University of Montana
ScholarWorks at University of Montana

Graduate Student Theses, Dissertations, & Professional Papers

Graduate School

2023

Characterization of Protein Folding Pathways and Structural Stability

Michael Tyler Rothfuss

Follow this and additional works at: https://scholarworks.umt.edu/etd

Let us know how access to this document benefits you.

Recommended Citation
Rothfuss, Michael Tyler, "Characterization of Protein Folding Pathways and Structural Stability" (2023). *Graduate Student Theses, Dissertations, & Professional Papers*. 12217.
https://scholarworks.umt.edu/etd/12217

This Dissertation is brought to you for free and open access by the Graduate School at ScholarWorks at University of Montana. It has been accepted for inclusion in Graduate Student Theses, Dissertations, & Professional Papers by an authorized administrator of ScholarWorks at University of Montana. For more information, please contact scholarworks@mso.umt.edu.

Characterization of Protein Folding Pathways and Structural Stability

by

Michael Tyler Rothfuss

B.S., Biochemistry, University of Montana, Missoula, MT, USA, 2017

Dissertation

presented in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in Biochemistry & Biophysics

The University of Montana
Missoula, MT

December 2023

Approved by:
Dr. Ashby Kinch, Dean of the Graduate School

Dr. Bruce E. Bowler, Dissertation Advisor, Department of Chemistry & Biochemistry
Dr. Mark L. Grimes, Committee Member, Division of Biological Sciences
Dr. Stephen R. Sprang, Committee Member, Division of Biological Sciences
Dr. Travis J. Wheeler, Committee Member, University of Arizona
Dr. Travis S. Hughes, Committee Member, Department of Biomedical and Pharmaceutical Sciences

Abstract

Proteins are large, flexible molecules with an extremely large number of potential conformations. Proteins expressed in cells traverse available conformations to reach a consistent, thermodynamically stable, biologically active structure through a process known as protein folding. The atomic composition of the protein, defined by a sequence of amino acid residues encoded in DNA as a gene, determines the protein folding pathway and ultimate native structure of the protein molecule. Understanding the relationship between the sequence of amino acids and the resulting protein structure has been a central challenge in protein research for decades. To fill this knowledge gap, we test the hypothesis that the distribution of conformers observed for a short protein sequence across all known protein structures reflects that sequence's intrinsic structural properties. Qualitative and quantitative predictions based on our model are tested against experimental data for protein stability and folding pathways.

Replica-exchange Monte Carlo simulations, data mining of the Worldwide Protein Data Bank (wwPDB), analysis of published protein stability data, thermodynamic and kinetic folding experiments, and X-ray crystallography were used to characterize the structural properties of amino acid sequences. The role of turn sequences in guiding the protein folding process was extensively characterized by the combined methods. Turn composition, structural preferences, and cooperation with neighboring residues determined whether a turn had an active, passive, or counter-active role in a protein's folding process.

Proline-rich turns, NPSNP and KPSDP, from the two-helix bundles found in bacterial type III secretion system needle proteins form native-like structure early in the folding process. Each of these

turns are flanked by sequences with very high helix propensity that, when oriented by the turn, can actively nucleate the hydrophobic core of the protein. The hydrophobic turn, MGYE, from the three-helix bundle UBA(1) also forms native-like structure early in the folding process. This turn structure places the Met (M) and Tyr (Y) residues together, nucleating the hydrophobic core of UBA(1). These two residues can then stabilize the adjacent helices to form a Helix-Turn-Helix structure. The second, proline-containing turn in UBA(1), ASYNNP, forms non-native structure early in the folding process. This turn restructures late in the folding process when the third helix docks to the previous Helix-Turn-Helix structure. Each of the active turns characterized (NPSNP, KPSDP, and MGYE) direct the folding process by nucleating the protein's hydrophobic core.

A general purpose computational method to model the local structural properties of protein sequences was developed from data mined from the wwPDB. Turn mechanisms can be rapidly characterized using the tool, EmCAST, in conjunction with a PDB structure of the protein of interest. The impact of surface mutations on protein stability can also be scored by EmCAST. Models and calculations were extensively validated against experimental data for multiple protein and peptide systems. Calculations for stabilizing mutations at well-structured positions in UBA(1) produced a near perfect correlation with experimental measurements (R^2 = 0.97). A user-friendly web interface to the software was developed to share the method with other protein researchers. Our model provides key insights into the protein sequence/structure relationship that can be used to characterize protein surface stability, identify regions with dynamic structure, and predict protein folding intermediates.

Acknowledgements

Bruce Bowler

I had one key interest when I first entered the field of biochemistry: how does genetic information become functionally active in a living system? I had no idea how to tackle the enormous challenge of understanding the relationship between protein sequence, structure, and function. Your insight, guidance, and enthusiasm have been invaluable. Discussing and exploring protein models with you has put me on a wonderful journey that I will always remember fondly.

The UM Research Community

I'd like to thank my committee for their interest in my research, knowledge, and feedback. Learning from members of the Bowler lab and other protein researchers at UM has been instrumental in my development as a protein scientist. Thank you all for working with me and sharing your knowledge.

Dustin Becht, Ariel Frederick, and Kassie Boshae

Working together has been a blast! From serious scientific discussions to crude humor, you three kept life in Missoula fun and interesting. Learning from each other, bouncing ideas back and forth, and laughing together made my time in lab especially enjoyable.

Chloé and Allison Öst Rothfuss

The two of you motivate me to always do my best. I cherish the time we've spent together. I already see my curiosity, kindness, and strong-willed nature in each of you. Allison, you are so sweet, thoughtful, and determined to keep up with your big sister. Chloé, after hearing from mom that I understand how nearly everything in this world works, your eyes widened and you asked me: do you know how life works? I'm excited to teach you anything you'd like to know about the world. I am so proud of both of you. Each of you continue to impress me. I love you both, I am truly lucky to have such wonderful daughters.

Blake and Laurie Rothfuss

The path I've taken through life has been dark, turbulent, and riddled with mistakes. I would not be where I am today without your unwavering support. Thank you for never giving up on me.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AMU: Atomic Mass Unit
CD: Circular Dichroism
CI-2: Chymotrypsin Inhibitor 2
C_α: Protein/Peptide Alpha Carbon
DHAP: 2,5-Dihydroxyacetophenone
EmCAST: Empirical C-Alpha Stability Tool
FPLC: Fast Protein Liquid Chromatography
FRET: Förster Resonance Energy Transfer
GST: Glutathione S-transferases
GdmCl: Guanidinium Chloride
GdnHCl: Guanidine Hydrochloride
HEPES: 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HHR23A: human homolog of Rad23
HPLC: High Performance Liquid Chromatography
IAED: Suffix indicating labeling by IAEDANS
IAEDANS: 5-({2-[(iodoacetyl)amino]ethyl}amino)naphthalene-1-sulfonic acid
IODO: Suffix indicating labeling by Iodoacetamide
M/Z: Mass to charge ratio
MALDI-ToF: Matrix-assisted laser desorption/ionization Time of Flight
MES: 2-(N-morpholino)ethanesulfonic acid
MOPS: 3-(N-morpholino)propanesulfonic acid
MSA: Multiple Sequence Alignment
MT: Mutant
MWCO: Molecular Weight Cut-Off
NMR: Nuclear magnetic resonance
NTL9: N-Terminal Domain of Ribosomal Protein L9
PDB: Protein Data Bank
PMSF: phenylmethylsulfonyl fluoride
RCSB: Research Collaboratory for Structural Bioinformatics
REMC: Replica-Exchange Monte-Carlo
SASA: Solvent Accessible Surface Area
SDS-PAGE: sodium dodecyl sulfate–polyacrylamide gel electrophoresis
TCEP: tris(2-carboxyethyl)phosphine
TEV: Tobacco etch virus protease
TFA: Trifluoroacetic acid
UBA(1): Ubiquitin Associated Domain 1 of HHR23A
WT: Wild-Type
wwPDB: World Wide Protein Data Bank

Chapter 1: Introduction

The central dogma of molecular biology describes the flow of genetic information in a biological system[1]. Biomolecules becomes progressively functional across the principal pathway of the central dogma: DNA → RNA → Protein. The functional properties of a protein are determined by the 3D structure adopted by the flexible protein molecule. The unique sequence of amino acid residues that compose a protein is the principal piece of information, originally encoded in DNA as a gene, that defines the 3D structure of a protein[2]. Attempts to rationally edit, design, or modulate protein function require understanding the relationship between amino acid sequence and protein structure.

Before the relationship between protein sequence and structure can be characterized, the 3D structures of protein sequences need to be experimentally determined at atomic resolution. Countless researchers have isolated proteins from living systems and determined protein structure using X-ray crystallography and NMR techniques. The thousands of solved structures are openly shared between researchers in the world wide protein data bank (wwPDB)[3]. Recent breakthroughs in AI technology have developed deep learning techniques to analyze the large dataset of known protein structures and predict protein structure from sequence[4]. The success of AI models to predict

protein structure are overshadowed by their failure to model the significant structural effects of small changes in protein sequence[5]. This indicates AI methods have relied too heavily on sequence homology models instead of accurately modeling the thousands of atomic forces involved in structuring a protein. A significant knowledge gap remains due to the uncharacterized forces involved in defining protein structure. How a protein adopts its native structure, changes shape while it is functionally active, and responds to mutations are key questions that remain unanswered. These key questions will be answered if the protein folding problem can be solved.

Protein folding is the process wherein an unstructured protein arranges into its thermodynamically stable structure. The timescale of the folding process is generally in the millisecond to second range[6]. Protein structures are flexible and have many degrees of freedom. A small, 50-residue protein will have 98 rotation points across its peptide backbone (φ and ψ dihedral angles[7]). Underestimating the number states available at each rotation point to three dihedral angle values highlights the extreme magnitude of possible conformations: 3^{98}, or ~$5.7 \cdot 10^{46}$, conformers. An unfolded protein randomly moving through all of these available conformations would take years to find its thermodynamically stable structure, not seconds[6]. This apparent paradox demonstrates that protein folding is

not a random process; the protein's distinct sequence of amino acids must bias and direct the molecule toward a much smaller subset of available conformations.

The goal of this dissertation is to characterize the structural preferences of amino acid sequences that bias and direct protein folding pathways. We will test the hypothesis that the distribution of conformers observed for a short protein sequence across all known protein structures reflects that sequence's intrinsic structural properties. Previous work has demonstrated that the structural properties of an amino acid residue are influenced by the identity of neighboring residues[8]. We have chosen to focus on the properties of 4-residue protein fragments, tetrads, to account for these observed context-dependent effects. Previous structural studies using protein fragments have had limited success. A common method has been to cluster fragments of similar structure and to analyze the sequence motif of matching samples[9][10]. This cluster/motif strategy neglects two key pieces of information: the context-dependent effects of amino acid identity on structural properties (sequence motifs only retain information on the relative probability of amino acids at each position in the dataset) and the tendency for amino acid sequences to adopt multiple conformers. Characterizing the different conformers that a 4-residue sequence adopts rectifies these two issues.

The thousands of experimentally determined protein structures in the RCSB wwPDB[3] were analyzed to characterize the structural properties of each observed 4-residue tetrad. Short protein sequences, taken from the turn regions of natural proteins, were introduced into a low complexity, helical peptide to study their impact on the folding process. Full-atom simulations, tetrad models from the wwPDB, and experimental methods were used to study structure in the selected sequences. Energy calculations derived from the tetrad models were tested against previous stability measurements reported in literature for validation. Stabilizing mutations predicted by our energy calculations were introduced into a three-helix bundle, UBA(1)[11], to further test model accuracy and probe the folding mechanism of UBA(1). Our general purpose model of protein stability provides key advancements in understanding the transiently formed structures that occur during protein folding, modeling structural dynamics, and predicting the effects of surface point mutations. Software developed in conjunction with this research provides fast and simple tools to analyze and rationally manipulate the sequence/structure/function relationship in proteins.

Chapter 2: Chain Reversal Sequences

2A: Introduction

The relationship between primary, secondary, and tertiary protein structure has been studied for decades. Anfinsen's thermodynamic hypothesis asserts that the primary structure specifies a protein's tertiary structure by encoding atomic interactions that arrange a protein into its most thermodynamically stable conformation[2]. Proteins that obey Anfinsen's thermodynamic hypothesis provide ideal systems to study the protein folding code. A small protein that obeys Anfinsen's hypothesis is exceptionally ideal for folding studies as it minimizes the complexity of interactions necessary to define a stable tertiary structure.

One of the simplest tertiary structures is a two-helix bundle. Composed of only two alpha-helices connected by a loop, it is distinguished from pure secondary structure by two key structural features between the helices: the interconnecting loop and the hydrophobic interface. Hydrophobic residues are widely found packed together in the core of a protein and provide for the bulk of its stability[12]. The hydrophobic sidechains of varying sizes have been proposed to fit together like a jigsaw puzzle to encode a precise tertiary structure[12]. Contrary to this model, as many as 10

hydrophobic residues in the core of T4 Lysozyme were simultaneously mutated without destroying the structure or function of the protein[13]. This suggests other parts of the protein's primary structure are important for specifying tertiary structure. In the case of a two-helix bundle, this would be the loop connecting the two helices.

The interhelical sequence in a two-helix bundle can be characterized as having a counter-active, a passive, or an active role in establishing tertiary structure. Loop sequences which favor a conformation that is different from the tertiary structure are counter-active. In this case, favorable interactions from the hydrophobic interface must overcome structural preferences encoded in the loop sequence to bend the loop into a strained conformation. A highly flexible loop sequence falls into the passive category; hydrophobic interactions establish tertiary structure without resistance or assistance. An active loop sequence favors the tertiary structure independent of any hydrophobic interactions. In this case, both the loop and the hydrophobic interface work cooperatively to establish tertiary structure. Active loop sequences may play key roles in guiding protein folding when multiple hydrophobic interfaces are present. In addition to encoding a specific helix-helix orientation, active loops may influence the order of folding events as tertiary structure is established.

Several interhelical sequences, referred to as loops or turns, have been selected from natural proteins for characterization. The selected turn sequences were introduced into a low complexity poly-alanine helix. This system will be used to assess the impact of different turn sequences on tertiary structure in the absence of a well-defined hydrophobic core. Replica Exchange Monte Carlo (REMC) simulations were performed to predict and analyze the effects of the introduced turn sequences upon the poly-alanine helix at atomic resolution. A subset of simulated peptides were synthesized and characterized *in vitro* to experimentally test the accuracy of simulations.

2B: Simulation Design and Data Analysis

2B.1: Sequence Designs

Alpha helices in natural proteins rely on a combination of intrinsic helix propensity from primary structure and on stabilizing interactions between side chains. Isolation of natural helical sequences has demonstrated that helix propensity often does not provide sufficient stability for the helical structure to form under physiological conditions[14]. Helical propensity studies have utilized alanine, the strongest helix former, to form isolated alpha-helices in aqueous conditions[15]. Regularly spaced lysines can be added to enhance solubility of the polyalanine sequence. We have utilized this design to form our low complexity AK42 sequence, $(AAAAAK)_7$, that hosts turn sequences in our structural studies. The lysines in this sequence were repositioned in several designs (AK42r2 through AK42r6) to prevent steric clashes in the expected helix-helix interface. An all-alanine host, A42, was used to assess turn specificity when multiple helix-helix interfaces are available.

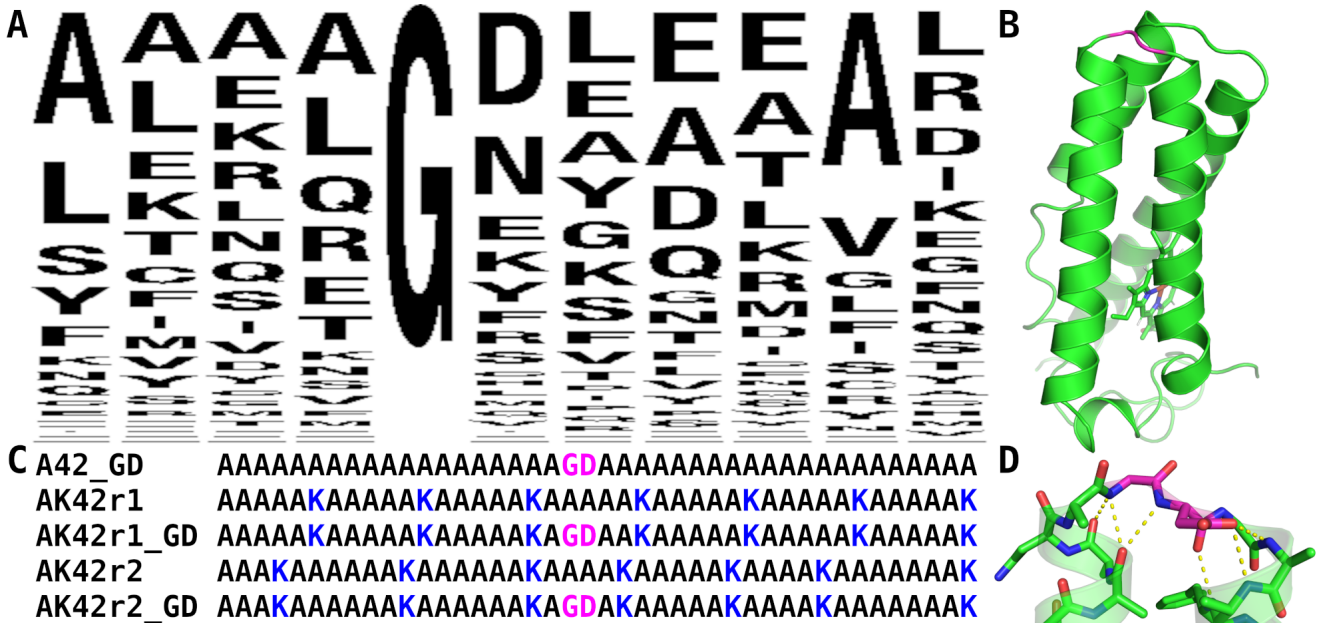

Figure 1: GD sequence design from the helical hairpin database

(A) Sequence logo for proteins in the helical hairpin database. (B) Structure of cytochrome c' from Rhodopseudomonas capsulata (PDB: 1CPQ). The GD loop sequence is rendered in magenta. (C) Sequence designs for AK42 and AK42_GD peptides. (D) Hydrogen bonds found in the GD loop of cytochrome c' (PDB: 1CPQ).

Compact two-helix bundles, called helical hairpins, have been collected into a database of structures[16]. Sequence alignment of database members reveals a prominent GX motif in the turn between the helices (Figure 1A). A representative structure, cytochrome *c*' from *Rhodopseudomonas capsulata* (PDB: 1CPQ), containing this motif is shown in Figure 1B. The consensus GD sequence encodes a helix stop/start in the primary structure: glycine is the most common residue found after an alpha-helix and aspartate is the most common N-terminal helix capping residue[17]. Glycine in cytochrome *c*' occupies its characteristic left-handed alpha-helical region of the Ramachandran plot, causing bifurcated hydrogen bonding from Gly and Asp backbone-amide NH's to partially cap the preceding helix (Figure

1D). Aspartate's sidechain forms bifurcated hydrogen bonds to cap the following helix (Figure 1D) and has the added benefit of stabilizing the helix macroscopic electrostatic dipole[18]. The GD motif was introduced into our AK42r1 sequence (Figure 1C) to test if the GD residues are sufficient to recreate the helix hairpin. A second pair of peptides were designed (AK42r2 and AK42r2_GD) wherein the lysine residues are repositioned to eliminate steric conflicts in the expected fold.

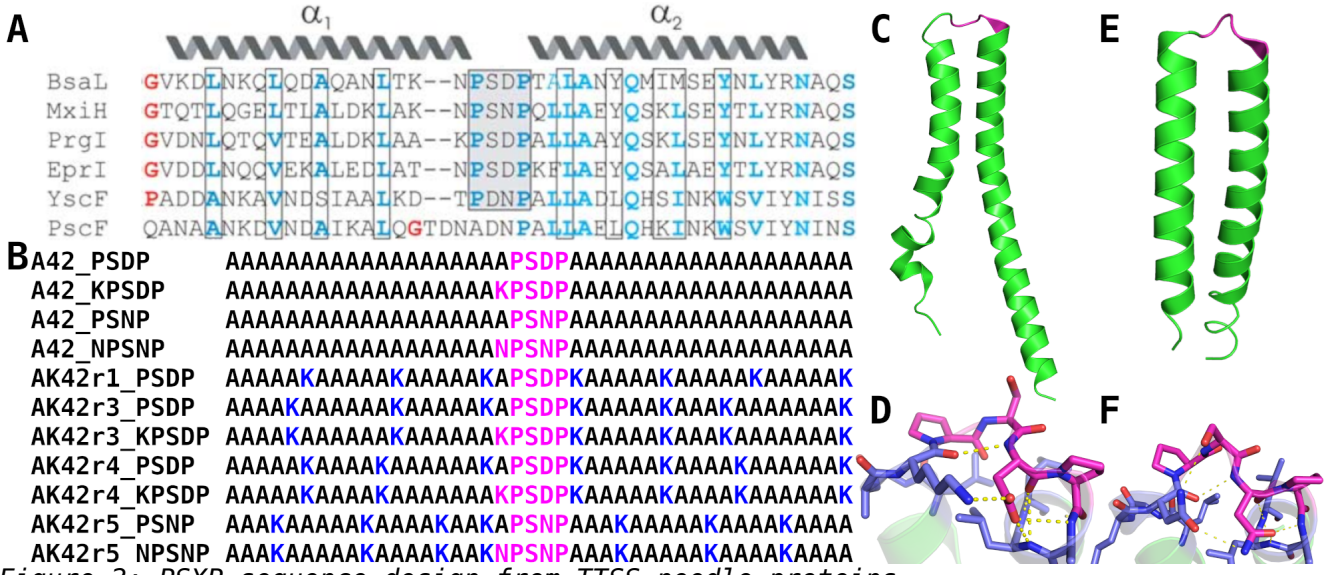

Figure 2: PSXP sequence design from TTSS needle proteins

(A) Sequence alignment of select TTSS needle proteins. (B) Sequence design for the PSDP and PSNP peptides based on PrgI and MxiH needle proteins, respectively. (C) NMR solution structure of the PrgI needle protein (PDB: 2JOW). (D) Hydrogen bonding network of the KPSDP loop (PDB: 2JOW). (E) Crystal structure of the MxiH needle protein (PDB: 2CA5, Chain B). (F) Hydrogen bonding network for the NPSNP loop (PDB: 2CA5).

Two helix bundles found in bacterial type III secretion system (TTSS) needle proteins show conserved features in both the hydrophobic interface and turn sequence (Figure 2A)[19]. Two prolines are well conserved at favorable positions relative to the end of

Helix-1 (C_{cap}) and the start of Helix-2 (N_{cap}). The first proline is positioned to terminate Helix-1 (C_{cap}^{+3}) and the second proline is optimal to initiate Helix-2 (N_{cap}^{+1}). The aspartate and asparagine residues in the PS(D/N)P motif are favorable N_{cap} residues wherein their sidechains cap helix 2 (Figure 2D,F)[17]. The (K/N)PS(D/N) residues form a type I beta-turn with a hydrogen bond between the backbone carbonyl (K/N) and the backbone amide nitrogen (D/N). This turn structure is reinforced by polar interactions between the K/N and D/N residues (Figure 2D,F). The NPSNP asparagine provides additional features not found in the KPSDP sequence; hydrogen bonding occurs between the asparagine carboxamide nitrogen and the nearby serine's (NPSNP) hydroxyl group (Figure 2F). These 2-3 hydrogen bonds that stabilize the type I beta-turn in the TTSS needle proteins may be sufficient to define the turn conformer locally – without aid from the hydrophobic interface.

The AK42r1_PSDP sequence was designed with only the PSDP motif inserted to assess if the rigidity and helix terminating/initiating properties of proline are sufficient to recreate the turn topology (Figure 2B). The AK42 sequence was revised to optimize charge distribution for the AK42r1_PSDP cluster-1 structure from simulation results (AK42r3_PSDP) or the PrgI NMR structure (AK42r4_PSDP). The position of one lysine near the turn was repositioned to form the AK42r3_KPSDP and AK42r4_KPSDP sequences. The AK42r5 sequence was

designed to optimize charge distribution for the MxiH fold and hosts the PSNP and NPSNP turn sequences. Each selected turn sequence was also included in the all-alanine host, A42.

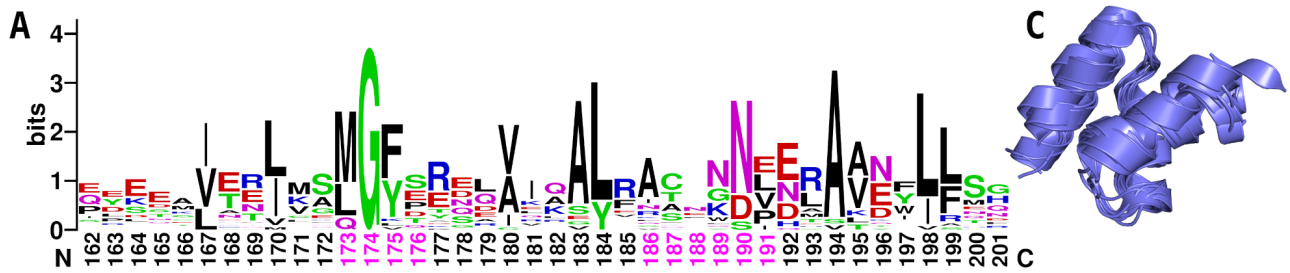

Figure 3: NCLoop sequence design from ubiquitin associated domains

(A) Sequence logo of various ubiquitin associated domains. (B) Sequences of the two HHR23A UBA domains and the designed AK42 peptides. (C) Structural overlay of select UBA domains (PDBs: 1IFY, 1DV0, 2QSF, 2D9S, 1VEG, 1YLA, 2DAK, 2LBC, 1WIV, and 2G3Q). (D) NMR solution structure of HHR23A UBA(1) (PDB: 1IFY).

Ubiquitin associated (UBA) domains adopt a consistent three-helix bundle fold (Figure 3C) but exhibit high sequence diversity (Figure 3A). Sequence conservation occurs in residues for both the hydrophobic core of the domain and in the turn regions (Figure 3A). The sequence from the first ubiquitin associated domain, UBA(1), of the human homologue of yeast Rad23A (HHR23A) was selected for characterization (Figure 3B). Two turn sequences are present: MGYE (turn 1) and ASYNNP (turn 2) (Figure 3D). In turn 1, the highly conserved glycine residue populates the left-handed helix region of the Ramachandran plot and positions the adjacent aliphatic Met and

Tyr residues to pack together. Only the Glycine residue is in a statistically favorable position relative to the C_{cap} or N_{cap} positions for helices 1 or 2, respectively[17]. The glutamate may stabilize helix 2 through polar i+3 interactions with a nearby Arg residue[18]. In turn 2, a buried serine forms trifurcated hydrogen bonds with nearby carbonyls that may or may not provide productive interactions for establishing turn structure. For the C_{cap} of helix 2, only the proline (C_{cap}^{+4}) is statistically favorable. The N_{cap} of helix 3 favors proline (N_{cap}^{+1}) and the two asparagines (N_{cap}^{-1}, N_{cap})[17]. The second asparagine's sidechain carboxamide forms bifurcated hydrogen bonds with helix 3's free backbone amide nitrogens. Together the SYNN residues form a type 2 beta-turn with backbone hydrogen bonding between the serine and second asparagine. These two turn sequences were introduced into our AK42 peptide to make the AK42r1_NCloops peptide. A variant with repositioned lysines, AK42r6_NCloops, was also created. A minimal hydrophobic core was designed for the AK42r6_NCloops2 sequence using PyMOL to assess whether support from a hydrophobic core is required for the correct helix topology to form. A peptide with only the MGYE sequence, AK42r6_Nloop, was designed to assess the folding mechanism of turn 1 in isolation. The double (NCloops) and single (Nloop) turn sequences were introduced into the all-alanine peptide to form the A42_NCloops and A42_Nloop sequences.

2B.2: Simulation Software

The CAMPARI molecular modeling software[20], version 2.0, was used to perform Replica Exchange Monte Carlo (REMC) simulations on our polyalanine peptides. Peptides start in a random conformation, generated by CAMPARI, with N-terminal acetylation and a C-terminal methylamide. An aqueous sphere with a 200 Ångstrom radius and an atom-based soft-wall boundary houses the simulated peptide. The system is charge balanced with 2.48 mM NaCl. Lennard-Jones parameters are provided for an implicit-solvation model by ABSINTH[21]. Charges and essential bond parameters are provided by OPLS[22]. A total of 100,000,000 steps are performed in each simulation; the first 6,000,000 steps are discarded as an equilibrium period. Sixteen temperatures are simulated: 285K, 305K, 315K, 323K, 331K, 339K, 347K, 355K, 363K, 371K, 379K, 390K, 405K, 425K, 445K, and 465K. Exchanges between adjacent temperatures are attempted every 2000 steps. Structures are exported every 2000 steps, providing 47,000 structures for each simulated temperature. A detailed list of the selected CAMPARI parameters is provided in Table 1.

FMCSC_PDBANALYZE = 0	FMCSC_SC_BONDED_T = 1.0	FMCSC_RIGIDFREQ = 0.05	FMCSC_SEQREPORT = 1
FMCSC_SHAPE = 2	FMCSC_SC_EXTRA = 0.0	FMCSC_CHIFREQ = 0.3	FMCSC_VDWREPORT = 1
FMCSC_SIZE = 200	FMCSC_SC_POLAR = 1.0	FMCSC_CRFREQ = 0.1	FMCSC_ELECREPORT = 1
FMCSC_BOUNDARY = 4	FMCSC_SC_IMPSOLV = 1.0	FMCSC_OMEGAFREQ = 0.3	FMCSC_INTERREPORT = 1
FMCSC_RANDOMIZE = 1	FMCSC_SAVPROBE = 2.5	FMCSC_PIVOTRDFREQ = 0.3	FMCSC_XYZOUT = 2000
FMCSC_NRSTEPS = 100000000	FMCSC_IMPDIEL = 78.2	FMCSC_PIVOTSTEPSZ = 10.0	FMCSC_ENOUT = 2000
FMCSC_EQUIL = 6000000	FMCSC_FOSTAU = 0.25	FMCSC_TRANSSTEPSZ = 10.0	FMCSC_PHOUT = 100000000000000
FMCSC_ENSEMBLE = 1	FMCSC_FOSMID = 0.1	FMCSC_ROTSTEPSZ = 20.0	FMCSC_TOROUT = 10000000000000

FMCSC_DYNAMICS = 1	FMCSC_SCRMODEL = 2	FMCSC_CLURBFREQ = 0.1	FMCSC_ACCOUT = 2000
FMCSC_REPLICAS = 16	FMCSC_SCRTAU = 0.5	FMCSC_CLURBMAX = 4	FMCSC_RSTOUT = 20000
FMCSC_REDIM = 1	FMCSC_SCRMID = 0.9	FMCSC_COUPLERIGID = 1	FMCSC_POLOUT = 2000
FMCSC_REMC = 1	FMCSC_INTERMODEL = 1	FMCSC_ROTFREQ = 0.1	FMCSC_RHCALC = 10000000000000
FMCSC_RESWAPS = 15	FMCSC_ELECMODEL = 2	FMCSC_RIGIDRDFREQ = 0.1	FMCSC_PCCALC = 10000000000000
FMCSC_RENBMODE = 2	FMCSC_CUTOFFMODE = 4	FMCSC_PKRFREQ = 0.05	FMCSC_SAVCALC = 1000000000000
FMCSC_REFREQ = 2000	FMCSC_NBCUTOFF = 10.0	FMCSC_PKRRDFREQ = 0.02	FMCSC_COVCALC = 1000000000000
FMCSC_UAMODEL = 0	FMCSC_ELCUTOFF = 14.0	FMCSC_PUCKERSTEP_DI = 4.0	FMCSC_ANGCALC = 1000000000000
FMCSC_SIGRULE = 1	FMCSC_CHECKFREQ = 50000	FMCSC_PUCKERSTEP_AN = 2.0	FMCSC_SEGCALC = 2000
FMCSC_EPSRULE = 2	FMCSC_USESCREEN = 1	FMCSC_NRCHI = 2	FMCSC_DIPCALC = 1000000000000
FMCSC_SC_IPP = 1.0	FMCSC_BARRIER = 10000.0	FMCSC_CHICYCLES = 4	FMCSC_POLCALC = 2000
FMCSC_SC_ATTLJ = 1.0	FMCSC_SC_ZSEC = 0.0	FMCSC_CHICOUPLE = 0	FMCSC_DSSPCALC = 100000000000
FMCSC_SC_WCA = 0.0	FMCSC_SC_DSSP = 0.0	FMCSC_CHIRDFREQ = 0.6	FMCSC_HOLESCALC = 10000000000
FMCSC_MODE_14 = 1	FMCSC_SC_TOR = 0.0	FMCSC_CHISTPESZ = 30.0	FMCSC_DIFFRCALC = 10000000000
FMCSC_FUDGE_ST_14 = 1.0	FMCSC_SC_DREST = 0.0	FMCSC_PHFREQ = 0.0	FMCSC_CONTACTCALC = 2000
FMCSC_FUDGE_EL_14 = 1.0	FMCSC_SC_TABUL = 0.0	FMCSC_PIVOTMODE = 1	FMCSC_CONTACTMIN = 3.5
FMCSC_SC_BONDED_B = 0.0	FMCSC_SC_POLY = 0.0	FMCSC_COUPLE = 0	FMCSC_XYZPDB = 4
FMCSC_SC_BONDED_A = 1.0	FMCSC_GHOST = 0	FMCSC_ALIGN = 4	FMCSC_XYZMODE = 1
FMCSC_SC_BONDED_I = 1.0			

Table 1: CAMPARI Parameters

Provided for simulation reproducibility.

2B.3: Structural Analysis

Clustering was used to identify dominant structures within each simulation. The GROMACS software suite[23] was used to cluster structures using the gromos method (gmx cluster -method gromos). All alpha carbons were used in the clustering procedure with a 0.15 angstrom RMSD cutoff. Only every 10th structure was used to reduce computation complexity. This clustering procedure was repeated using the clustered structures from all temperatures and replicates to determine the prevalence of structures across different simulations. This process identified consensus structures and determined their temperature dependence. Considerable structural information is missed

in this approach when the clustered structures only represent a small fraction of the sampled conformers, which is the case for many temperatures and simulations.

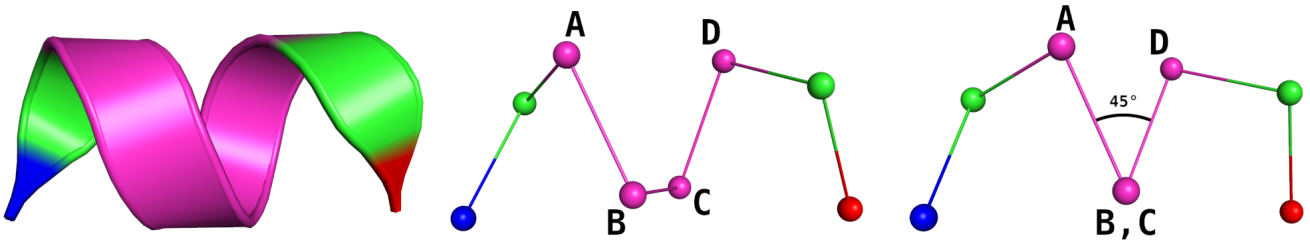

Figure 4: The four-residue alpha-carbon dihedral angle, τ.

Peptide dihedral angles (φ, ψ, and ω)[7] are widely used to precisely describe the backbone geometry in protein structures. Reduction to only φ and ψ allows convenient visualization of backbone geometry within 2D (φ, ψ) Ramachandran plots. A broader, less precise dihedral angle (τ)[24] can be measured between 4 sequential alpha carbons to describe the overall direction of the protein chain (Figure 4). Both alpha helices and beta sheets have distinct patterns of repeating τ centered around values of 50° and -145°, respectively; turns exhibit a wide range of τ combinations. This further simplification from 2D (φ, ψ) to 1D (τ) allows the backbone geometry of the whole protein to be visualized in a single plot. Data from thousands of structures can be rendered in this plot as a heatmap to show the τ conformational distribution of all structures produced by simulations to identify structural tendencies in the absence of a well-defined tertiary structure.

This strategy to visualize τ geometry across thousands of simulated structures for one protein can be adapted to visualize the structural tendencies of specific tetrads across many different proteins. Structural data harvested from the RCSB wwPDB[3] was used to generate analogous heatmaps for our simulated peptide sequences. These wwPDB heatmaps enable comparison of the tetrad conformer distributions found in reality to those of our simulated systems. Lower sample counts in the wwPDB require smoothing of the dataset, causing heatmap peaks to broaden. The methods used to generate these heatmaps from the wwPDB are detailed in Chapter 3.

In addition to overall backbone direction, the τ dihedral angle describes the relative orientation of sidechains for the middle two residues; B and C in the "ABCD" tetrad (Figure 4). The A-ψ and φ-D torsion angles can rotate freely without influencing the τ dihedral angle, offering a wide range of relative orientations between A and D side chains at a specific τ. Rotation at φ-B or C-ψ has a minor influence on τ and instead mostly effects the pseudo-bond angles between A-B-C or B-C-D, respectively. Rotation at either B-ψ or φ-C directly influences τ and the relative orientation of the B and C sidechains. However, equal and opposite rotation between the B-ψ and φ-C torsion angles preserves both τ and the relative orientation of B and C sidechains. This neutral effect of opposite rotation between B-

ψ and φ-C torsion angles is apparent when comparing type I and type II beta turns.

The alignment of B and C sidechains, ie: the torsion angle between B-C_β, B-C_α, C-C_α, and C-C_β, is similar to τ; extended conformers (τ≈180°) orient B and C sidechains in opposite directions whereas compact conformers (τ≈0°) orient the B and C sidechains in similar directions. When a consecutive series of τ are known (ie: for ABCD, BCDE, CDEF tetrads of the ABCDEF sequence), the approximate positions of C_α atoms for each residue can be modeled whereas sidechain orientation can be approximated for only the inner residues (BCDE).

2C: Materials and Methods

Name	Sequence	Sim. Analog
AK42_W	AAAAAKAAAAAKAAAAAKAAAAAKAAAAAKAAAAAKAAAAWK	AK42r1
PSDP_W	AAAAAKAAAAAKAAAAAKAPSDPKAAAAAKAAAAAKAAAAWK	AK42r1_PSDP
AK42_CW	AAAAKAAACAAKAAAAAKAAAAAKAAAAAKAAAKAAAAAAWK	AK42r3
PSDP_CW	AAAAKAAACAAKAAAAAKAPSDPKAAAAAKAAAKAAAAAAWK	AK42r3_PSDP

Table 2: Peptide sequences designed for in vitro characterization

The N-terminus of each peptide was acetylated. The C-terminus was methyl amidated (AK42_W, PSDP_W, AK42_CW) or amidated (PSDP_CW).

Blocked peptides analogous to simulated designs were ordered from Genscript (Table 2). A tryptophan was added at the C-terminus of the design to facilitate peptide concentration measurements. A cysteine residue was also introduced into the N-terminal region of two designs to enable fluorescent labeling. The fluorophore labeled

N-terminal Cys and C-terminal Trp enable FRET measurements which can estimate distances between these two residues.

2C.1: Peptide Purification

Peptides were resuspended in 0.065% TFA in water and purified by reverse phase HPLC. The Agilent Technologies 1200 Series HPLC was used with a Vydac C18 column. An elution gradient of 0.05% TFA in acetonitrile (5%→25%/10 mL, 25%→34.5%/9.5 mL; 1 mL/minute) was used to elute the peptides. Absorbance as 280nm was used to identify the eluting peptides. The collected fraction was flash frozen and lyophilized overnight in a VirTis Sentry lyophilizer. Samples were resuspended in either CD Buffer (20 mM MOPS, 100 mM NaCl, pH 7.0) or Labeling Buffer (100 mM $NaPO_4$, pH 7.2).

2C.2: Peptide Labeling

Peptide	Theoretical Mass
AK42_CW	3606.3
AK42_CW-IODO	3664.3
AK42_CW-IAED	3913.6
PSDP_CW	3704.3
PSDP_CW-IODO	3762.4
PSDP_CW-IAED	4011.7

Table 3: Expected masses of labeled peptides.

Theoretical masses were calculated using the Expasy ProtParam tool[25]. 42 amu was added for the N-terminal acetylation, 14 amu was added for the C-terminal methyl amidation. 58 amu was added for the iodoacetamide label. 307 amu was added for the 1,5-IAEDANS label.

The AK42_CW and PSDP_CW peptides were reacted with either iodoacetamide (184.964 g/mol) or 1,5-IAEDANS (434.25 g/mol). The peptides, TCEP, and labeling reactants were all dissolved in Labeling

Buffer (100 mM $NaPO_4$, pH 7.2). TCEP was added to the peptides in 5X molar excess and gently stirred for 1 hour under Argon gas. The labeling reactants were also kept under Argon gas during this hour. Labeling reagent (iodoacetamide or 1,5-IAEDANS) was added dropwise at 400X molar excess to the peptide/TCEP solution. The solution was mixed gently under Argon for 1-2 hours in the dark. The reaction was quenched by the addition of β-mercaptoethanol (4000X molar excess). Samples were then repurified by HPLC.

The masses of labeled peptides (Table 3) were confirmed by MALDI-ToF using a Bruker microflex mass spectrometer. MALDI samples were prepared using a DHAP/TFA matrix. Measured masses were all within 3 amu of the expected values (Figure 5).

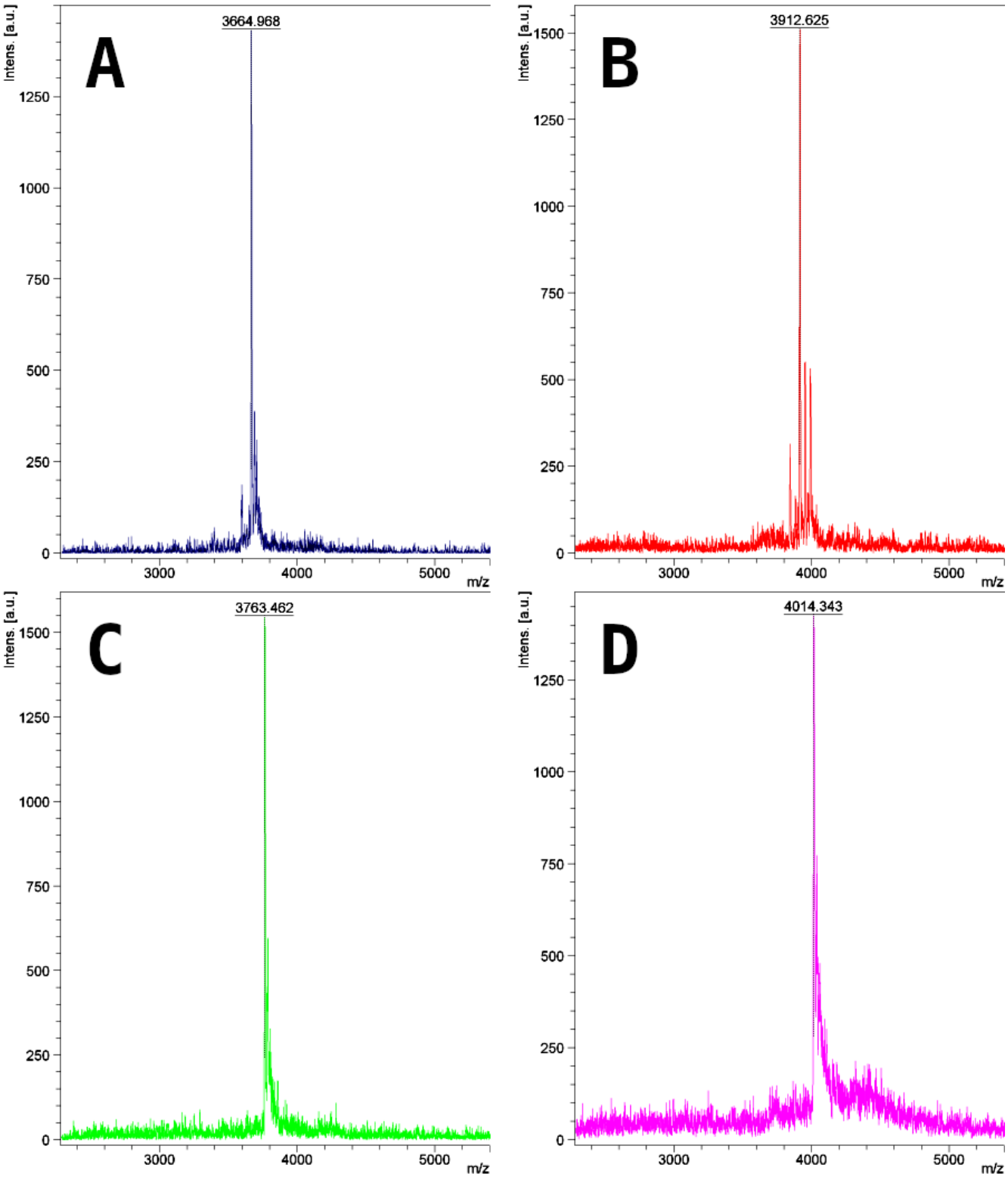

Figure 5: MALDI-ToF results for labeled peptides.

Results for the (A) AK42_CW-IODO, (B) AK42_CW-IAED, (C) PSDP_CW-IODO, and (D) PSDP_CW-IAED peptides are included. No peaks were present in the 5,000 to 10,000 m/z range.

2C.3: Spectroscopic Measurements

Peptide samples were prepared in CD Buffer (20 mM MOPS, 100 mM NaCl, pH 7.0) at 10-20 µM concentrations. An Applied Photophysics Chirascan CD Spectrophotometer was used to take circular dichroism (CD) and fluorescence measurements. The sample was held in a 4x4 fluorescence cuvette. CD measurements were taken between 250 nm and 200 nm with a 1 nm step and a 1 nm bandwidth. Data were collected for 3 seconds per point. Sample temperatures were kept at 4°C. Thermal melts monitored by CD at 222 nm were implemented using a Peltier temperature controller. Temperature was varied from 4°C to 90°C during the melt.

Steady-state FRET measurements were done for the AK42_CW and PSDP_CW peptides labeled with 1,5-IAEDANS. Fluorescence was measured 90° perpendicular to the 280 nm excitation beam with a 305 nm cutoff filter and a 5 nm bandwidth. Emissions were measured between 600 nm and 300 nm. The tryptophan has a peak emission at 350 nm. The 1,5-IAEDANS has a peak emission at 490 nm. Tryptophan fluorescence for the blocked peptides (AK42_CW-IODO and PSDP_CW-IODO) was used to deconvolute the overlapping emission spectra in the labeled peptides (AK42_CW-IAED and PSDP_CW-IAED). The isolated tryptophan emission spectra were scaled to match the 350 nm peak in the labeled peptides. This scaled spectrum was subtracted from the raw emission spectra of labeled peptides to isolate the 1,5-IAEDANS emission spectrum (Figure

6). The integrated intensity of the donor (D, tryptophan) and acceptor (A, 1,5-IAEDANS) emissions were used to calculate FRET efficiency (E):

$$E = A/(A+D)$$

The FRET efficiency (E) and Förster radius (R_0) are related to the distance between donor and acceptor fluorophores (r):

$$E = 1 / (1 + (r/R_0)^6)$$

The tryptophan (D) 1,5-IAEDANS (A) pair has a Förster radius of 22 Å[26]. The steady-state FRET measurements were used to estimate the distance between fluorophores using these equations.

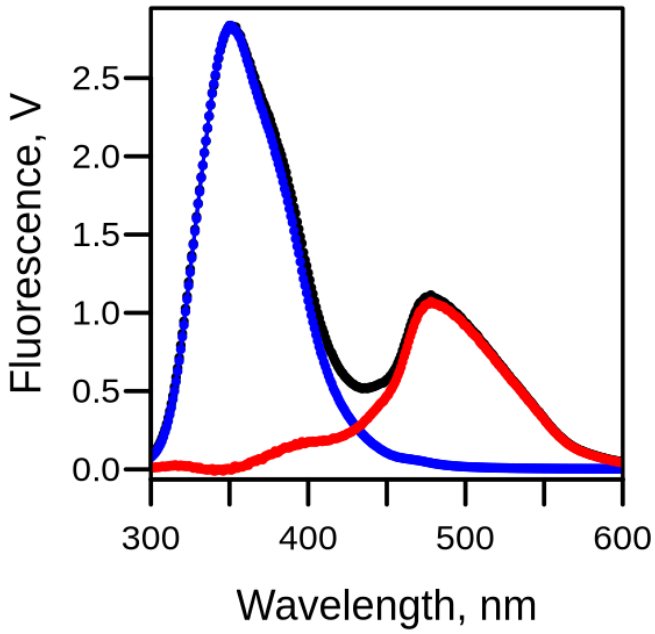

Figure 6: Example FRET deconvolution in R

Deconvolution for the PSDP_CW-IAED sample is shown. The original fluorescence spectrum is shown in black. The tryptophan emission (donor) is shown in blue. The 1,5-IAEDANS emission (acceptor) is shown in red.

2D: Results

2D.1: Polyalanine Helix

The simulated AK42 sequence displayed a high helical propensity and an inconsistent tendency to form turns within the central third of the peptide. The position of the turn varied both within each simulation and between simulations. The turns found in the top clusters for each simulation were typically short and unsupported by hydrogen bonding or sidechain interactions – three of nine simulations had 1-2 hydrogen bonds within the turn. Adjusting lysine positions did not influence the distribution of turns. The combined results from the nine simulations show that alpha helix geometry was the most frequently sampled conformation for each tetrad in the AK42 peptide (Figure 7A, red). Despite this, a full helix was not commonly found in the simulations. Only 8 full length helices were found in the combined dataset at 305K out of 423000 total structures; all other temperatures had 0-2 full length helices. Instead, transiently formed helix bundles (Figure 7F-I) with turns near the center of the peptide (Figure 7A, blue) dominate the simulations.

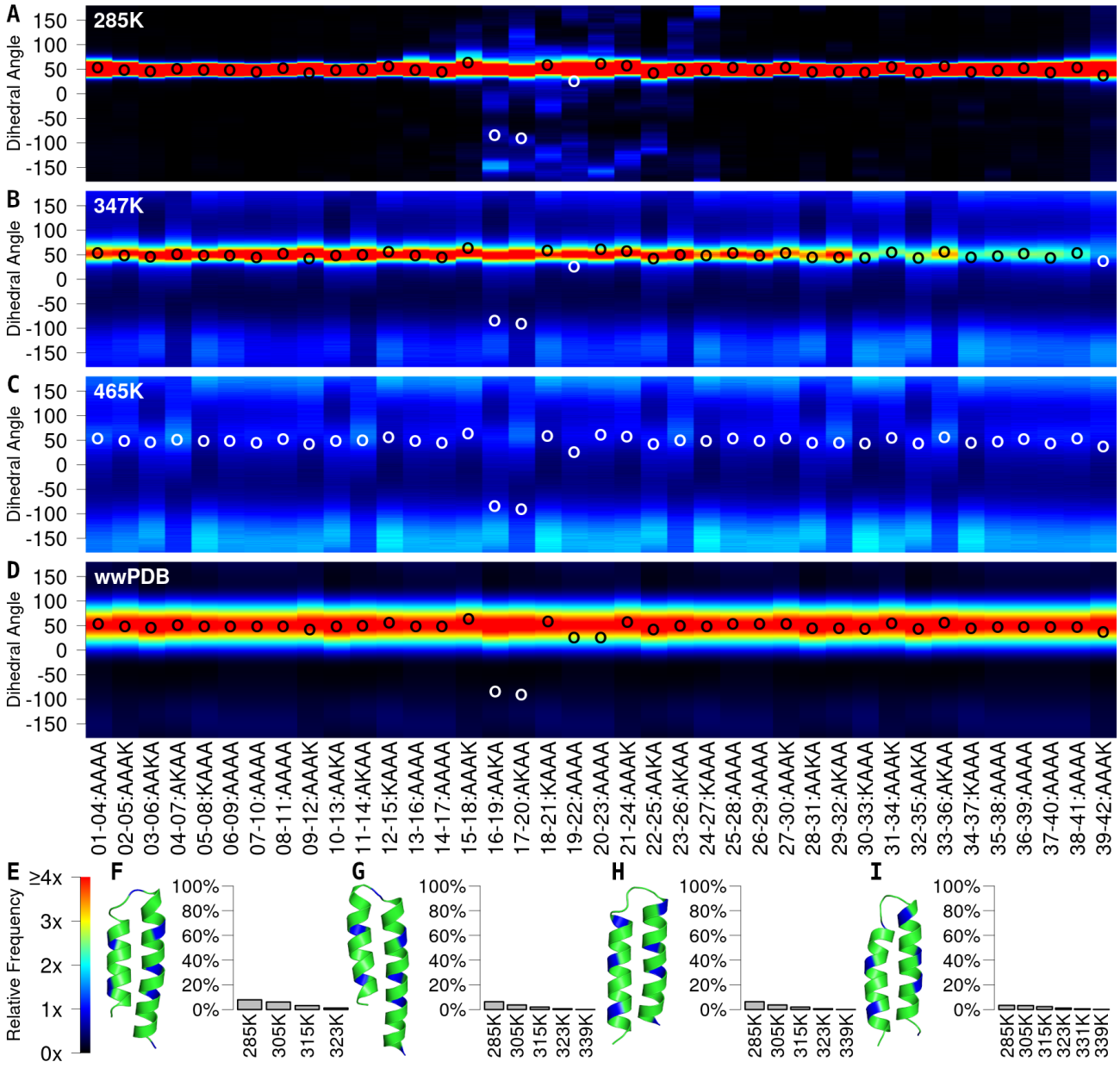

Figure 7: Combined simulation results for the AK42r3 peptide.

Dihedral angle distributions are rendered as heat for each tetrad in the AK42r3 peptide for 285K (A), 347K (B), and 465K (C) simulations. Smoothed dihedral angle distributions for the corresponding tetrad as found in all proteins in the wwPDB are shown for comparison (D). The series of dihedral angles found in top AK42r3 cluster structure (F) are shown as open circles (A-D). Heatmap colors depict the relative frequency of each dihedral angle (E). The top four cluster centers for AK42r3 are shown alongside their population distribution across temperature (F-I). The four structures are oriented with the N terminus on the left and rendered with lysine residues colored blue. Simulation data shown are the combined result of 9 replicates.

Substantial fraying of the helix at the C-terminus can be seen at elevated temperatures (Figure 7B), highlighting a bias for helix formation at the N-terminus of the peptide. The distribution of sampled tetrad conformations becomes more broad as temperature increases and shifts towards more extended conformations (Figure 7C, cyan). Weak differences in preferences can be seen between tetrads. The AKAA tetrads have a higher helical preference whereas KAAA has an extended preference (Figure 7C). Accordingly, helical structure preferentially nucleates at the AKAA segment as the simulation temperature lowers. Comparison of the lower temperature simulation data (Figure 7A) to the tetrad conformer distributions extracted from the wwPDB[3] (Figure 7D) shows complete agreement for the AK42 sequences.

2D.2: GD Loop Motif

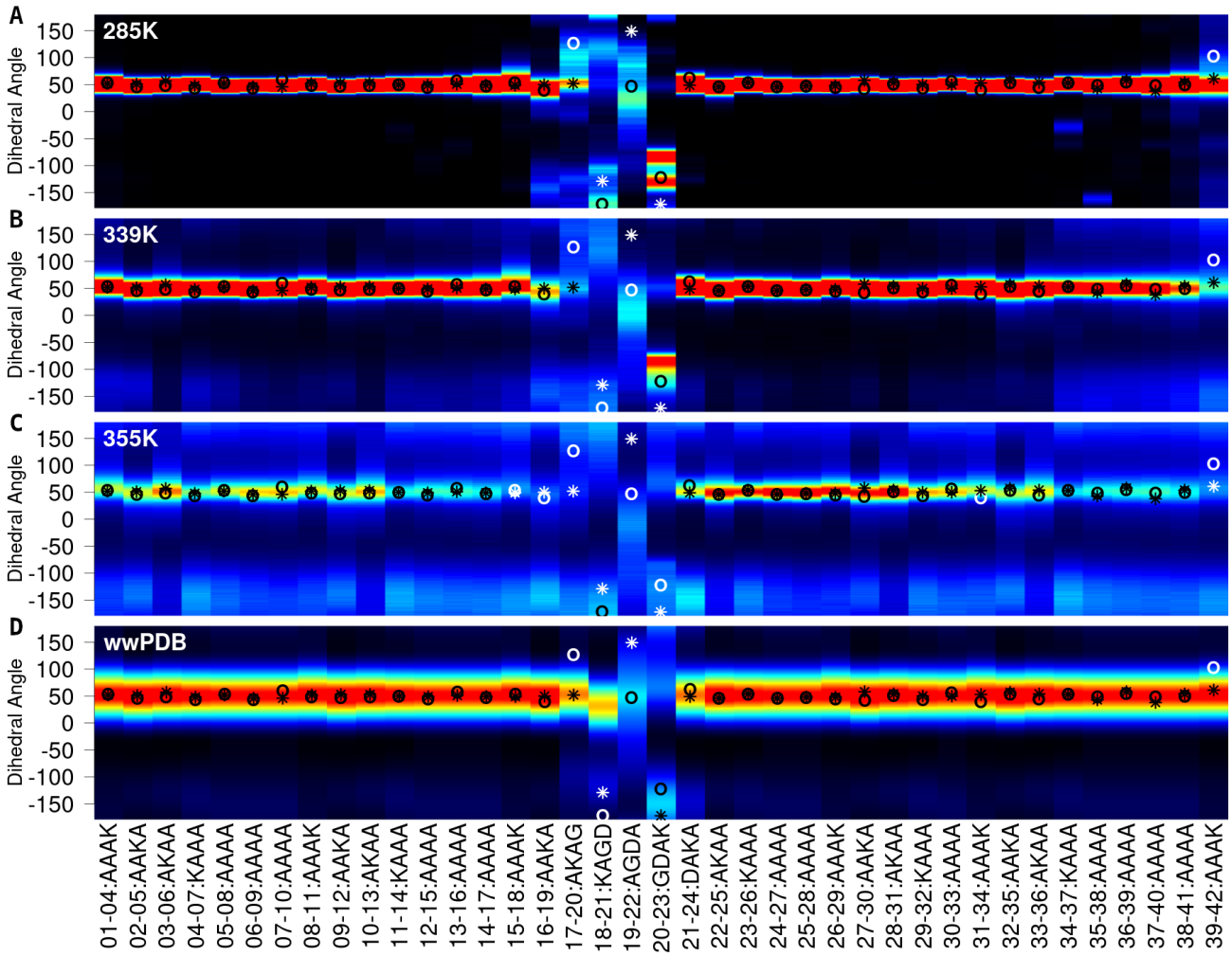

Figure 8: Sequence analysis for the AK42r2_GD peptide.

Tetrad heatmaps for the AK42r2_GD sequence using simulation data at 285K (A), 339K (B), and 355K (C). A tetrad heatmap for fragments extracted from the wwPDB is also included (D). Open circles indicate the geometry found in the top clustered structure from the simulations. Asterisks indicate the geometry found in the reference protein (pdb: 1CPQ). Symbols are colored black or white to maximize contrast. The scale used for heat is described in Figure 7E. Simulation data is the combined result of 11 replicates.

The GD sequence produced turns that were inconsistent between replicates. For all GD peptides, the boundary between Turn-1 and Helix-2 was consistent between replicates. The location of this boundary varied between the three designs due to differences in

conformational preferences of the 20-23:GDAA (A42_GD, AK42r1_GD) and 20-23:GDAK (AK42r2_GD) tetrads. The GD sequence with optimized charge distribution (AK42r2_GD) produced the most consistent results.

Full chain dihedral heatmaps provide a concise overview of how structure is formed in the AK42r2_GD peptide as it transitions to a lower energy state. Helical structure is the first to emerge in an N-to-C pattern (Figure 8C). Helix-2 forms before Helix-1; likely from the stabilizing features of aspartate: N_{cap} and helix dipole stabilization. The extended conformer of the 20-23:GDAK tetrad is the next to form (Figure 8B), establishing structure for the C-terminal side of Turn-1. Lastly, the C-terminal end of Helix-1 and the N-terminal half of Turn-1 form (Figure 8A). This order of events crudely matches the relative frequencies of tetrad dihedrals found in the wwPDB (Figure 8D): helical regions have the strongest dihedral preference, followed by the 20-23:GDAK and 19-22:AGDA tetrads, and lastly the 18-20:KAGD tetrad – which has a very weak preference for the extended conformer.

It is worth noting that while the distribution of dihedral angles is well dispersed for the 19-22:AGDA and 20-23:GDAK tetrads at high temperatures (Figure 8C), the peak of this distribution is similar across all temperatures (Figure 8A-C). Although a structural preference is established early in folding for this region, it is evidently not sufficient to define a consistent 2 helix bundle.

Upstream segments, 16-19:AAKA, 17-20:AKAG, and 18-21:KAGD, have dihedral distributions that shift or split during folding. For example, 17-20:AKAG samples a mixture of α-helix(τ≈50°)/3-10 helix(τ≈90°)/extended(τ≈-145°) conformations at high temperature (Figure 8C). As temperature lowers, the balance between these conformations varied between replicates. The α/3-10 helical conformers were the predominate result (Figure 8A), but a subset of simulation replicates sampled the extended conformer at higher frequencies. This split coincides with the emergence of clustered structures (~331K), suggesting non-specific interactions between the helices are responsible for the structural variations.

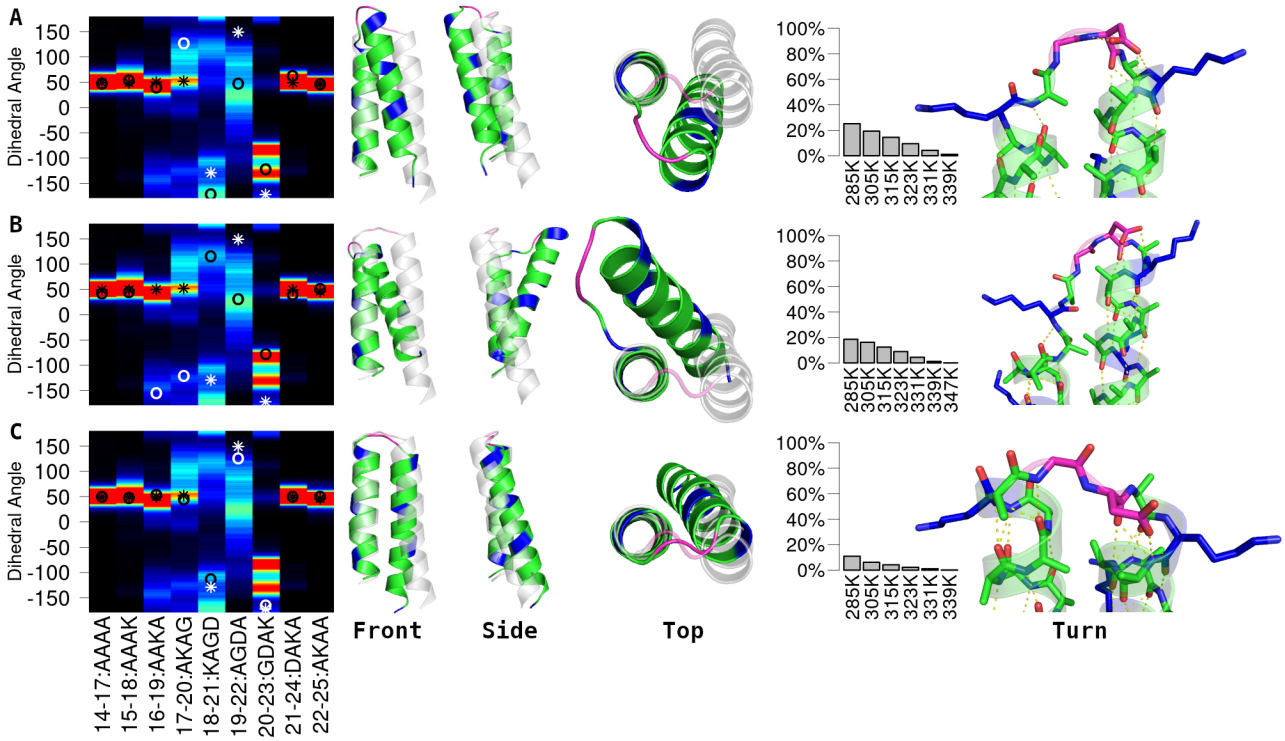

Figure 9: Clustered structures for AK42r2_GD.

Results are included for the top 3 (A: 1, B: 2, C: 3) most populated clusters. The dihedral angles for each cluster are shown as circles on the 285K tetrad heatmap. Asterisks mark geometry from the natural protein (pdb: 1CPQ). A front, side, and top view of the clustered structure is included (green) with the reference structure (pdb: 1CPQ, grey) aligned at Helix-1 to compare fold topology. The distribution of each cluster across different temperatures is included. A zoomed in image of the turn structure depicts polar interactions. The GD sequence is colored magenta, lysines are colored blue.

The top three clusters, which account for 55% of the structures at 285K, consistently utilized aspartate's sidechain carboxylate to provide a favorable N_{cap} for Helix-2 (Figure 9A-C, Turn view). The top two clusters form helix-helix interfaces different from the reference structure (Figure 9A,B, Top view). The third cluster produced a near perfect match for tetrad dihedrals and fold topology (Figure 9C). There is a slight downwards shift of Helix-2 relative to Helix-1

(Figure 9C, Front view), but the residues involved in the helix-helix interface are the same as in the reference structure.

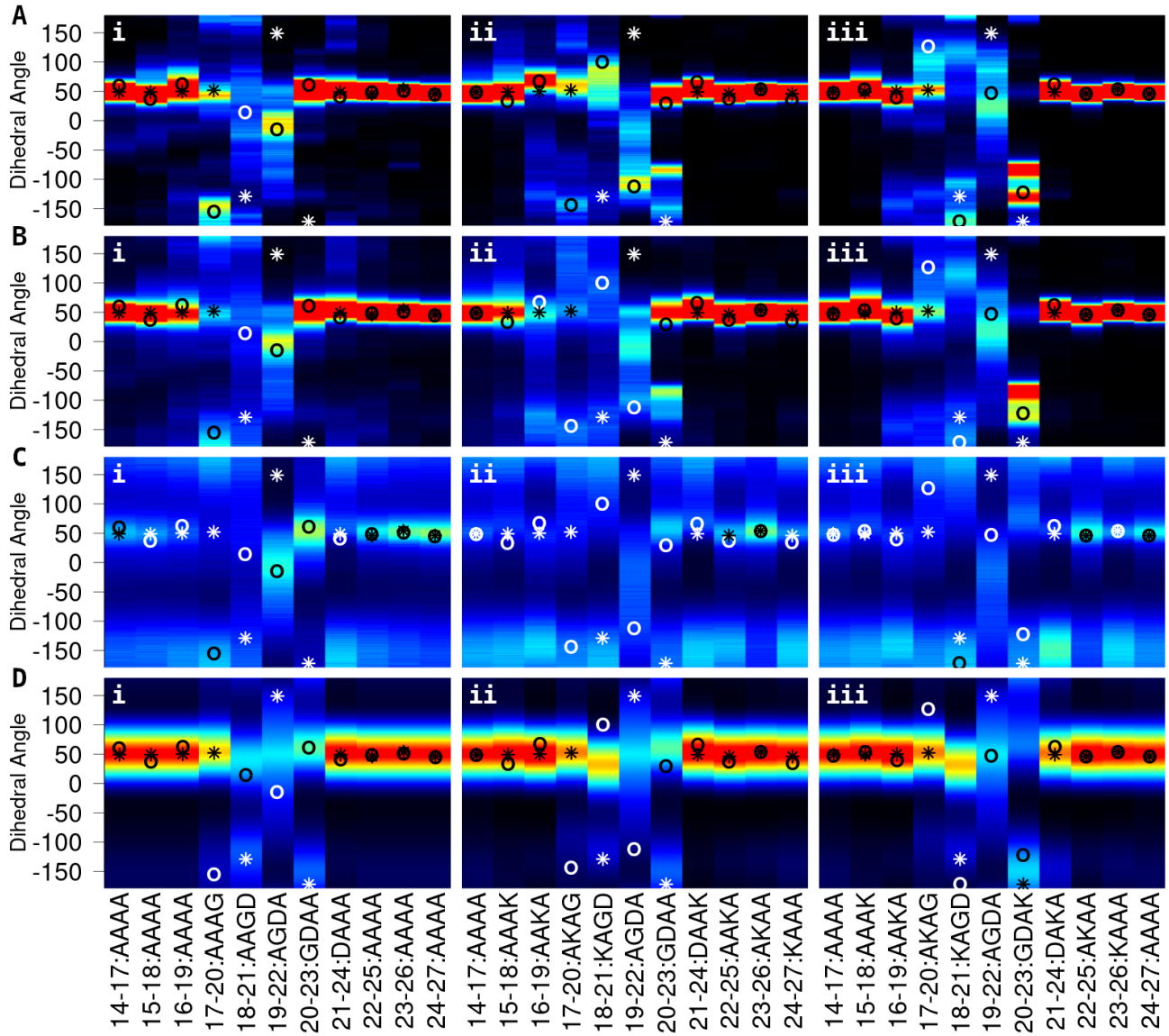

Figure 10: Turn analysis for GD peptides.

Tetrad heatmaps using simulation data at 285K (A), 323K (B), 371K (C), or the entire wwPDB (D) for the A42_GD (i), AK42r1_GD (ii), and AK42r2_GD (iii) peptides. Simulation results (A-C) are the combined results of 8-11 replicates. Open circles indicate the series of dihedral angles found in the top cluster found across all replicates. Asterisks represent the geometry found in the corresponding tetrads of the reference protein (pdb: 1CPQ). Symbols are colored black or white to maximize contrast. The tetrads for each segment shown are annotated on the X-axis. Tetrads excluded from these images were all at helical geometry (τ≈50°) at 285K. The scale used for heat is described in Figure 7E.

Significant differences in turn behavior occurred for the three GD sequences (Figure 10A). The context of the introduced GD sequence

influenced the conformer distribution of tetrads. Across the wwPDB, the 20-23:GDAA and 20-23:GDAK tetrads both sampled helical and extended conformers – but the ratio of these two conformers varies between the two tetrads. The 20-23:GDAA tetrad samples mostly helical conformers (Figure 10D_{i}; τ=59°, 63.15%) whereas the 20-23:GDAK tetrad samples mostly extended conformers (Figure 10D_{iii}; τ=-151°, 55.64%). The same trend was observed in simulation and lead to different folding outcomes. In AK42r2_GD 20-23:GDAK predominately sampled extended conformations at all temperatures (Figure 10A-C_{iii}) whereas in A42_GD 20-23:GDAA was predominately helical (Figure 10A-C_{i}). This shifted the location of the turn and caused a different pattern of tetrad dihedrals.

The AK42r1_GD peptide behaved similar to the A42_GD peptide at high temperatures (Figure 10$C_{i,ii}$) but deviated as clustered structures formed below 331K (Figure 10$B_{i,ii}$). The 19-22:AGDA and 20-23:GDAA tetrads, which were consistent across the A42_GD and AK42r2_GD simulations, varied between AK42r1_GD replicates. This different result is likely caused by steric clashes of the unoptimized lysine residues; mapping the lysines from AK42r1_GD onto A42_GD cluster-1 places K31 in the helix-helix interface. For both A42_GD and AK42r1_GD, there was significant variation in structure at the 17-20:AXAG and 18-21:XAGD tetrads between simulation replicates. The 18-21:AAGD (A42_GD) and 18-21:KAGD (AK42r1_GD) tetrads had nearly

uniform dihedral distributions at 323K (Figure 10B$_{i,ii}$). Without a specific (A42_GD) or compatible (AK42r1_GD) hydrophobic interface to guide folding, this segment adopted different geometries in each simulation replicate.

2D.3: PSXP Loop Motif

A near perfect match to the reference protein, MxiH, for the AK42r5_NPSNP peptide was observed for its top cluster (Figure 11A$_{iv}$). Even with no surface charges to guide the helix-helix interface, turn structure very similar to MxiH formed in the A42_NPSNP peptide (Figure 11A$_{iii}$). Peptides without the leading Asparagine residue in the turn sequence (A42_PSNP and AK42r5_PSNP) had inconsistent deviations from the MxiH turn structure (Figure 11A$_{i,ii}$). The two polar interactions from this Asparagine seen in the MxiH crystal structure (Figure 2E) were not formed in the top cluster for AK42r5_NPSNP. Replacing alanine with asparagine (APSNP) in the A42_PSNP cluster-1 structure causes steric clashes. These observations indicate that the leading asparagine in NPSNP guides turn structure formation by restricting access to alternate conformers through steric obstruction. A matching shift in dihedral distributions can be seen in data from the wwPDB between the 17-20:AAAP (Figure 11E$_{i}$) and 17-20:AANP (Figure 11E$_{iii}$) tetrads. In the AK42r5_NPSNP design, one of the lysines is placed within this tetrad. The resulting sequence (AKNP) is a perfect match to the sequence of

MxiH (Figure 2A). Introduction of this lysine into the 17-20:AKNP segment further shifts the wwPDB distribution to match the target structure (Figure 11E$_{iv}$). Polar interactions between the Lys and Asn residues in this tetrad likely favor the compact conformer this tetrad adopts.

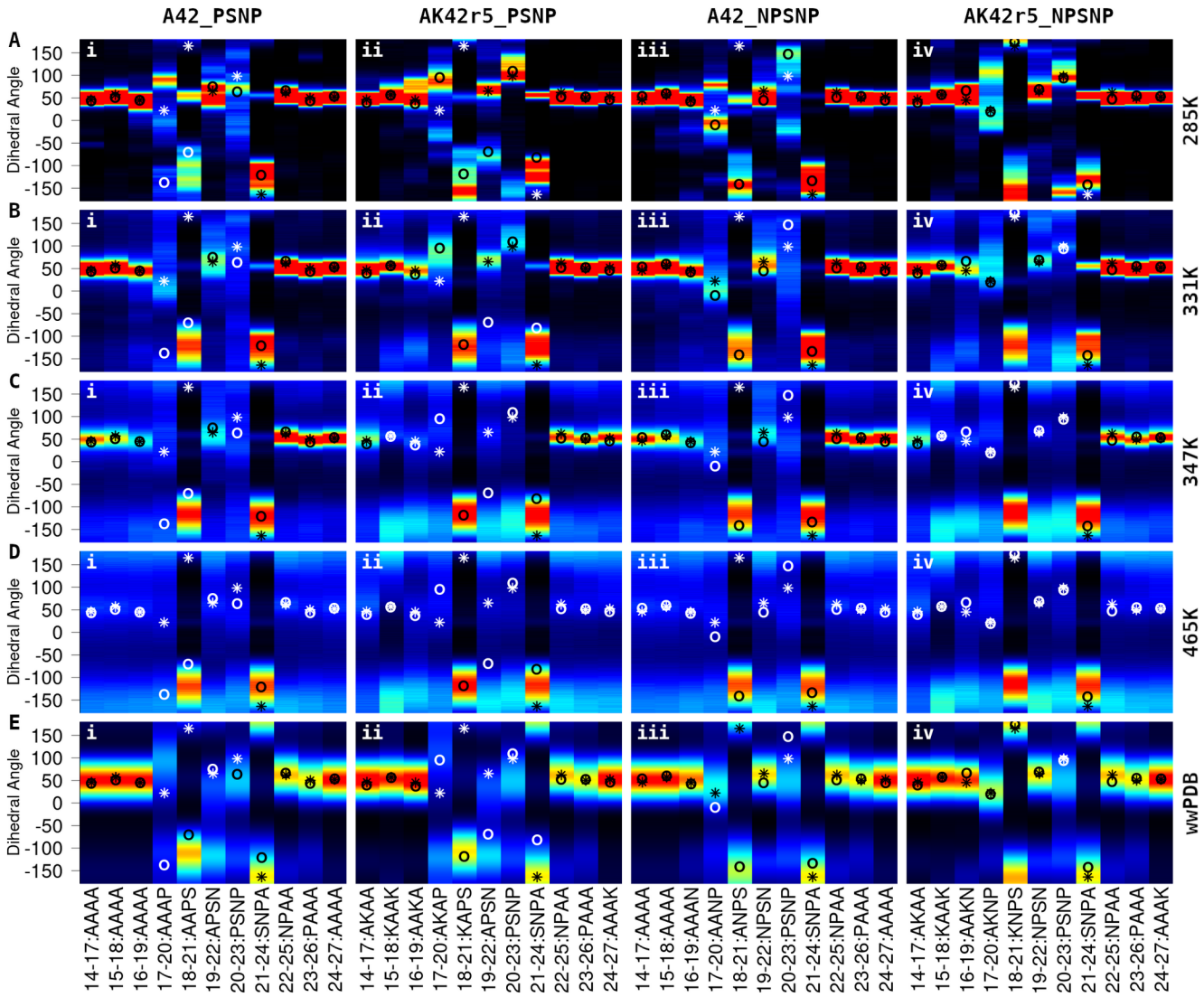

Figure 11: Turn analysis for PSNP peptides.

Tetrad heatmaps using simulation data at 285K (A), 331K (B), 347K (C), 465K (D), or the entire wwPDB (E) for the A42_PSNP (i), AK42r5_PSNP (ii), A42_NPSNP (iii), and AK42r5_NPSNP (iv) peptides. Simulation results (A-D) are the combined results of 8 replicates. Open circles indicate the series of dihedral angles found in the top cluster from all replicates. Asterisks represent the geometry found in the corresponding tetrads of the reference protein (pdb: 2CA5). Symbols are colored black or white to maximize contrast. The tetrads for each segment are annotated on the X-axis. Tetrads excluded from these images were all at helical geometry (τ≈50°) at 285K. The scale used for heat is described in Figure 7E.

On-target structure was present in all PSNP peptides at 465K for the 18-21 and 21-24 XXPX tetrads (Figure 11D). All other turn tetrads transition from an unproductive extended conformation towards more

compact, on-target geometries as temperature lowers. For these tetrads, on-target structure begins to emerge at 331K as clusters begin to form (Figure 11B) – except for the 17-20:AKAP tetrad in AK42r5_PSNP (Figure $11B_{ii}$), which forms an off-target structure. In AK42r5_NPSNP, part of the 20-23:PSNP distribution, which is the last tetrad to structure, remains in the previous extended conformation at 285K (Figure $11A_{iv}$). This extended population is coincident with the off-target helical population of the adjacent 21-24:SNPA tetrad.

Similar to the GD peptides, there is crude agreement between the geometry and order of structure formation in tetrads between simulation and wwPDB datasets. Both the helical segments and the XXPX segments have similar major populations (85-95%). The major XXPX dihedral population is spread across more angles than the helical populations, lowering the apparent heat (Figure 11E). Structure for the XXPX tetrads is established first (Figure 11D), possibly due to the overall tendency for extended conformers to be preferred at high temperatures. Helical structure forms next at 347K (Figure 11C). The 17-20:AXNP and 19-22:NPSN segments are the next most represented in wwPDB datasets (Figure $11E_{iii,iv}$) and next to form in the A42_NPSNP (Figure $11B_{iii}$) and AK42r5_NPSNP (Figure $11B_{iv}$) peptides at 331K. Structure in the 20-23:PSNP tetrad was the least represented in the wwPDB dataset (Figure $11E_{iii,iv}$) and the last to form in simulation at 285K (Figure $11A_{iii,iv}$).

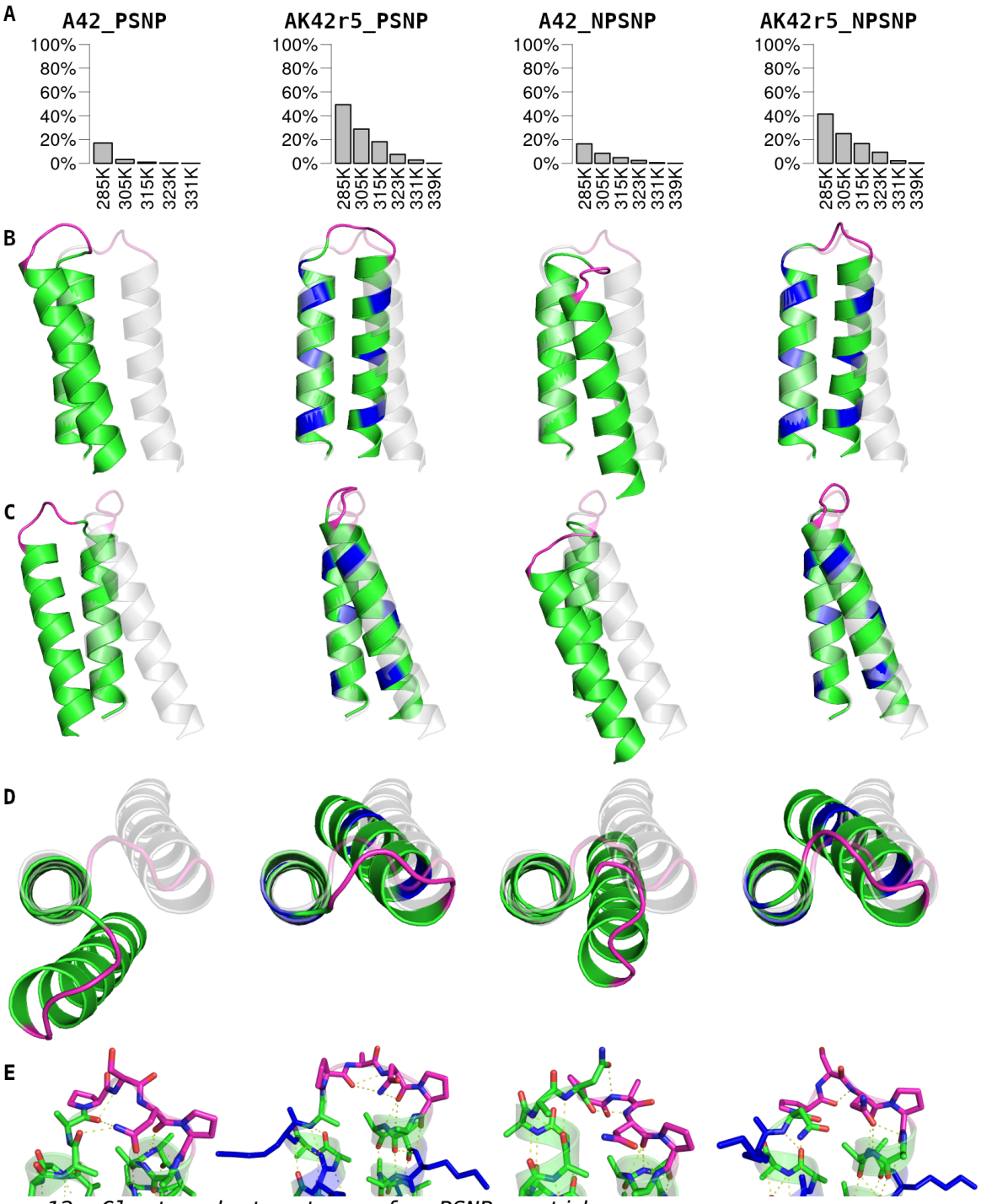

Figure 12: Clustered structures for PSNP peptides.

Population distributions (A) for the top cluster of each PSNP peptide. Structural overlays of each cluster (green) to the MxiH protein (grey) aligned at Helix-1 are provided for front (B), side (C), top (D), and turn (E) views. The PSNP sequence is colored magenta. Lysine residues are colored blue.

The MxiH two helix bundle utilizes hydrophobic sidechains larger than alanine to form the hydrophobic helix-helix interface. The cluster-1 structure for AK42r5_NPSNP places the two helices closer together and slightly skewed due to the smaller alanine residues at the interface (Figure 12B,D). Despite this adjustment, the turn structure aligns well with MxiH (Figure 11A_{iv}, 12B-D). A similar tetrad dihedral pattern to MxiH is also seen in the A42_NPSNP cluster-1 structure (Figure 11A_{iii}), but the compounded differences lead to a different fold topology (Figure 12B-D). Neither A42_PSNP nor AK42r5_PSNP matched the turn structure of MxiH, but remarkably the helix topology in AK42r5_PSNP is nearly identical to AK42r5_NPSNP despite significant differences in turn structure (Figure 12B-D).

The backbone hydrogen bond in the type I beta-turn across the NPSN asparagines (Figure 2E) was poorly formed in the AK42r5_NPSNP simulation even though the C_α orientations were similar (Figure 11A_{iv}, 12B-D). The distance between carbonyl-oxygen and amide-nitrogen is 3.8 Å in the cluster-1 structure, compared to 3.1 Å in the MxiH crystal structure (pdb: 2CA5). The Amide-nitrogen in the cluster-1 structure is not oriented towards the carbonyl-oxygen (Figure 12E). Other structures in the cluster did form a hydrogen bond similar to MxiH, but the angle between the CO and NH bonds was nearly perpendicular instead of aligned.

Incorrect modeling of the NPSN proline's C_α tetrahedron may be the underlying issue. The bond angle between backbone amide-N, C_α, and carbonyl-C is 117° for this proline in the MxiH crystal structure (pdb: 2CA5). This wide angle is also present in multiple PrgI structures for the KPSD proline: 124° (pdb: 2JOW; NMR), 119° (pdb; 2G0U, NMR), and 118° (pdb: 2X9C, 3ZQE, 3ZQB; X-ray). The second PSXP proline had a narrower angle (114±1°) in these structures. In CAMPARI, the range of N-C_α-C bond angles sampled is 109.5° to 111.5° for all residues and temperatures. Bending the MxiH structure to match a narrower angle disrupts the NPSN backbone hydrogen bond orientation, whereas in the AK42r5_NPSNP structures bending towards a wider angle improves hydrogen bond orientation. Evidently these missed stabilizing interactions were not required for AK42r5_NPSNP to match the MxiH turn structure, but may underpin differences in structure for the PSDP peptides.

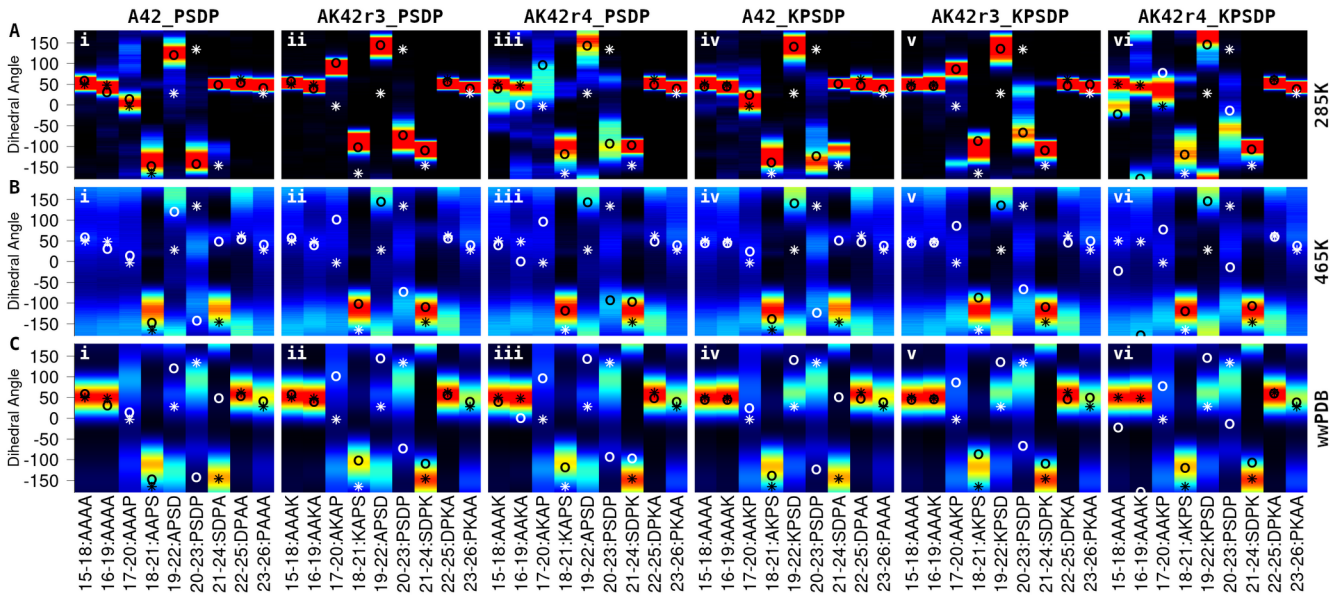

Figure 13: Turn analysis for PSDP peptides.

Tetrad heatmaps using simulation data at 285K (A), 465K (B), or the entire wwPDB (C) for the A42_PSDP (i), AK42r3_PSDP (ii), AK42r4_PSDP (iii), A42_KPSDP (iv), AK42r3_KPSDP (v), and AK42r4_KPSDP (vi) peptides. Simulation results (A-B) are the combined results of 8 replicates. Open circles indicate the series of dihedral angles found in the top cluster from all replicates. Asterisks represent the geometry found in the corresponding tetrads of the reference protein (pdb: 2JOW). Symbols are colored black or white to maximize contrast. The tetrads for each segment shown are annotated on the X-axis. Tetrads excluded from the images were all at helical geometry (τ≈50°) at 285K. The scale used for heat is described in Figure 7E.

None of the PSDP designs matched the PrgI reference structure. Two different charge optimization strategies were tested, one (AK42r3) targeting the AK42r1_PSDP cluster-1 structure (Figure 14A) and a second (AK42r4) targeting the PrgI NMR structure (pdb: 2JOW). In the AK42r1_PSDP cluster-1 structure, two are lysines adjacent to the helix-helix interface. These lysines were repositioned to face outwards in the AK42r3 design. The AK42r3_PSDP peptide behaved exceptionally well *in silico*, but poorly *in vitro*. Compared to MxiH (Figure 2E), the type I beta turn in PrgI (Figure 2D) has better hydrogen bond alignment. Consequently, the simulated PSDP peptides

will lose more stability at the target structure compared to the PSNP peptides due to the incorrect bond geometry at the PSXP proline.

Major deviations in structural preferences occurred for the XPSD tetrad (Figure 13A). In the wwPDB, the APSD tetrad samples two populations: τ=57° (36%) and τ=-129° (64%) (Figure 13$C_{i\text{-}iii}$). In the PSDP peptides the APSD tetrad preferentially sampled extended conformations (τ≈170°) at both high (Figure 13$B_{i\text{-}iii}$) and low (Figure 13$A_{i\text{-}iii}$) temperatures. A minor population is present for compact conformations (τ≈0°) at high temperatures (Figure 13$B_{i\text{-}iii}$). The KPSD tetrad has an enhanced preference for compact conformations in the wwPDB, τ=60° (65%), over the extended conformation, τ=-117° (35%) (Figure 13$C_{iv\text{-}vi}$). The compact conformer places the Lys and Asp residues together where electrostatic interactions between sidechains can stabilize the tetrad (Figure 2D). Contrary to this shift observed in the wwPDB, the KPSD tetrad in KPSDP peptides sampled compact conformers less than the PSDP peptides at all temperatures (Figure 13$A_{iv\text{-}vi}$,$B_{iv\text{-}vi}$). The adjacent PSDP tetrad also deviated considerably from the PrgI structure, but sampling was more consistent with fragments in the wwPDB. Two conformers were sampled at high temperatures (Figure 13B), but ultimately the minor population in wwPDB fragments (Figure 13C; τ=-79°, 27%) was adopted at low temperatures (Figure 13A).

Several peptides analogous to simulated designs were characterized *in vitro*: AK42_W (AK42r1), PSDP_W (AK42r1_PSDP), AK42_CW-IODO/AK42_CW-IAED (AK42r3), and PSDP_CW-IODO/PSDP_CW-IAED (AK42r3_PSDP). Circular dichroism measurements confirmed each peptide was helical in solution. The AK42_W peptide was considerably more helical than the corresponding PSDP_W peptide (Figure 14D). The PSDP_W peptide was expected to be slightly less helical due to the loss of helicity in the expected turn region. The observed 37% loss in 222 nm signal for PSDP_W, relative to AK42_W, exceeds the expected 10-12% loss and indicates that parts of the expected helix-helix bundle are unstructured. Thermal denaturation of the peptides revealed PSDP_W to be less stable than AK42_W (Figure 14E); helix-helix interactions were expected to stabilize PSDP_W relative to AK42_W. A stronger 222 nm signal was observed for PSDP_W at high temperatures (Figure 14E), which suggests more helical structure is retained in the denatured state. This interpretation is supported by *in silico* results; Helix-2 is partially formed at 371K in the PSDP peptides (data not included).

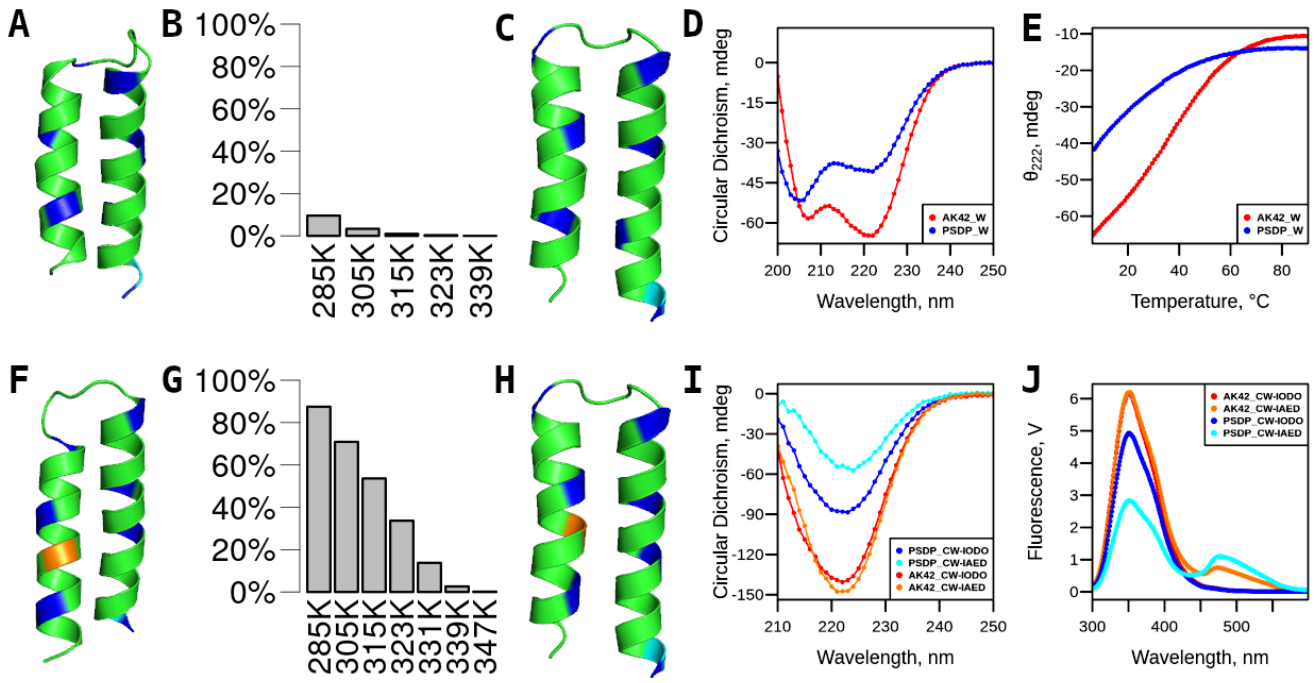

Figure 14: PSDP simulation cluster and in vitro characterization results.

Top cluster structures and populations are included for the AK42r1_PSDP (A, B) and AK42r3_PSDP (F, G) peptides. The position of select residues in the corresponding synthetic peptides are shown for lysines (blue), tryptophans (cyan), and cysteins (orange) for AK42_W/PSDP_W (A) and AK42_CW/PSDP_CW (F). The matching locations in the PrgI protein (pdb: 2JOW) are included for the AK42_W/PSDP_W (C) and AK42_CW/PSDP_CW (H) designs. CD scans (D) and thermal melts monitored by CD (E) are included for the AK42_W and PSDP_W peptides at 10μM. CD scans (I) and fluorescence scans (J) are included for the labeled AK42_CW/PSDP_CW peptides at 20μM.

Repositioning lysines near the helix-helix interface of the AK42r1_PSDP cluster-1 structure (Figure 14A), which had poor clustering consistency (Figure 14B), yielded the AK42r3_PSDP design which produced a new structure (Figure 14F) with very high clustering consistency (Figure 14G). Experimental characterization of the corresponding peptides, AK42_CW-IODO and PSDP_CW-IODO, did not show any improvements. The same 37% loss in signal for PSDP_CW-IODO, relative to AK42_CW-IODO, was observed (Figure 14I) as in the PSDP_W/AK42_W pair (Figure 14D). Labeling with the IAEDANS

fluorophore had a varied effect on the two peptides. AK42_CW-IAED and AK42_CW-IODO peptides produced very similar CD spectra (Figure 14I). Another loss in signal at 222 nm occurred for the PSDP_CW-IAED (Figure 14I); 64% loss relative to AK42_CW-IAED, 39% loss relative to PSDP_CW-IODO. Mapping lysine positions onto the PrgI structure (pdb: 2JOW) for the PSDP_W (Figure 14C) and PSDP_CW (Figure 14H) sequences highlights potential conflicts. For both designs, two lysine residues are positioned in the helix-helix interface where steric clashes will occur. The labeled Cys residue in the PSDP_CW design is also positioned in the helix-helix interface.

Steady-state FRET characterization of the labeled peptides shows higher resonance between fluorophores in the PSDP_CW-IAED peptide than in the AK42_CW-IAED peptide (Figure 14J). The FRET efficiencies for PSDP_CW-IAED (0.34) and AK42_CW-IAED (0.15) provide approximate distance measurements of 25Å (PSDP_CW-IAED) and 29Å (AK42_CW-IAED). The expected α-carbon distances between the FRET residues is 16Å for PSDP_CW-IAED (PrgI, pdb: 2JOW) and 48Å for AK42_CW-IAED (42-residue α-helix, simulated structure). An ensemble of nonspecific folds, similar to what was observed in simulated AK42r3 peptides, may be responsible for the lower approximate average distance measured for AK42_CW-IAED. Breakage of Helix-1 in PSDP_CW-IAED due to steric clashes may leave ~5 residues of Helix-1's C-terminus structured with the remainder forming a disordered structure nearby. The distance

between the 41-Trp C_α and this proposed helical segment is 24Å, similar to the FRET distance approximation.

The experimental results suggest the structures predicted by CAMPARI are not occurring in the designed peptides. Instead, the expected turn structure from PrgI appears to form. Steric clashes of lysines in the helix-helix interface will disrupt one or both helices in the PSDP_W and PSDP_CW peptides. The IAEDANS label had no effect on structure in AK42_CW, but further destabilized PSDP_CW. This is consistent with the Cys residue being positioned at the helix-helix interface in PSDP_CW (Figure 14H) due to the PrgI turn structure; labeling with IAEDANS will introduce additional steric clashes. FRET distance approximations and circular dichroism measurements are compatible with a partially formed PrgI fold with a disordered N-terminal region. Structure in Helix-2 is assumed to be favored over Helix-1 due to the stabilizing N_{cap} features of Asp and Pro, which stabilized Helix-2 *in silico* at even high temperatures.

2D.4: Ubiquitin Associated Domain 1 of HHR23A

The AK42r6_NCloops peptide formed a 3-helix bundle in simulation with a partial match to the UBA(1) structure. The presence of a minimal hydrophobic core in AK42r6_NCloops2 did not improve fold consistency. Steric clashes in the AK42r1_NCloops design led to inconsistent turn structure between replicates. The charge optimized AK42r6_NCloops design created a fold with on-target Turn-1 geometry

but off-target geometry for Turn-2 with an elongated Helix-2. Tetrad geometries from simulations were consistent with populations found in the wwPDB with one exception in Turn-2 for both simulated peptides and *in vitro* UBA(1) protein.

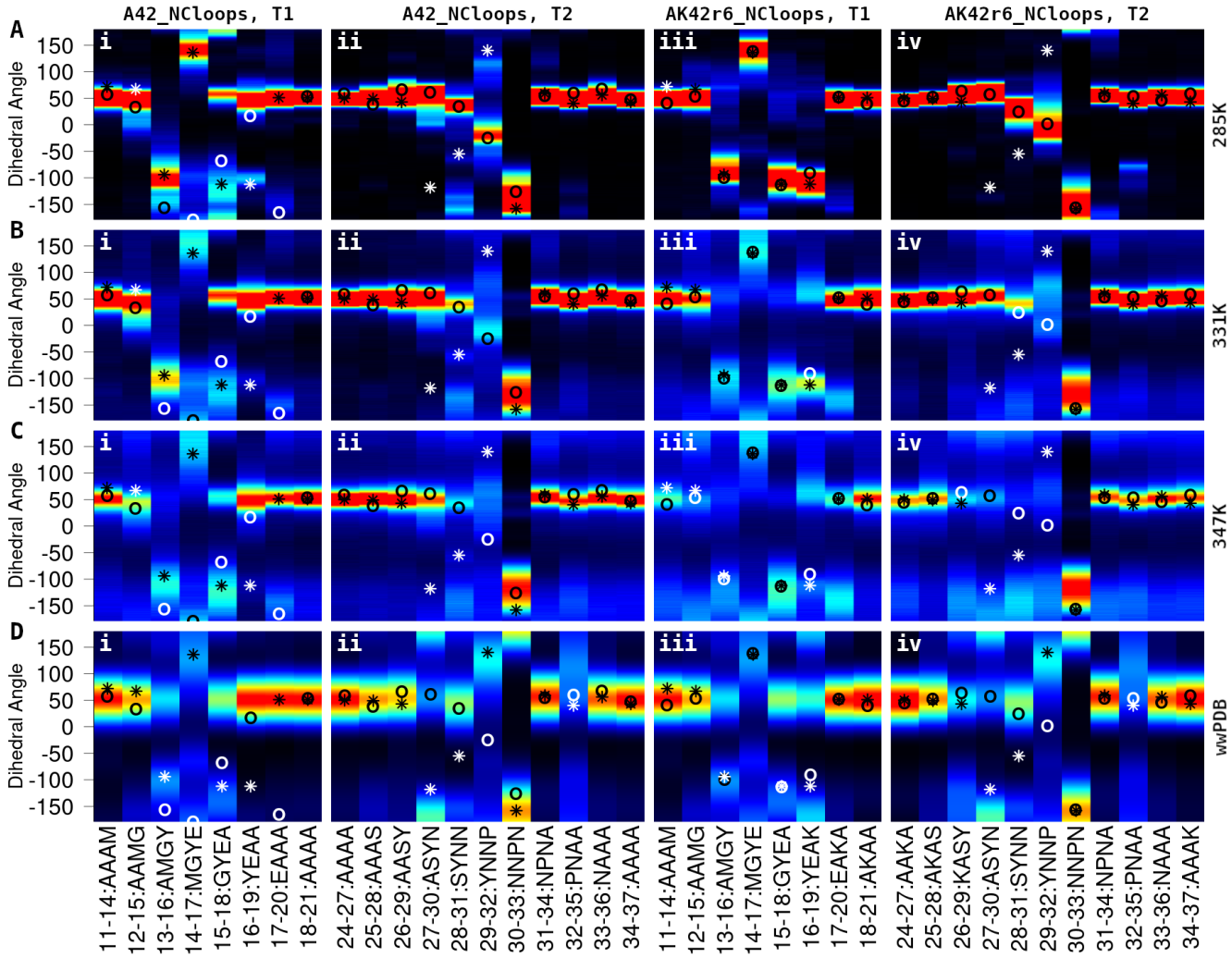

Figure 15: Turn analysis for NCloops peptides.

Tetrad heatmaps using simulation data at 285K (A), 331K (B), 347K (C), or the entire wwPDB (D) for the A42_NCloops Turn-1 (i), A42_NCloops Turn-2 (ii), AK42r6_NCloops Turn-1 (iii), and AK42r6_NCloops Turn-2 (iv) peptide regions. Simulation results (A-C) are the combined results of 8 replicates. Open circles indicate the series of dihedral angles found in the top cluster from all replicates. Asterisks represent the geometry found in the corresponding tetrads of the reference protein (pdb: 6W2H). Symbols are colored black or white to maximize contrast. The tetrads for each segment are annotated on the X-axis. Tetrads excluded from these images were all at helical geometry (τ≈50°) at 285K. The scale used for heat is described in Figure 7E.

Three of four tetrads in AK42r6_NCloops Turn-1 established on-target preferences early in folding: 13-16:AMGY, 14-17:MGYE, and 15-18:GYEA (Figure 15C_{iii}). The extended conformers of these tetrads orient the MGYE sidechains in alternating directions, causing Met and Tyr sidechains to face similar directions. Hydrophobic interactions between Met and Tyr sidechains can simultaneously stabilize the extended conformation found in this turn and nucleate the hydrophobic core of the helical bundle. The Glu residue faces the opposite direction, creating a bias for solvent interactions on one side of Turn-1. The last tetrad in Turn-1, 16-19:YEAK, shifted from a helical conformation (Figure 15C_{iii}) to an extended conformation (Figure 15B_{iii}) during folding. The final structure of Turn-1 (Figure 15A_{iii}) uses geometry that is well represented in the wwPDB for each tetrad (Figure 15D_{iii}): 13-16:AMGY (τ=-105°, 45.1%), 14-17:MGYE (τ=126°, 61.41%), 15-18:GYEA (τ=-142°, 35.14%), and 16-19:YEAK (τ=-174°, 45.27%). Each tetrad has two conformers with nearly balanced (50/50) populations. There is no clear relationship between the order of structure formation in Turn-1 and conformer preferences found in the wwPDB.

Only the 30-33:NNPN tetrad formed on-target structure for Turn-2 in the AK42r6_NCloops peptide (Figure 15A_{iv}). Similar to other simulated XXPX tetrads, on-target extended conformers were sampled across all temperatures (Figure 15A_{iv},B_{iv},C_{iv}) and in the wwPDB (Figure

15D$_{iv}$; τ=-160°, 97.75%). Both the cluster-1 structure from simulations and the UBA(1) crystal structure (pdb: 6W2H) have Turn-2 tetrads that form geometry outside of populations in the wwPDB. These tetrads are 28-31:SYNN for UBA(1) and 29-32:YNNP for AK42r6_NCloops (Figure 15D$_{iv}$). This suggests that Turn-2 is poorly optimized for the target fold and is partially structured through helix-helix interactions. The lack of a well-defined hydrophobic core in AK42r6_NCloops is likely responsible for the different distortion pattern in Turn-2 between AK42r6_NCloops and UBA(1). For both systems, only one tetrad is distorted; all other tetrads in Turn-2 form geometry commonly found across the wwPDB. Only one off-target tetrad in AK42r6_NCloops samples the minor population in the wwPDB instead of the on-target major population, 27-30:ASYN (Figure 15D$_{iv}$; τ=54°, 40.88%; τ=-172°, 59.12%).

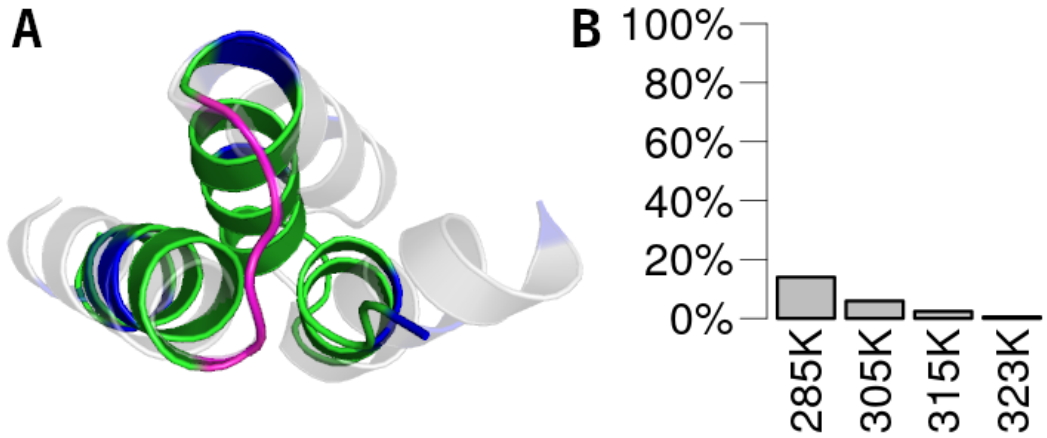

Figure 16: Clustering results for the AK42r6_NCloops peptide.

(A) Structural alignment at Turn-1 for AK42r6_NCloops cluster-1 (green) with UBA(1) (grey, pdb: 6W2H). The turn sequence is colored magenta. Lysine residues are blue. (B) Population distribution for the cluster-1 structure.

Structural variations in Turn-1 (Figure 15A$_{iii}$) lead to poor clustering results for AK42r6_NCloops (Figure 16B). Alignment of cluster-1 Turn-1 shows strong structural agreement to the UBA(1) X-ray structure (Figure 16A). The correct helix orientations are formed in AK42r6_NCloops, but at incorrect angles. Differences in Helix-1/Helix-2 skew are caused by deviations away from 3-10 geometry in Helix-1 (Figure 15A$_{iii}$, 11-14:AAAM) and extended geometry in Turn-1 (Figure 15A$_{iii}$, 16-19:YEAK). The different Turn-2 structure causes the different placement of Helix-3 relative to Helices 1 and 2.

The lysine-free A42_NCloops design behaved similar to AK42r6_NCloops at Turn-2. A deviation from wwPDB populations was again observed for the 29-32:YNNP tetrad, but utilizing a slightly more negative τ (Figure 15D$_{ii}$). Other tetrads in A42_NCloops Turn-2 matched the behavior of AK42r6_NCloops, but with poorer consistency (Figure 15A$_{ii}$). Tetrads in Turn-1 sampled on-target geometry in A42_NCloops (Figure 15A$_{i}$), but more often formed off-target structure. The strongest deviation in the 16-19:YEAA tetrad (Figure 15A$_{i}$) is consistent with shifts in wwPDB distributions due to the replacement of lysine with alanine. The wwPDB distribution for 16-19:YEAA is largely helical (Figure 15D$_{i}$; τ=50°, 90.26%; τ=-172°, 9.74%) whereas 16-19:YEAK has a balanced preference for helical and extended conformers (Figure 15D$_{iii}$; τ=52°, 54.73%; τ=-174°, 45.27%). This suggests off-target structure in Turn-1 is due to context-

dependent effects causing higher helix propensity in the end of Turn-1, not due to the loss of charge patterning.

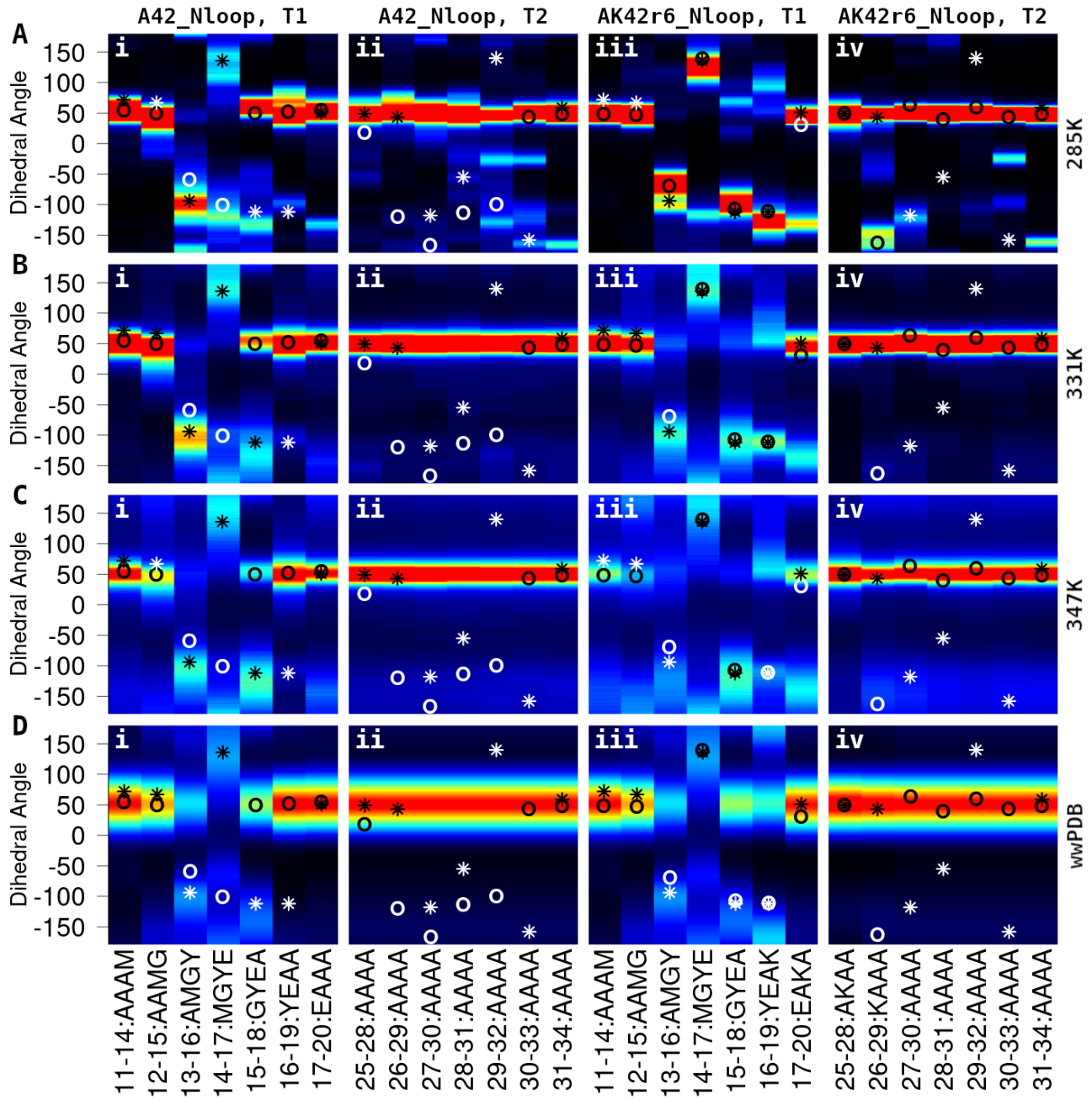

Figure 17: Turn analysis for the Nloop peptides.

Tetrad heatmaps using simulation data at 285K (A), 331K (B), 347K (C), or the entire wwPDB (D) for the A42_Nloop Turn-1 (i), A42_Nloop Turn-2 (ii), AK42r6_Nloop Turn-1 (iii), and AK42r6_Nloop Turn-2 (iv) peptide regions. Simulation results are the combined results from 8 replicates. Open circles indicate the series of dihedral angles found in the top cluster from the combined simulations. Asterisks represent the geometry found in the corresponding tetrads from the reference protein (pdb: 6W2H). Symbols are colored black or white to maximize contrast. The tetrads for each segment shown are annotated on the X-axis. Tetrads excluded from these images were all at helical geometry (τ≈50°) at 285K. The scale used for heat is described in Figure 7E.

The structure in Turn-1 was on-target even in the absence of Turn-2 in the AK42r6_Nloop peptide (Figure 17A$_{iii}$). The structural behavior of Turn-1 in A42_Nloop/AK42r6_Nloop (Figure 17A$_{i,iii}$) was analogous to A42_NCloops/AK42r6_NCloops (Figure 15A$_{i,iii}$). Presence of Turn-1 caused the formation of random turns in the Turn-2 region of the peptide (Figure 17A$_{iv}$). The results show that Turn-1 can adopt the correct structure from its primary sequence alone, independent of stabilizing interactions between the three helices. The random distribution of folds in the Turn-2 region (Figure 17A$_{iv}$) reinforces the observation that charge patterning alone is insufficient to favor a target fold in AK42 peptides.

2E: Conclusions

The structural properties of five turns were characterized utilizing computational methods applied to both simulated and empirical data. The selected sequences demonstrated counter-active, passive, and active turn behaviors. Charge patterning across the peptide was found to be important to enable target folds, but was insufficient to favor the target fold without the matching turn sequence. Comparison of simulated data (CAMPARI) to empirical data (wwPDB) offered partial validation of simulated behavior and highlighted differences due to proposed folding-induced strain or inaccuracies in the simulated proline C_α sp^3 tetrahedron. Assessment

of the natural sequences containing these turn sequences offers additional context to understand how the different turn mechanisms are utilized in proteins.

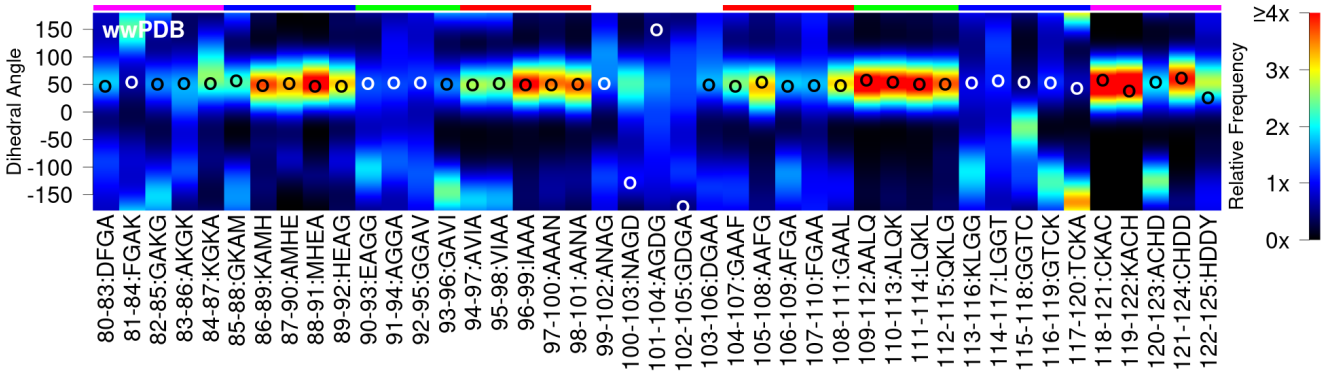

Figure 18: Sequence analysis for Cytochrome c' protein.

Tetrad heatmaps for wwPDB fragments matching the sequence from the Cytochrome c' protein's GD bundle. Open circles indicate the series of dihedral angles found in Cytochrome c' (pdb: 1CPQ). Symbols are colored black or white to maximize contrast. Regions of the heatmap are color coded at the top of the diagram to highlight aligned regions within the bundle.

Computational results characterize the GD sequence as a passive turn with many compatible folds. Structure in the turn region was established in a C-to-N pattern (Figure 8A-C). The behavior at the C-terminal end of the GD turn was consistent but context-dependent (Figure $10A_{i,iii}$); neighboring residues had a strong influence on turn structure. The N-terminal end of the GD turn formed late during folding and was highly variable (Figure $10A_{iii}$). Variability at the N-terminal side of the GD turn lead to many different fold topologies (Figure 9), one of which matched the GD two helix bundle from Cytochrome c' (Figure 9C).

The two-helix bundle from Cytochrome c' is unusual; there are 7 glycines spread across the two helices. Helical structure is poorly

represented in the wwPDB for the corresponding glycine containing tetrads. Examination of wwPDB heatmaps for the full sequence reveals an interesting pattern; each region of poor helical character is aligned with a region of strong helical character in the opposite helix (Figure 18). This suggests interactions between helices are the driving mechanism for this fold, not the turn sequence geometry. Unstructured segments flanking a turn, such as the N-terminal region of Helix 2 (Figure 18, 103-106:DGAA to 108-111:GAAL), will be able to undo any structural guidance turn geometry may provide. To overcome this, an active turn will need to nucleate secondary structure, a hydrophobic core, or both in addition to encoding a specific turn geometry before it can actively direct protein fold topology.

A mixture of computational and experimental techniques characterize the PSNP and PSDP sequences as active turns in their respective proteins, MxiH and PrgI. Incorrect modeling of the proline C_α tetrahedron by CAMPARI likely lead to lost stability in the simulated PSNP turn and incorrect results for the simulated PSDP turn. Experimental results for analogous PSDP peptides and empirical preferences derived from the wwPDB contradict CAMPARI results for the PSDP containing sequences. Incorporation of the leading polar residue, NPSNP or KPSDP, produced wwPDB heatmaps more consistent with the experimental structures for the two sequences; simulation results improved only for the NPSNP peptide.

The AK42r5_NPSNP sequence produced a near perfect match to the MxiH turn structure in CAMPARI (Figure 11A$_{iv}$). Even with no lysines to favor the correct hydrophobic interface, the A42_NPSNP turn sequence adopted geometry similar to MxiH (Figure 11A$_{iii}$). Comparison to results for the PSNP sequences indicate the leading Asn residue favors the turn structure of MxiH by restricting access to alternate turn conformations. Loss of this asparagine produced a different turn structure (Figure 11A$_{ii}$) consistent with shifts in wwPDB distributions (Figure 11E$_{ii}$). Remarkably, the same fold topology was established (Figure 12, AK42r5_PSNP). This result hints that well conserved turn sequences may resist deleterious mutations by having nearby variants that favor an alternate geometry that is still compatible with the original fold topology. Experimental measurements indicate the PSDP sequences behave similar to wwPDB distributions and the PrgI structure, not CAMPARI simulations. Loss of helical structure for synthetic PSDP peptides (Figure 14D,I) was consistent with the steric clashes expected from a PrgI turn structure (Figure 14C,H).

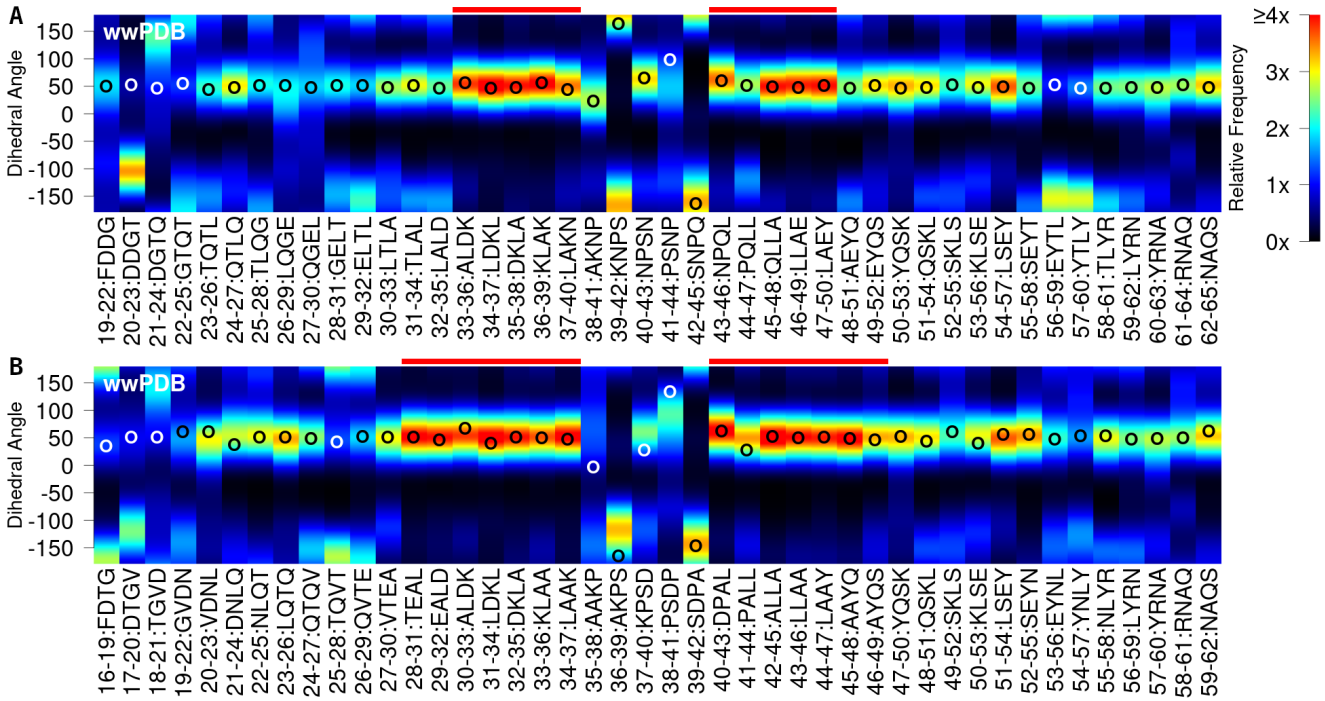

Figure 19: Sequence analysis for PrgI and MxiH proteins.

Tetrad heatmaps for wwPDB fragments matching the sequence from MxiH (A) and PrgI (B) proteins. Open circles indicate the series of dihedral angles found in MxiH (pdb: 2CA5) and PrgI (pdb: 2JOW). Symbols are colored black or white to maximize contrast. Regions of the heatmap are highlighted at the top of the diagram to align to emphasize symmetry about the turn region.

Structure in the NPSNP turn was nucleated first around the proline residues, then at the C-terminus of the turn (Helix-2), and finally at the N-terminus and center of the turn. The NPSNP sequence produces a turn larger than what is optimal for an all-alanine helix-helix interface; variability at the N-terminal side of the turn (Figure 11A$_{iv}$) may result from suboptimal hydrophobic packing. Heatmaps for the MxiH and PrgI proteins exhibit an interesting trend shared between the two proteins; the most helical regions of the proteins directly flank the turn (Figure 19). Furthermore, the two helical regions of high propensity that flank the turn are nearly

identical in length within each protein. This pattern suggests the turn directly orients the two flanking helical segments and triggers zipper-folding across the remainder of the helix bundle. A compensatory effect for differences in turn optimization may be visible between the two proteins. The geometry for the NPSNP segment of MxiH is more commonly found in the wwPDB (Figure 19A) than the corresponding geometry for the KPSDP segment of PrgI (Figure 19B). The less optimized turn of PrgI may be compensated by the longer region of highly helical tetrads flanking the turn; PrgI has 7 tetrads whereas MxiH has 5 tetrads.

Both active and counter-active turns were found in the UBA(1) three helix bundle. Turn-1 formed an early on-target extended conformation (Figure $15C_{iii}$) that placed the hydrophobic Met and Tyr residues close together. The C-terminal end of Turn-1 initially formed a premature Helix-2 (Figure $15C_{iii}$), but later transitioned to the correct extended structure (Figure $15B_{iii}$). The structure of Turn-1 was preserved in our single-turn AK42r6_Nloop design (Figure $17A_{iii}$) and even caused random turns to form in the Turn-2 region (Figure $17A_{iv}$). With consistent geometry in Turn-1 established, the behavior of flanking residues determines the role of the turn in the folding process. For AK42 designs, high helix propensity flanks the turn and enables the turn sequence to orient the two helices. In UBA(1), hydrophobic interactions between Turn-1 (Met-173 and Tyr-175), Ile-

170, and Val-180 stabilize the flanking helical segments and nucleate the hydrophobic core of the protein. In both systems, these features enable the turn sequence to have an active role in folding. Experimental measurements of a UBA(1) variant support this interpretation. Stabilization of the Turn-1 sequence caused stabilization of UBA(1)'s folding transition state, indicating native-like structure for Turn-1 is formed early in the folding process[27].

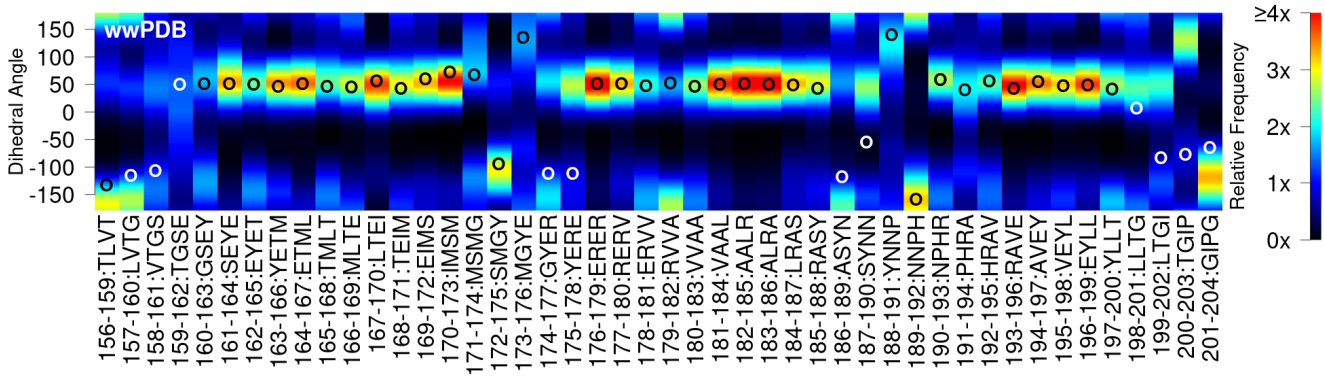

Figure 20: Sequence analysis for UBA(1) protein.

Tetrad heatmaps for wwPDB fragments matching the sequence from UBA(1). Open circles indicate the series of dihedral angles found in UBA(1) (pdb: 6W2H). Symbols are colored white or black to maximize contrast.

The second turn in UBA(1) was found to be counter-active in both AK42 and UBA(1) systems. All previous turn geometries – excluding the PSDP simulations – utilized geometries similar to populations in the wwPDB. Both UBA(1) and the AK42r6_NCloop cluster-1 structures had one tetrad that was outside populations found in the wwPDB. This suggests the preferred geometry of Turn-2 is in conflict with the tertiary structure of UBA(1). Consequently, CAMPARI simulations for AK42r6_NCloop did not form Turn-2 geometry similar to UBA(1).

Instead, it appears the structure of Turn-2 is partially restructured by tertiary interactions, notably for the 187-190:SYNN tetrad of UBA(1) (Figure 20). The utilized combination of active and counter-active turns in UBA(1) may provide a necessary mechanism to order protein folding events for consistent folding.

Overall, there was widespread agreement between simulation structures and tetrad populations in the wwPDB. Differences in simulation results due to rearranged alanine/lysine residues were matched with corresponding shifts in wwPDB distributions. This highlights the importance of amino acid context and the value of information contained in the wwPDB tetrad distributions. Each examined tetrad sampled 2 populations of conformers in the wwPDB. Simulations forming off-target structure often sampled the minor population of the wwPDB distribution instead of the on-target, major population. Data for the SYNN tetrad (Figure 20) shows that the heatmap is resistant to distortion from outliers; the uncommon geometry of SYNN in UBA(1) is included in the wwPDB dataset. This is, of course, dependent on the representation of each tetrad within the wwPDB. Assessment of heatmaps for Cytochrome c', MxiH, PrgI, and UBA(1) proteins suggests turn structure and folding mechanisms may be sufficiently characterized by comparing experimental structures to wwPDB tetrad distributions.

Turn structure was formed early in folding near proline residues (MxiH, PrgI, UBA(1):Turn-2) and near interacting hydrophobic residues (UBA(1):Turn-1). These turns were either active or counter-active, depending on their structural context. Hydrogen bonds expected to stabilize turns were poorly formed in CAMPARI, suggesting these interactions are supplementary and do not actively guide structure. The steric effects and attractive forces between different sidechains had a more significant role in establishing backbone geometry in turn sequences.

Chapter 3: Empirical C-Alpha Stability Tool (EmCAST)

Research covered in this chapter has been published in the Journal of the American Chemical Society[27].

3A: Introduction

At the earliest stages of protein folding the nucleation of structure should be dominated by protein-water and localized protein-protein interactions. The work presented in this chapter seeks to define the relationship between energy and structure under such conditions. Free energy (G) can be explicitly described in terms of enthalpy (H), entropy (S), and temperature (T):

$$\Delta G = \Delta H - T\Delta S$$

Modeling energy in this approach requires meticulously accurate summation of the many enthalpic and entropic forces involved – a non-trivial process. Atomic force fields used in molecular models for proteins rely on useful approximations to evaluate enthalpic and entropic contributions to free energy. Summation over the thousands of approximations is prone to accumulated systematic error,

restricting model accuracy[28]. An alternative approach to model free energy can be achieved using a population equilibrium (K):

$$\Delta G = -RT \cdot \ln(K)$$

This approach only needs to know the populations of different states to resolve free energy. The enthalpic and entropic forces involved are implicitly represented within the population distribution.

The two strategies can be contrasted in the context of identifying the lowest energy conformer of a 4-residue peptide. The explicit model needs to evaluate polar interactions, hydrogen bonding, Van der Waals forces, conformational entropy of the peptide, and the entropy of the solvation shell around the peptide. The energy equation must be coupled with a downhill, random, or exhaustive search through possible peptide conformations. In the implicit approach, the lowest energy conformer is simply the most common conformer. The apparent difference in complexity for the two approaches is balanced by differences in the experimental datasets the methods are based on. The explicit method relies on the experimental characterization of fundamental forces that are independent of peptide composition or environment. The implicit approach requires experimental measurements of each peptide in the environment of interest.

To build the ideal dataset for the implicit method it is technically feasible to use NMR techniques to identify peptide structures and populations in an aqueous solution – but doing so for every possible peptide is infeasible. Moving beyond this ideal dataset, an experimentally measured distribution of peptide conformers can be observed in protein fragments taken from the wwPDB[3]. Peptide conformers sampled in the wwPDB will be influenced by both the complex interactions from a protein's tertiary fold and by the local preferences of the peptide sequence. The varied influence of different tertiary folds on a peptide's conformer should produce random deviations in conformer preference. In contrast, the innate local preferences of the peptide should provide consistent biases within every sample. With sufficient sampling, the local preferences will emerge while the influences of tertiary structure will form a noisy background.

Software was developed to implement this strategy to model the free energy of peptides using conformer distributions extracted from the wwPDB. The key parameters involved in the calculation are the peptide's sequence and the peptide's conformer. Varying either or both of these parameters produces a difference in free energy calculations (ΔΔG) that can be tested experimentally. A straightforward test is to mutate a single residue in the peptide's sequence while keeping the peptide's conformer constant. This can be

realized by making mutations at a solvent accessible site in a well-structured protein. A solvent accessible site is chosen to limit interactions of the mutation site to the sequence-local interactions we are modeling. Tertiary interactions supporting flanking residues help restrict the fragment of interest to a single conformer. Differences in free energy between the two variants can be measured and compared to the calculated energy difference. Calculations were tested using previous data published in literature and by experimentally measuring the change in stability for new mutations calculated to be stabilizing. Free energy calculations were also used to generate 3D models, folding funnels, and folding pathways for short peptides.

3B: Database and Software Design

3B.1: Heatmap Generation

All possible peptide conformers must be equitably enumerated in order to evaluate ΔG based on population equilibrium. Effectively fixed sample sizes from the wwPDB constrain the potential resolution of conformer enumeration for which meaningful statistics can be derived. For a four-residue peptide (tetrad), the principal peptide dihedral angles phi (φ) and psi (ψ) produce 360^7 ($7.83 \cdot 10^{17}$) possible conformers when using a 1° bin. This greatly exceeds the number of samples per tetrad in the wwPDB (mean = 532, median = 324, range = 0-

10559; 20180101 snapshot). The four residue C_α dihedral angle (τ) captures the overall shape of the peptide backbone and reduces enumeration to 360 (1° bins). The two excluded angles (θ_1 and θ_2) are well correlated with τ[29] and can be considered implicitly represented within τ. The different amino acid sidechains create distinct τ distributions for the 20^4 (160,000) possible tetrads. This demonstrates that the effects of sidechain structure are well represented within τ and can also be considered implicitly.

Protein fragments were extracted from the wwPDB using the 20180101 snapshot to develop a conformer distribution for every tetrad using the τ dihedral angle. Over 80% of the tetrads had at least 100 samples; 0.37% (590 tetrads) had no samples in the wwPDB. Every protein in the database was split into each possible 12-residue fragment.

xxx**ABCD**xxx

Fragments were grouped by the amino acid identity of the central 4 residues (**ABCD**). Fragments identical at all 12 residues had their sample contributions weighted to account for sampling biases due to redundancy in the wwPDB. This approach was chosen over non-redundant databases to incorporate information from the many sequence variants of proteins found within the complete wwPDB. A second weighting parameter was added to the sampling algorithm to bias data towards solvent-exposed fragments. Solvent-exposed sites in a protein are

expected to better reflect the intrinsic structural preferences of the primary structure because they are less likely to be affected by long-range tertiary contacts[30]. Solvent accessibility of the central 4 residues was calculated using STRIDE[31]. Fragments containing covalent modifications were discarded, except for disulfide bonded cysteines within the central 4 residues (**ABCD**). The four-residue C_α dihedral angle (τ) was measured for the three center-most segments within each fragment (X**ABC**, **ABCD**, **BCD**X).

The collection of fragments for each possible tetrad were used to generate population heatmaps based on τ. Two heatmaps were generated for each tetrad to consider the influence of neighboring τ: an upstream heatmap [τ(X**ABC**) vs. τ(**ABCD**) vs. frequency] and a downstream heatmap [τ(**ABCD**) vs. τ(**BCD**X) vs. frequency]. During heatmap generation, samples were weighted by their fractional solvent accessibility, f_{SA}. The data from STRIDE at each residue of the 4-residue sequence (**ABCD**) was used to calculate solvent accessible surface area for each tetrad, with fraction solvent accessibility calculated using the sum of the solvent accessibility of the four residues in Gly-X-Gly peptides[32] as the maximum possible solvent accessibility for that tetrad. Heatmap samples were smoothed across a 90° radius about the τ angle pair (τ_x, τ_y) using a quadratic decay. The weighting for each point within the circle is calculated as:

$$w(\tau_{xs}, \tau_{ys}) = (1 - (D/90°))^2 \qquad \text{(Eq. 3.1)}$$

where D is the radial distance between smoothed point (τ_{xs}, τ_{ys}) and the central τ angle pair (τ_x, τ_y). Smoothing is done using a grid of τ pairs with τ_x and τ_y having integer values ranging from -179° to 180°. Radial distance is calculated as $D = (|\tau_{xs} - \tau_x|^2 + |\tau_{ys} - \tau_y|^2)^{1/2}$ and $w(\tau_{xs}, \tau_{ys})$ is evaluated for $D \leq 90$. Thus, $w(\tau_{xs}, \tau_{ys})$ is effectively set to 0 for all points outside the circle. An example of the resultant smoothing for a single fragment sample is shown in Figure 21A. The net equation for each sample contribution to the heatmap is:

$$\text{sample} = w * f_{SA} * n^{-1} \qquad \text{(Eq. 3.2)}$$

where w is the smoothing weight, f_{SA} is the fractional solvent accessibility, and n is the number of fragments with identical 12-residue sequences.

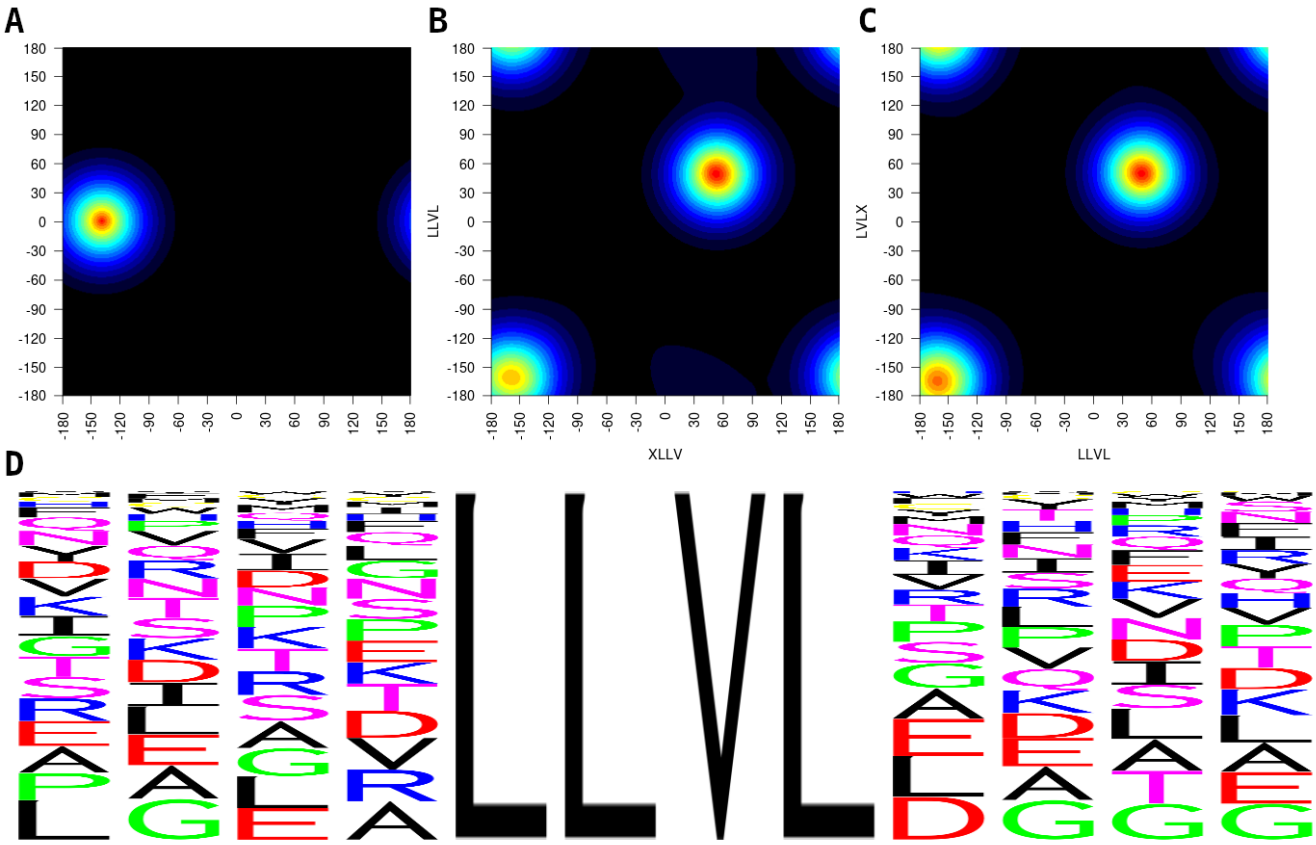

Figure 21: Example fragment heatmaps and sequence motif.

(A) Visualization of the smoothing process for a single datapoint centered at -140°, 0°. (B-D) Data is included for the LLVL fragment collection composed of 308 unique fragments (3510 total). Upstream τ distributions (B, XLLV and LLVL) and downstream τ distributions (C, LLVL and LVLX) are included. A sequence logo for the weighted collection of 12-residue fragments in the LLVL dataset is included (D).

The values from all tetrad samples are summed into 2D heatmaps that yield smooth population landscapes (Figures 21B-C, 22). Example heatmaps are included for the LLVL fragment collection (Figure 21B-C). LLVL heatmaps show a preference for either helical (τ ≈ 55°) or extended (τ ≈ -160°) geometry that is sensitive to structural context; it prefers helix geometry in helical contexts and extended geometry in extended contexts. A sequence logo of weighted LLVL

fragments visualizes the variety of sequences sampled within the fragment collection (Figure 21D). The generated 2D heatmaps are saved as binary files to be used in runtime-optimized energy calculations. The 2D heatmaps were also condensed into 1D heatmaps to simplify both data visualization and peak searching algorithms. The 2D data for either upstream (**ABCD** vs. X**ABC**) or downstream (**BCD**X vs. **ABCD**) heatmaps can be converted to a 1D dataset for the **ABCD** tetrad by computing the population sum for each **ABCD** τ angle across all adjacent (X**ABC** or **BCD**X) τ angles. The 1D heatmaps were saved as both colorized bitmaps and as binary files.

3B.2: Free Energy Calculations

Multiple heatmaps from our fragment database are used to calculate the difference in native state free energy between two sequence/structure pairs. Unless otherwise stated, the same native state structure is used in the compared sequence/structure pairs when modeling protein variants; only the sequence is changed. Thus, an important assumption is that the structure of the protein does not change when the mutation is made. Each 4-residue segment containing the mutation site is used in the calculation, each segment contributing 2 heatmaps (upstream and downstream). For a proposed UBA(1) Y188G mutation, the segments used are RASY, ASYN, SYNN, and YNNP for WT and RASG, ASGN, SGNN, and GNNP for the mutant (MT) – providing a total of 16 heatmaps. Four of these heatmaps (for SYNN

and SGNN) are shown in Figure 22. Together this scores the mutation in the context of a 9-residue target structure (zBBB**X**BBBz) where sequence and structure are modeled at the mutation site (**X**) and the 6 residues flanking it (B). Only backbone geometry is modeled for the outermost residues (z). This only scores residue-residue interactions between the mutation site and the 6 residues surrounding it in sequence (BBB**X**BBB).

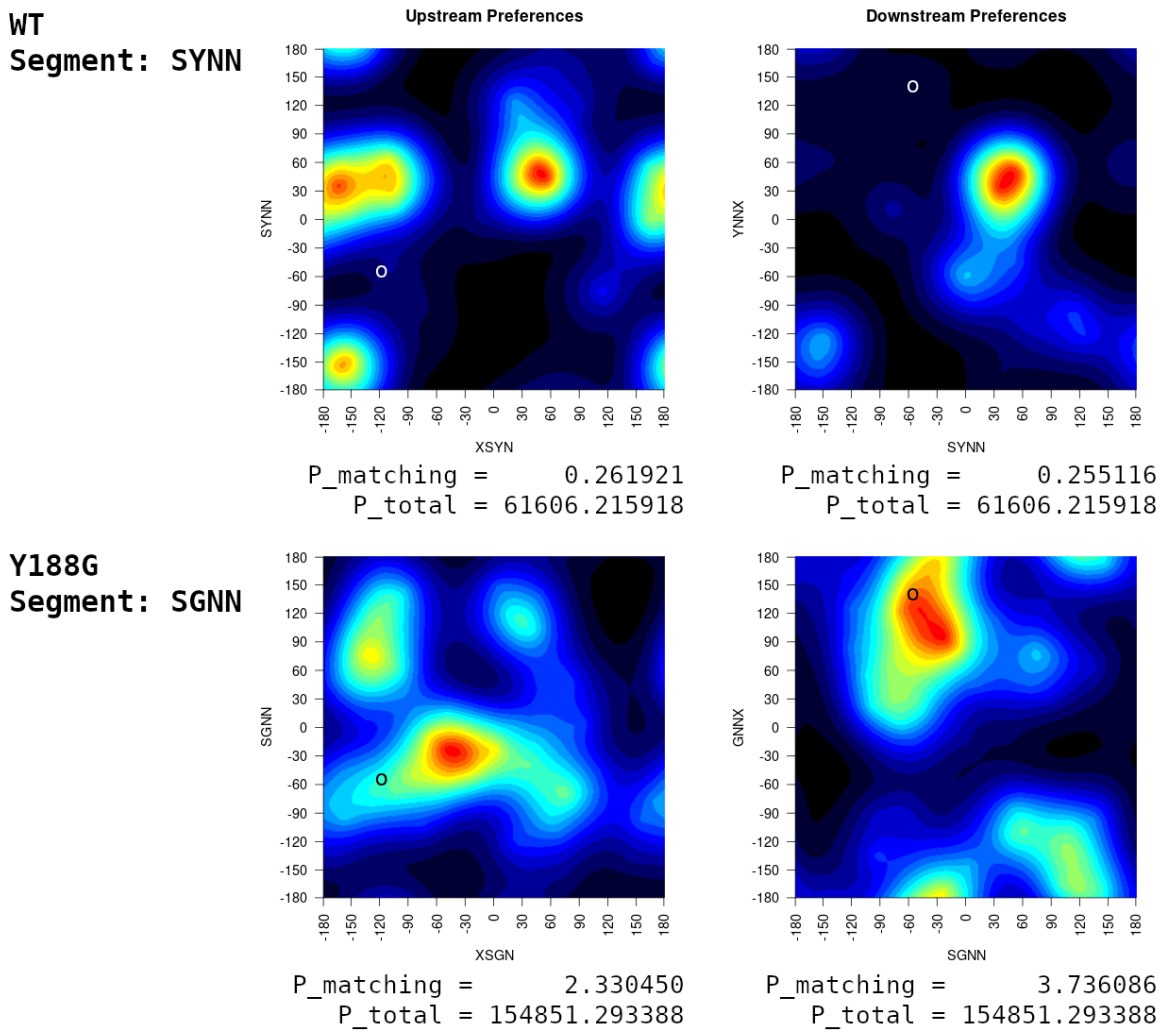

Figure 22: UBA(1) fragment heatmaps and population calculations for a mutation

Select fragment τ,τ population heatmaps used to predict the UBA(1) Y188G mutation. Heatmaps and values used in the SYNN → SGNN segment calculation are shown. The central τ(SYNN/SGNN) is plotted on the X-axis of downstream maps and the Y-axis of upstream maps. Upstream τ(XSYN/XSGN) is plotted on the X-axis of upstream maps. Downstream τ(YNNX/GNNX) is plotted on the Y-axis of downstream maps. Heat values show the fragment populations of τ,τ pairs found in the SYNN (WT) and SGNN (Y188G) fragment collections. The corresponding τ,τ pairs found in the WT native structure (pdb: 6W2H) are shown with circles. Circles are colored black or white to maximize contrast.

The population data in each heatmap at a target τ,τ pair ($P_{matching}$) and the summed total of heatmap values (P_{total}) are used to calculate $\Delta G_{heatmap}$ (Eq. 3.3, Figure 22). Temperature (T) is set to

match the experimental conditions of interest (typically 298.15 K) unless otherwise noted. The free energy difference between WT and MT versions of the same heatmap (ie: $SYNN_{upstream}$ and $SGNN_{upstream}$) is $\Delta\Delta G_{heatmap}$ (Eq. 3.4). The two upstream/downstream $\Delta\Delta G_{heatmap}$ values are averaged and scaled (divided by three to account for overlap during summation because each dihedral angle is sampled three times: central, preceding and following positions during the summation) to determine $\Delta\Delta G_{segment}$ (Eq. 3.5). Finally, $\Delta\Delta G_{variant}$ is determined by summing $\Delta\Delta G_{segment}$ for all segments containing the mutation site (Eq. 3.6).

$$\Delta G_{heatmap} = RT \cdot \ln\left(\frac{P_{matching}}{P_{total} - P_{matching}}\right) \quad \text{(Eq. 3.3)}$$

$$\Delta\Delta G_{heatmap} = \Delta G_{heatmap}(MT) - \Delta G_{heatmap}(WT) \quad \text{(Eq. 3.4)}$$

$$\Delta\Delta G_{segment} = \frac{0.5 \cdot \left(\Delta\Delta G_{heatmap}(Upstream) + \Delta\Delta G_{heatmap}(Downstream)\right)}{3} \quad \text{(Eq. 3.5)}$$

$$\Delta\Delta G_{variant} = \sum \Delta\Delta G_{segment} \quad \text{(Eq. 3.6)}$$

This approach predicts changes in native state free energy (ΔG_N) due to changes in local interactions (Figure 23). Theoretically, the calculated $\Delta\Delta G_{variant}$ will only match changes in the free energy of unfolding ($\Delta\Delta G_u$) when two conditions are met. Only mutations that do not interact with residues outside the model window (BBBXBBB) will have $\Delta\Delta G_{variant}$ equivalent to ΔG_N. Additionally, only mutations that do not alter the free energy of the denatured state ($\Delta G_D = 0$) will have ΔG_N equivalent to $\Delta\Delta G_u$. Other complications may arise if incorrect assumptions are made about the structures used in the calculation

(i.e., that WT and MT native structure backbones are identical) or if the structural probe used to measure $\Delta\Delta G_u$ is insensitive to structural changes at the mutation site.

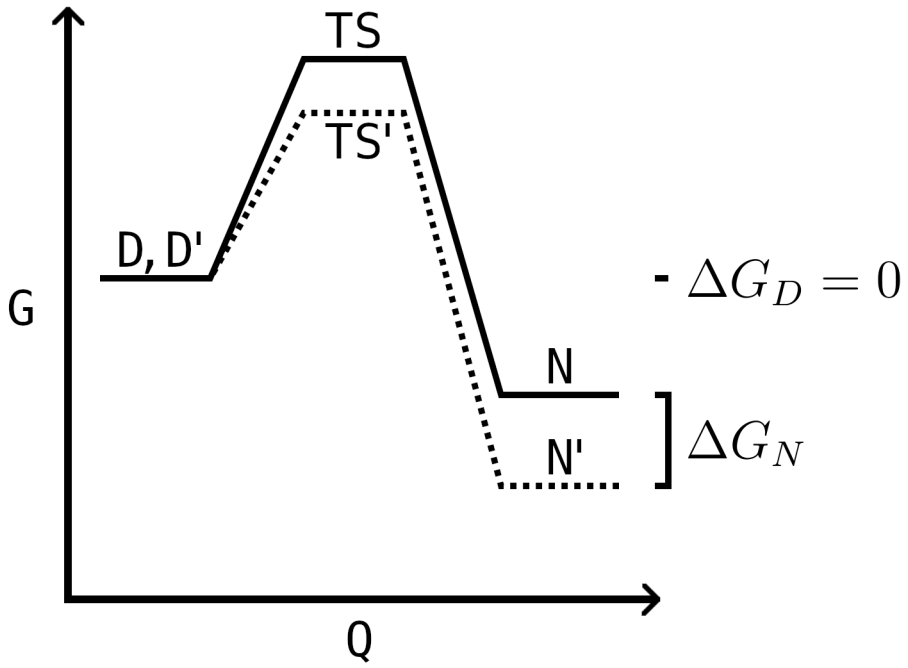

Figure 23: Two-state protein folding reaction

Denatured (D), transition (TS), and native (N) states are visualized on a free energy (G) vs. reaction coordinate (Q) diagram. A stabilized variant is depicted by the dashed line and D', TS', and N' states. The mutation induced change in native state energies is highlighted ($\Delta G_N = G_N - G_{N'}$). An ideal mutation for native state structural analysis will not alter the free energy of the denatured state ($\Delta G_D = G_D - G_{D'} = 0$).

Initially, double-precision floating-point values (8 byte) were used in the 360x360 heatmaps to store the population values for each τ,τ pair. To reduce memory requirements, single-precision floating-point (4 byte), unsigned 16-bit integer (2 byte), and unsigned 8-bit integer (1 byte) formats were tested. To convert to integer values, floating-point values were scaled to utilize the full range of each

integer data type; population totals were recalculated. Free energy calculations were consistent up to the hundredth of a kcal/mol between double-precision floating-point, single-precision floating-point, and unsigned 16-bit integer data types – unsigned 8-bit integers had large deviations. Switching to the lower resolution 16-bit datatype reduced the total heatmap storage requirements to 78 GB (from 308 GB for double-precision floating-point). A software daemon was written in C to cache all available heatmaps in memory, listen for calculation requests, and quickly respond with the calculation result and sampling information. Client software was written in JavaScript to process protein PDB structures, send sequences/dihedrals to the daemon, and process/visualize the results. A web application, the Empirical C-Alpha Stability Tool (EmCAST), facilitating access to the method has been deployed and is available at https://emcast.org for academic use.

3B.3: Sequence to Structure Modeling

The free energy calculations developed can be used to compare different protein structures with identical sequence. This leads to evaluation of ΔG using the same 2D heatmaps but at different τ,τ coordinates. For a select protein sequence, a series of τ dihedral angles can be selected, scored, and used to reassemble a protein backbone. To explore possible conformers, the major and minor τ peaks were selected from each tetrad's 1D heatmap. For each tetrad, there

are typically 2 to 4 major and minor peaks in the τ conformational distribution. A list of all possible combinations of tetrad dihedrals is then generated for the sequence of interest corresponding to the set of all possible structures for the sequence.

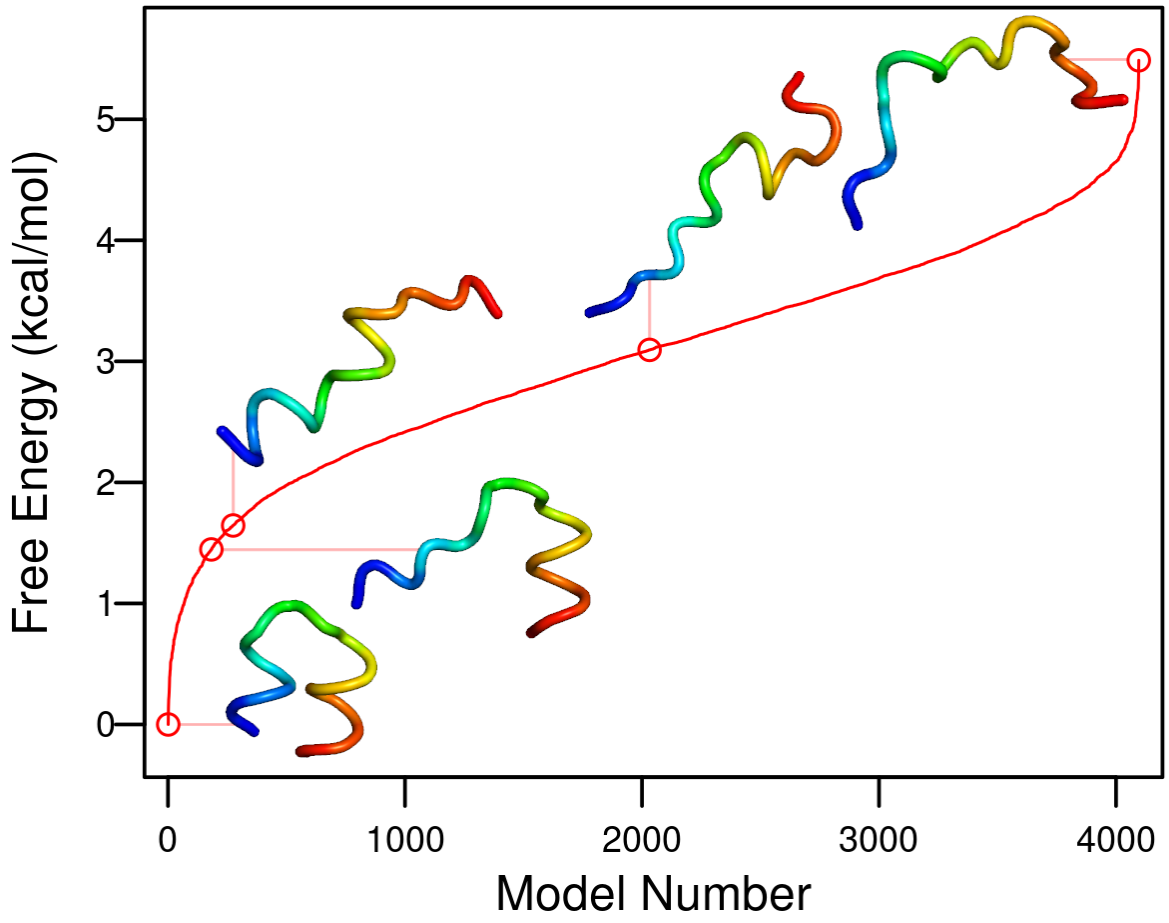

Figure 24: Example folding funnel from sequence to structure modeling

The computed folding funnel for a segment of WT UBA(1) is shown. The sequence modeled includes turn 1 (T1) and portions of helix 1 (H1) and helix 2 (H2) that flank it (TEIMSMGYERERVVAA). Select models are shown, colored from N (blue) to C (red).

For the set of structures generated, the bond angle between three adjacent C_αs is set to 91.72°. Although this bond angle can

vary[29], we use the average value for this bond angle for α-helical structure in our fragment database as a simplification; most of the structures modeled in this dissertation are α-helical. The C_α-C_α bond length is set to 3.84 Å. Positions of C_α atoms for each possible structure are computed using these geometric parameters; no other atoms are modeled. Free energy values are calculated using Eq. 3.3. $\Delta G_{heatmap}$ is used in place of $\Delta\Delta G_{heatmap}$ in Eq. 3.5. For each conformer, the summation in equation 3.6 is then over all dihedral angles for that conformer. The conformer with the lowest energy is used as a ground state (ΔG = 0 kcal/mol) on the energy scale; less stable conformers have positive values for ΔG. The ensemble of structures generated produce a relatively smooth distribution of energy states available for the peptide sequence (Figure 24).

3C: Testing Data from Literature

Solvent exposed point mutations in proteins reported in literature provided the initial means of validation for energy calculations during software development. Surface mutations in the FF Domain from HYPA/FBP11[33] were the first dataset EmCAST calculations were tested against. The stabilities of FF Domain variants were characterized by urea unfolding experiments monitored by fluorescence in 50 mM sodium acetate (pH 5.7) and 100 mM sodium chloride. Two different NMR structures exist for the FF Domain under native

conditions: a native state structure (pdb: 1UZC, Figure 25D) and a transiently populated intermediate (pdb: 2KZG, Figure 25A)[34].

$$N\ (1UZC) \rightleftharpoons I\ (2KZG) \rightleftharpoons D$$

We suspect stability data taken from unfolding experiments reports on the $I \rightleftharpoons D$ transition, with the $N \rightleftharpoons I$ transition occurring earlier with less change to the fluorescence signal. This may be evident in the slight deviations from sigmoid unfolding behavior between 1 and 2 M urea (Reference [33], Figure 2, 320 nm).

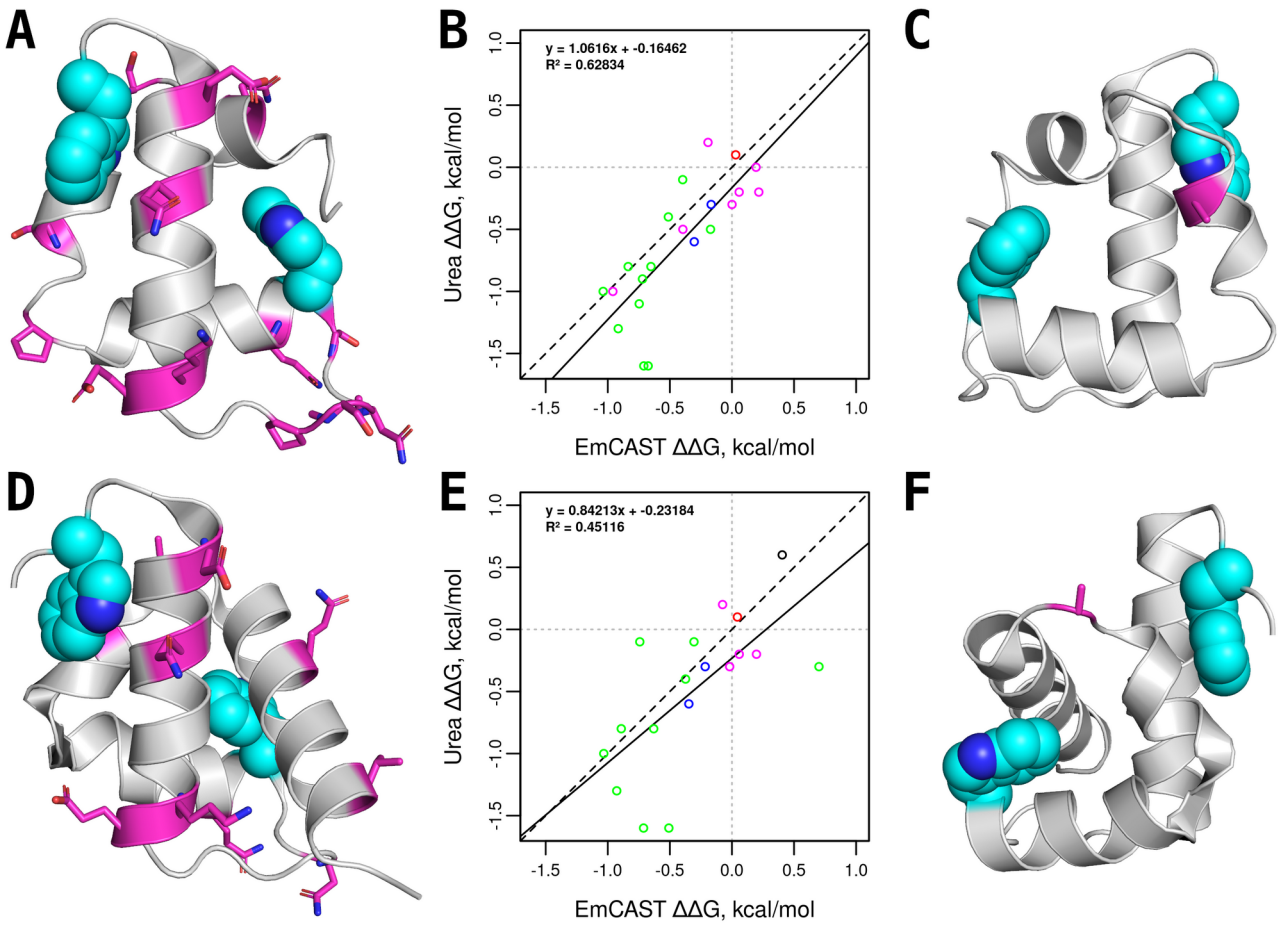

Figure 25: EmCAST predictions for FF Domain NMR structures.

Predictions were made using the intermediate (A-C, pdb: 2KZG) or native (D-F, pdb: 1UZC) NMR structures. Experimental data from literature were taken from fluorescently monitored urea melts performed in 50 mM sodium acetate (pH 5.7) and 100 mM sodium chloride. Selected mutation sites are colored magenta in the protein structure (A, D). Fluorescent probes used to monitor unfolding are colored cyan. The solid black line in the correlation plots (B, E) is the line of best fit; the dashed black line marks the location of a perfect fit. Datapoint colors classify the type of mutation using the first matching category: gain/loss of a glycine or proline (green), gain of a negative charge (red), gain of a positive charge (blue), gain/loss of a polar sidechain (magenta), default (black). The A53G mutation site, which produces different ΔΔG predictions between the two structures, is highlighted in panels C (intermediate) and F (native).

We have opted to use the intermediate structure (2KZG) in our stability tests. Further evidence may lie in the destabilizing A53G mutation, which is found in either a turn (N, 1UZC, Figure 25F) or a helix (I, 2KZG, Figure 25C). EmCAST predicts A53G to be stabilizing

in the native state structure (+0.738 kcal/mol) and destabilizing in the transition state intermediate structure (-0.334 kcal/mol). The experimental stability measured for this mutation is -0.30±0.08 kcal/mol, consistent with the EmCAST prediction using the intermediate structure. Correlation between experimental stability values and EmCAST calculations using the intermediate structure were reasonable (Figure 25B), producing a slope near 1 (m = 1.06), an intercept near 0 (b = 0.16 kcal/mol), and an R^2 correlation coefficient of 0.63.

To further test the generality of EmCAST to model mutations at surface-exposed sites, we searched the ProThermDB[35] and the folding literature for mutation sets at surface-exposed sites. Data were limited to monomeric proteins with two-state unfolding and at least 10 surface mutations. We felt it was important to have at least 10 mutations in each protein to determine if there were qualitative differences between EmCAST's accuracy across different protein types. Datasets were found for the B-Domain of Staphylococcal Protein A (pdb: 1SS1), barnase (pdb: 1BNI), the src SH3 domain (pdb: 1SRL), Chymotrypsin inhibitor 2 (pdb: 2CI2), the N-terminal domain of Ribosomal Protein L9 (pdb: 2HBB), Staphylococcal nuclease (pdb: 1STN), T4 Lysozyme (pdb: 1L63), and RNase T1 (pdb: 9RNT). Stability correlations were found to be sensitive to experimental conditions (salt concentrations and buffer pH) and whether the protein has

thermodynamically significant residual structure in the denatured state.

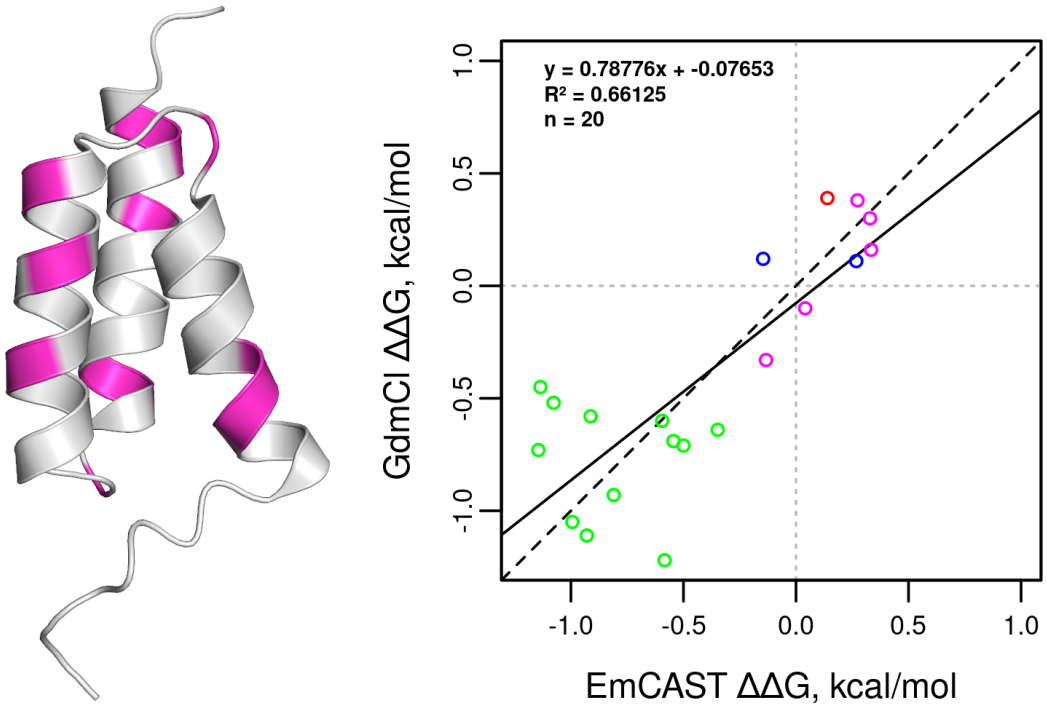

Figure 26: EmCAST predictions for the B-Domain of Staphylococcal Protein A.

Predictions were made using a solution NMR structure (pdb: 1SS1). Experimental data from literature were taken from CD monitored GdnHCl melts performed in 50 mM sodium acetate (pH 5.5) and 100 mM sodium chloride. Selected mutation sites are colored magenta in the protein structure. Datapoint colors classify the type of mutation using the first matching category: gain/loss of a glycine or proline (green), gain of a negative charge (red), gain of a positive charge (blue), gain/loss of a polar sidechain (magenta), default (black).

Twenty surface exposed variants of the B-Domain of Staphylococcal Protein A had stability measurements determined from CD monitored GdnHCl melts performed in 50 mM sodium acetate (pH 5.5) and 100 mM sodium chloride[36]. Correlations between EmCAST calculations and B-Domain experimental measurements (Figure 26) were similar in quality to the previous FF domain correlation. In the B-

Domain, the slope is slightly skewed away from 1 (m = 0.79) by a group of destabilizing variants. The intercept is near 0 (b = -0.08 kcal/mol) and the R^2 correlation coefficient is 0.66.

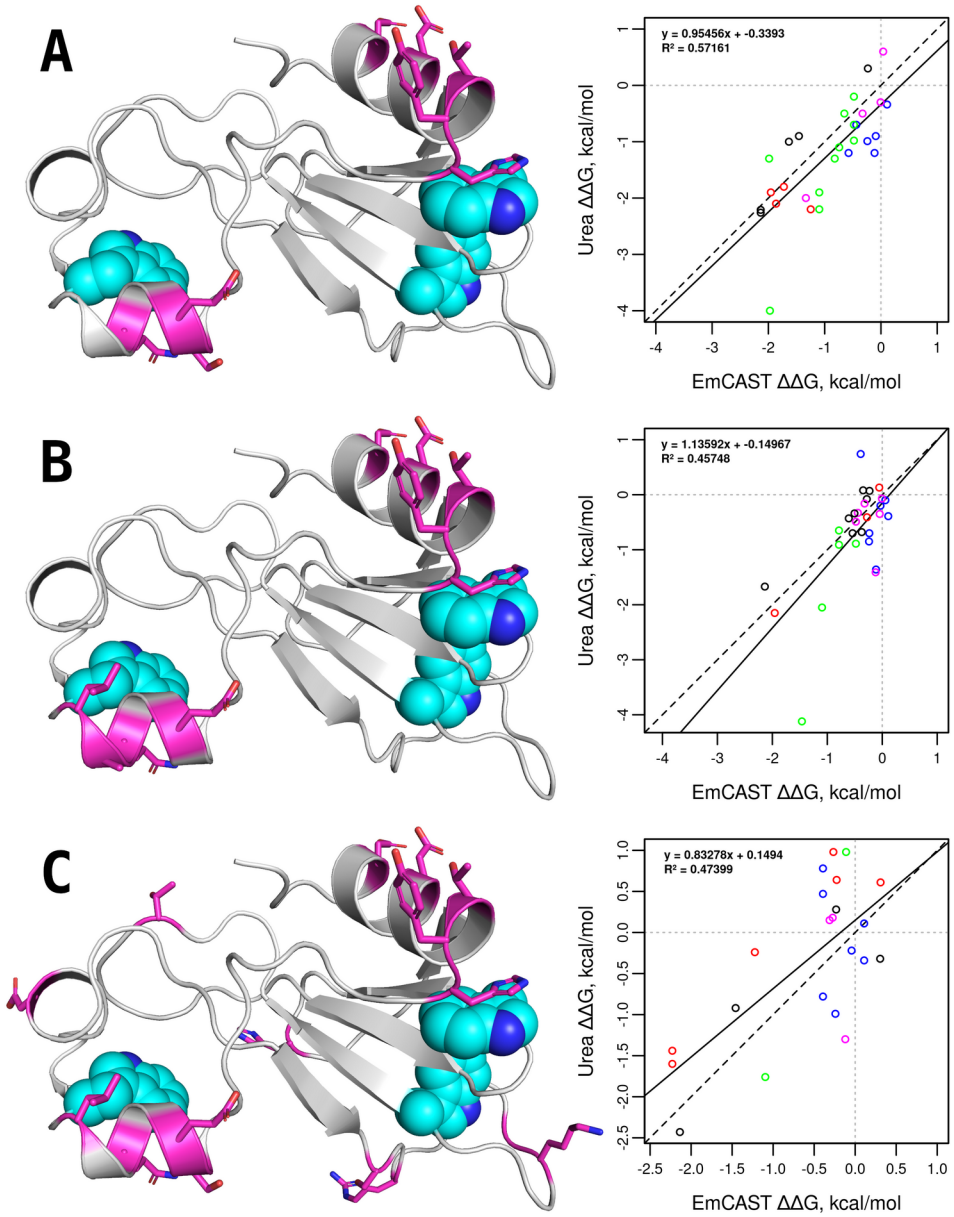

Figure 27: EmCAST predictions for barnase.

Predictions were made using an X-ray structure (pdb: 1BNI). Experimental data from literature were taken from fluorescently monitored urea melts performed under various conditions: (A) 1.0 M MES (pH 6.3) and 613 mM sodium chloride, (B) 450 mM to 1.0 M MES (pH 6.3), and (C) 50 µM to 50 mM MES (pH 6.3). The selected mutation sites are colored magenta. The fluorescent probes used to monitor unfolding are colored cyan. Datapoint colors classify the type of mutation using the first matching category: gain/loss of a glycine or proline (green), gain of a negative charge (red), gain of a positive charge (blue), gain/loss of a polar sidechain (magenta), default (black).

Stability measurements for barnase variants exist for various experimental conditions, namely at different buffer and salt concentrations[37][38][39][40][41][42]. Stability values were taken from fluorescently monitored urea melts at pH 6.3. Data was sorted into three different groups of conditions: 1.0 M MES / 613 mM NaCl (Figure 27A), 450 mM to 1.0 M MES / No NaCl (Figure 27B), and 50 µM to 50 mM MES / No NaCl (Figure 27C). EmCAST calculations correlated best with the data collected with NaCl present (Figure 27A), producing a slope near 1 (m = 0.95) and a reasonable correlation coefficient (R^2 = 0.57). The intercept was below 0 (b = -0.34 kcal/mol), influenced by a destabilizing outlier. For data collected without salt (Figure 27B-C), high errors occurred for charged mutations that were predicted by EmCAST to have a minimal influence on stability. Comparison of stability measurements for WT barnase and T16R barnase[43] highlights the sensitivity of stability measurements to salt content (Table 4). The calculation for the T16R variant by EmCAST is -0.39 kcal/mol (pdb: 1BNI), which is similar to measurements made in the 50-500 mM NaCl range.

Buffer	Salt	WT ΔG, kcal/mol	T16R ΔG, kcal/mol	T16R ΔΔG, kcal/mol
10 mM Na-MES	None	10.01	9.23	-0.78
50 mM Na-MES	None	10.15	10.63	0.48
50 mM Na-MES	50 mM NaCl	10.37	10.21	-0.16
50 mM Na-MES	500 mM NaCl	10.88	10.55	-0.33
50 mM Na-MES	900 mM NaCl	12.90	12.90	0.00

Table 4: Comparison of barnase T16R stability measurements

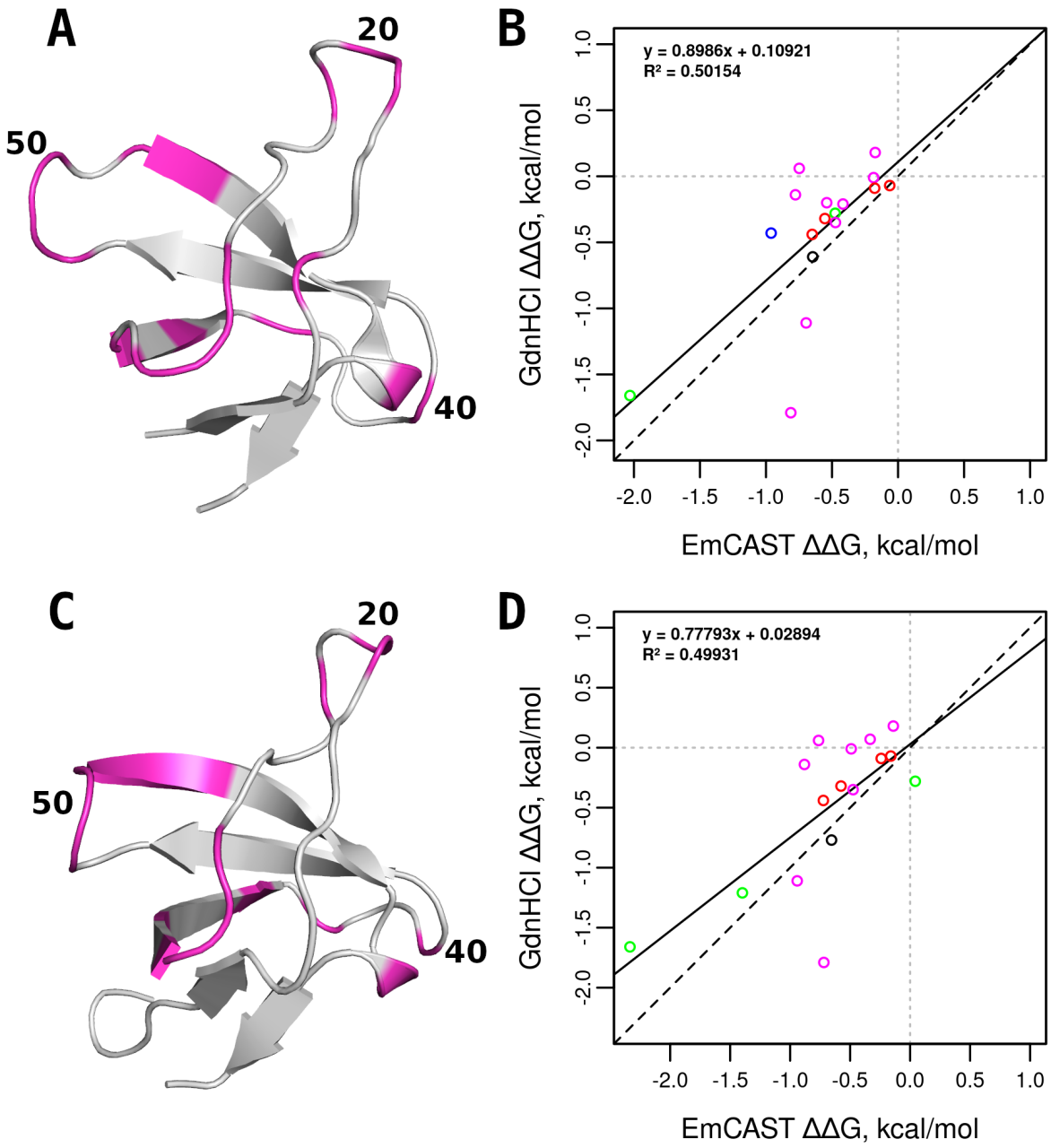

Figure 28: EmCAST predictions for the src SH3 domain.

Predictions were made using either a solution NMR structure (A-B, pdb: 1SRL) or an X-ray structure (C-D, pdb: 4JZ4). Experimental data from literature were taken from GdnHCl stopped flow folding kinetics experiments monitored by fluorescence in 50 mM sodium phosphate (pH 6) at 295K. Selected mutation sites are colored magenta in the protein structure. The approximate positions of residues 20, 40, and 50 are indicated. Datapoint colors classify the type of mutation using the first matching category: gain/loss of a glycine or proline (green), gain of a negative charge (red), gain of a positive charge (blue), gain/loss of a polar sidechain (magenta), default (black).

Data on surface exposed mutations in β-sheet structure are available for the src SH3 domain. Experimental values were measured using GdnHCl stopped flow folding kinetics experiments monitored by fluorescence in 50 mM sodium phosphate (pH 6) at 295K[44]. One datapoint was manually excluded from the dataset, Y14A; this residue forms long-range hydrophobic interactions with Y60 despite being solvent-accessible. Calculations by EmCAST are sensitive to backbone geometry and do not account for structural rearrangements in proteins caused by the introduced mutation(s). Calculations are only relevant at surface exposed positions where long range sidechain-sidechain interactions are absent. These considerations proved problematic for the src SH3 domain: the backbone geometry, solvent accessible surface area, and EmCAST stability predictions varied across different experimental structures for a subset of the available mutations.

Notable variation can be seen in the loop regions in the available NMR and X-ray structures (Figure 28A,C). Mutations in these two regions were predicted to be destabilizing by EmCAST, but were measured to be neutral experimentally (Figure 28B,D: datapoints near y = 0). Together, this suggests these two loops are prone to rearrangements and have a minor role in defining src SH3's structure. Predictions were more accurate for the 50's loop. Stability loss in the most destabilizing mutation included, T50A (Figure 28B,D: y = -1.79), was notably under predicted by EmCAST. This may be a sampling

issue within EmCAST; unfavorable sequence/structure pairs have lower representation within the underlying fragment database. Higher EmCAST prediction error is commonly seen for the most destabilizing mutations in previous proteins as well (Figures 25B, 26, 27A). Overall, predictions were reasonable with a slope close to 1 (m = 0.78-0.90), an intercept near 0 (b = 0.03-0.11 kcal/mol), and an acceptable correlation coefficient (R^2 = 0.50).

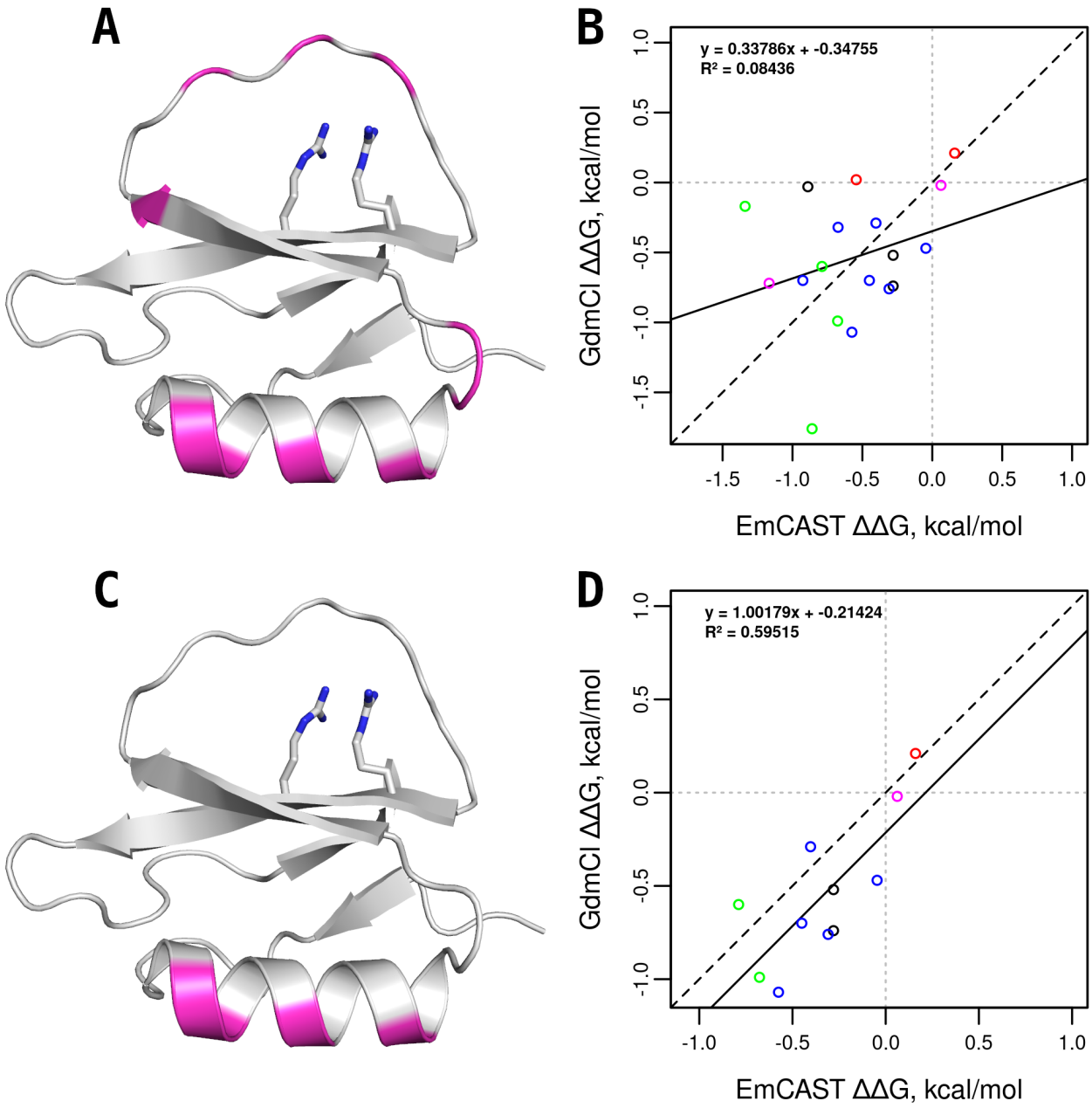

Figure 29: EmCAST predictions for Chymotrypsin Inhibitor 2.

Predictions were made using an X-ray structure (pdb: 2CI2). Experimental data from literature were taken from GdmCl folding equilibrium experiments monitored by fluorescence in 50 mM MES (pH 6.25) at 25C. Selected mutation sites are colored magenta in the protein structure. Correlations are shown for the full set of solvent exposed mutation sites (A-B) or a selected subset (C-D). Datapoint colors classify the type of mutation using the first matching category: gain/loss of a glycine or proline (green), gain of a negative charge (red), gain of a positive charge (blue), gain/loss of a polar sidechain (magenta), default (black).

Data exists for mutations in a mixed α/β protein, chymotrypsin inhibitor 2 (CI-2). Measurements were taken from GdmCl folding equilibrium experiments monitored by fluorescence in 50 mM MES (pH 6.25) at 25°C[45]. EmCAST calculations for the full set of available surface-exposed mutations did not produce a correlation (Figure 29A,B). High prediction error was localized to the C-terminal region of the protein. This region forms a unique sheet-loop-sheet structure wherein two buried arginine residues support a broad loop (Figure 29A,C). This atypical structure, not supported by a hydrophobic core, may be susceptible to rearrangements caused by internal or flanking mutations. The loop in this region contains the active site of CI-2 and is known to exhibit large thermal motions[46]. A reasonable correlation is found when the C-terminal region is excluded (Figure 29C,D) with a slope of 1 (m = 1.00), a near zero intercept (b = -0.21 kcal/mol), and a decent correlation coefficient (R^2 = 0.60).

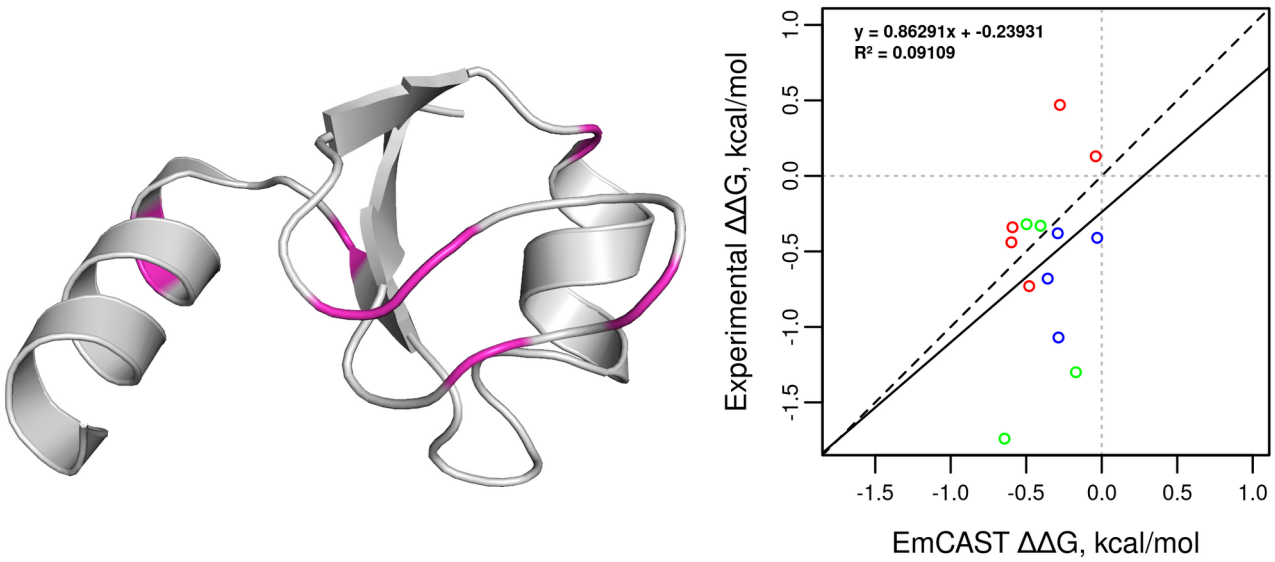

Figure 30: EmCAST predictions for the N-terminal domain of ribosomal protein L9.

Predictions were made using an X-ray structure (pdb: 2HBB). Experimental data from literature were taken from either urea or GdnHCl equilibrium folding experiments monitored by CD in 20 mM sodium acetate (pH 5.5) and 100 mM NaCl at 25C. Selected mutation sites are colored magenta in the protein structure. Datapoint colors classify the type of mutation using the first matching category: gain/loss of a glycine or proline (green), gain of a negative charge (red), gain of a positive charge (blue), gain/loss of a polar sidechain (magenta), default (black).

Surface mutations in the N-terminal domain of ribosomal protein L9 (NTL9) provides an opportunity to test EmCAST calculations in a protein known to have thermodynamically significant residual structure in the denatured state[47][48][49][50][51]. Non-native charge-charge interactions in the denatured state of NTL9 cause deviations in the expected pH dependence of stability; removal of the implicated charges by mutation resolves the deviation in the pH dependence of stability as modeled by the protein's pK_a[50]. Two key assumptions for EmCAST calculations are that both the denatured state and the native structure are unperturbed by the mutation. Experimental data from either urea or GdnHCl equilibrium folding experiments monitored by

circular dichroism in 20 mM sodium acetate (pH 5.5) and 100 mM NaCl at 25°C[52] were compared to EmCAST predictions. The full dataset did not produce a correlation (Figure 30). No common structural features across erroneous calculations were found. Each residue with mutations producing large errors had other mutations with accurate predictions. The only two mutations that did not alter charge (A22G and A42G) produced accurate predictions (Figure 30: x,y ≈ -0.5,-0.4); the other two green datapoints in Figure 30 are for K14G and K15G. Inaccuracy for the remaining mutations involving gain/loss of charge is consistent with electrostatic interactions in the denatured state of NTL9[47][48][50] being unaccounted for by EmCAST.

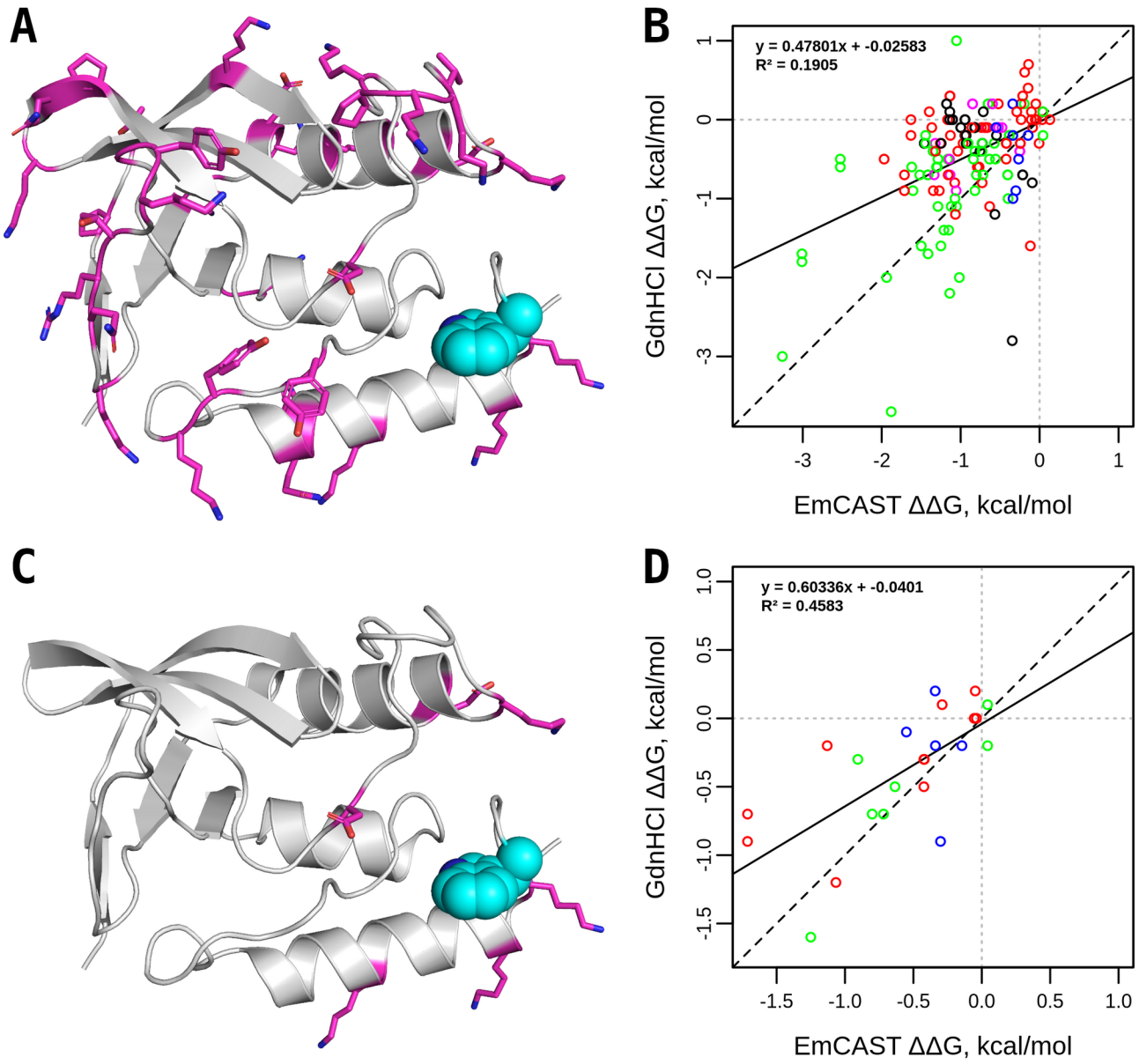

Figure 31: EmCAST predictions for staphylococcal nuclease

Predictions were made using an X-ray structure (pdb: 1STN). Experimental data from the ProThermDB were from GdnHCl equilibrium folding experiments monitored by fluorescence in 25 mM sodium phosphate (pH 7) and 100 mM NaCl at 20C. Selected mutation sites are colored magenta in the protein structure. The fluorescent probe used to monitor experiments is rendered cyan. Correlations are shown for the full set of solvent exposed mutation sites (A-B) or a selected subset near the fluorescent probe (C-D). Datapoint colors classify the type of mutation using the first matching category: gain/loss of a glycine or proline (green), gain of a negative charge (red), gain of a positive charge (blue), gain/loss of a polar sidechain (magenta), default (black).

A large set of mutations are available for Staphylococcal nuclease in the ProThermDB[35]. Data was selected from GdnHCl equilibrium folding experiments monitored by fluorescence in 25 mM sodium phosphate (pH 7) and 100 mM NaCl at 20°C. The full set of surface exposed mutations correlated with EmCAST calculations poorly (Figure 31A,B); the intercept was near 0 (b = -0.03 kcal/mol) but the slope was not near 1 (m = 0.48) and the correlation coefficient was low (R^2 = 0.19). High prediction error was found to be correlated with the distance between each mutation site and the fluorescent probe used to measure stability. The subset of mutations near the fluorescent tryptophan (Figure 31C) produced an improved correlation (Figure 31D) with a similar slope and intercept, but a higher correlation coefficient (R^2 = 0.46). This suggests the experimental probe is not sensitive to distant changes in structural stability. Many of the distant mutations predicted to be destabilizing by EmCAST were characterized as neutral experimentally (Figure 31B, datapoints near y = 0). This interpretation is further supported by the observed inconsistency of stability data for mutations measured by both tryptophan fluorescence and circular dichroism (Table 5)[53]. Furthermore, mutations in Staphylococcal nuclease are known to affect the denatured state[54] – which further obfuscates assessment of prediction error.

Variant	ΔΔG (Circular Dichroism), kcal/mol	ΔΔG (Fluorescence), kcal/mol	Discrepancy
K70W	-0.15 ± 0.14	-0.84 ± 0.14	0.7 ± 0.2
G88W	-0.69 ± 0.14	-1.22 ± 0.14	0.5 ± 0.2

Table 5: Comparison of Staphylococcal Nuclease Stabilities

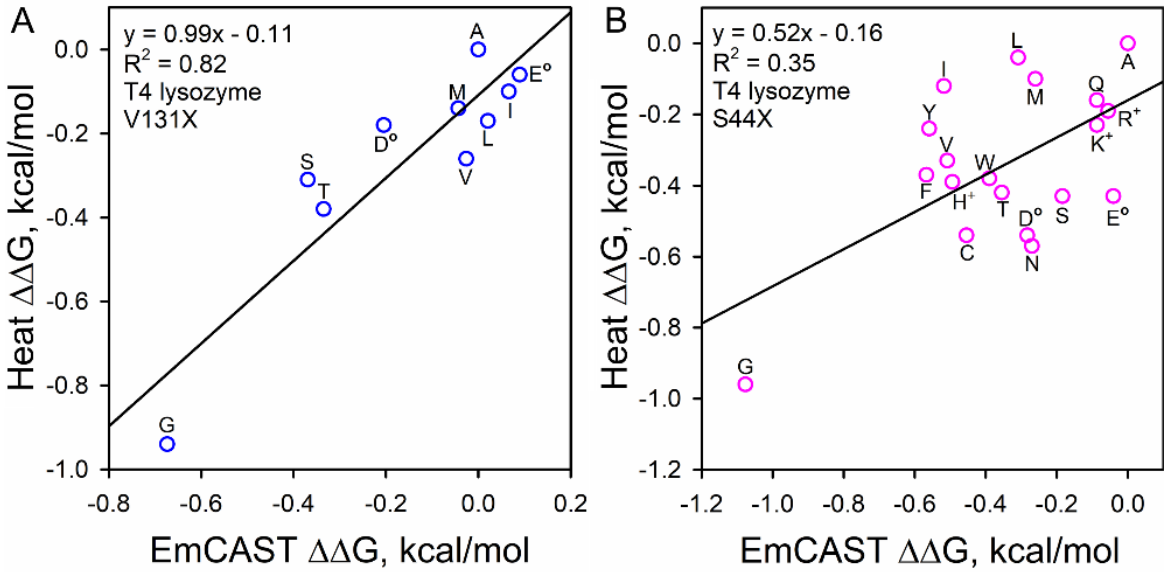

Figure 32: EmCAST predictions for T4 Lysozyme Helix Mutations

Predictions were made using an X-ray structure (pdb: 1L63). Experimental data was taken from thermally induced unfolding experiments monitored by CD in 25 mM KCl, 3 mM H_3PO_4, 17 mM KH_2PO_4 (pH 3.01). Data are shown for the two helix sites, V131 (A) and S44 (B).

Helical propensity studies in T4 Lysozyme allow a wider range of mutations to be scored by EmCAST at a single site. This enables structural and amino-acid contexts to be held constant while the correct ranking of mutant residues is assessed. Stability data is taken from thermally induced unfolding experiments monitored by circular dichroism in 25 mM KCl, 3 mM H_3PO_4, 17 mM KH_2PO_4 (pH 3.01)[55]. EmCAST predictions correlated very well for mutations at residue 131 (Figure 32A) but extremely poorly at residue 44 (Figure 32B); both mutation sites are well exposed. Data at position 131 produces one of

the best correlations for destabilizing mutations with a slope near 1 (m = 0.99), an intercept near 0 (b = -0.11 kcal/mol), and a high correlation coefficient (R^2 = 0.82). Correlation statistics at site 44 are misleading; without the glycine datapoint there would be no correlation. The protein environment EmCAST models is empirical and assumed to reflect physiological conditions. Stability data for T4 Lysozyme was collected under acidic conditions that likely aren't being correctly assessed by EmCAST. Difference in prediction accuracies may be caused by the different amino-acid contexts the two mutation sites are in. Site S44 contains two acidic residues within EmCAST's i±3 interaction window (AAKS**E**L**D**), one of which is adjacent to the mutation site. Site V131, by contrast, has only 1 acidic residue at the edge of this interaction window (**E**AAVNLA). Empirical modeling in EmCAST likely has these acidic residues charged, not neutral; the most common protonation state in the wwPDB is what is modeled by EmCAST. This will cause a mismatch in modeling between calculation and experiment. Incorrect modeling of a residue adjacent to a mutation likely impairs predictions for the S44 site.

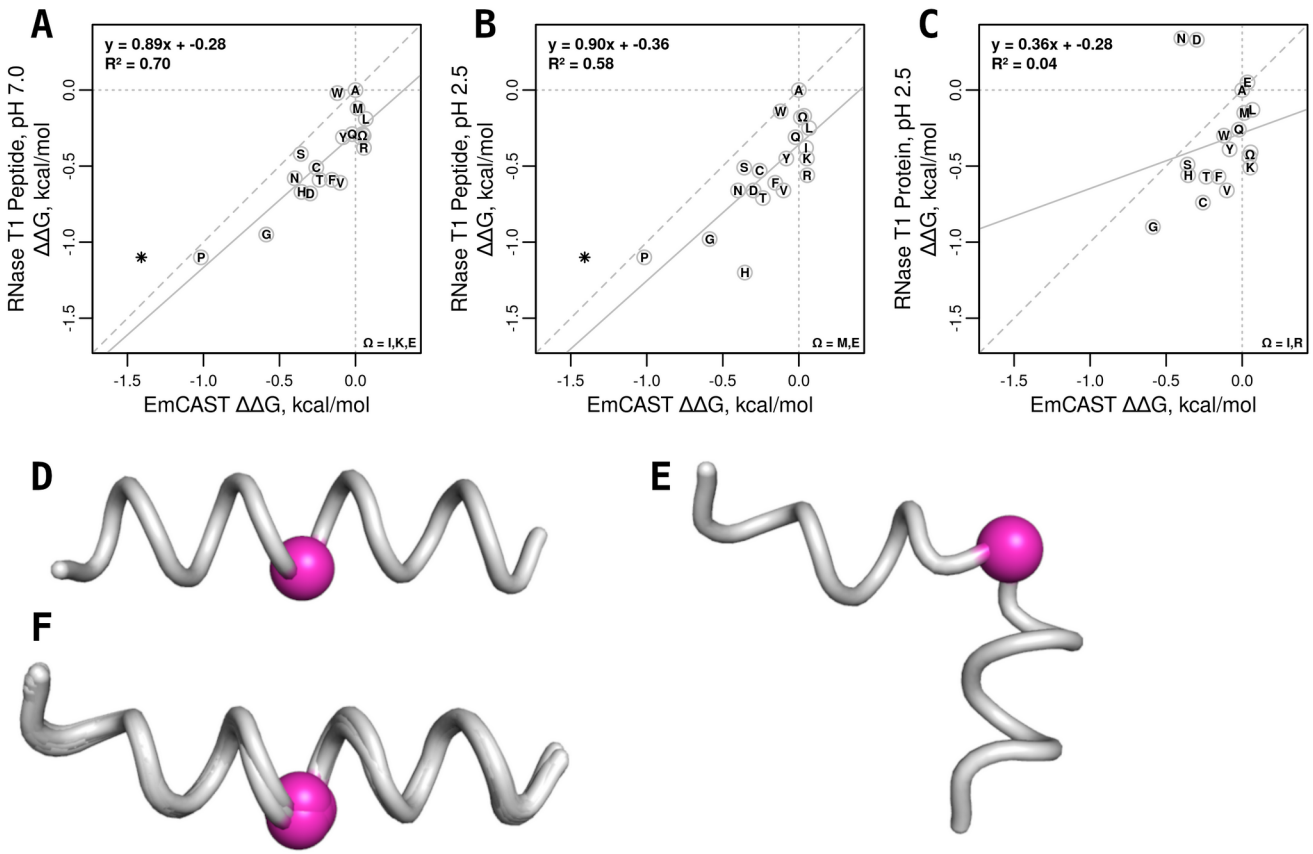

Figure 33: EmCAST predictions for RNase T1 Helical Mutations

*Predictions were made using the RNase T1 crystal structure (D, pdb: 9RNT) or using structures modeled by EmCAST (E, F). Experimental values are from urea folding equilibrium experiments monitored by circular dichroism at 0 °C in either 30 mM glycine pH 2.5 (B,C) or 30 mM MOPS pH 7 (A). Datapoint letters indicate the mutation type. Overlapping residues are marked with a Ω symbol and are described in the bottom right corner of each correlation plot. The EmCAST prediction for A21P using the crystal structure (pdb: 9RNT) is marked with a * symbol and is not included in the correlation line. The A21P mutation labeled 'P' uses the EmCAST relaxed structure for the helix (E). The three correlation plots show results for RNase T1 peptide at pH 7 (A), peptide at pH 2.5 (B), and protein at pH 2.5 (C).*

Helix propensity studies in RNase T1 provide stability data for A21 mutations in protein and peptide systems for different pH conditions. The helical RNase T1 peptide, which has no long range interactions, enables us to assess mutation induced structural changes predicted by EmCAST. Experimental data were obtained from urea folding equilibrium experiments monitored by circular dichroism

at 0°C in either: 30 mM glycine pH 2.5 (Figure 33B,C) or 30 mM MOPS pH 7 (Figure 33A)[56][57]. Stability data for the RNase T1 peptide at pH 7 correlated well with EmCAST calculations (Figure 33A) with a slope near 1 (m = 0.89), an intercept near 0 (b = -0.28 kcal/mol), and a strong correlation coefficient (R^2 = 0.70). Structural relaxation of the non-proline A21 mutations did not significantly change the structure of the helix (Figure 33D,F) or influence stability calculations. A structural rearrangement was predicted by EmCAST for the A21P variant (Figure 33E). Scoring the A21P mutation using the relaxed structure in the mutant sequence/structure pair lowered prediction error for A21P from 0.31 kcal/mol to 0.01 kcal/mol (Figure 33A). The A21P variant for the RNase T1 protein did not express *in vitro*, supporting the proposed structural distortion.

Stability predictions by EmCAST are the same across the three correlation plots (Figure 33A-C), but the experimental values vary. Correlation for the RNase T1 peptide under acidic conditions (Figure 33B) are slightly worse than the peptide data at neutral pH (Figure 33A); the most notable shift occurs for the A21H mutation. Data for the RNase T1 protein at pH 2.5 produced a poor correlation due to large shifts in stability for the A21N and A21D mutations (Figure 33C). These two mutations may create long-range polar interactions with residue N84 that don't exist in the peptide system and aren't modeled by EmCAST.

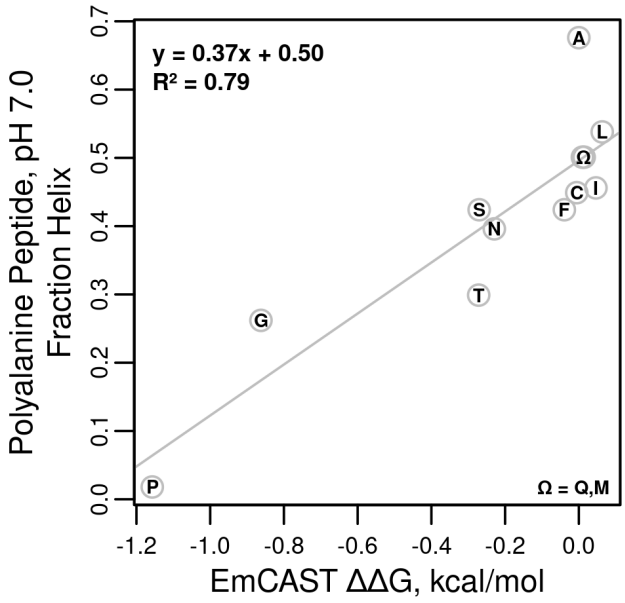

Figure 34: EmCAST predictions for polyalanine peptide

Peptide structures predicted by EmCAST were used in the calculations. Helix propensity data were measured by circular dichroism in 1 M NaCl pH 7.0 at 0 °C. Data for Q and M residues overlap and is represented by Ω.

Helical studies in a polyalanine peptide, $YGG(KAAAA)_3K\text{-}CONH_2$, let us test residue helix propensity modeled by EmCAST with minimal contributions from residues at the i±4 positions – which are beyond EmCAST's modeled interaction window. The helicities of different polyalanine peptide variants were measured by circular dichroism in 1 M NaCl pH 7.0 at 0 °C. Structures modeled by EmCAST for the polyalanine peptide and variants containing guest residues were used for the energy calculations, similar to the RNase T1 peptide. Measurements for fraction helix correlated well with energy predictions by EmCAST (Figure 34), the R^2 correlation coefficient is

0.79. The most notable prediction error occurs for alanine, the strongest helix former. EmCAST underestimates the stability of alanine in this system by approximately 0.4 kcal/mol, ranking it below isoleucine. This error is not unique to the polyalanine helix; alanine was incorrectly scored lower than isoleucine by EmCAST in the previous helical systems as well (Figures 32A and 33A). This trend suggests a systematic error within EmCAST is under-representing alanine's helix propensity across all tetrad fragments.

3D: Testing Stabilizing Mutations

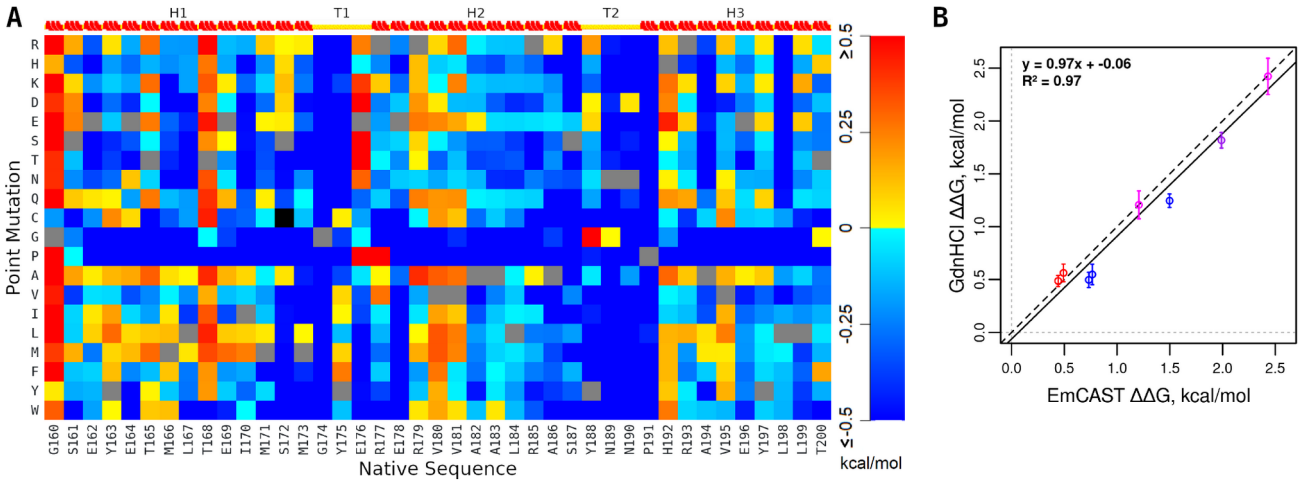

Figure 35: EmCAST predictions for UBA(1).

Predictions were made using the UBA(1) crystal structure (pdb: 6W2H). (A) Saturation mutagenesis heatmap for UBA(1). A scale bar that matches color to the degree of stabilization (positive values) or destabilization (negative values) is included on the right. Grey squares represent WT residues. Mutations with inadequate sampling are black. (B) Correlation plot between EmCAST predictions and stability data obtained from GdnHCl unfolding experiments. The line of best fit is shown as a solid black line. The dashed black line marks the position of a perfect fit. The red data points are single site mutations in helical regions and the blue data points are single or double mutations in turn regions. Magenta data points have equal numbers of mutations in helical and turn regions. The purple data point has two turn mutations and one helical mutation.

The majority of mutations found in literature were destabilizing (Figures 25B, 26, 27A, 28B, 29B, 30, 31D, 32A-B, 33A). This is to be expected; natural selection has already optimized protein sequences for stability, altering the sequence is generally destabilizing. EmCAST's ability to predict new, stabilizing mutations was tested using the small 3-helix bundle, UBA(1). The domain is one of two UBA domains found in the human homolog of *Saccharomyces cerevisiae* Rad23,

HHR23A, DNA excision repair protein[11]. EmCAST stability calculations were rapidly (sub-second) performed for all possible 779 mutations in UBA(1) to search for stabilizing mutations at surface exposed positions. A saturation mutagenesis heatmap was formed to navigate the calculated mutations (Figure 35A).

Four predicted stabilizing UBA(1) mutations were selected for experimental verification: two turn mutations (E176T and Y188G) and two helical mutations (T168R and H192E). The selected mutations sites are free of interactions outside of EmCAST's i±3 evaluation window, well represented within our fragment database, and are each predicted to stabilize UBA(1) by at least 0.5 kcal/mol. Experimental stability measurements of eight UBA(1) variants using the selected mutations correlated exceptionally well with EmCAST calculations (Figure 35B). The correlation slope is near 1 (m = 0.97), the intercept is near 0 (b = -0.06 kcal/mol), and the R^2 correlation coefficient is 0.97. The experimental methods and structural effects of the mutations are discussed in detail in Chapter 4: UBA(1) Folding Studies.

3E: Comparison to Existing Methods

3E.1: Consensus Sequence Approach

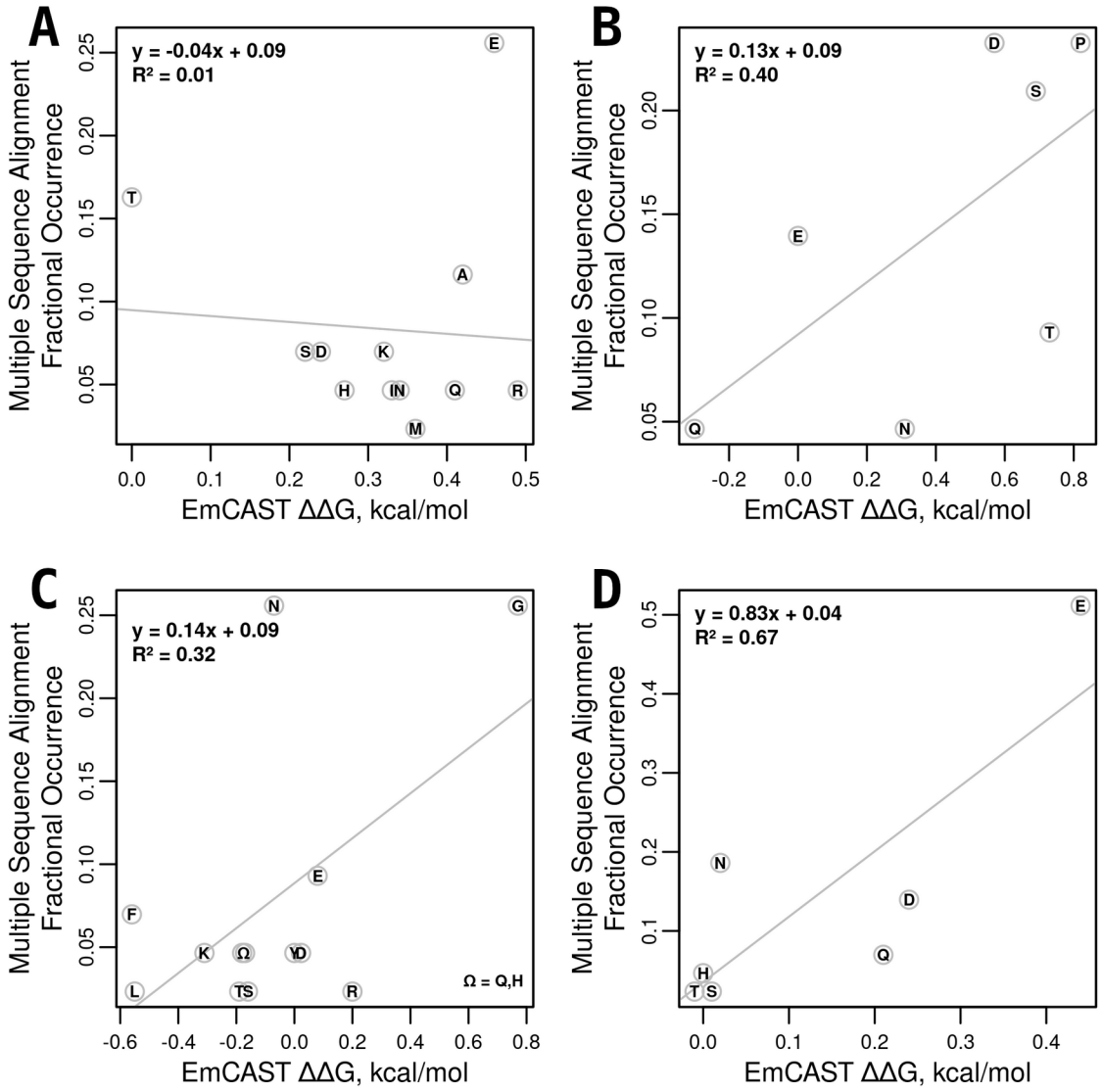

Figure 36: Comparison of EmCAST to UBA(1) Multiple Sequence Alignments

Plots of MSA fractional occurrences (43 samples) versus EmCAST calculations (T = 298.15K) for the four selected mutation sites in UBA(1). Correlation lines are shown as grey lines. Mutation sites included are (A) T168, (B) E176, (C) Y188, and (D) H192. In panel C, overlapping variants (Q and H) are represented with Ω *.*

EmCAST UBA(1) calculations were compared and contrasted against another stabilization strategy, the consensus sequence approach[58]. A

multiple sequence alignment (MSA) for UBA(1) using 43 sequences provided by Mueller and Feigon[11] was used to examine the correlation between stabilization in kcal/mol predicted by EmCAST and the fractional occurrence of an amino acid at the corresponding position in the MSA. The R^2 values for the four correlation lines ranged from 0.01 to 0.67 (Figure 36). For positions Y188 and H192, the mutations we chose based on EmCAST were the same mutations predicted by the MSA. There is notable disagreement at position 188; the MSA models Y188G and Y188N to be equally viable while EmCAST predicts Y188N to be slightly destabilizing. Apart from Y188N, the most frequent MSA variants (T168E, E176D, E176P, Y188G, and H192E) were all predicted to be stabilizing by EmCAST (Figure 35A). For positions T168 and E176, the mutations selected by EmCAST would not have been predicted as favorable from the MSA in Mueller and Feigon[11]. The stability of the consensus sequence by MSA was not characterized.

3E.2: Other Stability Prediction Tools

Several other mutation prediction tools were tested to assess whether EmCAST offers any benefits compared to existing prediction methods. The selected tools include FoldX, Rosetta-ddG, mCSM, SDM, DUET, INPS-3D, and PopMuSiC. Stability datasets were selected for their compatibility with EmCAST: surface mutations measured near physiological conditions (pH 5.5-8.5, salt present). Proteins with thermodynamically significant residual structure in the denatured

state were excluded. Stability measurements for the B-Domain of Staphylococcal Protein A, the FF Domain, barnase, and UBA(1) were used to compare the selected methods. The UBA(1) dataset provides an important metric for comparison. Each of the tools selected for comparison were trained and fitted to experimental stability data; only the UBA(1) dataset is guaranteed to be excluded from the training data.

The necessary software was obtained for Rosetta-ddG and FoldX; the other tools are only available as web services. Rosetta-ddG was run using the provided scripts for "Protein stability protocol 1: ddg_monomer, row 16" (https://github.com/Kortemme-Lab/ddg) with Rosetta (source version 2016.02.58402, compiled by gcc 4.8.5 20150623). A more recent version of Rosetta was found incompatible with the provided Rosetta-ddG scripts. For FoldX, FoldX 5.0 (win32) was the software release used. The RepairPDB command was used to prepare WT structures for analysis. The PositionScan command was used on the repaired PDB structure to predict single point mutations within the WT structure. Calculations for PopMuSiC (https://soft.dezyme.com/), INPS-3D (https://inpsmd.biocomp.unibo.it/inpsSuite/default/index3D), and DUET (http://biosig.unimelb.edu.au/duet/stability) were collected on November 3rd 2021 from their respective online websites. Predictions for SDM and mCSM were taken from the output provided by the DUET web

service. Stability values for Rosetta-ddG, FoldX, and PopMuSiC had their signs flipped to make positive values stabilizing. Predictions for UBA(1) variants with multiple mutations were taken as the sum of the individual mutations in every method, except for Rosetta-ddG.

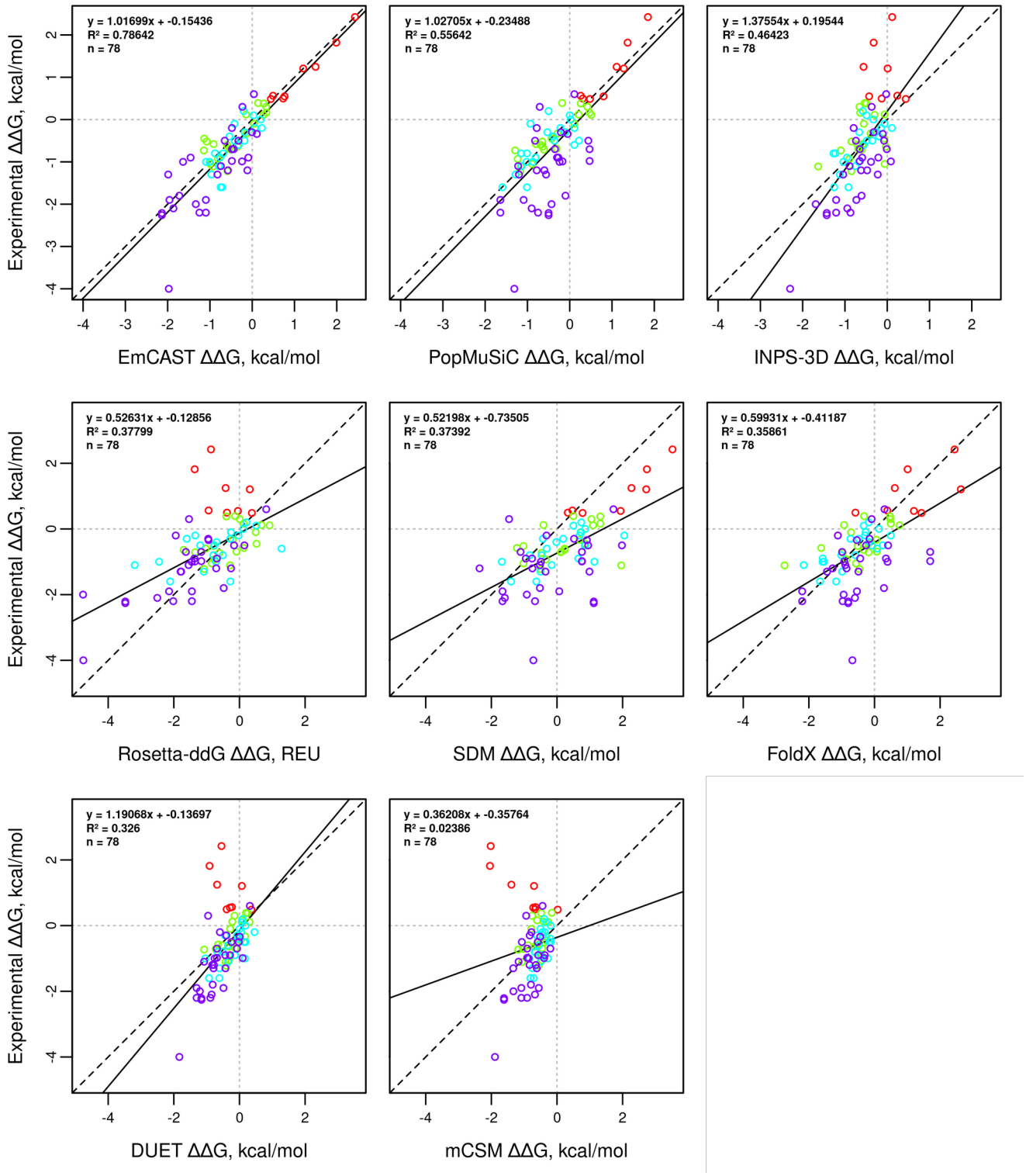

Figure 37: Comparison of different prediction methods for mutations

Ranking by prediction correlation was as follows: EmCAST (R^2 = 0.79), PopMuSiC[59][60] (R^2 = 0.56), INPS-3D[61] (R^2 = 0.46), Rosetta-

ddG[62] (R^2 = 0.38), SDM[63] (R^2 = 0.37), FoldX[64] (R^2 = 0.36), DUET[65] (R^2 = 0.33), and mCSM[66] (R^2 = 0.02) (Figure 37). Many of the methods tested struggled to predict our UBA(1) mutations as stabilizing. Only EmCAST, PopMuSiC, SDM, and FoldX predicted stabilizing ΔΔG values for the majority of the UBA(1) mutations.

3F: Conclusions

The energy calculations implemented in EmCAST were largely successful when tested against experimental data for surface mutations from 10 proteins and 2 peptides. Accuracy was inconsistent for proteins with thermodynamically significant residual structure in the denatured state (NTL9 and Staphylococcal nuclease) and for stability measurements taken under acidic conditions (T4 Lysozyme and RNase T1). Both residual structure in the denatured state and non-physiological conditions were expected to conflict with our energy calculations. Accuracy was also limited in the poorly supported loop structures of two proteins (src SH3 domain and Chymotrypsin inhibitor 2). Stability changes in these regions were consistently overestimated by EmCAST compared to experimental measurements. This trend suggests changes in local stability are poorly correlated with changes in global stability for protein regions unsupported by the protein's hydrophobic core. Excluding data from these flexible regions, calculations were well correlated for the remaining 6

proteins (FF Domain, B-Domain from Staphylococcal Protein A, barnase, src SH3 domain, Chymotrypsin inhibitor 2, and UBA(1)) and 2 peptides (RNase T1 and polyalanine). Correlation plots consistently produced slopes near 1, intercepts near 0, and R^2 correlation coefficients exceeding 0.57.

Calculation accuracy for destabilizing mutations in RNase T1 and polyalanine peptides was improved by relaxing the peptide structures according EmCAST's energy equation. Other structures without tertiary interactions, such as early folding conformers or isolated peptides, may be modeled using the same approach. Assessment of multiple helix propensity studies revealed a consistent inaccuracy of alanine's helix propensity in EmCAST. Alanine was incorrectly scored as less favorable than isoleucine for helical structure in multiple amino acid sequences, hinting at an underlying systematic error in EmCAST's design. The maximum residue solvent accessible surface area (SASA), taken from a Gly-X-Gly peptide, may create biases when weighing fragment samples by SASA percentage; different tetrad conformers likely have different maximum SASA values. A β-sheet conformer, for example, likely has a maximum SASA% close to 50% whereas the maximum SASA% for an α-helix conformer may be closer to 100%. This would artificially enhance the helix propensity of tetrads that favor β-sheet geometry. Introduction of a conformer dependent maximum SASA% value may be an adjustment worth investigating. Removal of SASA%

weighting improved the modeling of helix propensity when comparing alanine and isoleucine, but lead to worse stability correlations overall.

Stability calculations for mutations taken from literature had R^2 correlation coefficients typically in the 0.60-0.70 range, but the stabilizing mutations introduced to UBA(1) produced a practically perfect correlation (R^2 = 0.97). Multiple factors may be responsible for the exceptional accuracy observed for the UBA(1) mutations. Only stabilizing mutations were tested in UBA(1). This is advantageous over other datasets, which were almost exclusively destabilizing, due to better statistical coverage in the associated EmCAST heatmaps and by the native structure not being disrupted by the mutation. We also had the advantage of selecting positions and mutations ideal for EmCAST: sites absent of long range interactions and mutations with sufficient fragment sampling. The structure of UBA(1) may offer its own advantages compared to other proteins. Predictions for mutations in the flexible regions of the src SH3 domain correlated poorly with experimental stability measurements. One interpretation of this result is that structure in the flexible regions has a reduced influence on global stability of the protein. In UBA(1), the hydrophobic core holding the protein together involves residues from all three helices and the two turns connecting them. This creates an inter-dependence of stability between the three helices and two

turns. Consequently, local stability in each of these structural elements is likely directly tied to the global stability of UBA(1). Mutation sites sampled in other proteins might not have the same 1:1 correlation between local stability and global stability.

Multiple advantages are apparent for EmCAST over other prediction methods. Sequence redesign using multiple sequence alignment (MSA) to stabilize UBA(1) by reaching a consensus sequence predicted some, but not all, of the stabilizing mutations found by EmCAST. Fundamentally, the MSA method requires stabilizing mutations to have already been found and selected for by nature in homologous proteins. Predictions by EmCAST go beyond this limitation, mapping stabilizing mutations nature has yet to explore. Furthermore, stability changes are accurately quantified by EmCAST whereas MSA offers no estimated stability change. EmCAST outperformed every prediction software tested for a set of surface exposed mutations in both speed and accuracy. A key advantage lies in the design of EmCAST; it produces 1:1 correlations with physical measurements without fitting any constants or parameters to a training dataset of stability measurements. Most of the methods tested appear to be overly reliant on training data and failed miserably with the new, stabilizing UBA(1) mutations.

Chapter 4: UBA(1) Folding Studies

Research covered in this chapter has been published in the Journal of the American Chemical Society[27]. Thermodynamic and kinetic results for wild-type UBA(1) were performed by Dustin C. Becht. Data from X-ray crystallography of UBA(1) variants were collected and solved by Baisen Zeng (Y188G) and Levi J. McClelland (E176T/Y188G).

4A: Introduction

Characterization of protein structural stability and folding pathways in a small 3-helix bundle, UBA(1), was guided by stabilizing mutations predicted by EmCAST. UBA(1) is one of two UBA domains found in the human homolog of *Saccharomyces cerevisiae* Rad23, HHR23A, DNA excision repair protein[11]. The stability, folding kinetics, denatured state properties, and X-ray structure of wild type (WT) UBA(1) have been previously characterized[67][68]. The domain is of modest stability (2.4 kcal/mol), providing a good candidate for rational stabilization. Structural studies assess the magnitude and mechanism of stabilizing mutations as the UBA(1) sequence is optimized by EmCAST. Stabilizing point mutations provide high resolution probes into the folding landscape of UBA(1) and help characterize its folding pathway.

4B: Materials and Methods

4B.1: Preparation of Site-directed Mutations

The pGEX-2T(TEV) plasmid containing the UBA(1) gene was used as a template for site-directed mutagenesis[67]. Site-directed mutagenesis was carried out using the QuikChange Lightning PCR-based mutagenesis kit (Agilent). Primers for mutagenesis were obtained from Invitrogen (Table 6). DNA isolated from transformed XL-10 Gold *Escherichia coli* using the QIAprep Spin Miniprep Kit (QIAGEN) was sequenced to confirm mutations (Eurofins Genomics).

Application(s)	Forwards Primers	Reverse Primers
WT→Y188G	CCCTGAGAGCCAGCGGCAACAACCCCCACC	GGTGGGGGTTGTTGCCGCTGGCTCTCAGGG
WT→H192E	CAGCTACAACAACCCCGAACGAGCCGTGGAGTATC	GATACTCCACGGCTCGTTCGGGGTTGTTGTAGCTG
WT→E176T and Y188G→E176T/Y188G	GAGATCATGTCCATGGGCTATACGCGAGAGCGGG	CCCGCTCTCGCGTATAGCCCATGGACATGATCTC
WT→T168R and E176T/Y188G→T168R/E176T/Y188G	GAGTATGAGACGATGCTGAGGGAGATCATGTCCA	TGGACATGATCTCCCTCAGCATCGTCTCATACTC
Y188G→Y188G/H192E and T168R/E176T/Y188G→T168R/E176T/Y188G/H192E	CGGCAACAACCCCGAACGAGCCGTGGAGT	ACTCCACGGCTCGTTCGGGGTTGTTGCCG

Table 6: UBA(1) Mutagenic Primers

4B.2: Protein Expression and Purification

The pGEX-2T(TEV) plasmid[67] containing the WT or mutant UBA(1) gene fused to Glutathione-S-transferase (GST) was used to transform BL21(DE3) *E. coli* cells (New England Biolabs) followed by selection on ampicillin plates. A single colony was used to inoculate 5 mL of LB media containing 500 µg of ampicillin and grown for 16 hours with

shaking (150 rpm) at 37 °C. The 5 mL cultures were used to inoculate Fernbach flasks holding 1 L of sterile LB media containing 100 mg of ampicillin. The 1 L cultures were grown with shaking (150 rpm) at 37 °C until reaching an OD_{550} of 0.8. Protein expression was induced using IPTG at a final concentration of 1 mM. Incubation temperature was lowered to 30 °C and the cultures were allowed to grow for an additional 3 hours. Cultures were harvested and cell pellets were frozen at -80 °C.

WT and variant forms of UBA(1) were extracted from *E. coli* cell pellets with BugBuster Protein Extraction Reagent (EMD Millipore) using 5 mL of reagent per 1 g of cells. RNase and DNase were added to degrade RNA and DNA. 100 mM PMSF was added (50 µL per gram of cells) to the lysis solution to inhibit serine proteases. The clarified lysate was purified by GST affinity chromatography as previously described[67]. The fusion protein was cleaved using 30 µg of TEV protease per mg of protein. The GST-UBA(1) and TEV solution was gently shaken overnight at 4 °C. The cleaved sample was concentrated to 1-2 mL by centrifuge ultrafiltration using a 3,000 molecular weight cut off (MWCO) membrane (EMD Millipore). UBA(1) released from the GST fusion protein was separated from GST and TEV protease by size exclusion chromatography using a Superdex Peptide 10/300 GL high performance column (GE Healthcare) coupled to an AKTA FPLC (GE Healthcare), as previously described[67]. Separate but partially

overlapping peaks were observed for GST and UBA(1). Fractions for UBA(1) were repeatedly collected, concentrated, and re-injected until the GST peak ceased to overlap with the UBA(1) peak. The purity of the UBA(1) fractions was confirmed by SDS-PAGE and the identity of the UBA(1) variants confirmed by MALDI-ToF mass spectrometry.

4B.3: Guanidine Hydrochloride Denaturation

An Applied Photophysics Chirascan Circular Dichroism (CD) Spectrophotometer interfaced with a Hamilton Microlab 500 Titrator was used to carry out GdnHCl titrations at 25 ºC in the presence of CD buffer (20 mM MES, 40 mM NaCl, pH 6.5). Protein concentration was evaluated using absorbance at 280 nm and extinction coefficients determined by the Expasy ProtParam tool[25]. A "Native UBA(1)" sample was prepared by diluting UBA(1) into CD buffer to a final concentration of 5 µM. 7 M guanidine hydrochloride (GdnHCl) in CD buffer was used as chemical denaturant. A "Denatured UBA(1)" sample was prepared by diluting UBA(1) into 7 M GdnHCl CD Buffer to a final concentration of 5 µM. Refractive indices of the CD buffer and the "Denatured UBA(1)" sample were measured using a refractometer (Fisher Scientific). The Nozaki equation for the dependence of refractive index on GdnHCl concentration[69] was used to determine the final concentration of GdnHCl in the "Denatured UBA(1)" sample. A volume of 2 mL of the "Native UBA(1)" sample was loaded into a 1 cm fluorescence cuvette (Hellma, Art. No. 101-10-40) in an Applied

Photophysics Chirascan CD Spectrophotometer with temperature controlled at 25 ºC. The "Denatured UBA(1)" sample was titrated into the "Native UBA(1)" sample using the Hamilton Microlab 500 Titrator. Ellipticity was measured at 222 nm using 250 nm as background (θ_{222}). Eq. 4.1 was fit to plots of θ_{222} vs. [GdnHCl][70][71] to obtain the parameters, *m*, the rate of change of ΔG_u with respect to GdnHCl concentration and $\Delta G_u^{o'}(H_2O)$, the free energy of unfolding extrapolated to 0 M GdnHCl. In Eq. 4.1, θ_N and m_N are the intercept and slope of the native state baseline, θ_D and m_D

$$\theta_{222} = \frac{\left(\theta_N + m_N \cdot [GdnHCl]\right) + \left(\theta_D + m_D \cdot [GdnHCl]\right) \cdot e^{\left(\frac{m \cdot [GdnHCl] - \Delta G_u^{*'}(H_2O)}{RT}\right)}}{1 + e^{\left(\frac{m \cdot [GdnHCl] - \Delta G_u^{*}(H_2O)}{RT}\right)}} \qquad \text{(Eq. 4.1)}$$

are the intercept and slope of the denatured state baseline. Reported parameters are the average and standard deviation of at least three technical repeats.

4B.4: Folding Kinetics

Purified UBA(1) (220 µM) in CD buffer with or without GdnHCl (7.0 M) was mixed 1:10 with CD buffer containing various concentrations of GdnHCl using an Applied Photophysics SX20 stopped-flow spectrophotometer. Folding and unfolding reactions were monitored at 4 ºC through changes in UBA(1) tyrosine fluorescence. Excitation was at 280 nm with total fluorescence measured at 90° using a PM tube after passage through a 295 nm cut-off filter. Five

kinetic traces were collected for each final GdnHCl concentration. To account for the deadtime (1.62 ± 0.06 ms), 1.6 ms was added to all time points before a single exponential function was fit to the fluorescence versus time data to obtain observed rates constants, k_{obs}. Eq. 4.2 was fit to Chevron plots of the natural log of k_{obs} versus the final GdnHCl concentration to

$$\ln(k_{obs}) = \ln\left(k_f(H_2O)\cdot e^{\left(\frac{-m_{TS-D}\cdot[GdnHCl]}{RT}\right)} + k_u(H_2O)\cdot e^{\left(\frac{m_{TS-N}\cdot[GdnHCl]}{RT}\right)}\right) \qquad \text{(Eq. 4.2)}$$

determine folding and unfolding rate constants in the absence of denaturant, $k_f(H_2O)$ and $k_u(H_2O)$, respectively and $m_{TS\text{-}D}$ and $m_{TS\text{-}N}$, the *m*-values for the denatured and native states with respect to the transition state, respectively. $\Delta G_u^{o'}(H_2O)$ (Eq. 4.3), *m* (Eq. 4.4) and the Tanford β-value (β_T, Eq. 4.5) were calculated for each variant.

$$\Delta G_u^{o'}(H_2O) = RT\cdot\ln\left(\frac{k_f(H_2O)}{k_u(H_2O)}\right) \qquad \text{(Eq. 4.3)}$$

$$m_{eq} = (m_{TS-D} + m_{TS-N}) \qquad \text{(Eq. 4.4)}$$

$$\beta_T = \frac{m_{TS-D}}{(m_{TS-D} + m_{TS-N})} \qquad \text{(Eq. 4.5)}$$

4B.5: X-ray Crystallography

UBA(1) variants were purified as described above and concentrated to 20 mg/mL in 50 mM HEPES, 150 mM NaCl, pH 8.0. Commercially available screening kits were used in conjunction with a GRYPHON liquid-handling crystallization robot (Art Robbins

Instruments). Crystals were obtained by vapor diffusion at 20 °C from a sitting drop containing a 1:1 mixture of protein and reservoir solution (Y188G, 0.1 M phosphate-citrate pH 4.2, 0.2 M ammonium sulfate, 40%(v/v) ethylene glycol for PDB file 6W2G and 2.0 M ammonium sulfate for PDB file 6W2I; E176T/Y188G, 4.0 M sodium formate). X-ray diffraction data were collected at the Stanford Synchotron Radiation Lightsource beamline 9-2 or 12-1 with a DECTRIS PILATUS 6M detector. The data were indexed, integrated, and scaled using XDS[72] and Aimless[73]. The 1.45 Å Y188G structure (6W2I) was solved by sulfur single-wavelength anomalous diffraction (SAD) phasing. The other two structures were solved by molecular replacement using PHENIX/PHASER[74] with 6W2I (1.10 Å Y188G structure; 6W2G) or 6W2G (E176T/Y188G structure, 7TGP) as the search model. Model building was accomplished in PHENIX[74] and the structures were refined through iterative cycles of manual adjustment in Coot[75] and refinement of atomic positions, real space, occupancy, and thermal parameters in PHENIX[74]. Statistics for the Y188G crystal structures are provided in Tables 7 and 8, and for the E176T/Y188G crystal structure in Table 9. Structures have been deposited in the PDB.

Data Collection[a]	
Beamline	SSRL-SMB-9-2
Wavelength (Å)	1
Resolution range (Å)	32.88 - 1.1 (1.139 - 1.1)
Space group	P 1 2_1 1
Unit cell dimensions	
a, b, c (Å)	29.49, 40.08, 32.893
α, β, γ (◦)	90, 90.389, 90
Total reflections	184329 (11805)
Unique reflections	29568 (2521)
Multiplicity	6.2 (4.7)
Completeness (%)	94.78 (81.10)
Mean I/σ(I)	56.13 (16.63)
$CC_{1/2}$	1 (0.997)
CC*	1 (0.999)
Wilson B-factor (Å^2)	6.64
R_{merge}[b]	0.01716 (0.07858)
R_{meas}[b]	0.01868 (0.0884)
R_{pim}[b]	0.007273 (0.0396)
Refinement[a]	
Reflections used in refinement	29568 (2515)
Reflections used for R_{free}	1763 (148)
R_{work}[c]	0.1380 (0.1592)
R_{free}[d]	0.1514 (0.1698)
Number of total atoms	844
protein molecule	751
ligands	16
solvent	77
Total protein residues	99
RMS (bonds, Å)[e]	0.008
RMS (angles, ◦)[e]	1.04
Ramachandran favored (%)[e]	100.0
Ramachandran outliers (%)[e]	0.0
Clashscore	1.32
Average B-factor (Å^2)	9.33
Macromolecules (Å^2)	8.43
Ligands (Å^2)	14.28
Solvent (Å^2)	20.03

Table 7: X-ray Crystallography Data for UBA(1) Y188G (pdb: 6W2G)

[a]Data for the highest resolution shell are given in parenthesis. [b]R_{merge}=$\sum_{hkl} \sum_i$ | $I_i(hkl)$-⟨$I(hkl)$⟩|/$\sum_{hkl} \sum_i I_i(hkl)$, R_{pim} = $\sum_{hkl} \sqrt{(1/n - 1)} \sum_i$ | I_i (hkl)-⟨I (hkl)⟩|/ $\sum_{hkl} \sum_i$ I_i (hkl) where I_i (hkl) is the i^{th} observation of the intensity of the reflection hkl. R_{meas} is the same as R_{pim} except the prefactor is$\sqrt{(n/n - 1)}$. [c]R_{work} =$\sum_{hkl}$ || F_{obs}|-|F_{calc}||/ $\sum_{hkl}$ |F_{obs}|, where F_{obs} and F_{calc} are the observed and calculated structure-factor amplitudes for each reflection hkl. [d]R_{free} was calculated with 6% of the diffraction data that were selected randomly and excluded from refinement. [e]Calculated using MolProbity[76].

Data Collection[a]	
Beamline	SSRL-SMB-9-2
Wavelength (Å)	0.9795
Resolution range (Å)	31.16 - 1.45 (1.502 - 1.45)
Space group	$P4_3$
Unit cell dimensions	
a, b, c (Å)	31.164, 31.164, 40.53
α, β, γ (◦)	90, 90, 90
Total reflections	92391 (8525)
Unique reflections	6875 (680)
Multiplicity	13.4 (12.5)
Completeness (%)	98.95 (97.84)
Mean I/σ(I)	22.37 (4.34)
$CC_{1/2}$	1 (0.926)
CC*	1 (0.98)
Wilson B-factor (Å^2)	11.23
R_{merge}[b]	0.07658 (0.5901)
R_{meas}[b]	0.07961 (0.615)
R_{pim}[b]	0.02153 (0.1703)
Refinement[a]	
Reflections used in refinement	6875 (680)
Reflections used for R_{free}	697 (67)
R_{work}[c]	0.150 (0.1554)
R_{free}[d]	0.183 (0.1855)
Number of total atoms	432
protein molecule	377
ligands	6
solvent	49
Total protein residues	49
RMS (bonds, Å)[e]	0.011
RMS (angles, ◦)[e]	1.10
Ramachandran favored (%)[e]	100.0
Ramachandran outliers (%)[e]	0.0
Clashscore	0.00
Average B-factor (Å^2)	15.14
Macromolecules (Å^2)	14.10
Ligands (Å^2)	26.40
Solvent (Å^2)	24.29

Table 8: X-ray Crystallography Data for UBA(1) Y188G (pdb: 6W2I)

[a]Data for the highest resolution shell are given in parenthesis. [b]R_{merge}=$\sum_{hkl} \sum_i$ | $I_i(hkl)$-⟨$I(hkl)$⟩|/$\sum_{hkl} \sum_i I_i(hkl)$, R_{pim} = $\sum_{hkl} \sqrt{(1/n - 1)} \sum_i$ | I_i (hkl)-⟨I (hkl)⟩|/ $\sum_{hkl} \sum_i$ I_i (hkl) where I_i (hkl) is the i^{th} observation of the intensity of the reflection hkl. R_{meas} is the same as R_{pim} except the prefactor is$\sqrt{(n/n - 1)}$. [c]R_{work} =$\sum_{hkl}$ || F_{obs}|-| F_{calc}||/ $\sum_{hkl}$ |F_{obs}|, where F_{obs} and F_{calc} are the observed and calculated structure-factor amplitudes for each reflection hkl. [d]R_{free} was calculated with 10% of the diffraction data that were selected randomly and excluded from refinement. [d]Calculated using MolProbity[76].

Data Collection[a]	
Beamline	SSRL-SMB-12-1
Wavelength (Å)	0.9795
Resolution range (Å)	31.12 - 1.4 (1.45 - 1.4)
Space group	$P4_3$
Unit cell dimensions	
a, b, c (Å)	31.12, 31.12, 40.21
α, β, γ (◦)	90, 90, 90
Total reflections	103849 (9344)
Unique reflections	7624 (754)
Multiplicity	13.6 (12.3)
Completeness (%)	99.82 (99.21)
Mean I/σ(I)	10.0 (1.45)
$CC_{1/2}$	0.997 (0.604)
CC*	0.999 (0.868)
Wilson B-factor (Å^2)	14.78
R_{merge}[b]	0.1471 (1.964)
R_{meas}[b]	0.1529 (2.05)
R_{pim}[b]	0.04116 (0.5799)
Refinement[a]	
Reflections used in refinement	7616 (754)
Reflections used for R_{free}	360 (32)
Rwork[c]	0.1605 (0.2845)
Rfree[d]	0.1747 (0.3622)
Number of total atoms	413
protein molecule	384
solvent	29
Total protein residues	51
RMS (bonds, Å)[e]	0.004
RMS (angles, ◦)[e]	0.50
Ramachandran favored (%)[e]	100.00
Ramachandran outliers (%)[e]	0.00
Clashscore	3.92
Average B-factor (Å^2)	21.61
Macromolecules (Å^2)	20.94
Solvent (Å^2)	30.59

Table 9: X-ray Crystallography Data for UBA(1) E176T/Y188G (pdb: 7TGP).

[a]Data for the highest resolution shell are given in parenthesis. [b]R_{merge}=$\sum_{hkl} \sum_i$ | I_i(hkl)-⟨I(hkl)⟩|/$\sum_{hkl} \sum_i I_i$(hkl), R_{pim} = $\sum_{hkl} \sqrt{(1/n - 1)} \sum_i$ | I_i (hkl)-⟨I (hkl)⟩|/ $\sum_{hkl} \sum_i$ I_i (hkl) where I_i (hkl) is the i^{th} observation of the intensity of the reflection hkl. R_{meas} is the same as R_{pim} except the prefactor is$\sqrt{(n/n - 1)}$. [c]R_{work} =$\sum_{hkl}$ || F_{obs}|-| F_{calc}||/ $\sum_{hkl}$ |F_{obs}|, where F_{obs} and F_{calc} are the observed and calculated structure-factor amplitudes for each reflection hkl. [d]R_{free} was calculated with 4.7% of the diffraction data that were selected randomly and excluded from refinement. [e]Calculated using MolProbity[76].

4C: Results

4C.1: Structure Stabilization

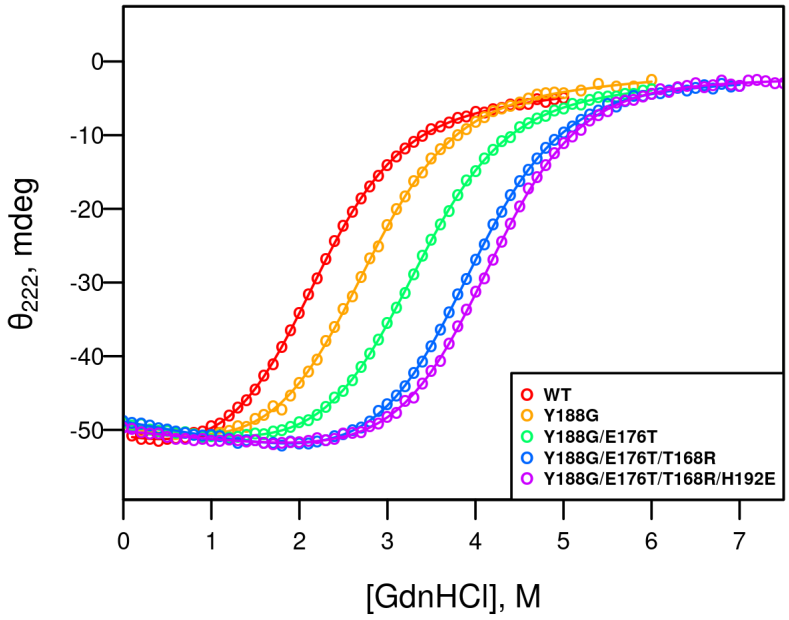

Figure 38: GdnHCl titrations for UBA(1) variants.

Representative unfolding curves for progressively stabilized UBA(1) variants. Unfolding was induced by GdnHCl titration and monitored by CD at 222 nm using a 250 nm baseline.

Guanidine hydrochloride (GdnHCl) unfolding experiments, monitored by circular dichroism (CD), were used to measure changes in protein stability (Figure 38). Eight UBA(1) variants were tested and matched predicted stability changes exceptionally well (Table 10 and Figure 35B) with a 0.16 kcal/mol standard error of the estimate. Stability enhancements notably were over-predicted for variants composed mainly of turn mutations (Table 10: E176T, Y188G, Y188G/E176T, Y188G/E176T/T168R). Combining these variants with nearby

stabilizing mutations in helices 1 or 3 abolished the energy discrepancy (Table 10: Y188G/H192E, Y188G/E176T/T168R/H192E). This observation may indicate that local dynamics at the mutation site can negate a portion of the predicted stability. Altogether, the four selected mutations double UBA(1)'s stability from 2.4 to 4.8 kcal/mol as predicted.

Variant	$\Delta G_u°'(H_2O)$, kcal/mol	*m*, kcal mol^{-1} M^{-1}	ΔΔG, kcal/mol	EmCAST ΔΔG, kcal/mol
WT	2.39 ± 0.05	1.16 ± 0.02	0.00	0.00
T168R	2.95 ± 0.07	1.13 ± 0.03	0.56 ± 0.08	0.49
E176T	2.89 ± 0.05	1.11 ± 0.01	0.50 ± 0.07	0.73
Y188G	2.94 ± 0.08	1.13 ± 0.02	0.55 ± 0.10	0.77
H192E	2.878 ± 0.003	1.145 ± 0.003	0.49 ± 0.05	0.44
Y188G/H192E	3.60 ± 0.12	1.13 ± 0.03	1.21 ± 0.13	1.21
Y188G/E176T	3.64 ± 0.04	1.11 ± 0.02	1.25 ± 0.06	1.50
Y188G/E176T/T168R	4.21 ± 0.05	1.10 ± 0.01	1.82 ± 0.08	1.99
Y188G/E176T/T168R/H192E	4.81 ± 0.16	1.18 ± 0.04	2.42 ± 0.17	2.43

Table 10: Parameters from GdnHCl Unfolding Experiments for UBA(1) Variants.

EmCAST predictions were made using the crystal structure of WT UBA(1) (pdb: 6W2H)[67]. Errors in $\Delta G_u°'(H_2O)$ and m are the standard deviations of the parameters obtained from separate fits of Eq. 4.1 to three GdnHCl titrations for each protein. The error in ΔΔG is obtained from standard propagation of the error in the $\Delta G_u°'(H_2O)$ values.

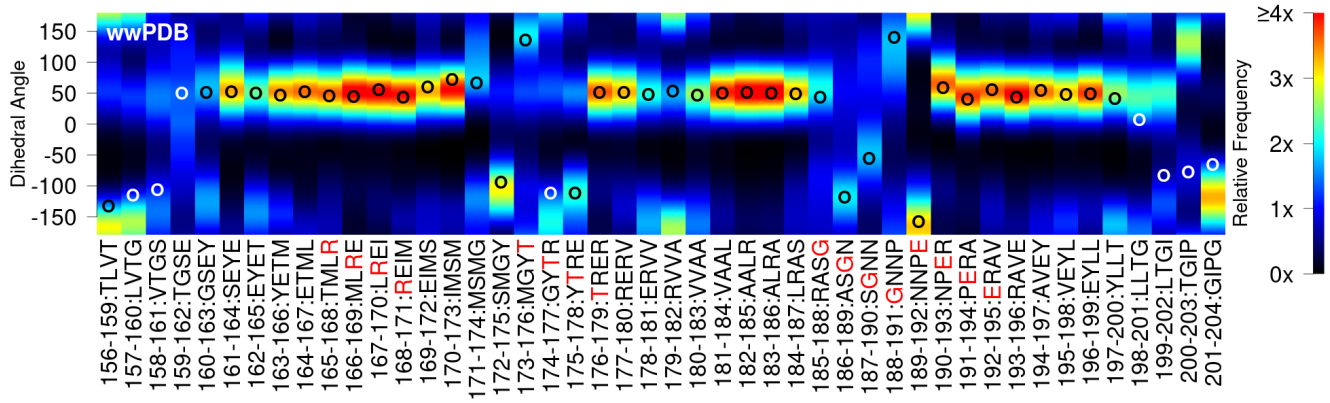

Figure 39: Fragment heatmap for UBA(1) quadruple mutant.

The primary sequence of UBA(1) is represented as its sequence of overlapping tetrads (x-axis). Mutated residues are highlighted in red. The tertiary structure of UBA(1) (pdb: 6W2H)[67] is represented by the 4-residue Cα dihedral angle (y-axis, open circles). The distribution of samples in our fragment database is rendered as heat (red = most populated, black = zero population).

Sequence optimization by EmCAST adjusts the protein sequence to shift fragment heatmaps to better match the tertiary structure of the protein (Figure 39) compared to the wild-type sequence (Figure 20). The physical mechanism(s) behind stabilization are not revealed by EmCAST due to the empirical nature of the free energy potential. Further characterization of the optimized UBA(1) structure is necessary to elucidate the atomic interactions leading to stabilization. A critical assumption of EmCAST is that the mutations used to stabilize a protein do not alter the structure of the protein. We were able to crystallize and solve the structures of the Y188G and Y188G/E176T variants of UBA(1) using X-ray crystallography (Tables 7, 8, and 9). The structures confirmed that the two turn mutations E176T and Y188G were able to enhance stability without disturbing the tertiary structure of UBA(1) or the backbone conformation of the two turns (Figure 40).

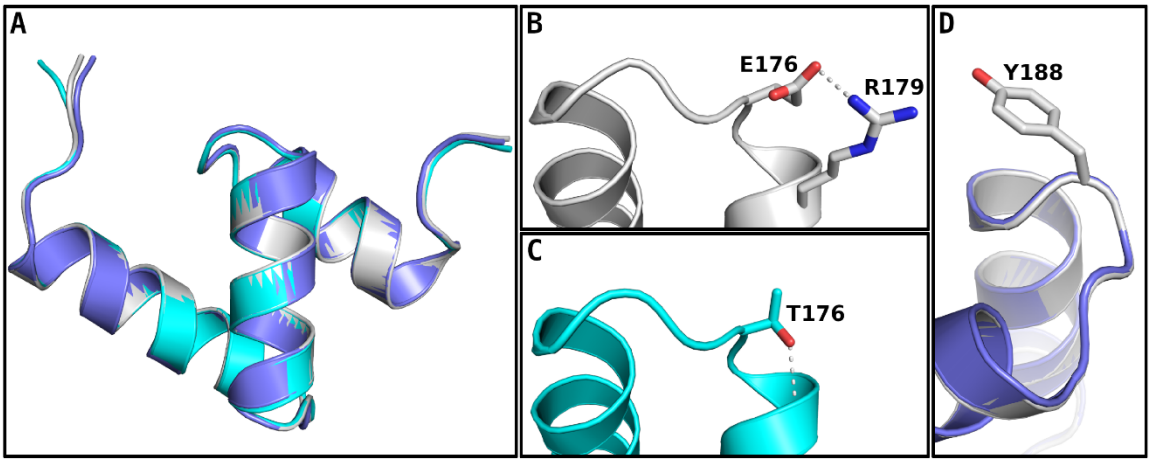

Figure 40: X-ray structures of WT UBA(1) and turn variants.

(A) Cartoon overlay of UBA(1) WT (grey, PDB file: 6W2H)[67], Y188G (cobalt, PDB file: 6W2G), and E176T/Y188G (cyan, PDB file: 7TGP) X-ray structures. (B) UBA(1) WT turn 1, a potential electrostatic interaction between E176 and R179 side chains is highlighted. (C) UBA(1) E176T/Y188G turn 1, hydrogen bonding between T176's gamma-hydroxyl and R179's backbone-amide NH is observed. (D) Cartoon overlay of UBA(1) turn 2 for WT (grey) and the Y188G (cobalt) variant with the Y188 side chain rendered.

Residue E176 provides two stabilizing features that are lost upon mutation to Thr: stabilization of helix 2's macroscopic electrostatic dipole[18] and a constructive electrostatic intrahelix (i, i+3) interaction[77] with R179 (Figure 40B). The E176T mutation more than compensates for these lost features by introducing a favorable Ncap[78] to helix 2 (H2), wherein T176's gamma-hydroxyl hydrogen bonds to R179's backbone-amide NH (Figure 40C). Other experimental[79] and database[78][80] analyses of proteins indicate that an E→T mutation at an α-helix Ncap should be stabilizing. Residue Y188 has φ,ψ angles that fall within the left-handed α-helix region of the Ramachandran plot (Figure 41A). Glycine is more commonly found in this backbone geometry (Figure 41B), suggesting that Y188G

stabilizes UBA(1) through backbone torsion angle optimization (Figure 40D).

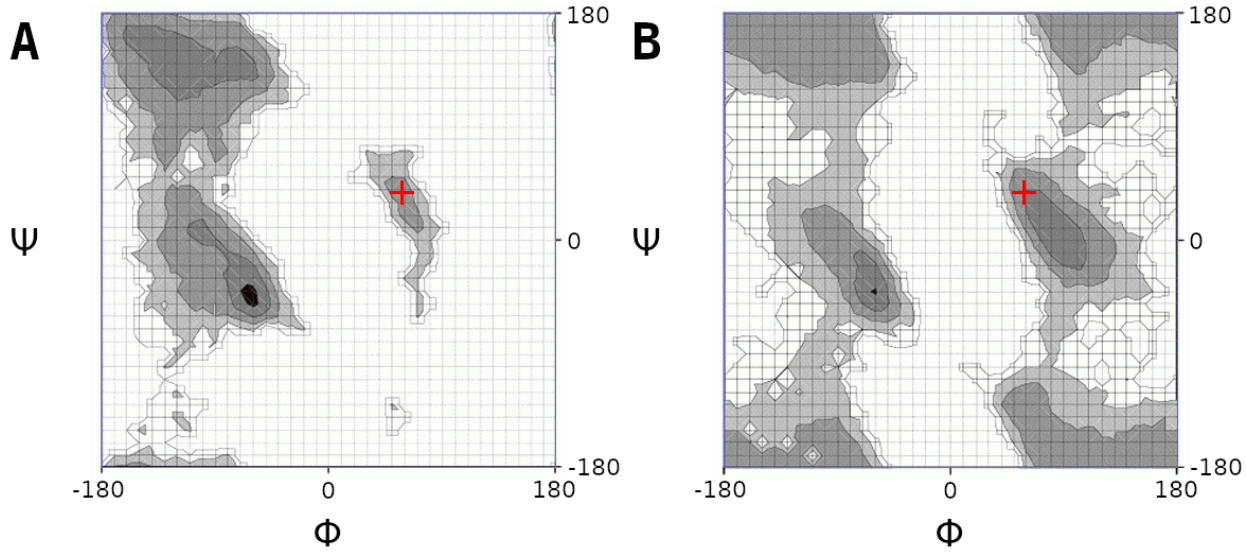

Figure 41: Ramachandran plots of UBA(1) Y188G.

The two plots show main-chain conformational tendencies for tyrosine (A) and glycine (B) amino acids[81]. The φ/ψ angles occupied by Y188 (pdb: 6W2H) are shown by the red mark.

The two helical mutations, T168R and H192E, are both favored over WT residues on empirical helix propensity scales[82]. H192E places a glutamate at the N2 position of helix 3, stabilizing the helix dipole[18]. Experimental[79] and database[78][80] analyses are also consistent with stabilization by a H→E mutation at the N2 position of an α-helix. Beyond intrinsic helical propensity, the features involved in our most stabilizing mutation, T168R, remain elusive. Introducing the opposite charge with T168E is predicted to add a similar level of stabilization (Figure 35A). The stabilizing mutagenic potential, predicted by EmCAST, for residues flanking this site drop after either mutation (Figure 42A-C). Conversely,

introducing nearby mutations T165E and E169A in EmCAST (+0.408 kcal/mol) removes about 0.3 kcal/mol of stabilization from the T168R and T168E mutations (Figure 42D). Taken together, these predictions suggest sequence-context-dependent effects play a significant role in the stabilization provided by T168R.

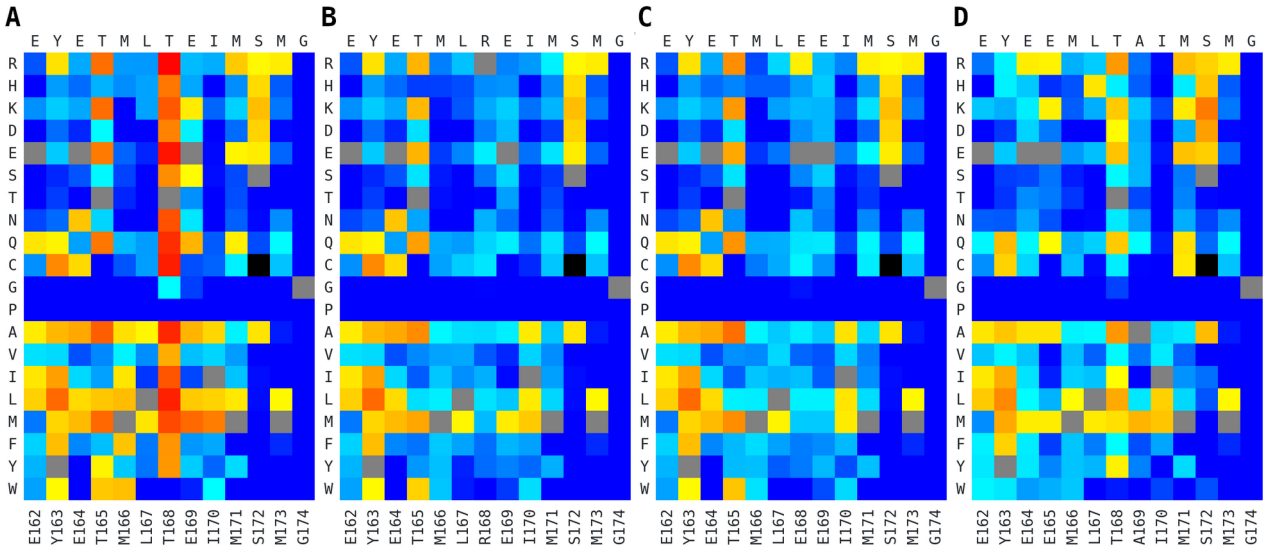

Figure 42: EmCAST predictions about position 168 for several UBA(1) variants. Variants include WT (A), T168R (B), T168E (C), and T165E/E169A (D). For color code see Figure 35A.

4C.2: Folding Kinetics and Mechanisms

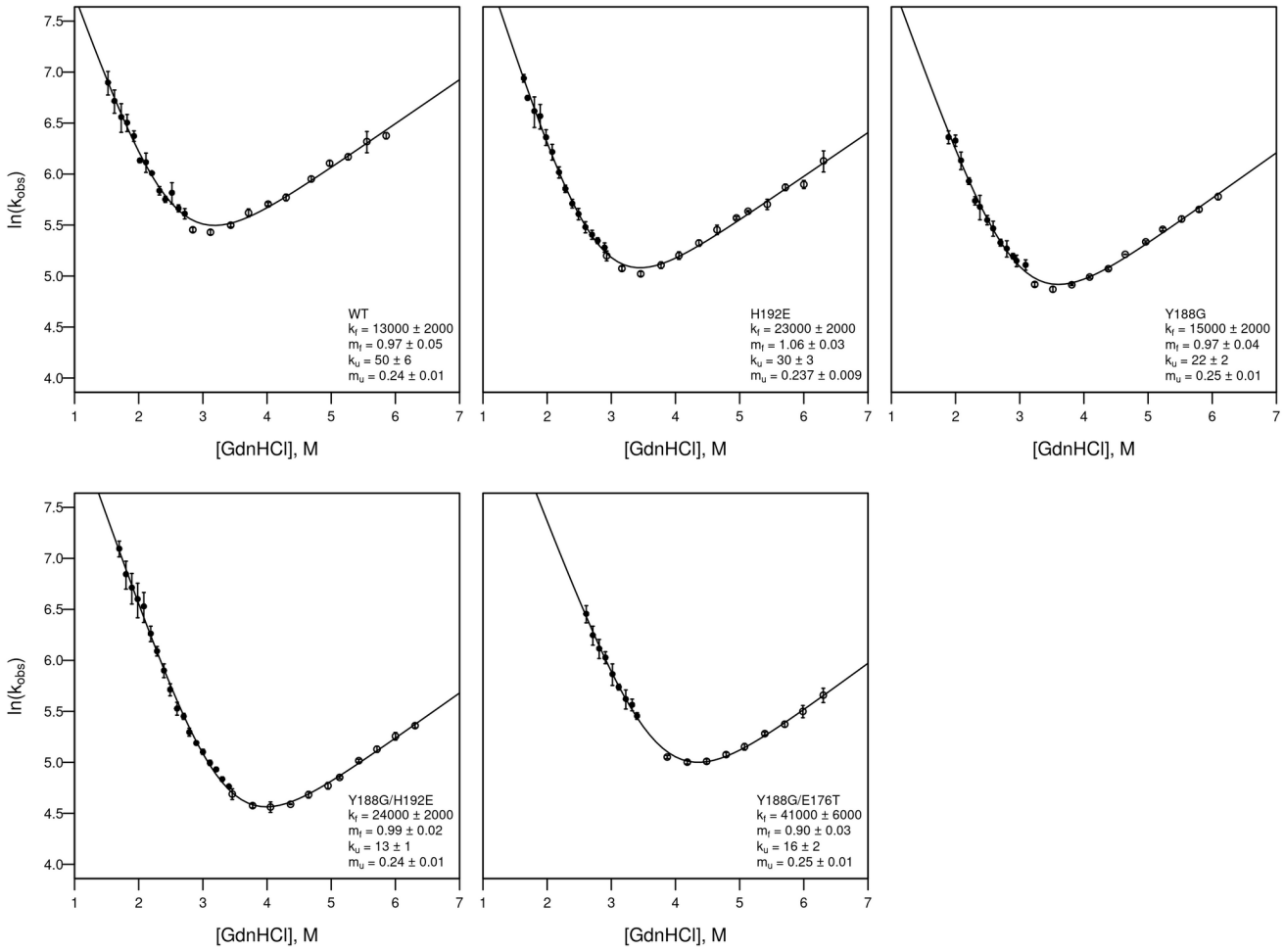

Figure 43: Chevron plots for UBA(1) variants.

Data is included for UBA(1) WT, H192E, Y188G, Y188G/H192E, and Y188G/E176T variants. Data from folding and unfolding stopped-flow experiments are shown as solid and open circles, respectively. The data for WT UBA(1) have been reported previousy[67] and are shown here for comparison with data for the variants.

Alterations to UBA(1)'s folding landscape were analyzed by stopped-flow experiments for several variants (Figure 43, Table 11). All variants exhibited decreases in unfolding rate consistent with the deliberate stabilization of the native state using EmCAST. The transition state was also stabilized in each variant as evidenced by

enhanced folding rates. Optimizing the native-state backbone torsion angle preference of turn 2 (Y188G) provided only minor increases in the folding rate, suggesting that turn 2 plays a passive role in UBA(1)'s folding process. Stabilizing helix 2 through N-capping (E176T) or helix 3 through helix dipole optimization (H192E) yielded dramatic increases in folding rates. These observations are consistent with a diffusion-collision model[67][83], wherein the helices form early in the folding process and subsequently dock onto each other. E176T, while nearly identical to H192E in terms of its effect on stability, provides a notably larger acceleration to the folding process. This difference may be attributed to the immediate availability of N-capping interactions by E176T, indicating that helix-capping interactions can promote efficient folding. Observations of helix capping residues promoting structure in the denatured state further support this interpretation[68]. In contrast, macroscopic dipole optimization by H192E will only be available after the formation of helix 3.

Variant	$k_f(H_2O)$, s^{-1}	$k_u(H_2O)$, s^{-1}	m_{TS-D}, kcal mol^{-1} M^{-1}	m_{TS-N}, kcal mol^{-1} M^{-1}	m_{eq}, kcal mol^{-1} M^{-1}	β_T
WT	13000 ± 2000	50 ± 6	0.97 ± 0.05	0.24 ± 0.01	1.21 ± 0.05	0.80 ± 0.01
Y188G	15000 ± 2000	22 ± 2	0.97 ± 0.04	0.25 ± 0.01	1.21 ± 0.04	0.80 ± 0.01
H192E	23000 ± 2000	30 ± 3	1.06 ± 0.03	0.237 ± 0.009	1.30 ± 0.03	0.818 ± 0.007
Y188G/H192E	24000 ± 2000	13 ± 1	0.99 ± 0.02	0.24 ± 0.01	1.23 ± 0.02	0.801 ± 0.008
Y188G/E176T	41000 ± 6000	16 ± 2	0.90 ± 0.03	0.25 ± 0.01	1.16 ± 0.03	0.78 ± 0.01

Table 11: Folding Kinetics Parameters of UBA(1) Variants

[a]The reported errors for $k_f(H_2O)$, $k_u(H_2O)$, m_{TS-D} and m_{TS-N} are the standard errors of the parameters obtained from fits of Eq. 4.2 to the Chevron plot data. The error in m_{eq} and β_T are from standard propagation of the errors in m_{TS-D} and m_{TS-N}.

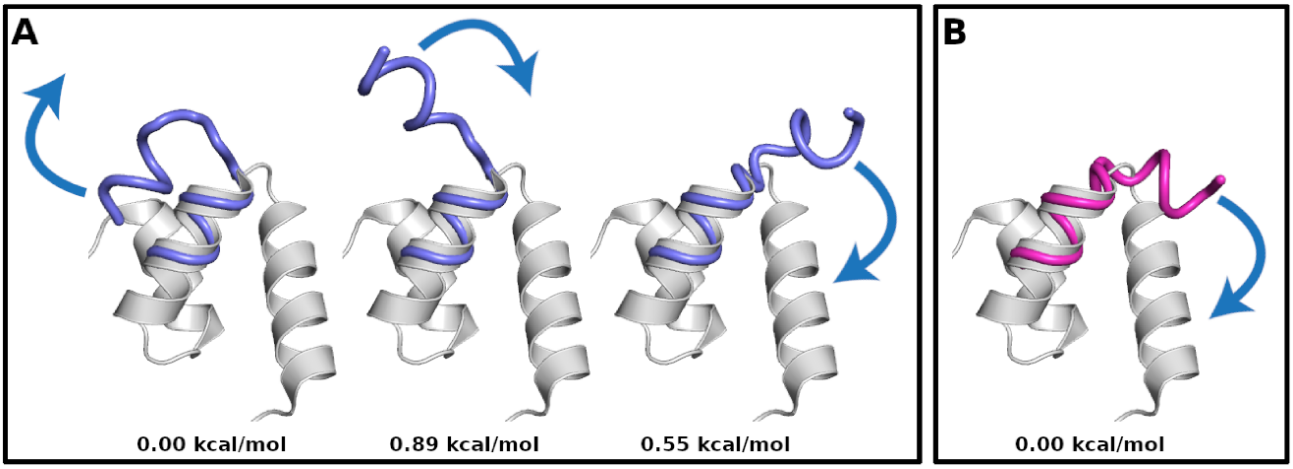

Figure 44: Modeled folding mechanisms of UBA(1) Turn 1

Proposed folding mechanisms for UBA(1) WT (A) and T168R/E176T (B) are shown. Available conformations for the H1-T1-H2 segment of UBA(1) are modeled and their relative energies scored by EmCAST. Select conformers (cobalt, magenta) are aligned to the crystal structure of UBA(1) (pdb: 6W2H, grey) using H2. Proposed movements to transition from local minima towards global minima are depicted. The energy state of each conformer, relative to the segment's local minimum, is included.

Modeling the energetic distribution of UBA(1) turn conformations with EmCAST provides additional insights into the folding kinetics of UBA(1) variants. The lowest energy conformation for WT UBA(1) T1 leads to a counter-productive helix-turn-helix fold. This transient helical bundle would need to be disrupted before T1 can restructure to accommodate the tertiary structure of UBA(1) (Figure 44A). In contrast, our optimized T1 variant only needs to slightly bend T1 to position H1 to form the native state structure of UBA(1) (Figure 44B). Disfavoring the formation of counter-productive folding intermediates may be the underlying mechanism through which the E176T mutation drastically enhances folding rates for UBA(1). For comparison, the lowest energy conformers of both WT and optimized T2

variants position the helices such that they can directly swing into place (Figure 45).

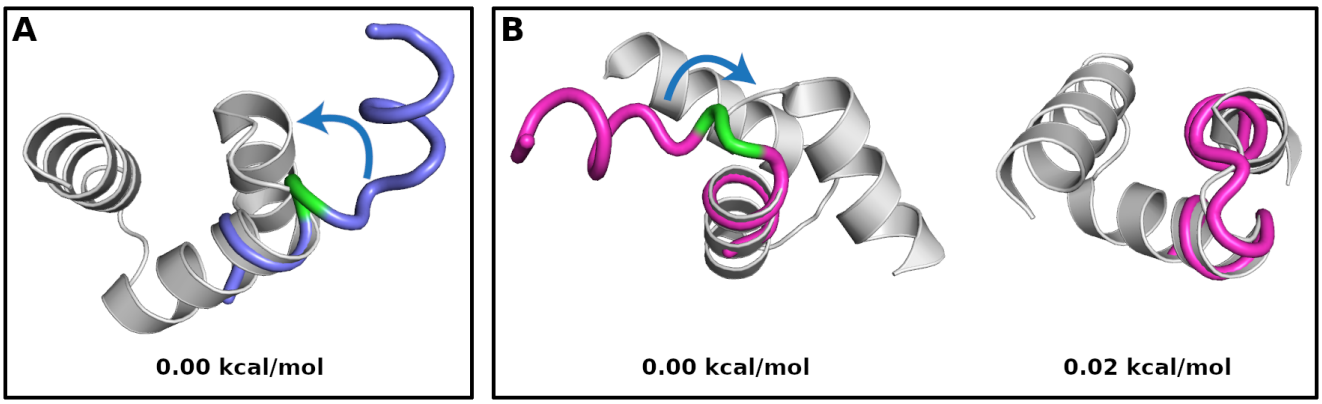

Figure 45: Modeled folding mechanism for UBA(1) Turn 2

Proposed folding mechanisms for UBA(1) WT (A) and Y188G/H192E (B) are shown. Available conformations for T2 and the flanking H2 and helix 3 (H3) segments of UBA(1) are modeled and their relative energies scored by EmCAST. Select conformers (cobalt, magenta) are aligned to the crystal structure of UBA(1) (pdb: 6W2H, grey) using H2. Proposed movements to transition from local minima towards global minima are depicted. The energy state of each conformer, relative to the segment's local minimum, is included. The central axis of the dihedral angle that needs rotation is highlighted in green (A: ASYN, B: SGNN). The C-terminal end of H2 is elongated in our model for T2 in WT UBA(1) (see panel A). The extra hydrogen bond in this model would need to be broken during the folding process, likely slowing the folding process. Two conformers are effectively tied for the lowest energy conformer of T2 in the Y188G/H192E variant (see panel B). Either rotation near G188 (B, left) or slight bending of T2 (B, right) are needed for these conformers to match the tertiary structure of UBA(1).

4D: Conclusions

Changes in local stability for mutations predicted by EmCAST correlated perfectly with measured changes in global stability of UBA(1). Structural enhancements were made that reduced backbone torsional strain (Y188G), added helix capping (E176T), stabilized the helix dipole (H192E), and optimized context-dependent effects (T168R). Combined, these mutations doubled the net stability of

UBA(1) from 2.4 to 4.8 kcal/mol. X-ray crystallography confirmed mutations in the two turns did not cause any deviations in backbone geometry. Additional surface mutations are predicted by EmCAST to further stabilize UBA(1): G160P (+1.013 kcal/mol), S161E (+0.245 kcal/mol), T165A (+0.225 kcal/mol), and R179Q (+0.31 kcal/mol). The G160P mutation optimizes structure in the N-terminal tail of UBA(1), which will likely have a minimal impact on the global stability of UBA(1) compared to other mutations. The tested mutations and additional predictions demonstrate that a considerable amount of stability can be gained by optimizing a protein's surface.

Measured enhancements in folding rates induced by stabilizing mutations reflect the formation of native-like structure in the protein's folding transition state. The E176T mutation, which optimizes Turn-1 geometry and caps Helix-2, has the strongest impact on the transition state among the measured variants. Simulations from section 2D.4: Ubiquitin Associated Domain 1 of HHR23A model WT Turn-1 to partially form early in folding with structure in the C-terminal side of Turn-1 forming last. The E176T mutation optimizes structure in this slow-to-form segment of Turn-1 in EmCAST heatmaps (Figure 20:YERE vs. Figure 39:YTRE). Structural modeling by EmCAST visualizes how forming the complete Turn-1 structure streamlines the folding process and avoids counter-active conformers (Figure 44). Results from the three methods complement each other well, characterizing

Turn-1 formation as the first step in UBA(1) folding with the E176T mutation optimizing the process. Simulation results model the stabilization of Helix-1 and Helix-2 to coincide with the formation of Turn-1 (Figure 15B_{iii}), forming an initial Helix-Turn-Helix structure. Folding kinetics characterize Helix-3 to form next (H192E), followed by Turn-2 (Y188G) – consistent with Helix-3 forming and then structuring Turn-2 as it docks to the H1-T1-H2 structure.

Chapter 5: Conclusions

The protein sequence/structure relationship was explored through peptide REMC simulations, data mining of the wwPDB, analysis of data from literature, and experimental methods. The structural preferences and folding pathways of multiple turn sequences were modeled by CAMPARI in REMC simulations. Structural features observed in CAMPARI simulations were consistent with the structural preferences extracted from the wwPDB by EmCAST for the GD, NPSNP, and MGYE turn sequences. Disagreement between CAMPARI and EmCAST models occurred for the KPSDP turn sequence; ultimately experimental results contradicted CAMPARI and favored modeling by EmCAST. Structural differences occurred for the ASYNNP turn sequence between CAMPARI, EmCAST, and experimental UBA(1) structures. We propose that structure modeled by EmCAST reflects the local structural preference of the ASYNNP turn, but tertiary interactions in the CAMPARI peptide and UBA(1) protein restructure the ASYNNP turn into a strained conformer. This interpretation is supported by folding kinetic studies in UBA(1) protein which characterize the ASYNNP turn as forming late in the folding process.

In the context of their source proteins, the selected turns were characterized as having passive (GD), active (NPSNP, KPSDP, and MGYE) and counter-active (ASYNNP) roles in directing folding. Heatmaps from

EmCAST provide sufficient information to establish these turn classifications independently. Experimental results for the KPSDP, MGYE, and ASYNNP turn sequences support these classifications. Additional data from EmCAST highlights how the rest of each protein is designed to cooperate with each of the characterized turn mechanisms. Active turn sequences need to be flanked by well-defined structures, otherwise any directivity from the turn can be negated by flexibility in the protein backbone.

The NPSNP and KPSDP turn sequences are flanked by symmetric regions of high helical propensity in their respective proteins. We propose these 8-10 residue long segments flanking NPSNP and KPSDP facilitate turn-mediated nucleation of the two-helix bundles through a zipper-folding mechanism. The MGYE turn sequence from UBA(1) places the Met and Tyr turn residues close together, nucleating the hydrophobic core of UBA(1) and stabilizing the flanking segments of Helix-1 and Helix-2 through hydrophobic interactions. The mechanism of structure propagation varies between the XPSXP and MGYE turn sequences, but both of these active turns work with flanking residues to nucleate a hydrophobic core. Data from CAMPARI, EmCAST, and folding kinetics experiments describe a consistent folding model for UBA(1): Turn-1 (T1, MGYE) forms early, stabilizing Helix-1 (H1) and Helix-2 (H2) in a H1-T1-H2 structure, then Helix-3 forms and docks

onto the H1-T1-H2 structure while restructuring the counter-active Turn-2 (ASYNNP) into a strained conformer.

Heatmaps from EmCAST provide a powerful method to succinctly visualize the local stability or strain in a protein's structure. Energy calculations derived from these wwPDB fragment heatmaps proved to be very effective at modeling protein stability changes induced by surface mutations and predicting structure for protein/peptide sequences free of any tertiary interactions. Good correlations between calculations and experimental data were obtained for numerous protein/peptide systems. Near perfect calculations were made for surface mutations in UBA(1) – a compact protein where changes in local and global stability may be exceptionally well correlated. In contrast, data for the src SH3 domain suggests changes in local stability are poorly correlated with global stability in flexible regions unsupported by the protein's hydrophobic core.

The accuracy of EmCAST calculations for stability changes in proteins and structural shifts in peptides across a wide range of amino acid sequences and structural types mark successful modeling of the intrinsic structural properties of amino acid sequences under physiological conditions. Results from helix propensity studies highlight potential room for improvement in the design of EmCAST's energy calculations. Supplementation with another energy potential to model the long-range tertiary interactions outside of EmCAST's i±3

window may produce an accurate and complete model for protein structure and stability. In its current state, EmCAST proved to be a useful method for analyzing and optimizing protein surfaces for stability.

Chapter 6: Potential Applications

The success of our sequence-local protein energy calculations offers many benefits to protein research and development. Several potential applications, both realized and speculated, are discussed and explored. Application of the method centers around characterization of protein movements in flexible regions and/or rational manipulation of the protein surface through mutagenesis. The speed and accuracy of our calculations greatly exceed previously attempted methods.

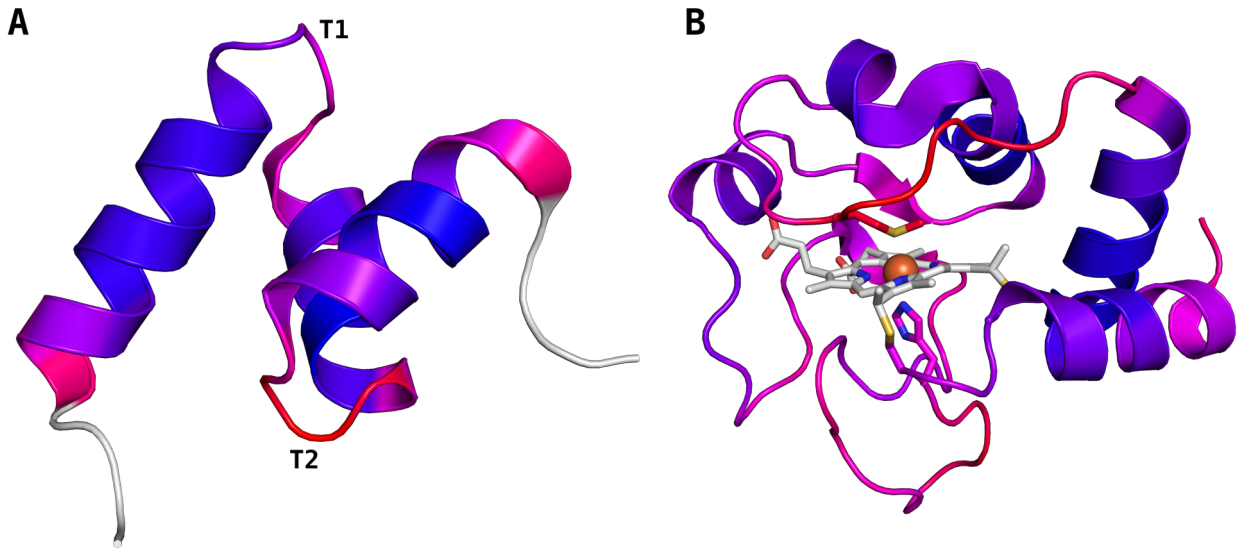

Figure 46: Stability calculations for UBA(1) and iso-1-cytochrome c

Local stability is visualized for UBA(1) (A, pdb: 6W2H) and iso-1-cytochrome c (B, pdb: 2YCC). Stability is colored from blue (most stable) to red (least stable). Calculations for the white protein segments were excluded. Calculations for iso-1-cytochrome c (B) were performed with cysteine residues replaced with serines to improve fragment sampling.

Visualization of our energy calculations for each residue in a protein provides a powerful tool to assess folding pathways,

structural dynamics, and functional sites on a protein. Energy calculations from Chapter 3: Empirical C-Alpha Stability Tool (EmCAST) are performed for each position in the protein. Equation 3.5 is used to calculate energy for each tetrad containing the residue of interest, but $\Delta G_{heatmap}$ (Eq. 3.3) is used in place of $\Delta\Delta G_{heatmap}$ (Eq. 3.4). The free energy values for each tetrad containing the residue of interest are averaged to provide a residue stability value. The range of values are used to color residues from most stable (blue) to least stable (red).

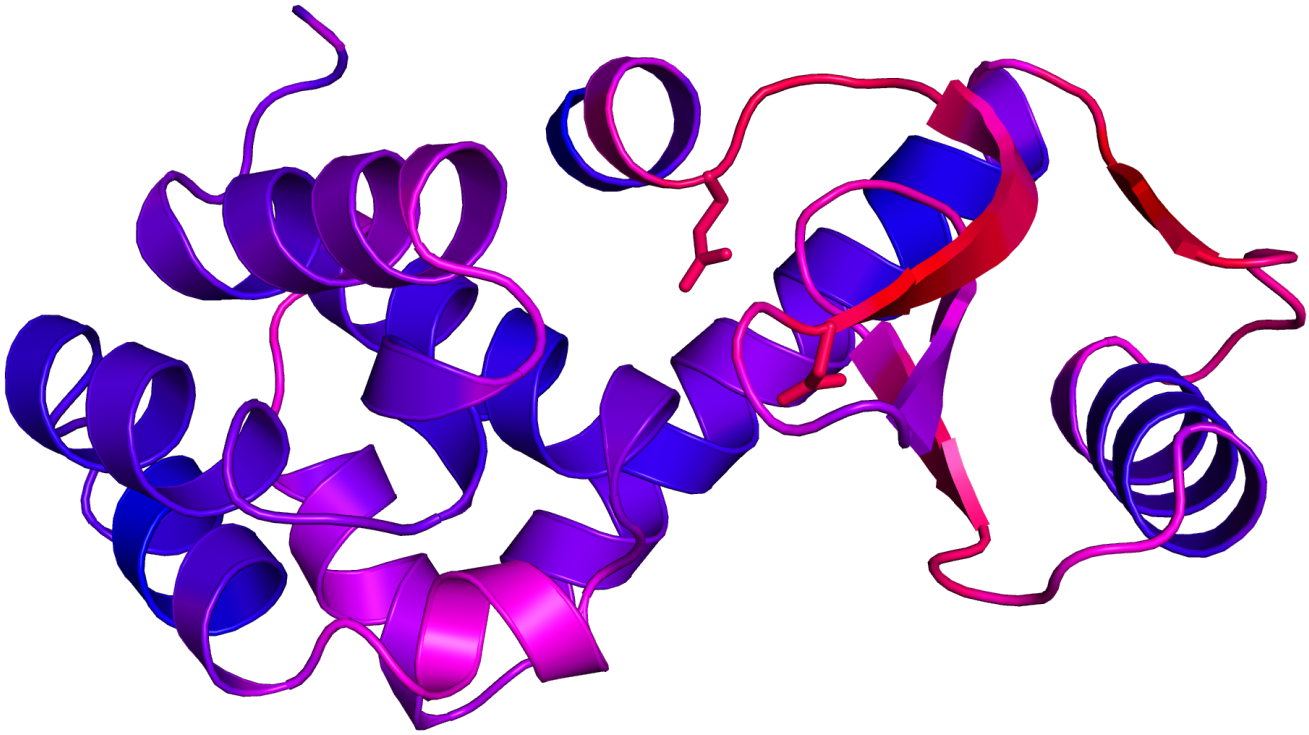

Figure 47: Stability calculations for T4 Lysozyme

Local stability is visualized for T4 Lysozyme (pdb: 2LZM). Stability is colored from blue (most stable) to red (least stable). The two residues involved in the catalytic site of T4 Lysozyme are rendered as sticks.

The local stabilities calculated in UBA(1) (Figure 46A) correspond to the order of folding events: the most stable regions form first, the least stable form last. The order of stability in iso-1-cytochrome *c* (Figure 46B) is consistent with NMR studies of its denatured state[84]. The functional regions of a protein, which often require a degree of flexibility, show up as the least stable parts of iso-1-cytochrome *c* (Figure 46B) and T4 lysozyme (Figure 47). This suggests the functional regions of a protein may be rapidly identified by our energy calculations.

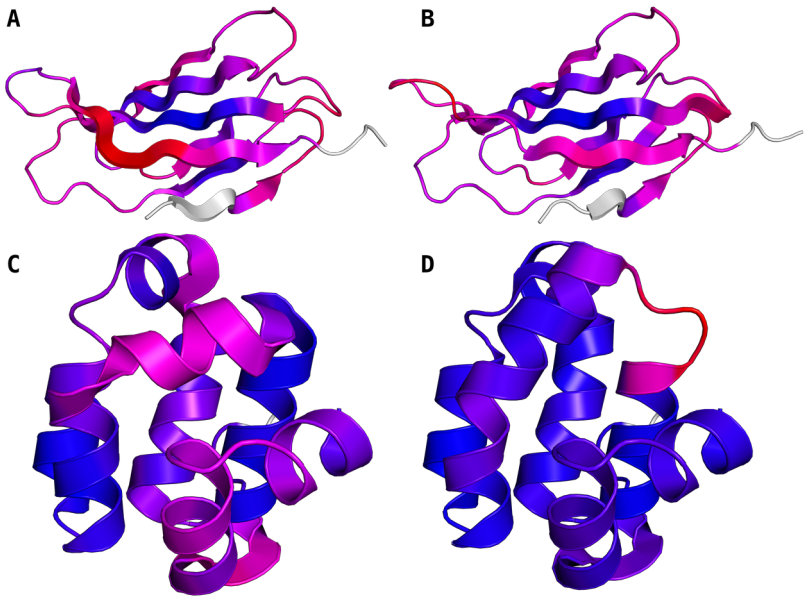

Figure 48: Stability calculations for 11FN3 and T4 Lysozyme

Local stability is visualized for the 11th FN3 (11FN3) domain (A, pdb: 5DFT; B, pdb: 6XAY) and the C-terminal domain of L99A T4 Lysozyme (C, pdb: 3DMV; D, pdb: 2LC9). Stability is colored from blue (most stable) to red (least stable). Structure colored white is not scored. Structures for 11FN3 as a single domain (A) or connected to other FN3 domains (B) are shown. Structures for the major (C) and minor (D) population structures for L99A T4 Lysozyme are shown.

Structure prone to rearrangements are also highlighted by our energy calculations. Calculations model the last strand in the 11th

FN3 domain (11FN3) to be the least stable (Figure 48A). In another crystal structure for 11FN3, this strand is restructured and the adjacent loop is modeled to be the least stable (Figure 48B). Similar behavior is seen in the cavity creating variant, L99A, of T4 Lysozyme (T4L). When a hydrophobic cavity is created in the C-terminal domain of T4L, the least stable helix (Figure 48C) rearranges to fill the cavity in a transiently formed alternate conformer (Figure 48D).

Stability calculations for surface mutations provide a high fidelity strategy to edit protein surfaces. As demonstrated with our work on UBA(1), the stability of a protein can be greatly enhanced by optimizing a protein's surface for stability. Enhancing protein stability provides key benefits to the shelf-life and immunogenicity of protein-based pharmaceuticals[85][86], the development of efficacious biocatalysts[87], the utility of protein-based scaffolds[88][89], and the directed evolution of new protein functions[90][91].

The functional properties of a protein may be rationally modulated by increasing/decreasing flexibility by strategically introducing destabilizing/stabilizing mutations. Neutral mutations may be deliberately selected in biophysical studies to leave the natural structure undisturbed while altering charge, sidechain size, or introducing cysteine residues for fluorescent labeling. Surface mutations that cause conformer-dependent changes in stability may be

designed to alter the balance of two conformers, or to isolate a specific conformer for biophysical studies or immunogenic purposes.

Additional software may be developed to remodel flexible regions in proteins according to our local energy calculations. This may reveal previously undetectable conformers that are transiently formed during a protein's function. When applied to a flexible binding pocket, this may reveal more productive targets for docking studies or small molecule designs in pharmaceutical development. Our energy calculations may also help guide machine learning methods for structure prediction by defining the initial structural biases of a sequence. Complementing our local energy calculations with an accurate long-range energy equation has the potential to provide a complete model of protein mechanics that may one day model entire folding pathways from start to finish at atomic resolution, solving the protein folding problem.

Bibliography

1: F. Crick, Central dogma of molecular biology. Nature 227 (5258), 561-563 (1970). https://doi.org/10.1038/227561a0

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

3: H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, The Protein Data Bank. Nucleic Acids Research 28, 235-242 (2000). https://doi.org/10.1107/s0907444902003451

4: J. Skolnick, M. Gao, H. Zhou, and S. Singh, AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function. J. Chem. Inf. Model 61 (10), 4827-4831 (2021). https://doi.org/10.1021/acs.jcim.1c01114

5: M. A. Pak, K. A. Markhieva, M. S. Novikova, D. S. Petrov, I. S. Vorobyev, E. S. Maksimova, F. A. Kondrashov, and D. N. Ivankov, Using AlphaFold to predict the impact of single mutations on protein stability and function. PLoS One 18 (3), e0282689 (2023). https://doi.org/10.1371/journal.pone.0282689

6: M. Karplus, The Levinthal paradox: yesterday and today. Folding and Design 2 (1), S69-S75 (1997). https://doi.org/10.1016/S1359-0278(97)00067-9

7: G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology 7(1), 95-99 (1963). https://doi.org/10.1016/S0022-2836(63)80023-6

8: S. Schwarzinger, G. Kroon, T. R. Foss, J. Chung, P. E. Write, and H. J. Dyson, Sequence-Dependent Correction of Random Coil NMR Chemical Shifts. Journal of the American Chemical Society 123 (13), 2970-2978 (2001). https://doi.org/10.1021/ja003760i

9: C. O. Mackenzie and G. Grigoryan, Protein structural motifs in prediction and design. Current Opinion in Structural Biology 44, 161-167 (2017). https://doi.org/10.1016/j.sbi.2017.03.012

10: C. Bystroff and D. Baker, Prediction of local structure in a protein using a library of sequence-structure motifs. Journal of Molecular Biology 281, 565-577 (1998). https://doi.org/10.1006/jmbi.1998.1943

11: T. D. Mueller and J. Feigon, Solution structures of UBA domains reveal a conserved hydrophobic surface for protein-protein interactions. Journal of Molecular Biology 319, 1243-1255 (2002). https://doi.org/10.1016/S0022-2836(02)00302-9

12: M. J. Behe, E. E. Lattman, and G. D. Rose, The protein-folding problem: the native fold determines packing, but does packing determine the native fold?. PNAS 88 (10), 4195-4199 (1991). https://doi.org/10.1073/pnas.88.10.4195

13: N. C. Gassner, W. A. Baase, and B. W. Matthews, A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. PNAS 93 (22), 12155–12158 (1996). https://doi.org/10.1073/pnas.93.22.12155

14: J. K. Myers, C. N. Pace, and J. M. Scholtz, Trifluoroethanol effects on helix propensity and electrostatic interactions in the helical peptide from ribonuclease T1. Protein Science 7, 383-388 (1997). https://doi.org/10.1002%2Fpro.5560070219

15: A. Chakrabartty, T. Kortemme, and R. L. Baldwin, Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. Protein Science 3, 843-852 (1994). https://doi.org/10.1002/pro.5560030514

16: G. D. Rose, Helical hairpin database. Personal Communication (2013).

17: R. Aurora and G. D. Rose, Helix capping. Protein Science 7 (1), 21-38 (1998). https://doi.org/10.1002%2Fpro.5560070103

18: L Serrano and A. R. Fersht, Capping and α-helix stability. Nature 342, 296-299 (1989). https://doi.org/10.1038/342296a0

19: L. Zhang, Y. Wang, W. L. Picking, W. D. Picking, and R. N. De Guzman, Solution Structure of Monomeric BsaL, the Type III Secretion Needle Protein of Burkholderia pseudomallei. JMB 359, 322-330 (2006). https://doi.org/10.1016/j.jmb.2006.03.028

20: A. Vitalis and R. V. Pappu, Methods for Monte Carlo Simulations of Biomacromolecules. Annual Reports in Computational Chemistry 5: 49-76 (2009). https://doi.org/10.1016/s1574-1400(09)00503-9

21: A. Vitalis and R. V. Pappu, ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. Journal of Computational Chemistry 30 (5), 673-699 (2009). https://doi.org/10.1002/jcc.21005

22: G.A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen, Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J. Phys. Chem. B 105, 6474-6487 (2001). https://doi.org/10.1021/jp003919d

23: H. J. C. Berendsen, D. van der Spoel, and R. van Drunen, GROMACS: A message-passing parallel molecular dynamics implementation. Computer Physics Communications 91(1-3), 43-56 (1995). https://doi.org/10.1016/0010-4655(95)00042-E

24: J. Lyons, A. Dehzangi, R. Heffernan, A. Sharma, K. Paliwal, A. Sattar, Y. Zhou, and Y. Yang, Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. Journal of Computational Chemistry 35, 2040-2046 (2014). https://doi.org/10.1002/jcc.23718

25: E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. Wilkins, R. Appel, and A. Bairoch, Protein identification and analysis tools on the ExPASy server. The Proteomics Protocols Handbook 571-607 (2005). https://web.expasy.org/protparam/

26: P. Wu and L. Brand, Resonance energy transfer: methods and applications. Analytical Biochemistry 218 (1), 1-13 (1994). https://doi.org/10.1006/abio.1994.1134

27: M. T. Rothfuss, D. C. Becht, B. Zeng, L. J. McClelland, C. Yates-Hansen, and B. E. Bowler, High-Accuracy Prediction of Stabilizing Surface Mutations to the Three-Helix Bundle, UBA(1), with EmCAST. Journal of the American Chemical Society 145 (42), 22979-22992 (2023). https://doi.org/10.1021/jacs.3c04966

28: D. Baker, What has de novo protein design taught us about protein folding and biophysics?. Protein Science 28 (4), 678-683 (2019). https://doi.org/10.1002/pro.3588

29: T. J. Oldfield and R. E. Hubbard, Analysis of Cα geometry in protein structures. Proteins 18(4), 324-337 (1994). https://doi.org/10.1002/prot.340180404

30: M. Blaber, X. Zhang, J. D. Lindstrom, S. D. Pepiot, W. A. Baase, B. W. Matthews, Determination of α-Helix Propensity within the Context of a Folded Protein: Sites 44 and 131 in Bacteriophage T4 Lysozyme. Journal of Molecular Biology 235 (2), 600-624 (1994). https://doi.org/10.1006/jmbi.1994.1016

31: M. Heinig and D. Frishman, STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Research 32 (2), W500-W502 (2004). https://doi.org/10.1093/nar/gkh429

32: S. Miller, J. Janin, A. M. Lesk, and C. Chothia, Interior and surface of monomeric proteins. Journal of Molecular Biology 196, 641-656 (1987). https://doi.org/10.1016/0022-2836(87)90038-6

33: P. Jemth, R. Day, S. Gianni, F. Khan, M. Allen, V. Daggett, and A. R. Fersht, The Structure of the Major Transition State for Folding of an FF Domain from Experiment and Simulation. Journal of Molecular Biology 350 (2), 363-378 (2005). https://doi.org/10.1016/j.jmb.2005.04.067

34: D. M. Korzhnev, T. L. Religa, W. Banachewicz, A. R. Fersht, and L. E. Kay, A Transient and Low-Populated Protein-Folding Intermediate at Atomic Resolution. Science 329 (5997), 1312-1316 (2010). https://doi.org/10.1126/science.1191723

35: R. Nikam, A. Kulandaisamy, K. Harini, D. Sharma, and M. M. Gromiha, ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. Nucleic Acids Research 49, D420-D424 (2021). https://doi.org/10.1093/nar/gkaa1035

36: S. Sato and A. R. Fersht, Searching for multiple folding pathways of a nearly symmetrical domain: temperature dependent f-value analysis of the B domain of protein A. Journal of Molecular Biology 372, 254-267 (2007). https://doi.org/10.1016/j.jmb.2007.06.043

37: L. Serrano, J. Sancho, M. Hirshberg, and A. R. Fersht, α-Helix stability in proteins: I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. Journal of Molecular Biology 227, 544-559 (1992). https://doi.org/10.1016/0022-2836(92)90906-Z

38: A. Matouschek, L. Serrano, and A. R. Fersht, The folding of an enzyme: IV. Structure of an intermediate in the refolding of barnase analysed by a protein engineering procedure. Journal of Molecular Biology 224, 819-835 (1992). https://doi.org/10.1016/0022-2836(92)90564-Z

39: R. Loewenthal, J. Sancho, and A. R. Fersht, Histidine-aromatic interactions in barnase: Elevation of histidine pKa and contribution to protein stability. Journal of Molecular Biology 224, 759-770 (1992). https://doi.org/10.1016/0022-2836(92)90560-7

40: A. Horovitz and A. R. Fersht, Co-operative interactions during protein folding. Journal of Molecular Biology 224, 733-740 (1992). https://doi.org/10.1016/0022-2836(92)90557-Z

41: L. Serrano, M. Bycroft, and A. R. Fersht, Aromatic-aromatic interactions and protein stability: Investigation by double-mutant cycles. Journal of Molecular Biology 218, 465-475 (1991). https://doi.org/10.1016/0022-2836(91)90725-L

42: A. Horovitz, L. Serrano, and A. R. Fersht, COSMIC analysis of the major α-helix of barnase during folding. Journal of Molecular Biology 219, 5-9 (1991). https://doi.org/10.1016/0022-2836(91)90852-W

43: L. Serrano, A. Horovitz, B. Avron, M. Bycroft, and A. R. Fersht, Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. Biochemistry 29, 9343-9352 (1990). https://doi.org/10.1021/bi00492a006

44: D. S. Riddle, V. P. Grantcharova, J. V. Santiago, E. Alm, I. Rczinski, and D. Baker, Experiment and theory highlights role of native state topology in SH3 folding. Nature Structural Biology 6, 1016-1024 (1999). https://doi.org/10.1038/14901

45: L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. Journal of Molecular Biology 254, 260-288 (1995). https://doi.org/10.1006/jmbi.1995.0616

46: C. A. McPhalen and M. N. James, Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds. Biochemistry 26 (1), 261-269 (1987). https://doi.org/10.1021/bi00375a036

47: B. Kuhlman, D. L. Luisi, P. Young, and D. P. Raleigh, pKa values and the pH dependent stability of the N-terminal domain of L9 as probes of electrostatic interactions in the denatured state. Differentiation between local and nonlocal interactions. Biochemistry 38, 4896-4903 (1999). https://doi.org/10.1021/bi982931h

48: J. H. Cho and D. P. Raleigh, Electrostatic interactions in the denatured state and in the transition state for protein folding: effects of denatured state interactions on the analysis of transition state structure. Journal of Molecular Biology 359, 1437-1446 (2006). https://doi.org/10.1016/j.jmb.2006.04.038

49: J. H. Cho, S. Sato, and D. P. Raleigh, Thermodynamics and kinetics of non-native interactions in protein folding: a single point mutant significantly stabilizes the N-terminal domain of L9 by modulating non-native interactions in the denatured state. Journal of Molecular Biology 338, 827-837 (2004). https://doi.org/10.1016/j.jmb.2004.02.073

50: J. H. Cho and D. P. Raleigh, Mutational analysis demonstrates that specific electrostatic interactions can play a key role in the denatured state ensemble of proteins. Journal of Molecular Biology 353, 174-185 (2005). https://doi.org/10.1016/j.jmb.2005.08.019

51: J. H. Cho, W. Meng, S. Sato, E. Y. Kim, H. Schindelin, and D. P. Raleigh, Energetically significant networks of coupled interactions within an unfolded protein. PNAS 111, 12079-12084 (2014). https://doi.org/10.1073/pnas.1402054111

52: S. Sato, J. H. Cho, I. Peran, R. G. Soydaner-Azeloglu, and D. P. Raleigh, The N-terminal domain of ribosomal protein L9 folds via a diffuse and delocalized transition state. Biophysical Journal 112, 1797-1806 (2017). https://doi.org/10.1016/j.bpj.2017.01.034

53: A. G. Gittis, W. E. Stites, and E. E. Lattman, The phase transition between a compact denatured state and a random coil state in staphylococcal nuclease is first-order. Journal of Molecular Biology 232, 718-724 (1993). https://doi.org/10.1006/jmbi.1993.1425

54: D. Shortle, Staphylococcal nuclease: a showcase of m-value effects. Advances in Protein Chemistry 46, 217-247 (1995). https://doi.org/10.1016/S0065-3233(08)60336-8

55: M. Blaber, X. J. Zhang, J. D. Lindstrom, S. D. Pepiot, W. A. Baase, and B. W. Matthews, Determination of α-helix propensity within the context of a folded protein: sites 44 and 131 in bacteriophage T4 lysozyme. Journal of Molecular Biology 235, 600-624 (1994). https://doi.org/10.1006/jmbi.1994.1016

56: J. K. Myers, A direct comparison of helix propensity in proteins and peptides. PNAS 94, 2833-2837 (1997). https://doi.org/10.1073/pnas.94.7.283

57: J. K. Myers, C. N. Pace, and J. M. Scholtz, Helix propensities are identical in proteins and peptides. Biochemistry 36, 10923-10929 (1997). https://doi.org/10.1021/bi9707180

58: M. Sternke, K. W. Tripp, and D. Barrick, The use of consensus sequence information to engineer stability and activity in proteins. Methods in Enzymology 643, 149-179 (2020). https://doi.org/10.1016/bs.mie.2020.06.001

59: F. Pucci, K. V. Bernaerts, J. M. Kwasigroch, and M. Rooman, Quantification of biases in predictions of protein stability changes upon mutations. Bioinformatics 34, 3659-3665 (2018). https://doi.org/10.1093/bioinformatics/bty348

60: Y. Dehouck, A. Grosfils, B. Folch, D. Gilis, P. Bogaerts, and M. Rooman, Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 25, 2537-2543 (2009). https://doi.org/10.1093/bioinformatics/btp445

61: C. Savojardo, P. Fariselli, P. L. Martelli, and R. Casadio, INPS-MD: a web server to predict stability of protein variants from sequence and structure. Bioinformatics 32, 2542-2544 (2016). https://doi.org/10.1093/bioinformatics/btw192

62: E. H. Kellogg, A. Leaver-Fay, and D. Baker, Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins: Structure, Function, and Bioinformatics 79, 830-838 (2011). https://doi.org/10.1002/prot.22921

63: C. L. Worth, R. Preissner, and T. L. Blundell, SDM—a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Research 39, W215-W222 (2011). https://doi.org/10.1093/nar/gkr363

64: J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, The FoldX web server: an online force field. Nucleic Acids Research 33, W382-W388 (2005). https://doi.org/10.1093/nar/gki387

65: D. E. V. Pires, D. B. Ascher, and T. L. Blundell, DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Research 42, W314-W319 (2014). https://doi.org/10.1093/nar/gku411

66: D. E. V. Pires, D. B. Ascher, and T. L. Blundell, mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30, 335-342 (2014). https://doi.org/10.1093/bioinformatics/btt691

67: D. C. Becht, M. J. Leavens, B. Zeng, M. T. Rothfuss, K. Briknarova, and B. E. Bowler, Residual structure in the denatured state of the fast-folding UBA(1) domain from the human DNA excision repair protein HHR23A. Biochemistry 61, 767-784 (2022). https://doi.org/10.1021/acs.biochem.2c00011

68: M. J. Leavens, L. E. Spang, M. M. Cherney, and B. E. Bowler, Denatured state conformational biases in three-helix bundles containing divergent sequences localize near turns and helix capping residues. Biochemistry 60, 3071-3085 (2021). https://doi.org/10.1021/acs.biochem.1c00400

69: Y. Nozaki, The preparation of guanidine hydrochloride. Methods in Enzymology 26, 43-50 (1972). https://doi.org/10.1016/S0076-6879(72)26005-0

70: M. M. Santoro and D. W. Bolen, Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl a-chymotrysin using different denaturants. Biochemistry 27, 8063-8068 (1988). https://doi.org/10.1021/bi00421a014

71: J. M. Scholtz, G. R. Grimsley, and C. N. Pace, Solvent denaturation of proteins and interpretations of the m value. Methods in Enzymology 466, 549-565 (2009). https://doi.org/10.1016/S0076-6879(09)66023-7

72: W. Kabsch, XDS. Acta Crystallographica Section D: Structural Biology 66, 125-132 (2010). https://doi.org/10.1107/S0907444909047374

73: P. R. Evans and G. N. Murshudov, How good are my data and what is the resolution?. Acta Crystallographica Section D: Structural Biology 69, 1204-1214 (2013). https://doi.org/10.1107/S0907444913000061

74: P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, and P. H. Zwart, PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallographica Section D: Structural Biology 66, 213-221 (2010). https://doi.org/10.1107/S0907444909052925

75: P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan, Features and development of Coot. Acta Crystallographica Section D: Structural Biology 66, 486-501 (2010). https://doi.org/10.1107/S0907444910007493

76: V. B. Chen, W. B. Arendall III, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica Section D: Structural Biology 66, 12-21 (2010). https://doi.org/10.1107/S0907444909042073

77: J. M. Scholtz, H. Qian, V. H. Robbins, and R. L. Baldwin, The energetics of ion-pair and hydrogen-bonding interactions in a helical peptide. Biochemistry 32, 9668-9676 (1993). https://doi.org/10.1021/bi00088a019

78: R. Aurora and G. D. Rose, Helix capping. Protein Science 7 (1), 21-38 (1998). https://doi.org/10.1002/pro.5560070103

79: L. Serrano, J. Sancho, M. Hirshberge, and A. R. Fersht, α-Helix stability in proteins: I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. Journal of Molecular Biology 227, 544-559 (1992). https://doi.org/10.1016/0022-2836(92)90906-Z

80: J. S. Richardson and D. C. Richardson, Amino acid preferences for specific Llocations at the ends of α helices. Science 240, 1648-1652 (1988). https://doi.org/10.1126/science.3381086

81: R. J. Anderson, Z. Weng, R. K. Campbell, and X. Jiang, Main-chain conformational tendencies of amino acids. Proteins 60, 679-689 (2005). https://doi.org/10.1002/prot.20530

82: K. Fujiwara, H. Toda, and M. Ikeguchi, Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. BMC Structural Biology 12, 18 (2012). https://doi.org/10.1186/1472-6807-12-18

83: M. Karplus and D. L. Weaver, Protein folding dynamics: the diffusion-collision model and experimental data. Protein Science 3, 650-668 (1994). https://doi.org/10.1002/pro.5560030413

84: T. A. Danielson, J. M. Stine, T. A. Dar, K. Briknarova, and B. E. Bowler, Effect of an Imposed Contact on Secondary Structure in the Denatured State of Yeast Iso-1-cytochrome c. Biochemistry 56, 6662-6676 (2017). https://doi.org/10.1021/acs.biochem.7b01002

85: M. C. Manning, D. K. Chou, B. M. Murphy, R. W. Payne, and D. S. Katayama, Stability of protein pharmaceuticals: an update. Pharm. Res. 27, 544-575 (2010). https://doi.org/10.1007/s11095-009-0045-6

86: M. Sauerborn, V. Brinks, W. Jiskoot, and H. Schellekens, Immunological mechanism underlying the immune response to recombinant human protein therapeutics. Trends Pharmacol. Sci. 31, 53-59 (2010). https://doi.org/10.1016/j.tips.2009.11.001

87: A. S. Bommarius and M. F. Paye, Stabilizing biocatalysts. Chem. Soc. Rev. 42, 6534-6565 (2013). http://dx.doi.org/10.1039/C3CS60137D

88: S. A. Jacobs, M. D. Diem, J. Luo, A. Teplyakov, G. Obmolova, T. Malia, G. L. Gilliland, and K. T. O'Neil, Design of novel FN3 domains with high stability by a consensus sequence approach. Protein Eng. Des. Sel. 25, 107-117 (2012). https://doi.org/10.1093/protein/gzr064

89: A. Koide, M. R. Jordan, S. R. Horner, V. Batori, and S. Koide, Stabilization of a fibronectin type III domain by the removal of u nfavorable electrostatic interactions on the protein surface. Biochemistry 40, 10326-10333 (2001). https://doi.org/10.1021/bi010916y

90: J. D. Bloom, S. T. Labthavikul, C. R. Otey, and F. H. Arnold, Protein stability promotes evolvability. PNAS 103, 5869-5874 (2006). https://doi.org/10.1073/pnas.0510098103

91: J. Zheng, N. Guo, and A. Wagner, Selection enhances protein evolvability by increasing mutational robustness and foldability. Science 370, eabb5962 (2020). https://www.science.org/doi/10.1126/science.abb5962