High-Accuracy Prediction of Stabilizing Surface Mutations to the Three-helix Bundle, UBA(1), with EmCAST

Michael T. Rothfuss[a], Dustin C. Becht[a], Baisen Zeng[b], Levi J. McClelland[b,c], Cindee Yates-Hansen[b], and Bruce E. Bowler[a,b,*]

[a]Department of Chemistry and Biochemistry, University of Montana, Missoula, MT 59812, United States;

[b]Center for Biomolecular Structure and Dynamics, University of Montana, Missoula, MT 59812, United States;

[c]Division of Biological Sciences, University of Montana, Missoula, MT 59812, United States.

*Corresponding author. Email: bruce.bowler@umontana.edu

ABSTRACT

The accurate modeling of energetic contributions to protein structure is a fundamental challenge in computational approaches to protein analysis and design. We describe a general computational method, EmCAST (Empirical Cα Stabilization), to score and optimize sequence to structure in proteins. The method relies on an empirical potential derived from a database of the Cα dihedral angle preferences for all possible four-residue sequences using data available in the Protein Data Bank. Our method produces stability predictions that naturally correlate one-to-one with experimental results for solvent-exposed mutation sites. EmCAST predicted four mutations that increased the stability of a three-helix bundle, UBA(1), from 2.4 to 4.8 kcal/mol by optimizing residues in both helices and turns. For a set of eight variants, the predicted and experimental stabilizations correlate very well ($R^2 = 0.97$) with a slope near 1 and with a 0.16 kcal/mol standard error for EmCAST predictions. Tests against literature data for the stability effects of surface-exposed mutations show that EmCAST outperforms existing stability prediction methods. UBA(1) variants were crystallized to verify and analyze their structures at atomic resolution. Thermodynamic and kinetic folding experiments were performed to determine the magnitude and mechanism of stabilization. Our method has the potential to enable rapid, rational optimization of natural proteins, expand analysis of the sequence/structure relationship, and supplement existing protein design strategies.

INTRODUCTION

The identification of stabilizing residues is central to protein structure analysis, design, and optimization. Enhancing protein stability provides key benefits to the shelf-life and immunogenicity of protein-based pharmaceuticals,[1,2] the development of efficacious biocatalysts,[3] the utility of protein-based scaffolds,[4,5] and the directed evolution of new protein functions.[6,7] Current computational methods to predict the effect of mutations on stability have standard errors that often exceed the magnitude of observed stabilization[8,9] when applied to literature data sets and struggle to identify stabilizing mutations due to their relative scarcity in stability datasets.[10-12]

When existing algorithms are used as tools to rationally stabilize proteins, the success rate of creating stabilized variants is often only about 30%.[13] Mutations predicted to stabilize a protein often turn out to be significantly destabilizing.[8,14,15] Mutations that do stabilize proteins also tend to lower protein solubility.[8,12] Optimizing surface electrostatics of proteins has proven to be a fruitful approach to stabilize proteins.[16,17] However, like other stability prediction methods the standard error of the prediction can be significant.[16] Use of a multiple sequence alignment (MSA) to generate a consensus sequence has proven to be an effective approach to significantly stabilize proteins.[4,18-20] However, the MSA approach does not predict the magnitude of the stabilization, it often requires a large number of mutations and because of sequence-context effects on stability, it is difficult to discern if a small subset of the mutations could accomplish the desired stabilization.[21] Thus, there is a need for new methods that can accurately predict the magnitude of stabilizing mutations so that proteins can be stabilized efficiently with a small number of mutations.

To bridge this knowledge gap, we have developed a method to accurately predict stability changes for mutations at solvent-exposed positions in proteins. Although, solvent-exposed sites have traditionally been considered to be non-perturbing,[22] recent studies show that surface residues can be very effective at stabilizing proteins.[21,23] Our strategy uses structures in the RCSB PDB[24] as our sole source of data on structural preferences of four-residue sequences and naturally produces predictions in kcal/mol without the use of any fitted constants or black-box machine learning techniques. We utilize our tool, EmCAST (Empirical Cα Stabilization), to optimize sequence to structure in a 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.[25] We have previously characterized the stability, folding kinetics and denatured state properties of wild type (WT) UBA(1) and determined its X-ray structure at a resolution of 1.60 Å.[26,27] The domain is of modest stability (2.4 kcal/mol), providing a good candidate for rational stabilization. Here we experimentally characterize the structure, stability, and folding landscape of stabilized variants and compare the stabilizations of single and multi-site variants to those predicted by EmCAST.

RESULTS

Visualizing the Sequence/Structure Relationship. EmCAST uses a fragment database (FDB) to analyze the relationship between sequence and structure in proteins. Traditional methods have used clustered structures and sequence motifs to model this relationship in a structure-to-sequence approach.[28,29] This cluster/motif strategy neglects two key pieces of information: the context-dependent effects of amino acid identity on structural preferences[30-32] (sequence motifs only retain information on the relative probability of amino acids at each site of

a clustered structure) and the existence of multiple populated conformers for a given sequence. EmCAST's FDB takes a sequence-to-structure approach to preserve the sequence-context dependence of structural preferences. The dihedral angle formed between 4 consecutive alpha-carbons (τ)[32,33] is used to quantitatively represent all conformers for each four-residue sequence. We note that the two bond angles, θ_1 and θ_2, for the three pseudobonds that define the Cα dihedral angle τ can vary from about 90 to 150 degrees. However, τ and the values of θ_1 and θ_2 are correlated. For α-helix, τ is near 50° and θ_1 and θ_2 are near 95° and for β-sheet, τ is near -145 and θ_1 and θ_2 are near 125°.[32] Thus, we have found it sufficient to use only τ in developing EmCAST. The τ conformational distribution provides a continuous and finite scale that can represent all possible fragment conformers for each four-residue sequence (tetrad) a feature critical to the success of our energy calculations. Furthermore, our database of tetrad structural preferences retains information about the structural preference of τ for the tetrads that precede and follow each tetrad (Supporting Experimental Procedures and Figure S2). The entire RCSB PDB[24] was used in our FDB rather than non-redundant representations to capture data from natural and artificial protein variants. We did not use a resolution cut-off because we are only using Cα coordinates, which are more reliable than side chains in lower resolution structures. The database was built using a twelve residue sequence window to allow data from 4-residue sequences with identical internal and flanking sequences to be averaged, preventing sampling biases for protein structures with multiple entries in the PDB. As has been noted by the Matthews lab,[34] solvent-exposed sites are expected to better reflect intrinsic structural preferences because they are less likely to be affected by long-range tertiary contacts. Therefore, we have weighted the statistics in our database of structural preferences by fragment solvent accessibility (linearly from a weighting of 1 for fully exposed to 0 for fully buried) to bias data

towards the innate structural preferences of the underlying sequence (see Supporting Information). While membrane proteins were included in our database, weighting by solvent-accessibility effectively removes data for transmembrane segments of membrane proteins.

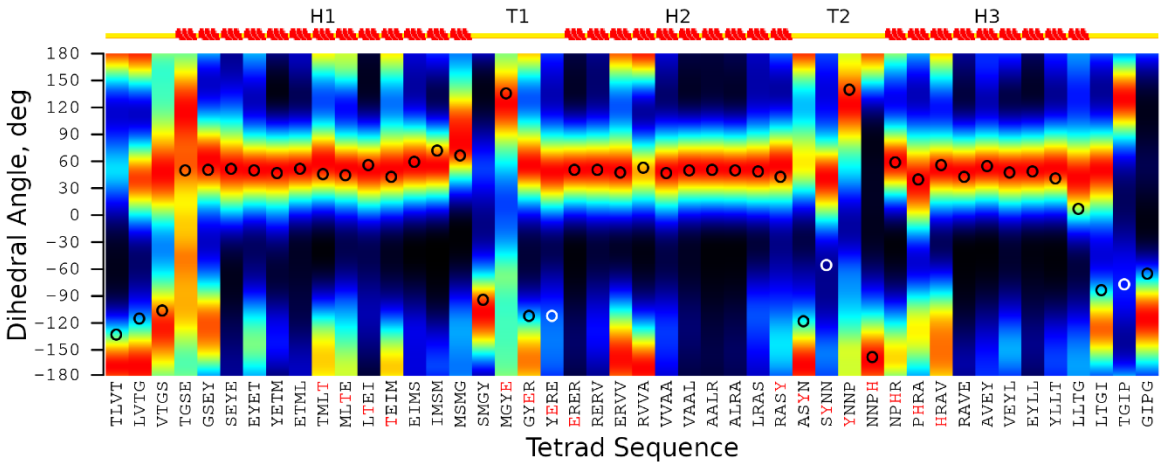

Figure 1. Fragment heatmap for WT UBA(1). The primary sequence of UBA(1) is represented as its sequence of overlapping tetrads (x-axis). The observed Cα dihedral angles (τ, y-axis) of UBA(1) (pdb: 6W2H)[26] are shown on the plot as black or white open circles. The distribution of samples in our fragment database is rendered as heat (red = most populated, black, zero population). The secondary structure of UBA(1) is visualized at the top to aid interpretation. α-helices and β-sheets have τ values centered around 50° and −145°, respectively.[32] The residues that are used for stabilizing mutations are colored red in each tetrad that contains them.

The primary structure of a protein can be broken down into a series of overlapping tetrad segments. Each tetrad has a collection of conformers in our FDB that can be represented as a population distribution across the dihedral angle τ. For each tetrad in our model system, UBA(1), we can plot this population distribution as a heatmap to compare the observed Cα dihedral angles for the tertiary structure of our protein against the collection of conformers found in our FDB (Figure 1). Most of the observed dihedral angles in UBA(1) fall within the dominant population of our FDB's corresponding tetrad fragments. This observation indicates that most of the primary structure of UBA(1) has local structural preferences that match its folded tertiary structure well. There are several exceptions, notably in the turn regions (T1 and T2, see Figure 1), suggesting

structure in these locations could be determined by non-local interactions to a significant extent. Alternatively, this observation indicates that the primary structure of UBA(1) in T1 and T2 could be rationally optimized to match and stabilize its tertiary structure.

Optimizing Sequence to Structure. Given a target structure, differences in fragment τ population distributions can be used to calculate an energy difference between two sequences. For a select tetrad segment and its corresponding fragments in our FDB, the population of fragments with matching τ ($P_{matching}$) and the population of total fragments (P_{total}) are used to calculate ΔG (Eq. 1). Examples of evaluations of $P_{matching}$ and P_{total} for wild type (WT) UBA(1) and the Y188G mutation for the tetrad SXNN (where X is Y or G) are shown in Figure S2. The ΔΔG between WT and mutant sequences provides an experimentally testable prediction of protein stability (Eq. 2), given the assumption that the mutation does not affect the structure of the protein.

$$\Delta G = RT \cdot \ln\left(\frac{P_{matching}}{P_{total} - P_{matching}}\right) \qquad \text{(Eq. 1)}$$

$$\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wildtype} \qquad \text{(Eq. 2)}$$

The τ distributions in Figure 1, are for isolated tetrads. However, because adjacent tetrads are expected to interact with each other, we use two 2D heatmaps for each tetrad to evaluate the effects of the preceding and following residues on the structural preferences of a given tetrad and then sum over all tetrads that contain the mutated residue (see Supporting Information for more details). Only interactions within 3 residues of the mutation site are modeled in this approach, leaving predictions at buried sites inherently unreliable. Our algorithm was prototyped and tested using surface mutations in the FF Domain from HYPA/FBP11[35] (Figure S3, R^2=0.63). After adapting the code for performance (see EmCAST Runtime Benchmarks in the SI), EmCAST was

used to predict ΔΔG values for 779 UBA(1) mutations (Figure 2) in less than 0.1 seconds (Tables S1 and S2).

Four stabilizing UBA(1) mutations were selected for experimental verification: two turn mutations (E176T and Y188G) and two helical mutations (T168R and H192E). Each mutation optimizes the local τ distribution to better match UBA(1)'s tertiary structure (Figure S4), producing a ΔΔG heatmap (Figure S5) with considerably fewer stabilizing mutations (red squares). The selected mutations sites are free of interactions outside of EmCAST's i±3 evaluation window, well represented within our fragment database, and are predicted to stabilize UBA(1) by at least 0.5 kcal/mol. We also used an MSA for UBA(1) with 43 sequences provided by Mueller and Feigon[25] to look at the correlation between stabilizations in kcal/mol predicted by EmCAST and the fractional occurrence of an amino acid at the sequence position of the mutation in the MSA (Figure S6). The R^2 values for the four correlation lines ranged from 0.01 to 0.67. 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 method predicted 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. For positions T168 and E176, the mutations selected by EmCAST would not have been predicted as favorable from the MSA in Mueller and Feigon.[25]

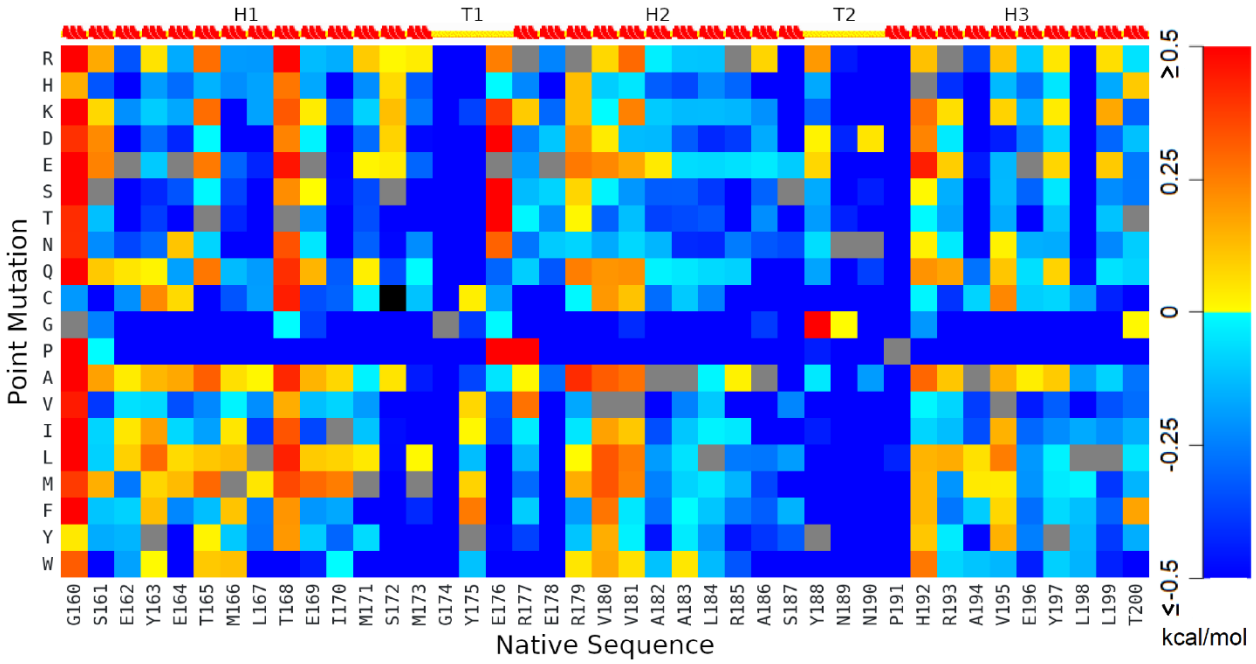

Figure 2. Saturation mutagenesis heatmap for WT UBA(1). A scale bar that matches color to degree of stabilization (positive values) or destabilization (negative values) is provided on the right. Grey squares represent WT residues. Mutations with inadequate fragment sampling are black, which corresponds to the 2D heatmap (Figure S2) having a value of zero at a τ pair for the mutation.

Guanidine hydrochloride (GdnHCl) unfolding experiments, monitored by circular dichroism (CD), were used to measure changes in protein stability (Figure 3A). Eight UBA(1) variants were tested and matched predicted stability changes exceptionally well (Table 1 and Figure 3B, slope ~ 1, $R^2 = 0.97$) with a 0.16 kcal/mol standard error of the estimate. Stability enhancements notably were over-predicted for variants composed mainly of turn mutations (Figure 3B, blue and purple points). Combining these variants with stabilizing mutations in helices 1 and 3 (H1, H3) abolished the energy discrepancy (Figure 3B, pink points). This observation may indicate that local dynamics at the mutation site can negate a portion of the predicted stability.

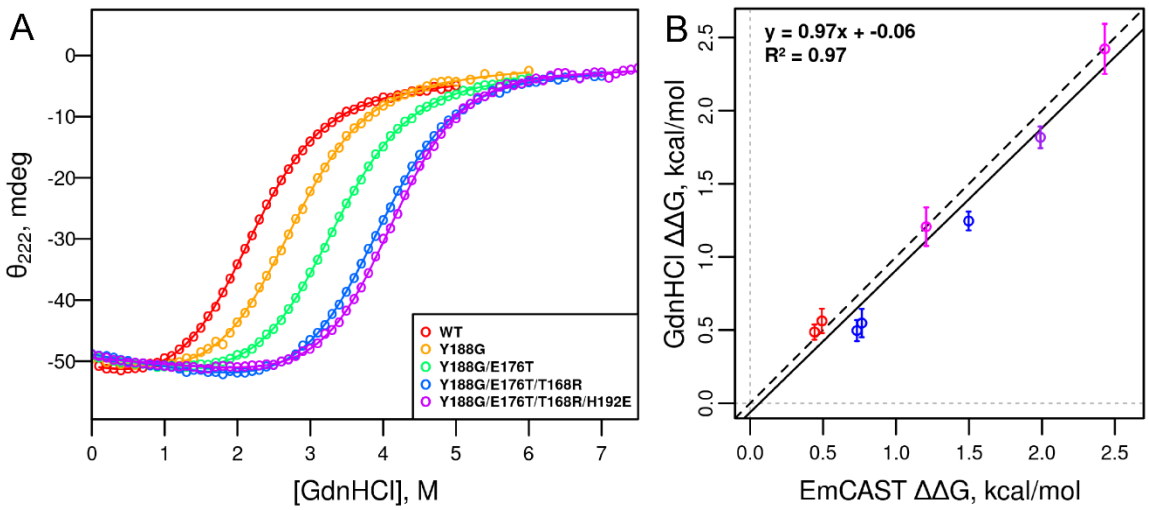

Figure 3. UBA(1) experimental stability data. (A) Representative unfolding curves for progressively stabilized UBA(1) variants. Unfolding was induced by GdnHCl titration and monitored by CD at 222 nm using a 250 nm baseline. (B) Correlation plot between EmCAST predictions and stability data obtained from GdnHCl unfolding experiments. The line of best fit is shown as a solid black line. The dashed black line indicates where a perfect fit would lie. The red data points are single site mutations in helical regions and the blue data points are single or double mutations in turn regions. Pink data points have equal numbers of mutations in helical and turn regions. The purple data point has two turn mutations and one helical mutation.

Table 1. Parameters from GdnHCl Unfolding Experiments for UBA(1) Variants and Corresponding ΔΔG Predictions from EmCAST.[a]

Variant	$\Delta G_u^{o'}(H_2O)$, 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	0
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	1.5
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

[a]EmCAST predictions were made using the crystal structure of WT UBA(1) (pdb: 6W2H).[26]
[b]Errors in $\Delta G_u^{o'}(H_2O)$ and *m* are the standard deviations of the parameters obtained from separate fits of Eq. 3 to three GdnHCl titrations for each protein. The error in ΔΔG is obtained from standard propagation of the error in the $\Delta G_u^{o'}(H_2O)$ values.

Altogether, the four selected mutations double UBA(1)'s stability from 2.4 to 4.8 kcal/mol as predicted. The mechanism(s) behind stabilization are not revealed within EmCAST due to the empirical nature of our free energy potential. Structural and folding kinetics data are provided below to further elucidate the atomic interactions leading to stabilization. We note that if mutations derived from an MSA had been used to stabilize UBA(1), a similar increase in stability would likely have been achieved based on the predictions of EmCAST. However, it is important to note that the MSA approach does not provide quantitative predictions and would not have predicted that two of the mutations we made would stabilize UBA(1).

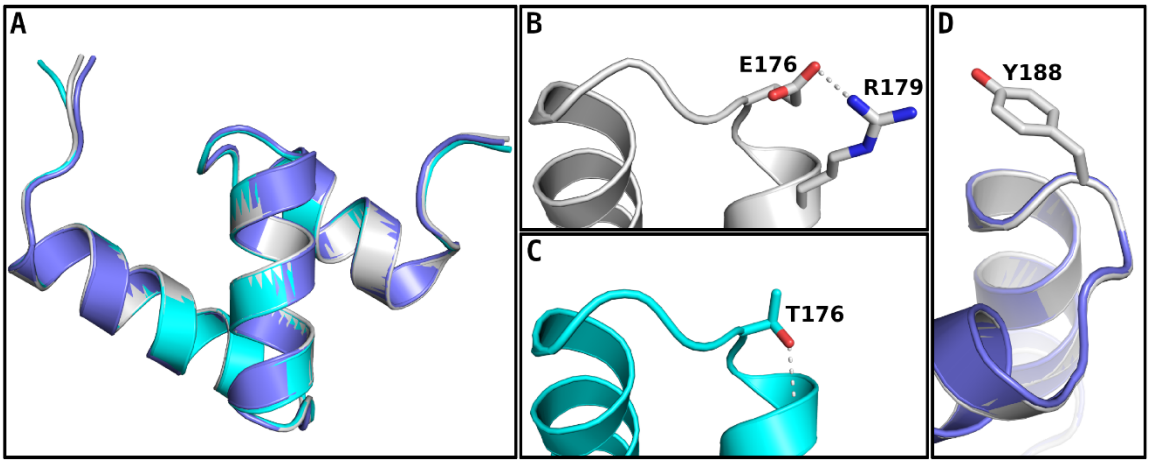

Figure 4. X-ray structures of WT UBA(1) and turn variants. (A) Cartoon overlay of UBA(1) WT (grey, PDB file: 6W2H),[26] Y188G (cobalt, PDB file: 6W2G), and E176T/Y188G (cyan, PDB file: 7TGP) X-ray structures. (B) UBA(1) WT turn 1, a potential electrostatic interaction between E176 and R179 side chains 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.

Analyzing the Stabilizing Effects of Mutations to UBA(1). A critical assumption of EmCAST is that the mutations used to stabilize a protein do not affect 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 S3, S4 and S5). 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 4). Residue E176 provides two stabilizing features that are lost upon mutation to Thr: stabilization of helix 2's macroscopic electrostatic dipole[36] and a constructive electrostatic intrahelix (i, i+3) interaction[37] with R179 (Figure 4B). The E176T mutation more than compensates for these lost features by introducing a favorable Ncap[38] to helix 2 (H2), wherein T176's gamma-hydroxyl hydrogen bonds to R179's backbone-amide NH (Figure 4C). Other experimental[39] and database[38,40] 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 S7A). Glycine is more commonly found in this backbone geometry (Figure S7B), suggesting that Y188G stabilizes UBA(1) through backbone torsion angle optimization (Figure 4D).

The two helical mutations, T168R and H192E, are both favored over WT residues on empirical helix propensity scales.[41] H192E places a glutamate at the N2 position of helix 3, stabilizing the helix dipole.[36] Experimental[39] and database[38,40] 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 2). The stabilizing mutagenic potential, predicted by EmCAST, for residues flanking this site drop after either mutation (Figure S8A-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 S8D). Taken together, these predictions suggest sequence-context-dependent effects play a significant role in the stabilization provided by T168R.

Alterations to UBA(1)'s folding landscape were analyzed by stopped-flow experiments for several variants (Table 2, Figure S9). 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,[26,42] 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.[27] In contrast, macroscopic dipole optimization by H192E will only be available after the formation of helix 3.

Table 2. Folding Kinetics Parameters of UBA(1) Variants.

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

[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 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}.

Modeling the energetic distribution of UBA(1) turn conformations provides additional insights into the folding kinetics of UBA(1) variants (see SI: Generating Protein Conformers). Briefly, combinations of likely dihedral angles from our heatmaps can be used to generate structures of the protein backbone for a given sequence in all accessible conformations. Each generated structure can then be scored by our energy equations. The set of possible conformations generates an energy landscape that resembles a folding funnel (Figure S10).[43] 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 5A). 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 5B). 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 S11).

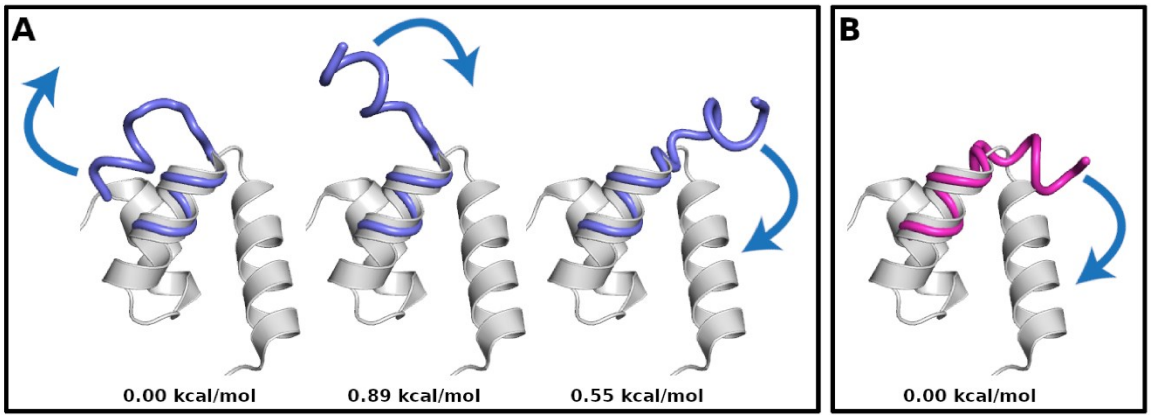

Figure 5. Modeled folding mechanisms of UBA(1) turn 1 for (A) WT and (B) T168R/E176T variants. Available conformations for the H1-T1-H2 segment of UBA(1) are modeled and their relative energies scored by EmCAST. Select conformers (cobalt, magenta) are aligned to the crystal structure of UBA(1) (pdb: 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.

Comparison to Existing Data and Methods. EmCAST provided high accuracy predictions of stabilizing mutations that led to substantial stabilization of UBA(1). To further test the generality of EmCAST to predict stabilizing mutations at surface-exposed sites, we searched the ProThermDB[44] 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 with isothermal stability measurements determined near physiological conditions. We felt it was important to have at least 10 mutations in each protein to determine if there were qualitative

difference between EmCAST's predictive abilities for different types of proteins. Three proteins were found: B-Domain of Staphylococcal Protein A (74% helix, 26% loop), FF Domain (71% helix, 29% loop), and barnase (56% loop, 23% helix, 21% sheet). Although, the match between predicted and experimental changes in stability did not achieve the exceptional accuracy observed for the UBA(1) mutations, the EmCAST predictions for these other proteins correlated well with the experimental data (Figure 6A). The standard error in EmCAST's prediction for the set of variants from the ProTherm database was 0.50 kcal/mol, a 3-fold increase relative to the standard error 0.16 kcal/mol in EmCAST's predictions for the UBA(1) variants. Unlike our engineered UBA(1) mutations, published surface mutations in other proteins were almost exclusively destabilizing. Our accuracy for destabilizing mutations may be affected by structural and/or statistical factors. Mutations that disfavor the native state structure may favor dynamic deviations from the native structure which are not accounted for by EmCAST because it assumes the structure is unaffected by mutation. Within EmCAST, destabilizing mutations have heatmaps that are less populated at the target geometry, leading to poorer statistical sampling/coverage. These factors likely limit prediction accuracy for destabilizing mutations and conversely may explain the high accuracy achieved for the stabilizing mutations made to UBA(1).

This dataset was also used to compare EmCAST against 7 other stability prediction methods (see SI Appendix). Ranking by prediction correlation was as follows: EmCAST ($R^2 = 0.79$, Fig. 6A), PopMuSiC[11,45] ($R^2 = 0.56$, Figure 6B), INPS-3D[46] ($R^2 = 0.46$, Figure 6C), Rosetta-ddG[47] ($R^2 = 0.38$, Figure S12A), SDM[48] ($R^2 = 0.37$, Figure S12B), FoldX[49] ($R^2 = 0.36$, Figure S12C), DUET[50] ($R^2 = 0.33$, Figure S12D), and mCSM[51] ($R^2 = 0.02$, Figure S12E). Many of the methods tested struggled to predict our UBA(1) mutations as stabilizing. Only EmCAST, PopMuSiC, SDM, and FoldX predicted stabilizing $\Delta\Delta G$ values for the majority of the UBA(1) mutations.

Although EmCAST does not perform as well for the mutations extracted from the ProTherm database, it still significantly outperforms the other methods even when the stabilizing UBA(1) variants are not included in the correlation (Figure S13). It is the only method that produces a slope of 1 with the data from the ProTherm database and it has the largest squared correlation coefficient (0.65) and the lowest standard error of the prediction among the methods tested (Table S6). EmCAST also has the smallest standard deviation of the average error of its predictions indicating that there are fewer large outliers in its predictions than most of the other methods.

In addition to outperforming all of the 7 tested methods in both speed and accuracy, several other features of EmCAST's design make it unique. Our stability predictions are free of any fitted constants, not based on experimentally determined stability values, and intrinsically antisymmetric with respect to the direction of a mutation. In other words, EmCAST will give ΔΔG for the T168R variant of UBA(1) that is equal in magnitude, but opposite in sign, to the reverse mutation back to the WT sequence, R168T, for the T168R variant of UBA(1).

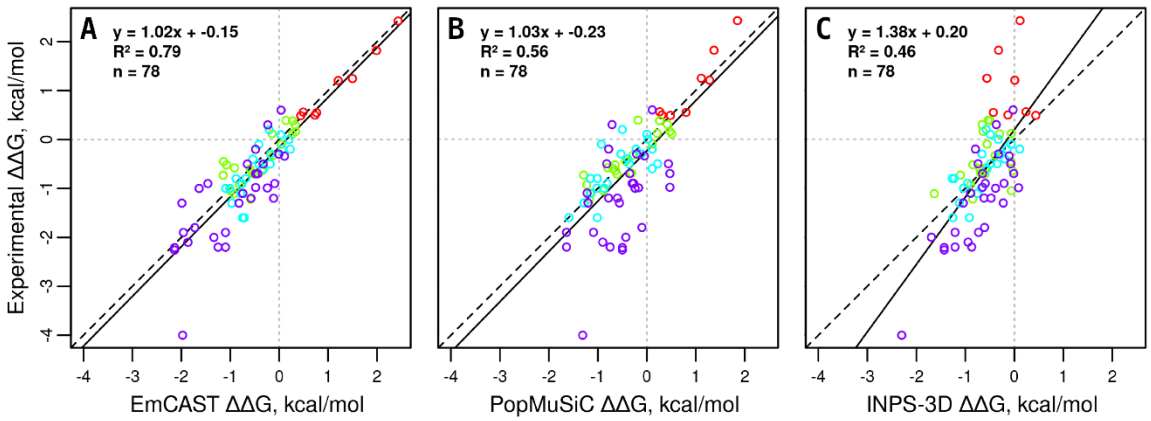

Figure 6. Surface mutation prediction correlations for the top three performing methods tested. The best predictors tested are (A) EmCAST, (B) PopMuSiC, and (C) INPS-3D. Proteins included are UBA(1) (red, pdb: 6W2H), B-Domain of Staphylococcal Protein A[52] (green, pdb: 1SS1), FF Domain[35] (teal, pdb: 2KZG), and barnase[39,53-57] (purple, pdb: 1BNI). The line of best fit is shown as a solid black line. The dashed black line indicates where a perfect fit would lie.

Application to Helical Propensity in Proteins and Peptides. The extensive studies on helical propensity enable assessment of EmCAST in specific sequence contexts for different mutations and experimental conditions.[34,58-64] Studies using helical peptides, which are not supported by long range tertiary interactions, enable structures to be relaxed according to our sequence-local energy calculations, as described above (see SI: Generating Protein Conformers). Previous work tested saturation mutations at surface position A21 of the RNase T1 helix in both peptide and protein models.[62,63] The A21P variant, which inhibited expression of RNase T1 protein, was the only mutation that caused the helix to distort in our peptide model based on the lowest energy conformer predicted by the EmCAST database (Figure S14). Accounting for this distortion reduced our error for the A21P variant in the RNase T1 peptide from 0.31 kcal/mol to 0.01 kcal/mol (Figure 7A, Figure S15A). Calculations for the other A21 variants were not significantly influenced by structural relaxation based on the structural preferences for the sequences from the EmCAST database. Overall stability predictions for RNase T1 peptide were reasonable at pH 7.0 (R^2 = 0.70, Figure 7A) and unreliable for the RNase T1 protein and peptide at pH 2.5 (R^2 = 0.04 and 0.58, respectively, Figure S15). Prediction of stability changes in helices from T4 Lysozyme produced inconsistent correlations at pH 3.0 (R^2 = 0.35 and 0.82, Figure S16). The presence of acidic/basic residues within EmCAST's i±3 interaction window may influence the consistency of correlations at non-neutral pH. The data from barnase in Figure 6 and Figure S12 include a set of A32X variants at a surface position in an α-helix.[61] These data were obtained in the presence of 50 mM MES buffer pH 6.3, conditions near neutral pH that are more favorable for predictions by EmCAST (most proteins are crystallized near pH 7). A slope of 0.9 and a R^2 of 0.59 are obtained with this set of mutations at the A32 helical site (Figure S16).

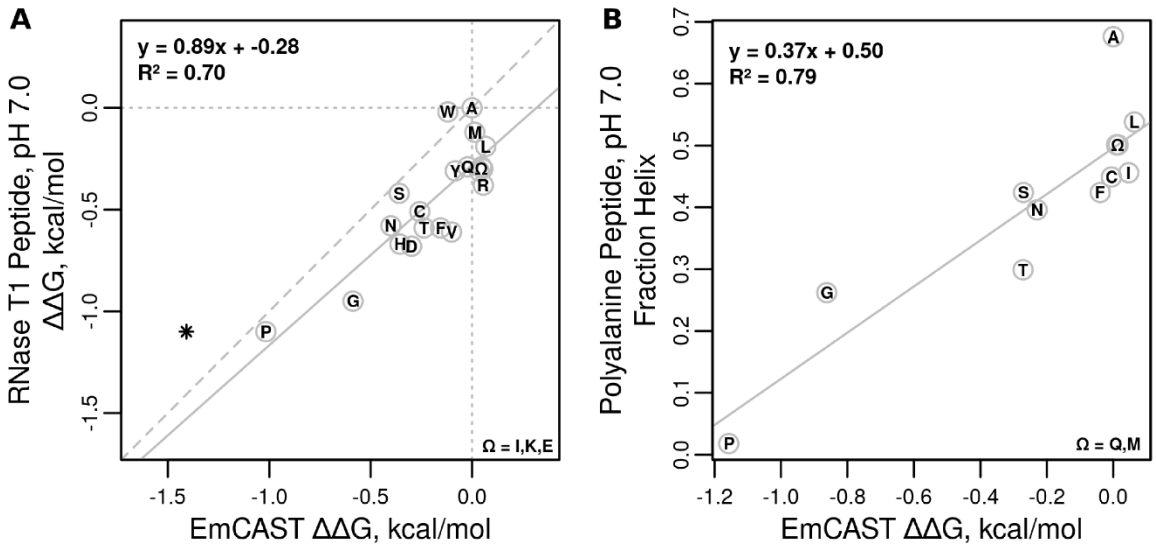

Figure 7. EmCAST stability predictions for helical sites. (A) Correlation plot for RNase T1 Peptide (30 mM MOPS, pH 7.0, 0°C). (B) Correlation plot for polyalanine peptide (1 M NaCl, pH 7.0, 0°C).[60] In panel A, the data points for the Ile, Lys and Glu variants overlap and are marked with the symbol Ω. In panel B, the Gln and Met data points overlap and are designated with the symbol Ω. All energies are relative to Ala, using the lowest energy structure predicted by the EmCAST database. Relaxing the RNase T1 peptide to the lowest energy structure predicted by EmCAST only significantly affected ΔΔG for proline. The ΔΔG for the unrelaxed structure predicted by EmCAST for the proline mutation in panel A is marked with an asterisk.

Information on sequence-to-structure relationships for residues at the i-4 and i+4 positions, which are known to affect helix stability,[65-68] is only retained at the level of sequence wild cards in the database used by EmCAST. Given that sequence information on these medium-range helix interactions is only partially cataloged in the EmCAST database, we tested the ability of EmCAST to predict changes in stability for host-guest studies in a polyalanine helix (Ac-YGG(KAAAA)$_3$K-CONH$_2$)[60] where i,i±4 interactions are minimal. The dataset produces a stronger correlation at pH 7.0 (R^2 = 0.79, Figure 7B) than that observed for RNase T1.

Regarding the correlation between EmCAST predictions and observed changes in stability at helical sites in proteins, we note that there is no obvious relationship between the method of denaturation and the quality of the correlation. The best and worst R^2 values are for the heat denaturation data for T4 lysozyme at positions 131 and 44, respectively. A better correlation is

observed for the urea denaturation data of barnase than for the GdnHCl denaturation data of RNase T1, while we observed excellent correlations for GdnHCl denaturation data for UBA(1). The average error of the EmCAST predictions were 0.12 kcal/mol for the V131X variants of T4 lysozyme, similar to what we observe for UBA(1). For the other datasets, the average error of the EmCAST predictions were larger than observed for the EmCAST predictions for UBA(1) (barnase, 0.26 kcal/mol, S44X variants of T4 lysozyme, 0.27 kcal/mol; RNase T1, 0.35 kcal/mol).

There are multiple structures for WT* (C54T, C97A) lysozyme. We used each of four different structures (1L63, 1.75 Å;[69] 219L, 1.66 Å;[70] 1LW9, 1.45 Å;[71] 5KHZ, 1.49 Å[72]) to predict the changes in stability for the S44X variants of T4 lysozyme using EmCAST. For each of the S44X mutations, the predicted change in stability was essentially independent of which of the four structures was used by EmCAST to predict the change in stability. For each of the 19 mutations, we calculated the average and standard deviation of the four predicted changes in stability. For the S44P mutation, the standard deviation of the predicted change in stability was largest ($\Delta\Delta G$ = -1.251 ± 0.014 kcal/mol, range of the prediction, -1.239 to -1.264 kcal/mol). For all other S44X mutations, the standard deviation of the EmCAST prediction was less than 0.006 kcal/mol. Thus, when multiple structures exist, the choice of the structure does not strongly affect the predicted change in stability if all are high quality structures.

Application to α/β and All-β Folds. The proteins that we evaluated from the ProTherm database were primarily helical or the mutations at surface-exposed sites were primarily at helical positions. To extend the validation of our method, we looked at three additional proteins with extensive mutational datasets and folds with more diverse secondary structure. As with the dataset from the ProTherm database, we limited analysis to sites with SASA of 50% or higher. A

large dataset comprising 18 mutations at surface-exposed sites measured with GdnHCl as denaturant is available for the src SH3 domain. SH3 is primarily composed of β-sheet secondary structure and large loops and the mutations in the dataset occur within both β-sheet and loop structures (Figure S17A).[73] We also analyzed surface mutations from two α/β folds CI2 (18 variants, GdnHCl denaturation, Figure S17B),[74] and NTL9 (13 variants, urea denaturation, Figure S17C).[75] The latter domain provides a test of our ability to predict stabilizing mutations for a protein known to have thermodynamically significant residual structure in the denatured state.[76-80]

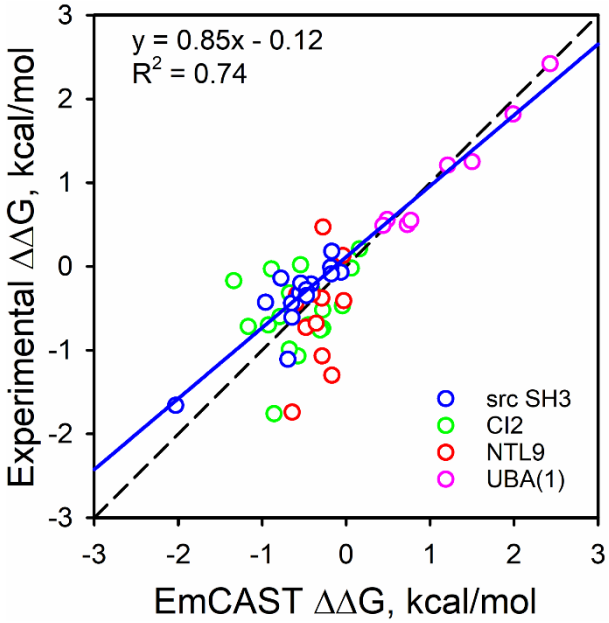

Figure 8. Plots of experimental ΔΔG versus EmCAST predictions for ΔΔG for the src SH3 domain, CI2 and NTL9. The data for UBA(1) are plotted for comparison. The blue solid line is the correlation between the src SH3 domain data and the EmCAST predictions. The equation for this correlation and squared correlation coefficient, R^2, is shown in the upper left corner. The dashed black line is for a prefect correlation.

Figure 8 shows the correlation plots for these three proteins. EmCAST provides excellent predictions for the SH3 data at solvent exposed sites (R^2 of 0.74 and a slope somewhat below 1). The standard error of the prediction is 0.31 kcal/mol. To test the importance of SASA we looked

at data with SASA from 40 – 49%, too. The standard error of the EmCAST prediction rose to 0.84 kcal/mol. The correlation line for the src SH3 data also would predict the ΔΔG values for the UBA(1) domain well. This result indicates that EmCAST can predict changes is stability well for mutations in both β-sheet and long loop structures.

The correlations are considerably poorer for CI2 and NTL9 (Figure S18, R^2 near 0.09 for both). The standard error of the prediction by EmCAST for CI2 is 0.51 kcal/mol and it is 0.56 kcal/mol for NTL9. As with the src SH3 domain, the standard error of EmCAST's predictions was also much larger for variants with SASA from 40 – 49% (1.1 kcal/mol for CI2 and 1.1 kcal/mol for NTL9). Thus, EmCAST's ability to predict stability changes drops off rapidly for residues with SASA less than 50%. For CI2 the two largest outliers are the P25A and P33A mutations with prediction errors of 1.2 and 0.9 kcal/mol, respectively. This result may indicate that EmCAST has difficulty predicting ΔΔG for proline mutations. The I37A mutation in the large loop is also a strong outlier (0.86 kcal/mol prediction error). This large loop is supported by hydrogen bonds to two buried arginines from the central β-sheet (Figure S17B). It is possible that the large errors in the P33A mutation and the I37A mutation result from small changes in the conformation of the loop that affect the hydrogen bonds to the arginines. Long-range interactions will not be predicted well by EmCAST and it will likely be important to look for effects on possible long-range interactions when designing stabilizing mutations. The poor results with NTL9 are not surprising. The stability of NTL9 is significantly affected by interactions in the denatured state.[76-80] EmCAST's predictions are based on stabilizing the native state structure. Similar poor results were obtained with Staphylococcal nuclease, another protein where mutations are known to affect the denatured state (Figure S19).[81]

We also note that the denatured state effects for NTL9 are due to electrostatic interactions. EmCAST predicts stability based on the conformation of the native state. EmCAST does not explicitly account for electrostatic interactions. Because of its i,i±3 window, it may implicitly capture some local electrostatic interactions. NTL9 stability was measure in urea, which unlike GdnHCl does not shield electrostatic interactions.[82] Thus, electrostatic interactions in both the native and the denatured state of NTL9 could be a component of the poor performance of EmCAST with respect to NTL9. Further experiments are necessary to probe EmCAST's ability to implicitly capture the effects of electrostatics when predicting stability changes.

Application of EmCAST to Vaccine Development. Analysis of the immunogenic mutations engineered into the SARS-CoV2 vaccines provides a promising demonstration of EmCAST's utility. A variant of the spike protein with two proline mutations (S-2P) has been previously shown to trap the SARS-CoV spike protein in its pre-fusion conformation and block post-fusion conformations, boosting its efficacy as an immunogen.[83] An analogous variant was used for the vaccines against Sars-CoV2.[84,85] EmCAST predicts the S-2P mutations to be mildly stabilizing to the pre-fusion conformer and severely destabilizing in the post-fusion conformer for both SARS-CoV and SARS-CoV2 spike proteins (Figure S20, Table S7). The net destabilization of the post-fusion conformer is 7.3 and 7.5 kcal/mol relative to the pre-fusion conformer for SARS-CoV and SARS-CoV2 spike proteins, respectively. This shows that the shift in spike protein conformational preference occurs not through stiffening of the pre-fusion conformer, but almost entirely through the destabilization of the post-fusion conformer.[86] A major strength of our method lies in our ability to predict mutations using only the coordinates of backbone Cα atoms. This property of EmCAST means that similar engineering strategies could

be implemented rapidly to generate useful vaccine components for use against emerging pathogens even if only low-resolution structures are available.

DISCUSSION

EmCAST utilizes a simple empirical energy potential that encapsulates all of the complex sequence-local interactions involved in structuring a protein. The work presented demonstrates accurate free energy calculations for UBA(1) mutations involving backbone torsional strain (Y188G), helix capping (E176T), helix dipoles (E176T, H192E), and context-dependent effects (T168R). Although not enumerated by EmCAST, multiple counteracting forces were correctly represented and scored in the E176T mutation. Our method also offers key insights on the nature of protein folding and stability. The four surface mutations engineered into UBA(1) increased its stability from 2.4 to 4.8 kcal/mol. The free energy of folding includes contributions from many weak forces (hydrogen bonds, van der Waals interactions, backbone angle preferences, electrostatic interactions and hydrophobic interactions) that are counterbalanced by chain entropy such that net protein stability is small (5 – 15 kcal/mol).[87] Because of this balance against chain entropy, our results with EmCAST show that it is possible to have a significant effect on net stability by optimizing local interactions. By focusing on surface accessible sites, where the impact of other weak forces is diminished because of exposure to water, EmCAST is able to make very accurate predictions of mutation-induced changes in stability.

Our results show that optimizing the local interactions of surface residues can be as effective and perhaps a more accurate approach than trying to optimize the hydrophobic core.[14,15] EmCAST compares the native state energies between two 7-residue target fragments to predict $\Delta\Delta G$, limiting evaluation of the free energy contributions to the aforementioned native-state

fragments. Accurately isolating these contributions to free energy may both enhance our interpretation of stability changes and aid the development of more accurate residue-residue interaction potentials.

Surface mutations found in several other proteins with diverse folds demonstrated the general applicability of our calculations, albeit with potentially somewhat lower accuracy for destabilizing mutations. Our results with large sets of mutations at the surface-exposed sites in helical proteins shows that even with the ability of EmCAST to include sequence context information out to three residues from the site of mutation, additional sequence context information may be necessary to better reproduce changes in stability even at surface sites. This issue also is evident when the experimental changes in stability at surface-exposed sites are compared between different proteins. The correlation between the ΔΔG values for the two mutation sites in T4 lysozyme is strong[34] (Figure S21A, $R^2 = 0.69$). However, the squared correlation coefficients, R^2, between ΔΔG values at the four surface-exposed sites studied in barnase, RNase T1 and T4 lysozyme range from 0.18 to 0.69 (Figure S21 and Table S8) indicating that there is significant context dependence even at surface-exposed site of helices in proteins. The R^2 values range from 0.35 to 0.82 for the correlations between observed ΔΔG values and EmCAST's predictions for these four surface-exposed sites, better than those between the four experimental datasets (Table S8). This observation indicates that EmCAST captures some, but not all, of the context dependence of ΔΔG at these surface-exposed helical sites.

Our results with CI2 and NTL9 demonstrate some limitations that must be accounted for when using EmCAST to design mutations. CI2 provides an example where long-range interactions may affect changes in stability predicted by EmCAST based on local conformational preferences. NTL9 and related results with Staphylococcal nuclease show that

EmCAST may not predict changes in stability reliably for proteins that have significant residual structure in the denatured state that is affected by mutations.

Finally, analysis of the S-2P mutations in coronavirus spike proteins suggest EmCAST may help researchers isolate specific conformers of proteins for experimental or immunogenic purposes. Given the relationship between structural dynamics and function,[88,89] use of EmCAST to selectively stabilize a particular conformer of a protein could be a useful means to manipulate protein function.

To share our work with a broad range of protein scientists we have designed a fast and easy to use web interface for EmCAST. The online version of our tool (www.emcast.org) is freely available to the research community.

EXPERIMENTAL PROCEDURES

Preparation of Site-directed Mutations. The pGEX-2T(TEV) plasmid containing the UBA(1) gene was used as a template for site-directed mutagenesis.[26] Site-directed mutagenesis was carried out using the QuikChange Lightning PCR-based mutagenesis kit (Agilent). Primers for mutagenesis were obtained from Invitrogen (Table S9). DNA isolated from transformed XL-10 Gold *Escherichia coli* using the QIAprep Spin Miniprep Kit (QIAGEN) was sequenced to confirm mutations (Eurofins Genomics).

Protein Expression and Purification. The pGEX-2T(TEV) plasmid[26] 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.[26] 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.[26] 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.

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.[90] 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[91] 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. 3 was fit to plots of θ_{222} vs. [GdnHCl][92,93] 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. 3, θ_N and m_N are the intercept and slope of the native state baseline, θ_D and m_D

$$\theta_{222} = \frac{(\theta_N + m_N \cdot [GdnHCl]) + (\theta_D + m_D \cdot [GdnHCl]) \cdot e^{\left(\frac{m \cdot [GdnHCl] - \Delta G_u^{o'}(H_2O)}{RT}\right)}}{1 + e^{\left(\frac{m \cdot [GdnHCl] - \Delta G_u^{o'}(H_2O)}{RT}\right)}} \quad \text{(Eq. 3)}$$

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

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 was fit to Chevron plots of the natural log of k_{obs} versus the final GdnHCl concentration to

$$\ln(k_{obs}) = \ln\left(k_f(H_2O) \cdot e^{\left(\frac{-m_{TS-D}\cdot[GdnHCl]}{RT}\right)} + k_u(H_2O) \cdot e^{\left(\frac{m_{TS-N}\cdot[GdnHCl]}{RT}\right)}\right) \qquad \text{(Eq. 4)}$$

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. 5), *m* (Eq. 6) and the Tanford β-value (β_T, Eq. 7) were calculated for each variant.

$$\Delta G_u^{o'}(H_2O) = RT \cdot \ln\left(\frac{k_f(H_2O)}{k_u(H_2O)}\right) \qquad \text{(Eq. 5)}$$

$$m_{eq} = (m_{TS-D} + m_{TS-N}) \qquad \text{(Eq. 6)}$$

$$\beta_T = \frac{m_{TS-D}}{(m_{TS-D}+m_{TS-N})} \qquad \text{(Eq. 7)}$$

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[94] and Aimless.[95] 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[96] with 6W2I (1.10 Å Y188G structure; 6W2G) or 6W2G (E176T/Y188G structure, 7TGP) as the search model. Model building was accomplished in PHENIX[96] and the structures were refined through iterative cycles of manual adjustment in Coot[97] and refinement of atomic positions, real space, occupancy, and thermal parameters in PHENIX.[96] Statistics for the Y188G crystal structures are provided in Tables S3 and S4, and for the E176T/Y188G crystal structure in Table S5. Structures have been deposited in the PDB.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free at http://pubs.acs.org. Supporting Information includes Supporting Experimental Methods that describe the production of EmCAST's fragment database; how EmCAST calculates changes in protein stability; how ensembles of protein conformers are produced and ranked with respect to free energy; the rationale for selecting data from the ProTherm database; how the EmCAST software was optimized and benchmarked; and the procedures used to implement other stability prediction software. Twenty-nine supporting figures which include an example of the effect smoothing a data point in a 2D τ,τ map, 2D τ,τ heatmaps; fragment heatmaps showing the effect of mutations on the match to observed Cα dihedral angles; saturation mutagenesis heatmaps showing the effect of stabilizing mutations on a saturation mutagenesis heatmap; a plot showing how stabilizing mutations predicted by

EmCAST and the MSA approach correlate; a Ramachandran plot showing why the Y188G mutation is favored; folding kinetics Chevron plots for UBA(1) variants; an example of a folding funnel generated by the EmCAST database for the sequence near turn 1 of UBA(1); figure of the most stable turn 2 conformers generated by EmCAST showing why the folding kinetics of turn 2 is less sensitive to mutation; correlation plots for other stability prediction methods; correlation plots for the data in Figure 6 without the UBA(1) data; a figure showing why proline is an outlier for the RNase T1 peptide stability changes if the structure is assumed to be invariant; correlation plots between observed stability changes at α-helical sites in proteins and EmCAST predictions; structures of all β and α/β domains used to test EmCAST showing mutation sites; correlation plots for observed changes in stability for CI2, NTL9 and Staphylococcal nuclease versus EmCAST predictions; structures of Sars-CoV2 spike protein showing sites where proline mutations stabilize the pre-fusion conformer relative to the post-fusion conformer; correlation plots for observed stability changes in α-helices from different proteins; reaction coordinate showing expected effects of EmCAST predictions on the free energy of the denatured state, transition state and native state; structure of staphylococcal nuclease illustrating the site of the Trp used to measure protein unfolding; structures showing sites of mutations and structural probes for proteins selected from the ProTherm database and associated correlation plots between observed changes in stability and EmCAST predictions (proteins include the FF domain, barnase, B-domain of staphylococcal protein A). Eleven supporting tables with information on EmCAST runtime benchmarks; X-ray crystallography data collection and refinement statistics for PDB deposits 6W2G, 6W2I and 7TGP; statistical errors of prediction of mutations from the ProTherm database; EmCAST predictions of stability changes caused by proline mutations for the pre- and post-fusion conformer of the Sars-CoV and Sars-CoV2 spike

protein; correlation coefficients for observed stability changes cause by mutations in α-helices of barnase, T4 lysozyme and RNase T1; primers used to prepare UBA(1) variants; discrepancies in stabilities of staphylococcal nuclease variants as measured by circular dichroism versus fluorescence spectroscopy; and the effects of salt concentration on the relative stability of WT and a T16R variant of barnase.

AUTHOR INFORMATION

Corresponding Author

*Bruce E. Bowler - Department of Chemistry and Biochemistry and Center for Biomolecular Structure and Dynamics, University of Montana, Missoula, Montana 59812, United States; orcid.org/0000-0003-1543-2466; Phone: (406) 243-6114; Email: bruce.bowler@umontana.edu; Fax: (406) 243-4227.

Authors

Michael T. Rothfuss - Department of Chemistry and Biochemistry University of Montana, Missoula, Montana 59812, United States; orcid.org/0000-0002-3664-6300.

Dustin C. Becht - Department of Chemistry and Biochemistry University of Montana, Missoula, Montana 59812, United States; orcid.org/0000-0003-3127-3418.

Baisen Zeng - Center for Biomolecular Structure and Dynamics, University of Montana, Missoula, Montana 59812, United States; orcid.org/0000-0002-7780-3501.

Levi. J. McClelland – Division of Biological Sciences and Center for Biomolecular Structure and Dynamics, University of Montana, Missoula, Montana 59812, United States; orcid.org/0000-0002-0868-8925.

Cindee Yates-Hansen - Center for Biomolecular Structure and Dynamics, University of Montana, Missoula, Montana 59812, United States; orcid.org/0000-0001-8116-2586.

Notes

The authors declare no competing interest.

ACKNOWLEDGMENT

This work was supported by grants from the NSF [CHE-1904895 (B.E.B)] and the NIGMS [GM148610 (B.E.B)]. The facilities of the Integrated Structural Biology Core Facility at the University of Montana, which is supported by a Centers of Biomedical Research Excellence award from the National Institute of General Medical Sciences [P30GM140963 (B.E.B)] were used for crystallization and structure determination.

REFERENCES

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

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

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

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

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

(6) Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H., Protein stability promotes evolvability, *Proc. Natl. Acad. Sci. U.S.A.* **2006,** *103*, 5869-5874. https://doi.org/10.1073/pnas.0510098103.

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

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

(9) Potapov, V.; Cohen, M.; Schreiber, G., Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details, *Protein Eng. Des. Sel.* **2009,** *22*, 553-560. https://doi.org/10.1093/protein/gzp030.

(10) Marabotti, A.; Del Prete, E.; Scafuri, B.; Facchiano, A., Performance of web tools for predicting changes in protein stability caused by mutations, *BMC Bioinf.* **2021,** *22*, 345. https://doi.org/10.1186/s12859-021-04238-w.

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

(12) Broom, A.; Trainor, K.; Jacobi, Z.; Meiering, E. M., Computational modeling of protein stability: quantitative analysis reveals solutions to pervasive problems, *Structure* **2020,** *28*, 717-726.e3. https://doi.org/10.1016/j.str.2020.04.003.

(13) Buß, O.; Rudat, J.; Ochsenreither, K., FoldX as protein engineering tool: better than random based approaches?, *Comput. Struct. Biotechnol. J.* **2018,** *16*, 25-33. https://doi.org/10.1016/j.csbj.2018.01.002.

(14) Kantaev, R.; Riven, I.; Goldenzweig, A.; Barak, Y.; Dym, O.; Peleg, Y.; Albeck, S.; Fleishman, S. J.; Haran, G., Manipulating the folding landscape of a multidomain protein, *J. Phys. Chem. B* **2018,** *122*, 11030-11038. https://doi.org/10.1021/acs.jpcb.8b04834.

(15) Yang, J.; Naik, N.; Patel, J. S.; Wylie, C. S.; Gu, W.; Huang, J.; Ytreberg, F. M.; Naik, M. T.; Weinreich, D. M.; Rubenstein, B. M., Predicting the viability of beta-lactamase: how folding and binding free energies correlate with beta-lactamase fitness, *PLoS One* **2020,** *15*, e0233509. https://doi.org/10.1371/journal.pone.0233509.

(16) Loladze, V. V.; Ibarra-Molero, B.; Sanchez-Ruiz, J. M.; Makhatadze, G. I., Engineering a thermostable protein via optimization of charge−charge interactions on the protein surface, *Biochemistry* **1999,** *38*, 16419-16423. https://doi.org/10.1021/bi992271w.

(17) Strickler, S. S.; Gribenko, A. V.; Gribenko, A. V.; Keiffer, T. R.; Tomlinson, J.; Reihle, T.; Loladze, V. V.; Makhatadze, G. I., Protein stability and surface electrostatics: a charged relationship, *Biochemistry* **2006,** *45*, 2761-2766. https://doi.org/10.1021/bi0600143.

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

(19) Sternke, M.; Tripp, K. W.; Barrick, D., Consensus sequence design as a general strategy to create hyperstable, biologically active proteins, *Proc. Natl. Acad. Sci. U.S.A.* **2019,** *116*, 11275-11284. https://doi.org/10.1073/pnas.1816707116.

(20) Tripp, K. W.; Sternke, M.; Majumdar, A.; Barrick, D., Creating a homeodomain with high stability and DNA binding affinity by sequence averaging, *J. Am. Chem. Soc.* **2017,** *139*, 5051-5060. https://doi.org/10.1021/jacs.6b11323.

(21) Sternke, M.; Tripp, K. W.; Barrick, D., Surface residues and nonadditive interactions stabilize a consensus homeodomain protein, *Biophys. J.* **2021,** *120*, 5267-5278. https://doi.org/10.1016/j.bpj.2021.10.035.

(22) Alber, T.; Sun, D. P.; Nye, J. A.; Muchmore, D. C.; Matthews, B. W., Temperature-sensitive mutations of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein, *Biochemistry* **1987,** *26*, 3754-3758. https://doi.org/10.1021/bi00387a002.

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

(24) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The protein data bank, *Nucleic Acids Res.* **2000,** *28*, 235-242. https://doi.org/10.1093/nar/28.1.235.

(25) Mueller, T. D.; Feigon, J., Solution structures of UBA domains reveal a conserved hydrophobic surface for protein-protein interactions, *J. Mol. Biol.* **2002,** *319*, 1243-1255. https://doi.org/10.1016/S0022-2836(02)00302-9.

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

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

(28) Mackenzie, C. O.; Grigoryan, G., Protein structural motifs in prediction and design, *Curr. Opin. Struct. Biol.* **2017,** *44*, 161-167. https://doi.org/10.1016/j.sbi.2017.03.012.

(29) Bystroff, C.; Baker, D., Prediction of local structure in proteins using a library of sequence-structure motifs, *J. Mol. Biol.* **1998,** *281*, 565-577. https://doi.org/10.1006/jmbi.1998.1943.

(30) Schwarzinger, S.; Kroon, G. J. A.; Foss, T. R.; Chung, J.; Wright, P. E.; Dyson, H. J., Sequence-dependent correction of random coil NMR chemical shifts, *J. Am. Chem. Soc.* **2001,** *123*, 2970-2978. https://doi.org/10.1021/ja003760i.

(31) Pappu, R. V.; Srinivasan, R.; Rose, G. D., The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding, *Proc. Natl. Acad. Sci. U.S.A.* **2000,** *97*, 12565-12570. https://doi.org/10.1073/pnas.97.23.12565.

(32) Oldfield, T. J.; Hubbard, R. E., Analysis of Cα geometry in protein structures, *Proteins: Struct. Funct. Genet.* **1994,** *18*, 324-337. https://doi.org/10.1002/prot.340180404.

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

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

(35) Jemth, P.; Day, R.; Gianni, S.; Khan, F.; Allen, M.; Daggett, V.; Fersht, A. R., The structure of the major transition state for folding of an FF domain from experiment and simulation, *J. Mol. Biol.* **2005,** *350*, 363–378. https://doi.org/10.1016/j.jmb.2005.04.067.

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

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

(38) Aurora, R.; Rose, G. D., Helix capping, *Protein Sci.* **1998,** *7*, 21-38. https://doi.org/10.1002/pro.5560070103.

(39) Serrano, L.; Sancho, J.; Hirshberg, M.; Fersht, A. R., 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, *J. Mol. Biol.* **1992,** *227*, 544-559. https://doi.org/10.1016/0022-2836(92)90906-Z.

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

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

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

(43) Wolynes, P. G., Energy landscapes and solved protein-folding problems, *Philos. Trans. R. Soc. London, Ser. A* **2005,** *363*, 453-467. https://doi.org/10.1098/rsta.2004.1502.

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

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

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

(47) Kellogg, E. H.; Leaver-Fay, A.; Baker, D., Role of conformational sampling in computing mutation-induced changes in protein structure and stability, *Proteins: Struct., Funct., Bioinf.* **2011,** *79*, 830-838. https://doi.org/10.1002/prot.22921.

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

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

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

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

(52) Sato, S.; Fersht, A. R., Searching for multiple folding pathways of a nearly symmetrical domain: temperature dependent ϕ-value analysis of the B domain of protein A, *J. Mol. Biol.* **2007,** *372*, 254-267. https://doi.org/10.1016/j.jmb.2007.06.043.

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

(54) Loewenthal, R.; Sancho, J.; Fersht, A. R., Histidine-aromatic interactions in barnase: Elevation of histidine pK_a and contribution to protein stability, *J. Mol. Biol.* **1992,** *224*, 759-770. https://doi.org/10.1016/0022-2836(92)90560-7.

(55) Horovitz, A.; Fersht, A. R., Co-operative interactions during protein folding, *J. Mol. Biol.* **1992,** *224*, 733-740. https://doi.org/10.1016/0022-2836(92)90557-Z.

(56) Serrano, L.; Bycroft, M.; Fersht, A. R., Aromatic-aromatic interactions and protein stability: Investigation by double-mutant cycles, *J. Mol. Biol.* **1991,** *218*, 465-475. https://doi.org/10.1016/0022-2836(91)90725-L.

(57) Horovitz, A.; Serrano, L.; Fersht, A. R., COSMIC analysis of the major α-helix of barnase during folding, *J. Mol. Biol.* **1991,** *219*, 5-9. https://doi.org/10.1016/0022-2836(91)90852-W.

(58) Rohl, C. A.; Chakrabartty, A.; Baldwin, R. L., Helix propagation and N-cap propensities of the amino acids measured in alanine-based peptides in 40 volume percent trifluoroethanol, *Protein Sci.* **1996,** *5*, 2623-2637. https://doi.org/10.1002/pro.5560051225.

(59) Baldwin, R. L., α-Helix formation by peptides of defined sequence, *Biophys. Chem.* **1995,** *55*, 127-135. https://doi.org/10.1016/0301-4622(94)00146-B.

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

(61) Horovitz, A.; Matthews, J. M.; Fersht, A. R., α-Helix stability in proteins: II. Factors that influence stability at an internal position, *J. Mol. Biol.* **1992,** *227*, 560-568. https://doi.org/10.1016/0022-2836(92)90907-2.

(62) Myers, J. K.; Pace, C. N.; Scholtz, J. M., A direct comparison of helix propensity in proteins and peptides, *Proc. Natl. Acad. Sci. U.S.A.* **1997,** *94*, 2833-2837. https://doi.org/10.1073/pnas.94.7.283.

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

(64) Blaber, M.; Zhang, X.-j.; Matthews, B. W., Structural basis of amino acid α helix propensity, *Science* **1993,** *260*, 1637-1640. https://doi.org/10.1126/science.8503008.

(65) Marqusee, S.; Baldwin, R. L., Helix stabilization by Glu-...Lys+ salt bridges in short peptides of *de novo* design, *Proc. Natl. Acad. Sci. U.S.A.* **1987,** *84*, 8898-8902. https://doi.org/10.1073/pnas.84.24.8898.

(66) Armstrong, K. M.; Fairman, R.; Baldwin, R. L., The (*i, i + 4*) Phe-His interaction studied in an alanine-based α-helix, *J. Mol. Biol.* **1993,** *230*, 284-291. https://doi.org/10.1006/jmbi.1993.1142.

(67) Huyghues-Despointes, B. M. P.; Martin Scholtz, J.; Baldwin, R. L., Helical peptides with three pairs of Asp-Arg and Glu-Arg residues in different orientations and spacings, *Protein Sci.* **1993,** *2*, 80-85. https://doi.org/10.1002/pro.5560020108.

(68) Padmanabhan, S.; Baldwin, R. L., Helix-stabilizing interaction between tyrosine and leucine or valine when the spacing is i,i + 4, *J. Mol. Biol.* **1994,** *241*, 706-713. https://doi.org/10.1006/jmbi.1994.1545.

(69) Nicholson, H.; Anderson, D. E.; Dao Pin, S.; Matthews, B. W., Analysis of the interaction between charged side chains and the α-helix dipole using designed thermostable mutants of phage T4 lysozyme, *Biochemistry* **1991,** *30*, 9816-9828. https://doi.org/10.1021/bi00105a002.

(70) Vetter, I. R.; Baase, W. A.; Heinz, D. W.; Xiong, J.-P.; Snow, S.; Matthews, B. W., Protein structural plasticity exemplified by insertion and deletion mutants in T4 lysozyme, *Protein Sci.* **1996,** *5*, 2399-2415. https://doi.org/10.1002/pro.5560051203.

(71) Gassner, N. C.; Baase, W. A.; Mooers, B. H. M.; Busam, R. D.; Weaver, L. H.; Lindstrom, J. D.; Quillin, M. L.; Matthews, B. W., Multiple methionine substitutions are tolerated in T4 lysozyme and have coupled effects on folding and stability, *Biophys. Chem.* **2002,** *100*, 325-340. https://doi.org/10.1016/S0301-4622(02)00290-9.

(72) Scholfield, M. R.; Ford, M. C.; Carlsson, A.-C. C.; Butta, H.; Mehl, R. A.; Ho, P. S., Structure–energy relationships of halogen bonds in proteins, *Biochemistry* **2017,** *56*, 2794-2802. https://doi.org/10.1021/acs.biochem.7b00022.

(73) Riddle, D. S.; Grantcharova, V. P.; Santiago, J. V.; Alm, E.; Ruczinski, I.; Baker, D., Experiment and theory highlight role of native state topology in SH3 folding, *Nat. Struct. Biol.* **1999,** *6*, 1016-1024. https://doi.org/10.1038/14901.

(74) Itzhaki, L. S.; Otzen, D. E.; Fersht, A. R., 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, *J. Mol. Biol.* **1995,** *254*, 260-288. https://doi.org/10.1006/jmbi.1995.0616.

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

(76) Kuhlman, B.; Luisi, D. L.; Young, P.; Raleigh, D. P., pK_a 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* **1999,** *38*, 4896-4903. https://doi.org/10.1021/bi982931h.

(77) Cho, J.-H.; Raleigh, D. P., 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, *J. Mol. Biol.* **2006,** *359*, 1437-1446. https://doi.org/10.1016/j.jmb.2006.04.038.

(78) Cho, J.-H.; Sato, S.; Raleigh, D. P., 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, *J. Mol. Biol.* **2004,** *338*, 827-837. https://doi.org/10.1016/j.jmb.2004.02.073.

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

(80) Cho, J.-H.; Meng, W.; Sato, S.; Kim, E. Y.; Schindelin, H.; Raleigh, D. P., Energetically significant networks of coupled interactions within an unfolded protein, *Proc. Natl. Acad. Sci. U.S.A.* **2014,** *111*, 12079-12084. https://doi.org/10.1073/pnas.1402054111.

(81) Shortle, D., Staphylococcal nuclease: a showcase of *m*-value effects, *Adv. Protein Chem.* **1995,** *46*, 217-247. https://doi.org/10.1016/S0065-3233(08)60336-8.

(82) Monera, O. D.; Kay, C. M.; Hodges, R. S., Protein denaturation with guanidine hydrochloride or urea provides a different estimate of stability depending on the contributions of electrostatic interactions, *Protein Sci.* **1994,** *3*, 1984-1991. https://doi.org/10.1002/pro.5560031110.

(83) Pallesen, J.; Wang, N.; Corbett, K. S.; Wrapp, D.; Kirchdoerfer, R., N.; Turner, H., L.; Cottrell, C., A.; Becker, M., M.; Wang, L.; Shi, W.; Kong, W.-P.; Andres, E., L.; Kettenbach, A., N.; Denison, M., R.; Chappell, J., D.; Graham, B., S.; Ward, A., B.; McLellan, J., S., Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen, *Proc. Natl. Acad. Sci. U.S.A.* **2017,** *114*, E7348-E7357. https://doi.org/10.1073/pnas.1707304114.

(84) Wrapp, D.; Wang, N.; Corbett, K. S.; Goldsmith, J. A.; Hsieh, C.-L.; Abiona, O.; Graham, B. S.; McLellan, J. S., Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation, *Science* **2020,** *367*, 1260-1263. https://doi.org/10.1126/science.abb2507.

(85) Hsieh, C.-L.; Goldsmith, J. A.; Schaub, J. M.; DiVenere, A. M.; Kuo, H.-C.; Javanmardi, K.; Le, K. C.; Wrapp, D.; Lee, A. G.; Liu, Y.; Chou, C.-W.; Byrne, P. O.; Hjorth, C. K.; Johnson, N. V.; Ludes-Meyers, J.; Nguyen, A. W.; Park, J.; Wang, N.; Amengor, D.; Lavinder, J. J.; Ippolito, G. C.; Maynard, J. A.; Finkelstein, I. J.; McLellan, J. S., Structure-based design of prefusion-stabilized SARS-CoV-2 spikes, *Science* **2020,** *369*, 1501-1505. https://doi.org/10.1126/science.abd0826.

(86) Byrne, P. O.; McLellan, J. S., Principles and practical applications of structure-based vaccine design, *Curr. Opin. Immunol.* **2022,** *77*, 102209. https://doi.org/10.1016/j.coi.2022.102209.

(87) Dill, K. A.; MacCallum, J. L., The Protein-folding problem, 50 years on, *Science* **2012,** *338*, 1042-1046. https://doi.org/10.1126/science.1219021.

(88) Henzler-Wildman, K.; Kern, D., Dynamic personalities of proteins, *Nature* **2007,** *450*, 964-972. https://doi.org/10.1038/nature06522.

(89) Fraser, J. S.; Clarkson, M. W.; Degnan, S. C.; Erion, R.; Kern, D.; Alber, T., Hidden alternative structures of proline isomerase essential for catalysis, *Nature* **2009,** *462*, 669-673. https://doi.org/10.1038/nature08615.

(90) Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M. R.; Appel, R. D.; Bairoch, A., Protein identification and analysis tools on the ExPASy server. In *The Proteomics Protocols Handbook*, Walker, J. M., Ed. Humana Press: Totowa, NJ, 2005; pp 571-607.

(91) Nozaki, Y., [3] The preparation of guanidine hydrochloride, *Methods Enzymol.* **1972,** *26*, 43-50. https://doi.org/10.1016/S0076-6879(72)26005-0.

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

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

(94) Kabsch, W., Integration, scaling, space-group assignment and post-refinement, *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010,** *66*, 133-144. https://doi.org/10.1107/S0907444909047374.

(95) Evans, P. R.; Murshudov, G. N., How good are my data and what is the resolution?, *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2013,** *69*, 1204-1214. https://doi.org/10.1107/S0907444913000061.

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

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

TOC graphic: for table of contents use only.