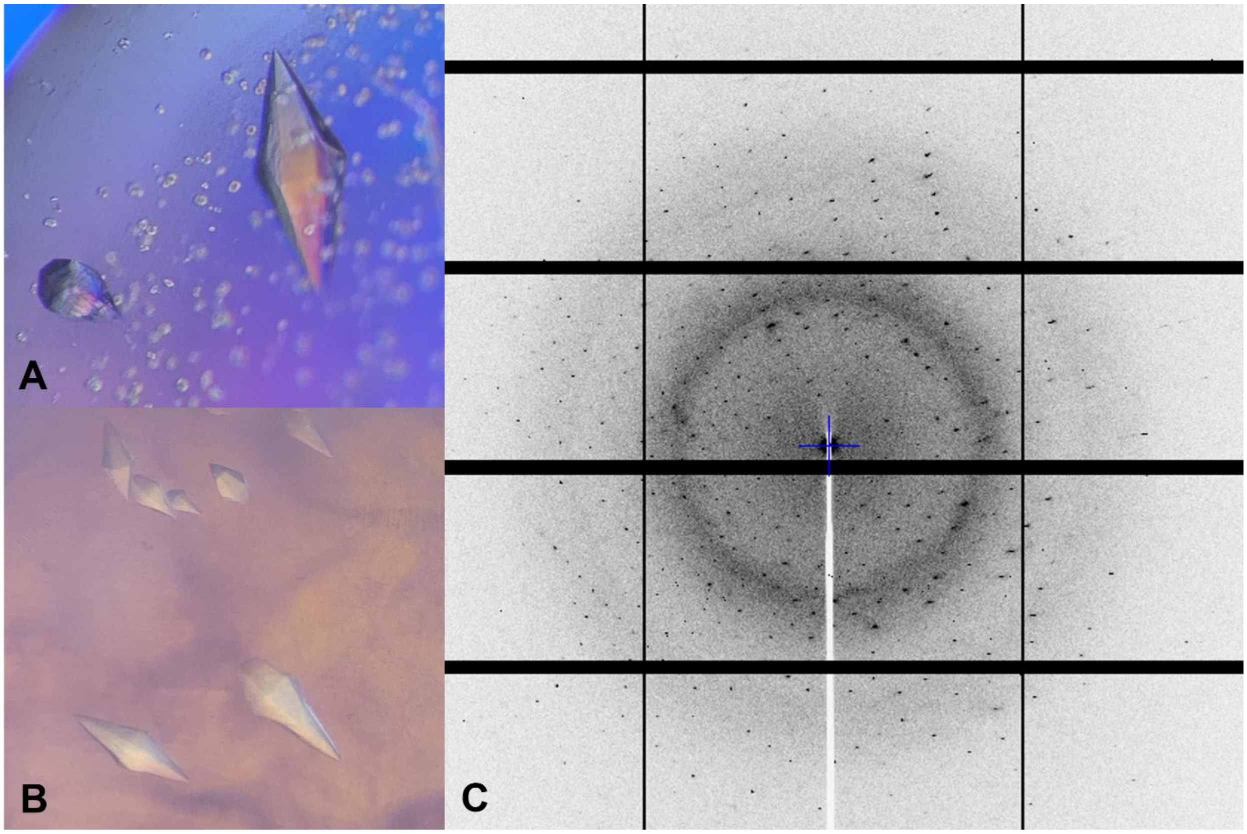


Figure C1. csHisH90G produces diffraction-quality crystals. (A, B) Diffraction-quality crystals for csHisH90G were grown in 0.1 M HEPES pH 8.1 and 17.6% PEG8000 with the additive praseodymium (III) acetate hydrate. Crystals were soaked in 20% PEG400 cryoprotectant and flash frozen in liquid nitrogen before being shot on the Canadian Macromolecular Crystallography Facility ID (O8ID) beamline at the Canadian Light Source (Saskatoon, SK). **(C)** Diffraction data were collected over 0.5° interval oscillations with 0.15 second exposure time with the detector set 400 mm away from the mounted crystal. Crystals diffracted to 2.85 Å, but data refinement revealed that the crystal was twinned, and the structure could not be solved.

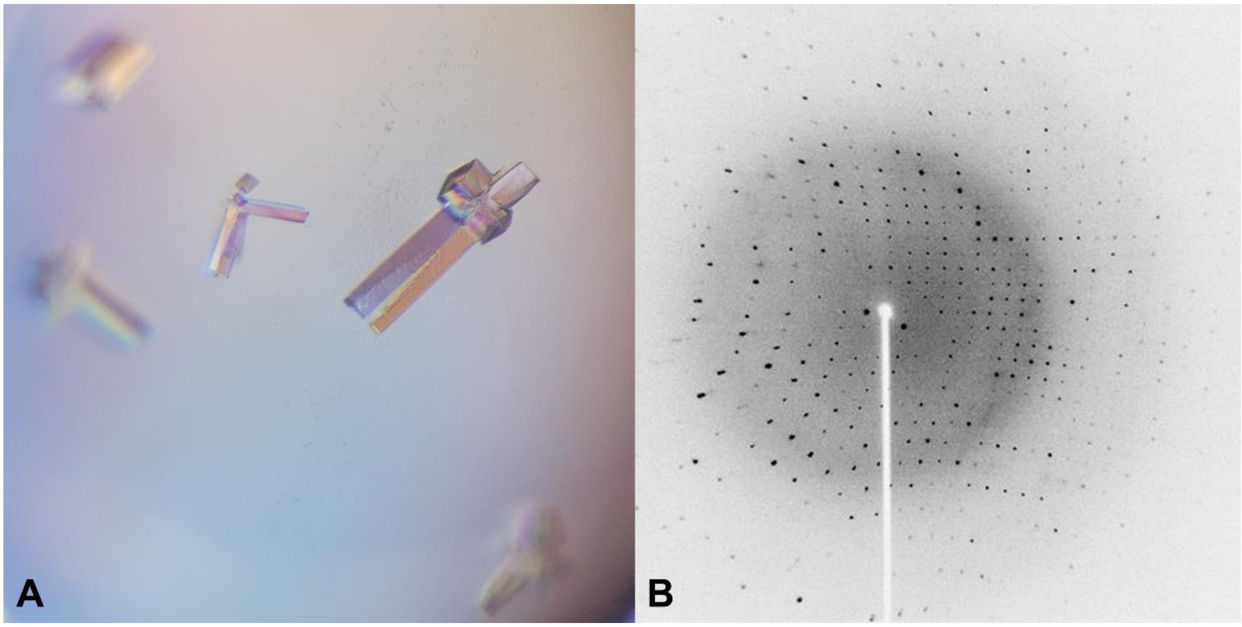
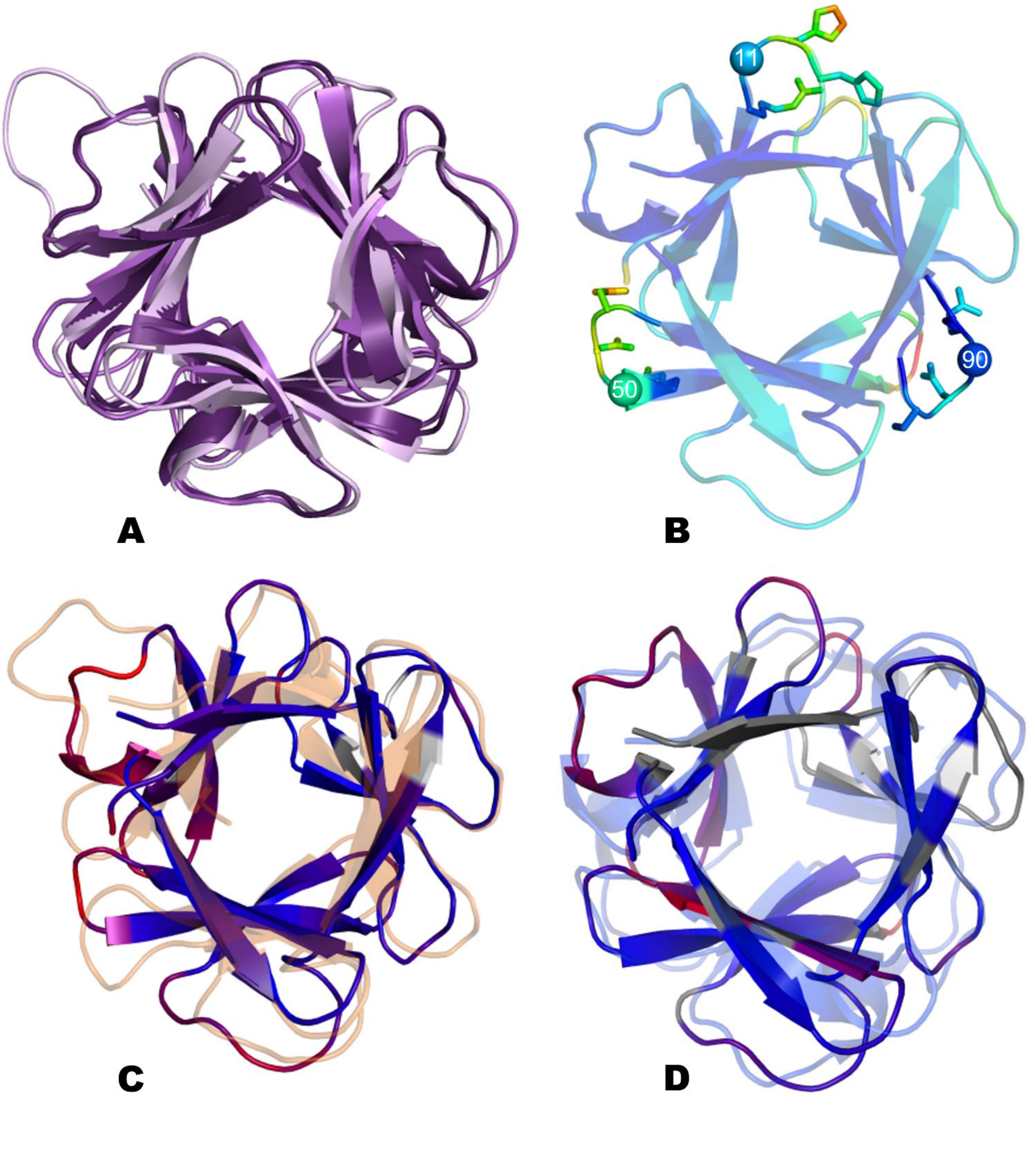


Figure C2. Rod-shaped fa-csHisH90G crystals diffract to 1.7 Å. (A) Diffraction-quality crystals for fa-csHisH90G were grown in 0.2 M $LiSO_4$ M, 0.1 M phosphate/citrate buffer pH 3.8, and 20 % PEG 1000. Crystals were soaked in 20% PEG400 cryoprotectant and flash frozen in liquid nitrogen. **(B)** Crystals were shot in-house using a Rigaku rotating copper anode X-ray generator (Cu K_α = 1.54 Å) and an R-axis IV++ detector with the detector set 150 mm away from the mounted crystal. Diffraction data were collected with a 10° 2θ offset and 120 second exposure time per 1° interval. Crystals diffracted to 1.70 Å.



A
11
50
90
B
C
D

Figure C3. Alignment of the fa-csHisH90G crystal structure to computational models, wtHis, and 3Foil. (A) Alignment of the Rosetta CM (Chivian et al., 2003; Song et al., 2013) (light purple) and ColabFold (Mirdita et al., 2022) (dark purple) models to the fa-csHisH90G crystal structure (medium purple) shows several deviations between the Rosetta CM model and the crystal structure, particularly in the hairpin cap and loops. In contrast, the ColabFold model for fa-csHisH90G shows good agreement to the crystal structure. **(B)** fa-csHisH90G is shown colored by B-factor, where lighter colors indicate higher B-factors and dark colors dark colors indicate low B-factors. The β9-β10 turn and symmetrically equivalent turns are shown in stick representation. G90 and symmetrically equivalent glycine residues are shown as spheres. **(C)** fa-csHisH90G aligns to the wtHis NMR structure with an RMSD of 2.25 Å. Regions with low deviation between structures are colored blue, while those with high deviation are in red. Residues not used in the alignment are given in white. **(D)** fa-csHisH90G aligns to the 3Foil crystal structure with an RMSD of 1.42 Å. Deviation between structures is colored as in (C).

Table C1. Backbone and sidechain angles for wtHis, csHisH90G, fa-csHisH90G, and 3Foil. Phi, psi, and chi angles are given for the NMR structure of wtHis, computational models of csHisH90G and fa-csHisH90G, and the fa-csHisH90G and 3Foil crystal structures. Core residue position is number relative to wtHis, and residue identities are indicated for all proteins listed. The Rosetta CM model for fa-csHisH90G is given as *fa-csHisH90G*, the ColabFold model is called *αfa-csHisH90G,* and the crystal structure is indicated in bold. Energy minimized structures are indicated with italics. Table compiled by Iain McDonald.

Core position	**Protein**	**Residue**	**Phi (°)**	**Psi (°)**	**Chi$_1$ (°)**	**Chi$_2$ (°)**	**Chi$_3$ (°)**	**Chi$_4$ (°)**
4	wtHis	R4	-164.758	-100.615	-155.42	160.618	-148.111	159.976
	csHisH90G	Y4	-141.878	150.796	-58.805	86.517		
	fa-csHisH90G	Y4	-137.644	148.37	-64.467	87.695		
	αfa-csHisH90G	Y4	-131.451	145.23	-69.122	-96.102		
	fa-csHisH90G	Y4	-125.452	152.63	-56.293	78.336		
	3Foil	Y5	-132.516	150.735	-71.11	83.992		
6	wtHis	F6	-115.245	115.225	-82.281	-57.164		
	csHisH90G	L6	-104.935	108.878	-61.927	-179.83		
	fa-csHisH90G	L6	-108.354	105.251	-61.208	177.844		
	αfa-csHisH90G	L6	-104.002	105.939	-72.409	80.531		
	fa-csHisH90G	L6	-116.55	108.61	-81.501	166.379		
	3Foil	L7	-116.155	117.371	-78.411	75.705		
14	wtHis	L14	-44.929	123.572	-58.618	-158.639		
	csHisH90G	L14	-77.217	116.322	-169.41	66.81		
	fa-csHisH90G	L14	-83.71	119.122	-174.543	68.57		
	αfa-csHisH90G	L14	-61.642	128.082	-167.948	72.817		
	fa-csHisH90G	L14	-56.16	132.667	-169.711	67.028		
	3Foil	L16	-65.182	124.326	-167.262	68.119		
21	wtHis	V21	-115.561	138.39	175.251			
	csHisH90G	V21	-89.812	134.063	178.924			
	fa-csHisH90G	V21	-97.469	138.196	173.618			
	αfa-csHisH90G	V21	-101.531	120.788	-171.902			
	fa-csHisH90G	V21	-106.75	120.966	-179.712			
	3Foil	V29	-104.519	131.696	169.849			
34	wtHis	F34	-118.279	147.595	-80.817	-63.107		
	csHisH90G	W34	-86.438	150.858	-68.531	87.596		
	fa-csHisH90G	W34	-80.371	157.924	-63.479	96.611		
	αfa-csHisH90G	W34	-115.243	121.911	-60.028	83.984		
	fa-csHisH90G	W34	-117.305	141.539	-64.913	85.929		
	3Foil	W42	-130.467	132.554	-60.901	85.514		
36	wtHis	V36	-115.516	122.549	86.031			
	csHisH90G	L36	-105.265	131.563	-178.528	69.645		
	fa-csHisH90G	L36	-98.299	132.049	-177.671	67.469		
	αfa-csHisH90G	L36	-90.2	118.033	-164.88	60.993		

	fa-csHisH90G	L36	-97.379	123.425	-172.43	67.359
	3Foil	L44	-92.982	99.715	-99.299	62.461
43	wtHis	V43	-134.285	150.777	-73.496	
	csHisH90G	Y43	-112.542	149.569	-73.166	86.769
	fa-csHisH90G	Y43	-126.928	144.719	-64.568	80.953
	αfa-csHisH90G	Y43	-131.594	138.768	-61.493	-93.364
	fa-csHisH90G	Y43	-129.306	142.466	-63.035	86.268
	3Foil	Y52	-137.114	152.461	-66.734	78.458
45	wtHis	L45	-116.67	128.191	-67.342	150.836
	csHisH90G	L45	-115.628	113.312	-74.41	91.001
	fa-csHisH90G	F45	-111.694	110.392	-70.556	-100.66
	αfa-csHisH90G	F45	-101.957	110.628	-64.403	-97.024
	fa-csHisH90G	F45	-117.578	110.858	-72.576	83.424
	3Foil	L54	-115.38	118.513	-80.28	80.04
53	wtHis	L53	-50.597	149.801	-93.981	-160.945
	csHisH90G	L53	-63.546	124.388	-178.177	65.42
	fa-csHisH90G	L53	-61.919	119.93	-178.439	63.793
	αfa-csHisH90G	L53	-61.698	117.61	-175.533	65.494
	fa-csHisH90G	L53	-63.273	123.635	-178.586	67.852
	3Foil	L63	-64.882	126.093	-167.636	67.202
61	wtHis	V61	-129.467	149.364	-172.275	
	csHisH90G	V61	-129.612	139.097	-178.479	
	fa-csHisH90G	V61	-137.248	132.966	-176.691	
	αfa-csHisH90G	V61	-99.776	133.144	-174.003	
	fa-csHisH90G	V61	-122.234	165.485	-65.959	
	3Foil	V76	-106.098	131.121	174.023	
74	wtHis	F74	-143.461	171.263	-73.011	82.242
	csHisH90G	W74	-126.845	141.155	-66.791	95.389
	fa-csHisH90G	W74	-114.477	136.545	-65.225	89.126
	αfa-csHisH90G	W74	-113.291	134.681	-59.023	83.388
	fa-csHisH90G	W74	-130.571	141.901	-63.719	84.449
	3Foil	W89	-130.603	128.135	-62.574	86.912
76	wtHis	L76	-129.207	145.184	127.478	67.458
	csHisH90G	L76	-86.597	129.416	-176.081	67.194
	fa-csHisH90G	L76	-81.184	119.351	179.374	65.144
	αfa-csHisH90G	L76	-95.305	108.38	-86.775	55.455
	fa-csHisH90G	L76	-89.077	124.978	-82.618	46.691
	3Foil	L91	-87.104	114.177	-109.146	-66.406
83	wtHis	V83	-148.77	168.747	43.712	
	csHisH90G	Y83	-100.459	155.498	-74.009	95.123
	fa-csHisH90G	Y83	-105.358	150.423	-68.232	93.703
	αfa-csHisH90G	Y83	-130.029	142.636	-61.271	-92.002
	fa-csHisH90G	Y83	-132.242	150.384	-79.057	85.405
	3Foil	Y99	-126.683	154.634	-69.191	80.318

85	wtHis	I85	-117.324	144.714	-54.32		
	csHisH90G	L85	-119.96	118.337	-81.819	82.572	
	fa-csHisH90G	L85	-119.998	116.596	-61.268	-176.19	
	αfa-csHisH90G	L85	-105.67	105.25	-81.76	66.882	
	fa-csHisH90G	L85	-110.737	111.522	-82.677	63.938	
	3Foil	L101	-119.04	117.881	-74.11	73.535	
93	wtHis	I93	-103.803	142.174	103.656		
	csHisH90G	L93	-63.081	143.886	-162.443	69.602	
	fa-csHisH90G	L93	-65.692	137.955	-171.826	73.51	
	αfa-csHisH90G	L93	-62.395	127.058	-177.78	69.4	
	fa-csHisH90G	L93	-61.591	135.602	-172.455	70.15	
	3Foil	L110	-60.623	129.129	-162.729	65.403	
101	wtHis	V101	-98.444	140.309	-162.728		
	csHisH90G	V101	-122.504	132.481	179.724		
	fa-csHisH90G	V101	-122.107	138.177	-69.003		
	αfa-csHisH90G	V101	-87.478	131.524	-179.61		
	fa-csHisH90G		-100.937	133.111	178.784		
	3Foil	V123	-101.062	131.757	170.104		
113	wtHis	F113	-116.25	179.548	-66.431	70.343	
	csHisH90G	W113	-121.633	153.651	-68.208	-91.519	
	fa-csHisH90G	W113	-127.57	148.454	-59.765	-81.632	
	αfa-csHisH90G	W113	-123.243	135.92	-57.575	87.36	
	fa-csHisH90G	W113	-127.856	145.556	-62.489	86.87	
	3Foil	W136	-132.511	134.178	-61.293	85.974	
115	wtHis	E115	-115.768	115.85	68.343	-168.364	-60.663
	csHisH90G	L115	-87.926	107.461	-175.225	63.74	
	fa-csHisH90G	L115	-81.655	106.453	-174.617	64.446	
	αfa-csHisH90G	L115	-92.138	109.594	-60.993	-173.056	
	fa-csHisH90G	L15	-100.873	110.97	-179.952	57.05	
	3Foil	L138	-93.797	104.358	-89.747	46.019	

Table C2. Backbone and sidechain angles for residues mutated by PROSS in wtHis and fa-csHisH90G. Phi, psi, and chi angles are given for the NMR structure of wtHis, the Rosetta CM model of fa-csHisH90G, and the fa-csHisH90G crystal structure. Core residue position is number relative to wtHis, and residue identities are indicated for all proteins listed. The Rosetta CM model for fa-csHisH90G is given in italics and the crystal structure is indicated in bold.

Core position	**Protein**	**Residue**	**Phi (°)**	**Psi (°)**	**Chi$_1$ (°)**	**Chi$_2$ (°)**	**Chi$_3$ (°)**	**Chi$_4$ (°)**
12	wtHis	H12	178.891	164.056	39.509	-100.091		
	fa-csHisH90G	R12	-82.442	156.788	62.603	-161.665	-64.703	166.934
	fa-csHisH90G	R12	-98.474	137.979	-68.739	179.9	-64.269	-103.116
20	wtHis	A20	-155.473	165.144				
	fa-csHisH90G	L20	-149.679	155.334	61.504	83.584		
	fa-csHisH90G	L20	-108.154	144.258	-52.279	-178.62		
38	wtHis	N38	-95.059	160.14	-27.625	-56.249		
	fa-csHisH90G	Q38	-77.267	130.638	-174.993	178.024	-15.644	
	fa-csHisH90G	Q38	-81.458	132.439	-64.252	-167.8	-7.875	
45	wtHis	L45	-116.67	128.191	-67.342	150.836		
	fa-csHisH90G	F45	-115.463	109.866	-72.143	-100.876		
	fa-csHisH90G	F45	-117.578	110.858	-72.576	83.424		
49	wtHis	C49	45.052	41.562	66.808			
	fa-csHisH90G	N49	45.284	56.977	-64.459	-39.189		
	fa-csHisH90G	N49	-98.341	16.055	58.609	8.186		
88	wtHis	H88	-72.5	-167.107	-161.716	42.749		
	fa-csHisH90G	S88	-76.851	158.532	66.55			
	fa-csHisH90G	S88	-65.489	-26.876	56.58			
89	wtHis	H89	42.246	23.328	-55.999	125.973		
	fa-csHisH90G	N89	64.775	13.509	-64.399	-41.229		
	fa-csHisH90G	N89	-85.377	7.082	59.582	10.229		
91	wtHis	H91	-93.651	157.384	51.261	-103.365		
	fa-csHisH90G	R91	-80.566	145.335	-65.033	176.811	67.923	85.571
	fa-csHisH90G	R91	-87.105	159.114	-61.295	179.645	69.053	85.891
112	wtHis	T112	-144.836	132.808	-103.503			
	fa-csHisH90G	L112	-105.454	140.604	-66.882	171.968		
	fa-csHisH90G	L112	-98.061	128.738	-70.512	172.974		

Table C3. C_α-SBM contacts made by β9-β10 turn residues in wtHis and the fa-csHisH90G crystal structure.

wtHis		fa-csHisH90G	
β9/β10 turn residue	**Contacting residue**	**β9/β10 turn residue**	**Contacting residue**
H88	I55	**S88**	I55
	G56		D70
	H71		H71
	L73		L73
H89		**N89**	I55
			G56
			D57
	H71		Q60
	E105		T103
H90	K86	**G90**	K86
	K104		K104
	E105		E105
H91	I55	**R91**	I55
	D57		K59
	H71		
	F74		F74
	I85		I85
	K86		K86
	G87		G87
	T103		T103
	K104		K104
	E105		E105

Appendix D: Hisactophilin core cavity volume

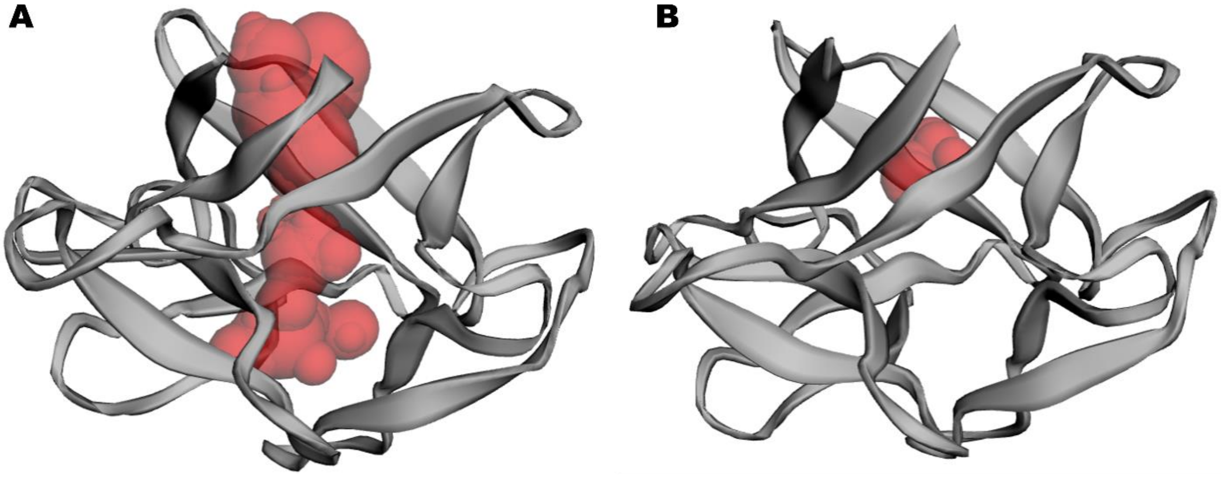


Figure D1. csHisH90G core cavity volume is significantly reduced compared to wtHis. Core cavity volumes were calculated for **(A)** wtHis and **(B)** csHisH90G using Computed Altas of Surface Topography of proteins (CASTp) (Tian et al., 2018). Cavity volumes were calculated using a 1.4 Å radius probe. CASTp identified a 64.8 Å^3 cavity that spans the entire wtHis core. csHisH90G exhibits a substantially smaller cavity of 2.57 Å^3, indicating that 3Foil core residues significantly reduce empty space in the hisactophilin core. No core cavity was detectable for 3Foil (data not shown).

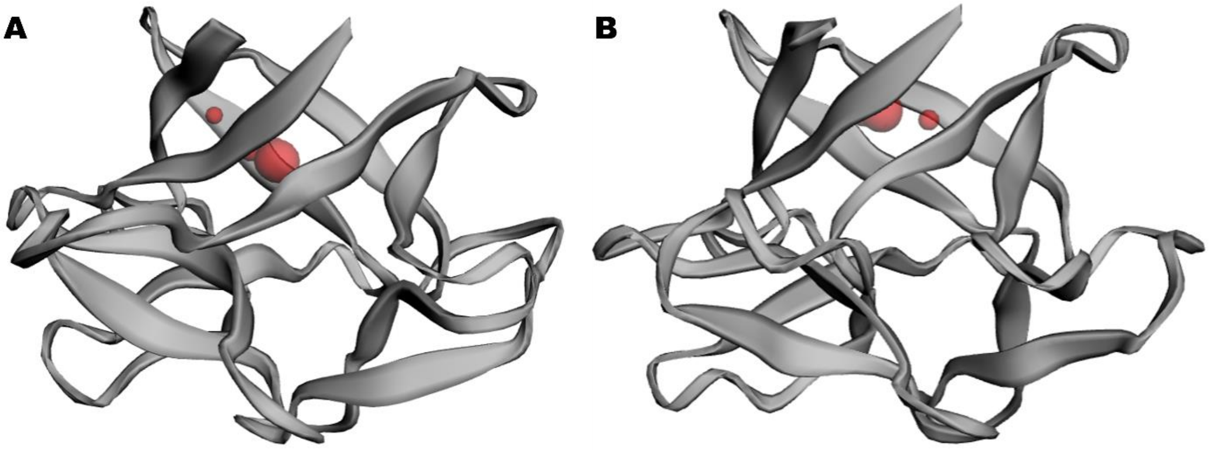


Figure D2. fa-csHisH90G has similar core cavity volumes in the computational and crystal structures. Core cavity volumes were calculated for **(A)** the Rosetta CM model and **(B)** crystal structure for fa-csHisH90G using Computed Altas of Surface Topography of proteins (CASTp) (Tian et al., 2018). Cavity volumes were calculated using a 1.4 Å radius probe. CASTp identified a 0.23 Å^3 cavity in the computational model for fa-csHisH90G and 0.34 Å^3 for the crystal structure.