Experimental and Computational Analysis of Protein Stabilization by Gly-to-D-Ala Substitution: A Convolution of Native State and Unfolded State Effects

Junjie Zou[1], Benben Song[1], Carlos Simmerling[1,2,*], Daniel Raleigh[1,*]

[1]Department of Chemistry and [2]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794-3400

* Authors to whom correspondence should be addressed: DPR email: daniel.raleigh@stonybrook.edu, phone: (631)632-9547; CS email: carlos.simmerling@stonybrook.edu, phone: (631)632-5424.

Keywords: Protein Stability, Protein Design, Unfolded State, Protein Folding, Thermodynamic Integration.

Running Title: Mirror Imaging Protein Design

Abstract

The rational and predictable enhancement of protein stability is an important goal in protein design. Most efforts target the folded state, however stability is the free energy difference between the folded and unfolded states thus both are suitable targets. Strategies directed at the unfolded state usually seek to decrease chain entropy by introducing cross-links or by replacing glycines. Cross-linking has led to mixed results. Replacement of glycine with an L-amino acid, while reducing the entropy of the unfolded state, can introduce unfavorable steric interactions in the folded state, since glycine is often found in conformations that require a positive φ angle such as helical C–capping motifs or type I' and II" β-turns. L-amino acids are strongly disfavored in these conformations, but D-amino acids are not. However, there are few reported examples and conflicting results have been obtained when glycines are replaced with D-Ala. We critically examine the effect of Gly-to-D-Ala substitutions on protein stability using experimental approaches together with molecular dynamics simulations and free energy calculations. The data, together with a survey of high resolution structures, show that the vast majority of proteins can be stabilized by substitution of C-capping glycines with D-Ala. Sites suitable for substitutions can be identified via sequence alignment with a high degree of success. Steric clashes in the native state due to the new sidechain are rarely observed, but are likely responsible for the destabilizing or null effect observed for the small subset of Gly-to-D-Ala substitutions which are not stabilizing. Changes in backbone solvation play less of a role. Favorable candidates for D-Ala substitution can be identified using a rapid algorithm based on molecular mechanics.

Introduction

A primary goal of protein design is to improve the stability of proteins since marginal stability can lead to loss of function, difficulty in formulating protein based pharmaceuticals, increased aggregation and degradation [1-5]. Small stable proteins are of interest as alternative scaffolds for presenting sequences in a defined structural context and as alternatives to antibodies for drug delivery, for targeting and as analytical tools [6-7]. Stabilizing small domains can be a challenge especially if the number of sites which can be targeted is limited by the need to preserve a subset of sites for functional reasons. Stability is dictated by the free energy difference between the unfolded state and the folded state. In order to increase the free energy difference, and thus improve stability, one can stabilize the folded state or destabilize the unfolded state, however the vast majority of approaches to rational design seek to manipulate folded state energetics by exploiting the known three-dimensional structure of the folded state [8-12]. The unfolded state is a dynamic ensemble, containing transient as well as longer lived elements of structure that can include both native and non-native interactions. The dynamic nature of the unfolded ensemble has made it difficult to target using rational design. Here we describe a general approach to rational protein design that exploits structurally conserved glycine residues and targets both the unfolded ensemble and the native state.

Folded state stabilization usually involves decreasing native state enthalpy, while unfolded state destabilization usually seeks to decrease its entropy. Increasing stability by decreasing the enthalpy of the folded state is more broadly studied, however, implementation of this strategy requires detailed structural information on the folded state [9, 11]. A decrease in the conformational entropy of unfolded states can be achieved by adding disulfide bonds or substituting glycine with non-glycine amino acids [8, 10, 12-17]. The former approach also requires tertiary structural information of the folded state, since disulfide bonds can introduce strain into the native state and have strict stereochemical requirements. In theory, the effect of adding a disulfide can be estimated using arguments based on loop entropy; the disulfide introduces a cross link in the chain and thereby reduces the configurational entropy of the unfolded state. However,

introduction of a disulfide can stabilize compact conformations in the unfolded state and lead to new unfolded state enthalpic interactions. These effects, together with native state strain, often result in engineered disulfides having only a modest or even unfavorable effect on protein stability [10, 18]. Complete cyclization of a protein by covalently linking the N and C termini has been employed in an attempt to enhance protein stability, but the same considerations come into play [19].

Targeting glycine residues is an attractive alternative strategy since introduction of a sidechain is a simple and effective way to decrease configurational entropy owing to the more restricted allowed region of the Ramachandran plot for an L or D amino acid relative to glycine. The approach should be effective provided that the addition of a sidechain does not lead to steric clashes in the folded state and provided the stereochemical constraints introduced by the sidechain are compatible with the native backbone geometry. The latter point is a significant issue since glycine is often located at sites which require a positive value of the backbone dihedral angle φ [20]. D-amino acids are the more attractive choice when targeting glycine residues that have positive values of φ, since these conformations are disfavored for L-amino acids, but allowed for D-amino acids [21-23]. Glycine residue with positive values of φ are commonly found in α-helical C-capping motifs and in type I' and II" β-turns, where a left-handed conformation (positive φ) is required [22, 24-26]. These glycines can often be identified using multiple sequence alignments since they are conserved for structural reasons; helical capping motifs have specific sequence requirements and there are well established sequence rules for type I' and II" β-turns [21-24, 26-27]. Glycines located at C-caps are often solvent exposed, thus any perturbation caused by substituting with a D-amino acid should be minimal since the new side chain is less likely to make steric clashes. This potentially opens the door to rational design in the absence of structural information, however conflicting results have been reported for D-Ala substitutions.

The effect of Gly-to-D-Ala substitutions has been reported for four different proteins: the N-terminal domain of the ribosomal protein L9 (NTL9), the C-terminal Ubiquitin associated domain of HHR23A (UBA), the mini-protein construct TC5b (Trp-cage) and human erythrocytic ubiquitin (ubiquitin) [12, 17, 28].

D-amino acids have also been used to stabilize small β-hairpin peptides [25]. The limited experimental measurements reveal several apparent contradictions: To first order, the entropic stabilization caused by Gly-to-D-Ala substitution is expected to be system independent, but not all proteins are stabilized by Gly-to-D-Ala substitutions and a significant range of ΔΔG˚ values have been reported for those that are. The stability of NTL9 and UBA are increased by a favorable 1.87 kcal/mol and 0.6 kcal/mol respectively when a C-capping Gly was replaced with D-Ala [12]. Note, in this manuscript, we report ΔG˚ values of unfolding, thus positive values of ΔΔG˚ indicate stabilization. The stability of Trp-cage was improved by 0.9 kcal/mol when G10 was substituted by D-Ala [17]. However, a G35D-Ala substitution at a helical C-capping position in ubiquitin was slightly destabilizing at pH=2.5 [28]. The lack of an effect was conjectured to be due to unfavorable contributions from backbone desolvation, caused by the introduction of a sidechain, that offset the decreased entropy of the unfolded state [28].

The limited data set indicates that replacement of glycines with positive φ-angles by D-Ala can be stabilizing, but it also leads to important questions: will the trend of an increase in stability be preserved if larger data sets are examined? What causes the range of values of $\Delta\Delta G^o$? Why does the replacement lead to no effect in ubiquitin? Can the energetic effects of a D-Ala substitution be quantitatively predicted? From a practical perspective, the key issues are whether or not it is possible to reliably and robustly predict*, a priori* which Gly to D-Ala replacements will be stabilizing, and by how much. This is critical since D-amino-acids must currently be introduced via solid phase synthesis or via chemical ligation methods.

In this study, we use a combined experimental and computational approach to systematically examine the consequences of replacing C-capping glycines with D-amino acids and develop a rapid algorithm for predicting when such substitutions will be stabilizing. Gly-to-D-Ala substitutions at the C-caps of α-helices in four additional proteins were examined, doubling the number of reported examples: the engrailed homeodomain (EH), the GA albumin-binding module (GA), the peripheral subunit-binding domain (PSBD) and the chicken villin subdomain (HP35) [29-32]. These proteins are all α-folds and each

contains a glycine C-capping residue with a positive φ angle (**Figure 1**). EH, GA and PSBD were randomly chosen and D-Ala replacements were found to be stabilizing. The small helical protein HP35 was predicted to be destabilized by Gly-to-D-Ala substitutions based on molecular modelling and serves as a negative control. Computational modelling successfully reproduced the experimental stability changes and indicates that intra-molecular van der Waals interactions in the folded state are the reason for the wide range of $\Delta\Delta G^{o}$ caused by Gly-to-D-Ala substitutions. Screening a database of representative high-resolution X-ray structures shows that 95% of C-capping Gly-to-D-Ala substitutions are predicted to be stabilizing and 80% of all substitutions are predicted to enhance stability by more than 1 kT. This work shows that Gly-to-D-Ala substitutions at C-caps of α-helices, under the guidance of molecular modelling, is a general strategy for rational protein design. This works reveals the rules for stabilizing proteins via D-Ala substitutions. This "mirror image" approach to protein design is widely applicable and sites suitable for substitution can be rapidly predicted.

Results

Proteins are usually stabilized by Gly-to-D-Ala substitution.

Published results on a limited set of proteins indicate a range of effects for Gly-to-D-Ala substitution at C-capping sites. However, the number of systems tested to date is too small to draw general conclusions. In order to gain better insight into the consequences of Gly-to-D-Ala substitutions at C-capping sites, Gly-to-D-Ala substitutions were examined in another four proteins (EH, GA, PSBD and HP35). All of these domains have been shown to fold reversibly in a 2-state fashion [29, 33-35]. Like NTL9, UBA and Ubiquitin, these proteins all have a C-capping glycine that is solvent exposed as judged by standard accessible surface area algorithms (**Figure 1**). The φ/ψ angles and the solvent accessibility of all of the glycine sites studied are provided in the supporting information (**Table S1)**. Thermal and denaturant induced unfolding curves of EH, GA, HP35, PSBD display sigmoidal transitions and all can be fit by standard methods to extract unfolding free energies (**Table 1**, **Figures S1** and **S2)**. The stability of EH G39D-Ala, GA G16D-Ala and PSBD G15D-Ala are 0.64 kcal/mol, 0.81 kcal/mol and 1.25 kcal/mol higher than the respective

wild-type. HP35 G11D-Ala is 0.38 kcal/mol less stable than wild-type HP35, but HP35 was intentionally selected as a negative control using the computational approach described below. The experimental measurements on these four additional proteins, especially the inclusion of an additional example (HP35) in which D-Ala substitution is destabilizing, provide a more robust test set for the computational studies described in the next several paragraphs.

Five of the six proteins which were randomly chosen without computational guidance exhibit enhanced stability when a C-capping Gly is replaced by D-Ala, suggesting that Gly-to-D-Ala substitutions at C-capping sites are likely to improve protein stability. Left unanswered are the questions why there is a significant range of $\Delta\Delta G^\circ$ values and why are HP35 and ubiquitin destabilized?

Table 1 Thermodynamic properties of EH, GA, HP35, PSBD and their D-Ala variants.

Protein	ΔG° of unfolding at 25 °C (kcal/mol)	m (kcal/mol M^{-1})	Tm (°C)	ΔH°(Tm) (kcal/mol)
EH	1.91 ± 0.03[1]	0.61 ± 0.01	55.6 ± 0.18	32.5 ± 0.74
EH G39D-Ala	2.55 ± 0.13[1]	0.66 ± 0.03	60.7 ± 0.38	33.1 ± 1.39
GA	4.71 ± 0.16[2]	1.00 ± 0.03	ND	ND
GA G16D-Ala	5.52 ± 0.19[2]	1.02 ± 0.04	ND	ND
PSBD	2.75 ± 0.07[1]	0.67 ± 0.01	52.5 ± 0.14	29.6 ± 0.51
PSBD G15D-Ala	4.00 ± 0.34[1]	0.73 ± 0.07	61.3 ± 0.23	31.9 ± 0.84
HP35	2.47 ± 0.12[1]	0.38 ± 0.03	76.1 ± 1.78	23.8 ± 0.57
HP35 G11D-Ala	2.08 ± 0.13[1]	0.45 ± 0.03	61.2 ± 1.34	21.7± 1.33

(1) Determined by urea denaturation; (2) Determined by GdnHCl denaturation; ND: Not determined. Uncertainties represent the standard error of the fit.

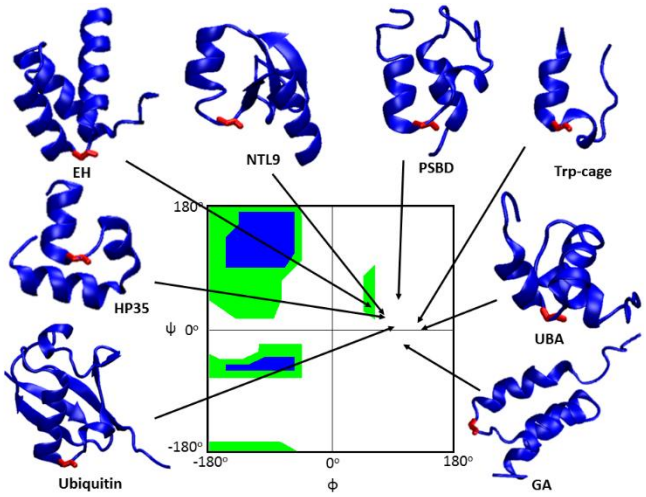


Figure 1. Ribbon representation of the proteins studied with the C-capping Gly colored red. φ/ψ angles of the C-capping glycines are indicated by arrows. The Ramachandran plot is colored green for broadly allowed and blue for most favored regions for L-amino acids, which is adopted from Ramaplot in VMD[36]. The Ramachandran plot for a D-amino acid is the mirror image about the central point (φ = 0° and ψ = 0°) of the plot shown above.

Gly-to-D-Ala substitutions can modulate ΔΔG° via other interactions in addition to entropic stabilization.

Recent computational work reported that Gly-to-L-Ala substitution entropically destabilizes the unfolded state by -TΔS = 0.3 kcal/mol when the unfolded states are modeled as tri and pentapeptides [37] , while earlier work provide estimates ranging from 0.05 to 0.72 kcal/mol [38-41]. The wide range of experimental unfolding free energy changes (0.39 kcal/mol destabilizing to 1.87 kcal/mol stabilizing) argues that interactions beyond entropic destabilization of the unfolded state play an important role in determining the change. A range of effects could counteract or supplement the entropic stabilization of replacing a C-capping Gly. Introduction of a sidechain at a C-capping Gly site can lead to increased desolvation of the polypeptide backbone, a process which is energetically unfavorable [28]. All else being equal, desolvation in the native state will destabilize a protein. However, desolvation of the backbone in the folded state is likely compensated by desolvation of the backbone in the unfolded state. Moreover, the desolvation penalty may also be compensated by new favorable intramolecular interactions such as buried hydrogen bonds or favorable van der Waals interactions. Desolvation of the backbone is thus unlikely to be the sole

reason for the wide range of experimental $\Delta\Delta G^{\circ}$ values. On the other hand, unfavorable van der Waals interactions, such as steric clashes between D-Ala and other residues in the folded state can offset the decrease of entropy in the unfolded states. These new folded state interactions will usually be alleviated upon unfolding and are less likely to perturb the unfolded state. We hypothesized that a significant contribution to the difference in $\Delta\Delta G^{\circ}$ values reflects differences in van der Waals interactions between the C-capping Gly/D-Ala and the rest of the protein in the folded state.

In order to test our hypothesis, molecular dynamics simulations (MD) of wild-type proteins and their D-Ala variants were conducted using the Amber ff14SB force field. MD simulations were also conducted for simplified unfolded state models to account for local unfolded state effects. Per-residue energy decomposition provided an estimate of the intramolecular van der Waals energy (E_{vdw}) contributed by C-capping Gly/D-Ala to the total potential energy of the protein. New unfavorable intramolecular van der Waals interactions in the folded state caused by the D-Ala sidechain lead to a negative value of $\Delta\Delta E_{vdw}$, while new favorable intramolecular van der Waals interactions in the folded state lead to a positive $\Delta\Delta E_{vdw}$ value. A good correlation between $\Delta\Delta E_{vdw}$ and $\Delta\Delta G^{\circ}$ is expected if the variation in $\Delta\Delta G^{\circ}$ values is determined by whether or not the D-Ala residue generates new contacts, and how strong these interactions are. $\Delta\Delta E_{vdw}$ can be calculated from snapshots derived from the MD simulations, while the contribution of backbone desolvation to $\Delta\Delta G^{\circ}$ can be studied by counting the number of water molecules that are blocked from interacting with the peptide backbone at the C-capping site in the folded and unfolded states using snapshots from the MD simulations. The difference provides an estimate of the net desolvation effect. It is important to validate the models used for these analyses and the applicability of the force field employed with more rigorous methods. Consequently, we first tested if our MD simulations were sufficiently converged and our force field accurate enough to reproduce the experimental data using thermodynamic integration (TI) free energy calculations.

Thermodynamic integration validates more approximate computational models and provides further insight into C-capping energetics

The model used for the unfolded states are tetrapeptides with neutral capping groups and the length of the MD simulations can only reach a time scale that is much smaller than the experimental time scale. A recently parametrized force field was chosen in this study, but, like all force fields, is still an approximate description of molecules [42]. Therefore, we tested our models by asking if we can reproduce the experimental values of $\Delta\Delta G^{\circ}$ using TI. 34 λ windows were simulated for 12 ns each. TI is computationally expensive and reaching complete ergodic convergence in each λ window is unlikely, thus two different starting structures of each protein were used for two independent TI calculations in order to evaluate precision. For each protein, one of the starting structures was the PDB structure, while the other one was the last frame of a 50 ns MD simulation. (supplemental information).

Similar values of $\Delta\Delta G^{\circ}$ were obtained for a given protein independent of the starting structure chosen, suggesting that the TI calculation has reached reasonable convergence during the time-scale of the simulations (**Figure 2A)**. The only significant difference between $\Delta\Delta G^{\circ}_{calc}$ values determined using the different starting structures occurs for EH. We believe the effect is due to the poorly resolved N-terminus of EH in the X-ray structure rather than issues with the computational models implemented here. Residues 1-4 are unresolved and not shown in the crystal structure, while residues 5-7 are resolved, but with low confidence [30]. We appended the 4 missing residues as an extended peptide to the crystal structure and conducted a MD simulation with restraints on all resolved residues to relax the four appended residues. The last frame of this restrained MD simulation was used as the starting structure for one of the TI calculations for EH (**Figure 2A** red bar). Following the restrained MD simulation, Gly 39 was changed to D-Ala and unrestrained MD simulation was carried out to fully relax the conformation. The last frame of this simulation was used as the starting structure for the other TI calculation of EH (**Figure 2A** cyan bar). During the unrestrained MD simulation, residues 1-7 formed contacts with Gly39 or D-Ala39; this was not observed during the restrained MD simulation. The difference in the calculated $\Delta\Delta G^{\circ}$ of EH may be caused by the difference in the extent of relaxation of the starting structures. Since residues 1-7 in the PDB structures are either unresolved or poorly resolved, the fully relaxed structure is likely a better

representation of the structure of EH. The better agreement between $\Delta\Delta G^{\circ}_{exp}$ and $\Delta\Delta G^{\circ}_{cal}$ when the fully relaxed structure was used as starting structure is consistent with this hypothesis.

A small root-mean-square error of 0.23 kcal/mol is obtained for the complete set of $\Delta\Delta G^{\circ}_{exp}$ and $\Delta\Delta G^{\circ}_{cal}$ values calculated using the last frames of 50 ns MD simulations as the starting structures (**Figure 2B**). This indicates that the simplified unfolded state model, sampling sufficiency and choice of force field provide accurate energetics for these systems. The good agreement also argues that the large span in experimental $\Delta\Delta G^{\circ}$ values is neither caused by complexity in the unfolded states nor by the different conditions and methods used for the experimental protein stability measurements since a simplified model for the unfolded states and a consistent computational approach were able to reproduce the experimental trends.

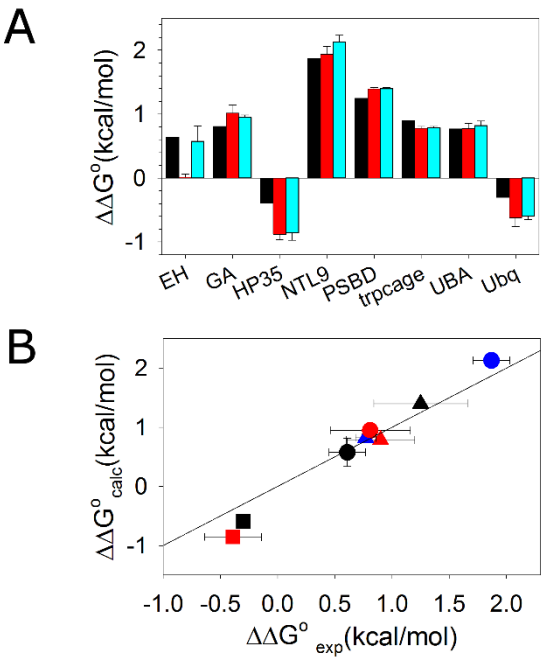


Figure 2. Thermodynamic integration reproduces experimental values of $\Delta\Delta G^{\circ}$. (A) Experimental $\Delta\Delta G^{\circ}$ values are shown in black. Calculated $\Delta\Delta G^{\circ}$ values using experimental structures as starting structures are shown in red. Calculated $\Delta\Delta G^{\circ}$ values using the last frames of 50 ns simulations as the starting structures are in cyan. (B) A scatter plot of experimental $\Delta\Delta G^{\circ}$ and calculated $\Delta\Delta G^{\circ}$ values using the last frames of a 50 ns simulation as the starting structure. Solid line represents $\Delta\Delta G^{\circ}_{exp} = \Delta\Delta G^{\circ}_{cal}$. EH ●; GA ●; HP35 ■; NTL9 ●; PSBD ▲; Trp-cage ▲; UBA ▲; Ubiquitin ■. The calculated value for the EH domain used in the plot was derived by using the unrestrained MD structure as the starting structure for the TI calculation. Positive $\Delta\Delta G^{\circ}$ values indicate stabilization.

.

The calculated change in van der Waals energy, $\Delta\Delta E_{vdw}$, is strongly correlated with $\Delta\Delta G^{\circ}$, but $\Delta\Delta G^{\circ}$ does not correlate with predicted desolvation effects.

To test our hypothesis that the entropic stabilization is modulated by variation in van der Waals interactions, $\Delta\Delta E_{vdw}$ values were calculated from the MD simulations. There is a strong correlation between $\Delta\Delta E_{vdw}$ and the $\Delta\Delta G^{o}$ values obtained experimentally or computationally with correlation coefficients of 0.89 in both cases (**Figure 3**). The results strongly support the hypothesis that van der Waals interactions between the D-Ala/Gly site and the rest of the protein play an important role in determining $\Delta\Delta G^{\circ}$.

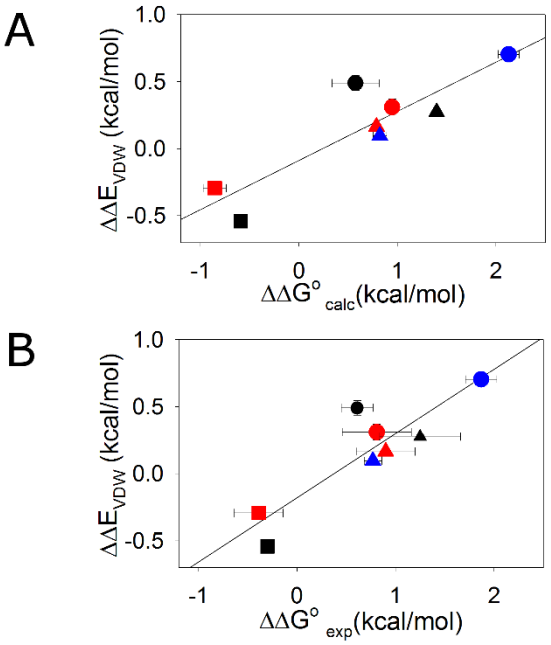


Figure 3. Scatter plot of $\Delta\Delta E_{vdw}$ and $\Delta\Delta G^{\circ}$ with solid line showing the linear fit. (A) Correlation of $\Delta\Delta E_{vdw}$ and $\Delta\Delta G^{\circ}$ values calculated by thermodynamic integration. r=0.89, p-value=0.0033 (B) Correlation of $\Delta\Delta E_{vdw}$ and experimental $\Delta\Delta G^{\circ}$ values. r=0.89, p-value =0.0033. EH ●; GA ●; HP35 ■; NTL9 ●; PSBD ▲; Trp-cage ▲; UBA ▲; Ubiquitin ■. Positive $\Delta\Delta G^{\circ}$ values indicate stabilization.

In order to examine potential correlations between the extent of backbone desolvation and the $\Delta\Delta G^\circ$ values, the first shell water molecules around backbone atoms in both the folded and unfolded states were counted. The difference in the number of water molecules blocked by D-Ala relative to Gly in the unfolded states and folded states (unfolded-folded) provides an estimate of the net desolvation effect of the new sidechain. Since the methyl group in D-Ala is non-polar, the mutation from Gly-to-D-Ala only changes the water accessibility of the backbone and counting the number of water around backbone is a reasonable metric for measuring desolvation effects. The calculations were performed by averaging over the last 160 ns of 12 independent MD simulations for the folded state and 144 ns of MD simulations for the unfolded state of each protein. No significant correlation is observed with $\Delta\Delta G^\circ$ values. The correlation coefficient for the number of waters blocked by D-Ala and $\Delta\Delta G^\circ_{calc}$ is only 0.16 and is just 0.17 for the correlation with $\Delta\Delta G^\circ_{exp}$ (**Figure 4**). If the desolvation effects in the unfolded state are disregarded and only the number of blocked waters in the folded state are counted, the correlation between $\Delta\Delta G^\circ_{calc}$ or $\Delta\Delta G^\circ_{exp}$ and the number of waters blocked by D-Ala relative to Gly is not improved, with correlation coefficients of 0.20 and 0.16 respectively. For three of the proteins (EH, HP35 and GA) the uncertainty, defined here as the standard deviation of the three sets of simulations with 4 independent simulations in each set, in the number of waters blocked by D-Ala in the unfolded and folded states is relatively large. However, this does not affect the conclusion that desolvation effects are not correlated with $\Delta\Delta G^\circ$. The good convergence in the $\Delta\Delta G^\circ_{calc}$ values in the absence of good convergence in the number of blocked waters reinforces that there is unlikely to be a significant net contribution of desolvation to $\Delta\Delta G^\circ$ for the systems studied here.

In principle, Poisson-Boltzmann (PB) based calculations could be used to estimate desolvation effects[43], however we observed during the 200 ns MD simulations of the folded states that subtle changes in conformation can lead to a significant change in the calculated PB desolvation energy of the backbone atoms owing to the long range nature of electrostatic interactions. This results in poor convergence for the PB calculations if the fluctuations in conformation are on the same time scale of the MD simulations

and leads to large error bars for PB based calculations of desolvation effects. EH and HP35 showed poor convergence in the PB calculations. The other six proteins have relatively good convergence, but no correlation between the desolvation effects calculated by PB and $\Delta\Delta G^{\circ}_{exp}$ was observed (r=0.28, p=0.58, slope=0.2) (**Figure S3**). The small slope indicates that differences in the PB desolvation energy do not make a contribution to the differences in $\Delta\Delta G^{\circ}$. The good convergence in the $\Delta\Delta G^{\circ}_{calc}$ values in the absence of convergence in the PB calculated solvation energy for all proteins further reinforces our conclusion that it is unlikely that desolvation makes a significant contribution to the range of $\Delta\Delta G^{\circ}$ values observed for the systems studied here.

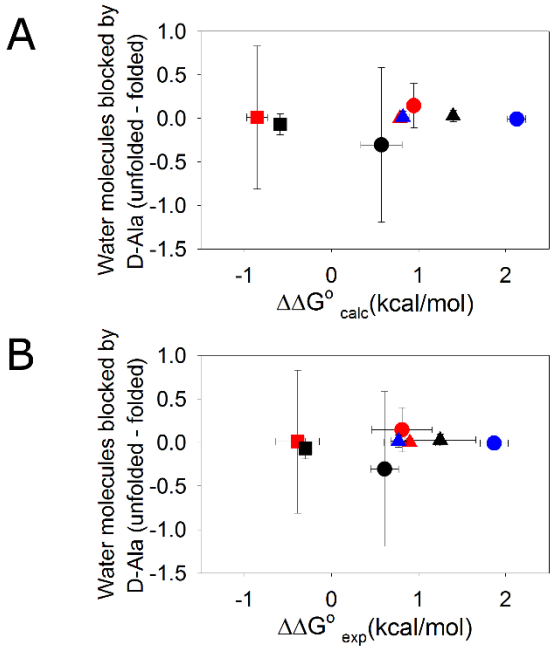


Figure 4. Changes in backbone solvation do not correlate with $\Delta\Delta G^{\circ}$. The difference in the number of water molecules blocked by D-Ala relative to Gly (Unfolded-folded) is plotted vs (A) calculated $\Delta\Delta G^{\circ}$ values (r=0.16). (B) experimental $\Delta\Delta G^{\circ}$ values (r=0.17). EH ●; GA ●; HP35 ■; NTL9 ●; PSBD ▲; Trp-cage ▲; UBA ▲; Ubiquitin ■. Positive $\Delta\Delta G^{\circ}$ values indicate stabilization.

The rapid screening of target proteins for D-Ala substitutions; a designed negative control helps to demonstrate proof of principle

It is prohibitively expensive to generate entire ensembles from an MD trajectory in explicit solvent in order to calculate $\Delta\Delta E_{vdw}$ values for a large set of proteins. Instead, a method which estimates $\Delta\Delta E_{vdw}$ in a time-efficient manner was developed in order to enable rapid screening of proteins for sites suitable for D-Ala substitution. The method was used to identify the HP35 D-Ala11 mutant as a negative control. The approach exploits the strong correlation between $\Delta\Delta E_{vdw}$ and $\Delta\Delta G^{\circ}$ identified above and uses a more rapid method to calculate $\Delta\Delta E_{vdw.}$ We calculated $\Delta\Delta E_{vdw\_gb}$, which like $\Delta\Delta E_{vdw}$, quantifies the contribution of the intramolecular van der Waals energy to $\Delta\Delta G^{\circ}$, but is obtained by running a short implicit-solvent simulation[44] instead of using a large ensemble from a long explicit-solvent MD simulation. The correlation between $\Delta\Delta E_{vdw}$ and $\Delta\Delta E_{vdw\_gb}$ is 0.84 (**Figure S4**) for the 8 systems in **Figure 3**. Although the implicit-solvent model is more coarse-grained than the explicit-solvent model and the length of simulation is significantly decreased, calculation of $\Delta\Delta E_{vdw\_gb}$ for a range of proteins should allow one to predict trends of $\Delta\Delta G^{\circ}_{exp}$ for hundreds of proteins in a time-efficient manner, provided the correlation between $\Delta\Delta E_{vdw\_gb}$ and the known $\Delta\Delta G^{\circ}_{exp}$ values is good. If desired, one can conduct further analysis of promising sites using longer MD simulations with explicit solvent or TI.

As shown in **Figure 5**, $\Delta\Delta E_{vdw\_gb}$ values (positive values represent net stabilization) are strongly correlated with the known values of $\Delta\Delta G^{\circ}_{exp}$ (r=0.94) (**Figure 5**). The strong correlation between $\Delta\Delta E_{vdw\_gb}$ and $\Delta\Delta G^{\circ}_{exp}$ further supports our hypothesis that the perturbation of van der Waals interactions are correlated with the effect of Gly-to-D-Ala substitutions on stability.

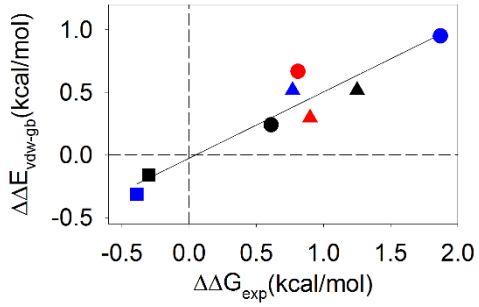


Figure 5. There is a strong correlation between $\Delta\Delta E_{vdw\text{-}gb}$ and $\Delta\Delta G^{\circ}_{exp}$. Positive values of $\Delta\Delta G^{\circ}$ indicate stabilizing effects. EH ●; GA ●; HP35 ■; NTL9 ●; PSBD ▲; Trp-cage ▲; UBA ▲; Ubiquitin ■; r=0.94 and p=0.0004. Positive $\Delta\Delta G^{\circ}$ values indicate stabilization.

The strong correlation between $\Delta\Delta E_{vdw\_gb}$ and $\Delta\Delta G^{\circ}_{exp}$ (**Figure 5**) indicates that linear regression can be used to predict the $\Delta\Delta G^{\circ}_{exp}$ values from their $\Delta\Delta E_{vdw\_gb}$ values using the empirical function:

$\Delta\Delta G^{\circ}_{exp}$ (kcal/mol) = 1.89 * $\Delta\Delta E_{vdw\text{-}gb}$ + 0.05

We examined a set of 120 monomeric proteins of less than 130 residues, which have structures determined at 2.0 Å resolution or better and at least one helix with a C-capping Gly. $\Delta\Delta E_{vdw\_gb}$ values were calculated for proteins with high sequence diversity. In all, 160 C-capping sites were analyzed (**Table S3**) and $\Delta\Delta E_{vdw\_gb}$ values ranging from -0.35 to 1.67 kcal/mol were obtained (**Figure 6A**). Here, negative values indicate a net destabilization and positive values reflect a net stabilization. The distribution of predicted $\Delta\Delta G^{\circ}$ values is plotted as a histogram in **Figure 6B**. Overall, 95% of the substitutions are predicted to lead to increased stability. Furthermore, ~80% of C-capping Gly-to-D-Ala substitutions in monomeric proteins will result in significant stabilization larger than 1kT.

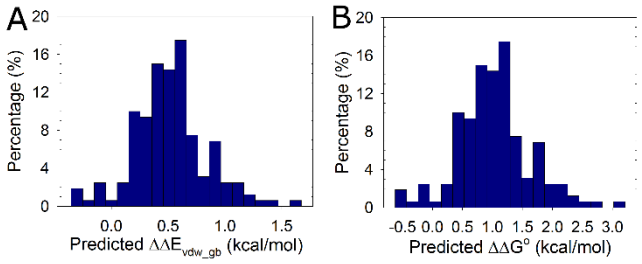


Figure 6. Proteins are stabilized by D-Ala substitutions. The distribution of $\Delta\Delta E_{vdw\_gb}$ and $\Delta\Delta G^{\circ}$ values for the 160 C-capping sites in the 120 non-redundant proteins is shown as a histogram. (A) Distribution of $\Delta\Delta E_{vdw\_gb}$ values. (B) Distribution of predicted $\Delta\Delta G^{\circ}$ values. Positive $\Delta\Delta G^{\circ}$ values represent a stabilizing effect.

From this distribution we selected the helical subdomain of the villin headpiece (HP35) as a negative test case, since it was one of the few proteins for which a D-Ala substitution was predicted to be destabilizing (**Figure S5**). $\Delta\Delta E_{vdw\_gb}$ for the replacement of Gly by D-Ala in HP35 was -0.31 kcal/mol (negative values represent net destabilization), which is comparable to the value for ubiquitin (**Figure 5**). As noted above, HP35 G11D-Ala has an experimentally measured stability 0.39 kcal/mol lower than wild-type HP35 (**Table 1**), confirming the computational prediction made prior to experiments.

Conclusions

Our analysis indicates that the energetics of C-capping interactions involve an interplay between two competing factors. Glycine residues are selected for such sites because they are able to adopt positive values of φ, but the choice of glycine introduces packing defects in the native state. The extremely high conservation of C-capping sites indicates that the evolutionary pressure to maintain the ability to adopt a positive value of phi at these sites leads to tolerance of packing defects in the structure. This highlights that protein stability includes compromises between competing interactions. Our results clearly show that Gly-to-D-Ala substitutions in C-capping motifs stabilize proteins when the folded state is not perturbed by unfavorable van der Waals interactions. The stability of EH, GA, NTL9, PSBD, Trp-cage and UBA were improved by 0.59 to 1.87 kcal/mol by Gly-to-D-Ala substitutions. Van der Waals interactions make a significant contribution to the observed spread in $\Delta\Delta G^{\circ}$ values. The fact that TI calculations quantitatively reproduced the experimentally observed effects, including the destabilization of HP35 and ubiquitin, argues that the range of reported $\Delta\Delta G^{o}$ values are not caused by variation in experimental protocols or complex effects in the unfolded state. The D-Ala variants of HP35 and ubiquitin were destabilized due to new unfavorable folded state van der Waals interactions that counteract the entropic stabilization. The systems studied here are two state folding but the general principles, unfolded state destabilization via entropic effects and native state stabilization by new favorable Van der Waals interaction also apply to proteins that fold via intermediates.

An important practical observation from this work is that steric clashes may still be generated by D-Ala substitution even if a C-capping glycine is identified as solvent exposed by measuring its solvent accessible surface area (SASA) (**Table S2**). The effect arises because the repulsive part of the van der Waals potential energy has a strong distance dependence, with the potential energy increasing rapidly as the distance between two atoms decreases. For example, moving a β-carbon from 3.2 Å to 2.8 Å from a carboxyl oxygen results in an increase in van der Waals potential energy of 1.9 kcal/mol using the Lennard-Jones potential in the Amber ff14SB force field [42]. This indicates that a more quantitative method than measuring SASA should be used when predicting the consequence of Gly-to-D-Ala substitutions at C-caps of α-helices.

Does the observation that the effects of the D-Ala substitutions can be predicted accurately using a highly simplified model of the unfolded state imply that the unfolded state is devoid of structure or long range contacts or residual structure? The answer is no; the data simply argues that the substitutions do not significantly impact the energetics of other unfolded state interactions; indeed residual structure has been detected in the unfolded states of several of the proteins studied [45-48].

In this study, experimental values of $\Delta\Delta G^\circ$ have been successfully reproduced by using molecular modelling for all proteins tested. These examples show that *in silico* molecular modelling and design serve as an excellent complement to experimental studies, and can allow one to rationally target unfolded state interactions. Predicted $\Delta\Delta G^\circ$ values of a large data set of structures indicate that most proteins will be stabilized by Gly-to-D-Ala substitutions at C-capping sites, opening the door to mirror image protein design.

C-capping glycines are strongly conserved in protein structures and can be identified by multiple sequence alignments, thus they can often be identified in the absence of structural information. The analysis presented here demonstrates that the replacement of such glycines is expected to be stabilizing 95% of the cases and to be significantly stabilizing 80% of the cases. This expected success rate is considerably better than has been observed with consensus method based on multiple sequence alignment and is

comparable to the most successful consensus method which take into account co-variation, suggesting that rational protein design is possible in the absence of structural information [49-50].

Associated Content

Supporting information

Experimental and computational methods. Additional figures and tables as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org

Acknowledgements

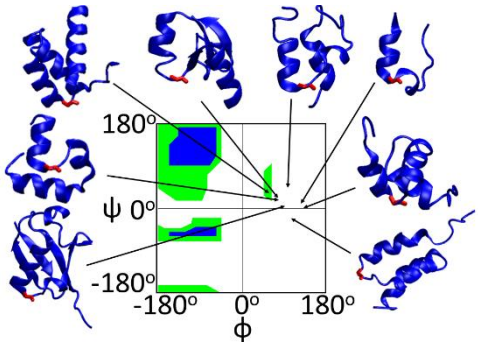
The authors gratefully acknowledge Prof. Rohit Pappu for numerous helpful discussions and Prof. Robert C. Rizzo for helpful suggestions. The authors also thank the Laufer Center for Physical and Quantitative Biology at Stony Brook University for access to computational resources and support, and Feng Zhang, Dr. James Maier and Koushik Kasavajhala for their administration of computational resources.

Funding sources. This work was supported by NSF grant MCB -1330259 to DPR and NIH grant GM107104 and an NSF Petascale Computational Resource (PRAC) Award from the NSF (OCI-1036208) to CS. We gratefully acknowledge support from Henry and Marsha Laufer. J.Z. was supported in part by a fellowship from the Laufer Center.

References

1. Chi, E. Y.; Krishnan, S.; Randolph, T. W.; Carpenter, J. F., *Pharm. Res.* **2003,** *20*, 1325-1336.
2. Daniel, R. M.; Cowan, D. A.; Morgan, H. W.; Curran, M. P., *Biochem. J.* **1982,** *207*, 641-644.
3. Parsell, D. A.; Sauer, R. T., *J. Biol. Chem.* **1989,** *264*, 7590-7595.
4. Chi, E. Y.; Krishnan, S.; Kendrick, B. S.; Chang, B. S.; Carpenter, J. F.; Randolph, T. W., *Protein Sci.* **2003,** *12*, 903-913.
5. McLendon, G.; Radany, E., *J. Biol. Chem.* **1978,** *253*, 6335-6337.
6. Binz, H. K.; Amstutz, P.; Pluckthun, A., *Nat. Biotechnol.* **2005,** *23*, 1257-1268.
7. Skrlec, K.; Strukelj, B.; Berlec, A., *Trends Biotechnol.* **2015,** *33*, 408-418.
8. Matthews, B. W.; Nicholson, H.; Becktel, W. J., *Proc. Natl. Acad. Sci. U. S. A.* **1987,** *84*, 6663-6667.
9. Nicholson, H.; Becktel, W. J.; Matthews, B. W., *Nature* **1988,** *336*, 651-656.
10. Matsumura, M.; Becktel, W. J.; Levitt, M.; Matthews, B. W., *Proc. Natl. Acad. Sci. U. S. A.* **1989,** *86*, 6562-6566.
11. Spector, S.; Wang, M.; Carp, S. A.; Robblee, J.; Hendsch, Z. S.; Fairman, R.; Tidor, B.; Raleigh, D. P., *Biochemistry* **2000,** *39*, 872-879.
12. Anil, B.; Song, B.; Tang, Y.; Raleigh, D. P., *J. Am. Chem. Soc.* **2004,** *126*, 13194-13195.
13. Sauer, R. T.; Hehir, K.; Stearman, R. S.; Weiss, M. A.; Jeitler-Nilsson, A.; Suchanek, E. G.; Pabo, C. O., *Biochemistry* **1986,** *25*, 5992-5998.
14. Wells, J. A.; Powers, D. B., *J. Biol. Chem.* **1986,** *261*, 6564-6570.
15. Wetzel, R.; Perry, L. J.; Baase, W. A.; Becktel, W. J., *Proc. Natl. Acad. Sci. U. S. A.* **1988,** *85*, 401-405.
16. Tidor, B.; Karplus, M., *Proteins* **1993,** *15*, 71-79.
17. Rodriguez-Granillo, A.; Annavarapu, S.; Zhang, L.; Koder, R. L.; Nanda, V., *J. Am. Chem. Soc.* **2011,** *133*, 18750-18759.
18. Betz, S. F.; Pielak, G. J., *Biochemistry* **1992,** *31*, 12337-12344.
19. Camarero, J. A.; Fushman, D.; Sato, S.; Giriat, I.; Cowburn, D.; Raleigh, D. P.; Muir, T. W., *J. Mol. Biol.* **2001,** *308*, 1045-1062.
20. Stites, W. E.; Meeker, A. K.; Shortle, D., *J. Mol. Biol.* **1994,** *235*, 27-32.
21. Richardson, J. S.; Richardson, D. C., *Science* **1988,** *240*, 1648-1652.
22. Aurora, R.; Rose, G. D., *Protein Sci.* **1998,** *7*, 21-38.
23. Gunasekaran, K.; Nagarajaram, H. A.; Ramakrishnan, C.; Balaram, P., *J. Mol. Biol.* **1998,** *275*, 917-932.
24. Hutchinson, E. G.; Thornton, J. M., *Protein Sci.* **1994,** *3*, 2207-2216.
25. Haque, T. S.; Gellman, S. H., *J. Am. Chem. Soc.* **1997,** *119*, 2303-2304.
26. Sibanda, B. L.; Thornton, J. M., *Nature* **1985,** *316*, 170-174.
27. Bystroff, C.; Baker, D., *J. Mol. Biol.* **1998,** *281*, 565-577.
28. Bang, D.; Gribenko, A. V.; Tereshko, V.; Kossiakoff, A. A.; Kent, S. B.; Makhatadze, G. I., *Nat. Chem. Biol.* **2006,** *2*, 139-143.
29. Chiu, T. K.; Kubelka, J.; Herbst-Irmer, R.; Eaton, W. A.; Hofrichter, J.; Davies, D. R., *Proc. Natl. Acad. Sci. U. S. A.* **2005,** *102*, 7517-7522.
30. Clarke, N. D.; Kissinger, C. R.; Desjarlais, J.; Gilliland, G. L.; Pabo, C. O., *Protein Sci.* **1994,** *3*, 1779-1787.
31. Johansson, M. U.; de Chateau, M.; Wikstrom, M.; Forsen, S.; Drakenberg, T.; Bjorck, L., *J. Mol. Biol.* **1997,** *266*, 859-865.
32. Kalia, Y. N.; Brocklehurst, S. M.; Hipps, D. S.; Appella, E.; Sakaguchi, K.; Perham, R. N., *J. Mol. Biol.* **1993,** *230*, 323-341.
33. Religa, T. L.; Johnson, C. M.; Vu, D. M.; Brewer, S. H.; Dyer, R. B.; Fersht, A. R., *Proc. Natl. Acad. Sci. U. S. A.* **2007,** *104*, 9272-9277.
34. Wang, T.; Zhu, Y. J.; Gai, F., *J. Phys. Chem. B* **2004,** *108*, 3694-3697.
35. Spector, S.; Kuhlman, B.; Fairman, R.; Wong, E.; Boice, J. A.; Raleigh, D. P., *J. Mol. Biol.* **1998,** *276*, 479-489.
36. Humphrey, W.; Dalke, A.; Schulten, K., *J. Mol. Graph.* **1996,** *14*, 33-38, 27-38.
37. Scott, K. A.; Alonso, D. O.; Sato, S.; Fersht, A. R.; Daggett, V., *Proc. Natl. Acad. Sci. U. S. A.* **2007,** *104*, 2661-2666.
38. Nemethy, G.; Leach, S. J.; Scheraga, H. A., *J. Phys. Chem.* **1966,** *70*, 998-&.
39. DAquino, J. A.; Gomez, J.; Hilser, V. J.; Lee, K. H.; Amzel, L. M.; Freire, E., *Proteins: Struct., Funct., Genet.* **1996,** *25*, 143-156.
40. Zaman, M. H.; Shen, M. Y.; Berry, R. S.; Freed, K. F.; Sosnick, T. R., *J. Mol. Biol.* **2003,** *331*, 693-711.
41. Baxa, M. C.; Haddadian, E. J.; Jumper, J. M.; Freed, K. F.; Sosnick, T. R., *Proc. Natl. Acad. Sci. U. S. A.* **2014,** *111*, 15396-15401.
42. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., *J. Chem. Theory Comput.* **2015,** *11*, 3696-3713.
43. Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., 3rd, *Acc. Chem. Res.* **2000,** *33*, 889-897.
44. Nguyen, H.; Roe, D. R.; Simmerling, C., *J. Chem. Theory Comput.* **2013,** *9*, 2020-2034.
45. Cho, J. H.; Meng, W.; Sato, S.; Kim, E. Y.; Schindelin, H.; Raleigh, D. P., *Proc. Natl. Acad. Sci. U. S. A.* **2014,** *111*, 12079-12084.
46. Meng, W.; Lyle, N.; Luan, B.; Raleigh, D. P.; Pappu, R. V., *Proc. Natl. Acad. Sci. U. S. A.* **2013,** *110*, 2123-2128.
47. Mok, K. H.; Kuhn, L. T.; Goez, M.; Day, I. J.; Lin, J. C.; Andersen, N. H.; Hore, P. J., *Nature* **2007,** *447*, 106-109.
48. Spector, S.; Rosconi, M.; Raleigh, D. P., *Biopolymers* **1999,** *49*, 29-40.
49. Magliery, T. J., *Curr. Opin. Struct. Biol.* **2015,** *33*, 161-168.
50. Steipe, B.; Schiller, B.; Pluckthun, A.; Steinbacher, S., *J. Mol. Biol.* **1994,** *240*, 188-192.

For Table of Contents Only



Supporting material for:

Experimental and Computational Analysis of Protein Stabilization by Gly-to-D-Ala substitution: A Convolution of Native State and Unfolded State Effects

Junjie Zou, Benben Song, Carlos Simmerling, Daniel Raleigh

Methods

Protein Solid Phase Synthesis

The proteins and their Gly-to-D-Ala variants were chemically synthesized using Fmoc chemistry [1]. Sequences of these proteins are provided below. EH, GA and PSBD have a free N-terminus and amidated C-terminus, while HP35 has a free N-terminus and free C-terminus. Peptide identity was confirmed using MALDI or ESI and purity was greater than 95%. EH, observed mass 7453.97, expected mass 7453.52; EH D-Ala, observed mass 7467.75, expected mass 7467.55; GA D-Ala, observed mass 5143.96, expected mass 5143.91; HP35, observed mass 4065.16, expected mass 4064.13; HP35 D-Ala, observed mass 4079.32, expected mass 4078.15. PSBD, observed 4400.72, expected 4402.10.

Sequences of the Proteins Synthesized for This Study

dA refers to D-Ala and L_N refers to nor-leucine.

EH: MDEKRPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKKS

EH-G39D-Ala: MDEKRPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELdALNEAQIKIWFQN

KRAKIKKS

GA: LKNAIEDAIAELKKAGITSDFYFNAINKAKTVEEVNALVNEILKAHA

GA-G16D-Ala: LKNAKEDAIAELKKAdAITSDFYFNAINKAKTVEEVNALVNEILKAHA

HP35: LSDEDFKAVFGMTRSAFANLPLWL_NQQHLKKEKGLF

HP35-G11D-Ala: LSDEDFKAVFdAMTRSAFANLPLWL_NQQHLKKEKGLF

PSBD: AMPSVRKYAREKGVDIRLVQGTGKNGRVLKEDIDAFLAGGA

PSBD-G15D-Ala: AMPSVRKYAREKdAVDIRLVQGTGKNGRVLKEDIDAFLAGGA

Backbone phi/psi Angles and Calculation of the Solvent Accessibility of the Gly Backbone

The φ/ψ angles of C-capping glycines were calculated by using VMD [2]. The same PDB structures used for molecular dynamics simulations were used and missing hydrogen atoms were added using tLeap in Amber [3]. The solvent accessible surface area (SASA) of C-capping glycines was calculated by using VMD with a water probe radii of 1.4 Å. The extended tetrapeptides were constructed using tLeap with the same local sequence as the respective full length proteins. The C-termini of the tetrapeptides were amidated and the N-termini were acetylated. Residues in the extended peptides all have φ and ψ angles equal to 180°. Fractional SASA is defined as the ratio between the SASA found for the PDB structure and the SASA found for the extended tetrapeptide.

Table S1. Backbone phi/psi and solvent accessibility of Gly

Protein	φ (°)	ψ (°)	SASA (Å^2)	SASA in extended tetrapeptide (Å^2)	Fractional SASA (%)
EH	51.8	35.8	64.0	88.5	72.4
GA	107.8	-21.7	55.5	120.8	45.9
HP35	75.7	19.8	66.9	73.1	91.5
NTL9	70.4	26.9	36.7	98.9	37.1
PSBD	84.0	48.1	63.5	94.0	67.6
Trp-cage	119.9	10.0	31.6	113.0	28.0
UBA	127.0	1.3	64.2	95.9	67.0
Ubiquitin	81.2	5.2	53.7	100.5	53.4

Thermal and Urea/Guanidine Denaturation

The unfolding free energy of each protein was measured by CD-monitored urea/guanidine hydrochloride denaturation at 222nm under the conditions listed in **Table S2**. Thermal denaturation experiments were also conducted at 222nm using the same buffer and pH employed for the urea/guanidine hydrochloride denaturation experiments. The concentration of urea/guanidine was determined by measuring the refractive index on a refractometer. Urea/guanidine denaturation experiments were carried out with a

titrator unit interfaced to the CD spectrometer. Unfolding curves for EH, GA, PSBD were recorded using Aviv model 62A DS and 202SF circular dichroism spectrophotometers. Unfolding curves for HP35 were recorded using an Applied Photophysics Chirascan instrument. ΔG^o of unfolding was determined by fitting the urea/guanidine denaturation curves to the following equation:

$$\theta[denaturant] = \frac{(a_n + b_n[denaturant]) + (a_d + b_d[denaturant])e^{-\left(\frac{\Delta G^o([denaturant])}{RT}\right)}}{1 + e^{-\left(\frac{\Delta G^o([denaturant])}{RT}\right)}} \quad (1)$$

$$\Delta G^o([denaturant]) = \Delta G^o(H_2O) - m[denaturant] \quad (2)$$

where θ is the measured ellipticity, a_n ,b_n ,a_d ,b_d are the parameters that define the signals of the native state and denatured state. $\Delta G^o([denaturant])$ is the free energy change upon unfolding as a function of denaturant and $\Delta G^o(H_2O)$ is the free energy change in the absence of denaturant. Thermal unfolding data was fit using standard methods and the Gibbs-Helmholtz equation to obtain the melting temperature T_m and ΔH^0 at T_m.

$$\theta[T] = \frac{(a_n + b_n T) + (a_d + b_d T)e^{-\left(\frac{\Delta G^o(T)}{RT}\right)}}{1 + e^{-\left(\frac{\Delta G^o(T)}{RT}\right)}} \quad (3)$$

$$\Delta G^o(T) = \Delta H^o(T_m)\left(1 - \frac{T}{T_m}\right) - \Delta C^o{}_p[T_m - T + Tln(\frac{T}{T_m})] \quad (4)$$

Where T_m is the melting temperature. $\Delta H^o(T_m)$ is the change of enthalpy upon unfolding at the melting temperature. $\Delta C^o{}_p$ is the change of heat capacity upon unfolding.

Thermal and Urea/Guanidine Denaturation Conditions

Table S2. Conditions for thermal and urea/guanidine denaturation experiments

Protein	Urea/guanidine hydrochloride	Buffer	pH	Temperature (˚C)
EH	Urea	50mM sodium acetate	5.7	5
GA	Guanidine hydrochloride	50mM phosphate	7.0	25
HP35	Urea	100mM sodium chloride and 20mM sodium acetate	4.8	25
PSBD	Guanidine hydrochloride	2mM sodium phosphate, 2mM sodium borate and 50mM sodium chloride	8.0	25

Molecular Dynamics Simulations Using an Explicit-water Model

The starting structures used for the simulations of EH, GA, HP35, NTL9, PSBD, Trp-cage, UBA and ubiquitin were obtained from the pdb files 1ENH [4] , 1PRB [5], 1WY4 [6], 2HBB [7], 2PDD [8], 1L2Y [9], 1DV0 [10] and 1UBQ [11] respectively. Residues not included in the sequences listed above were deleted from the pdb file and the actual missing residues were added by Swiss PDB [12] and equilibrated by MD simulations with restraints on all other residues. C-Terminal amidation and N-terminal acetylation was added if the studied proteins had these modifications. X-ray structures are available for EH, HP35, NTL9 and ubiquitin, while only NMR structures are available for GA, PSBD, Trp-cage and UBA. For proteins with multiple models from NMR studies, the RMSD of each model was calculated using the average conformation as the reference. The model with the lowest RMSD was chosen as the starting structure for MD simulations. Starting structures for D-Ala mutants were created using tLeap in Amber [3]. Four independent MD simulations were run for each protein and for the D-Ala variant with different initial velocities, which results in eight simulations in total. The length of the simulations were 200 ns with the stepsize set to 2 fs. All simulations were performed using the Amber software package with the Amber ff14SB force field [13] and TIP3P water [14]. Parameters for nor-leucine were obtained from Forcefield_NCAA [15]. No ions were included in the system. All simulations were conducted under constant pressure conditions at 298K using Berendsen barostat to control pressure [16]. Temperature was controlled using a weak-coupling algorithm with the coupling constant set to 1 ps [16]. Truncated octahedron boxes with periodic boundary condition were used. Particle mesh Ewald methods were used to calculate electrostatic energies [17]. Hydrogen atoms were constrained using the SHAKE algorithm [18]. The cutoff of non-bonded interactions was set to 8 Å. The N-terminus was acetylated and C-terminus was amidated for proteins which had free termini and in which the termini were calculated to be neutral since deprotonated N-terminus and protonated C-terminus are not currently available in the Amber force field [13]. Regular terminal residues defined in the Amber force field [13] were used for cases where the N and C termini were charged.

Local effects in the unfolded state were modeled as blocked tetrapeptides with sequence ACE-Xaa_1-Gly/dAla-Xaa_2-NH2. Xaa_1 and Xaa_2 are the two residues adjacent to the C-capping Gly/dAla in the full length protein sequences. This approach provides a model of purely local interactions and is not meant to mimic the actual unfolded chain. In order to enhance sampling, the tetrapeptides were simulated at 500K for 0.4ns, followed by cooling from 500K to 298K in 0.4ns and 0.4ns at 298K. This annealing cycle was repeated 120 times. Only data from 298K was collected for all cycles. These procedures were repeated thrice with different initial velocities which resulted in 3 sets of 4 independent folded state simulations and 3 sets of 120 annealing cycles of unfolded state simulations. A total of 96,000 frames from the folded state simulations and 144,000 frames from the unfolded state simulations at 298K were saved for analysis.

Starting Structures of PSBD, Trp-cage and UBA used for MD Simulations

PSBD, Trp-cage and UBA have multiple models obtained through NMR experiments. For each model, the backbone RMSD was calculated using VMD [2]. The reference coordinates are the averaged coordinates of all the models. The models used as starting structures are as follows:

Protein	PDB code	Model number
PSBD	2PDD	Model 32
Trp-cage	1L2Y	Model 32
UBA	1DV0	Model 15

Assignment of Protonation States of Titratable Residues during MD Simulations

Protonation states of titratable residues were set to reflect the pH at which thermodynamic properties of proteins were measured. The H++ server was used to determine the protonation state [19]. Experimental $\Delta\Delta G^\circ$ have been reported for the ubiquitin variants over the pH range of 2.5 to 3.5 [20]. The value of $\Delta\Delta G^\circ$ at pH 2.5 was compared to the calculated value since the TI approach only allows fixed protonation states. By fixing all the acidic residues and the C-terminus to be protonated, the system resembles that expected at pH=2.5.

Protonation states for titratable residues and terminus are listed in the table below. Asp, Glu, and C-termini which are not listed were fixed in the deprotonated state. Lys, Arg and N-termini which are not listed were fixed in the protonated state.

Protein	pH	Asp and Glu	His	C-terminus and N-terminus
EH	5.7			
GA	7.0		52, doubly protonated	
HP35	4.8		68, doubly protonated	
NTL9	5.5			
PSBD	8.0			Deprotonated N-terminus
Trp-cage	7.0			
UBA	6.5			
Ubiquitin	2.5	All Asp and Glu are protonated	68, doubly protonated	Protonated C-terminus

Free Energy Calculations

Free energy calculations were performed using non-softcore thermodynamic integration implemented in Amber [3, 21]. Gly was turned into D-Ala in three stages. In the first stage, partial charges on the CA/HA2/HA3 of Gly were turned off. In the second stage, three dummy atoms were added to the disappearing glycine and van der Waals interaction of these dummy atoms were turned on so a D-Ala with no partial charges on the CA/HA/CB/HB1/HB2/HB3 atoms appeared. In the third stage, partial charges on the CA/HA/CB/HB1/HB2/HB3 atoms of D-Ala were turned on. The first and third stages have λ evenly distributed from 0.0 to 1.0 with an interval of 0.1 including 0.0 and 1.0. In order to avoid singularity at $\lambda = 0.0$ and $\lambda = 1.0$ and have more sampling at where $dV/d\lambda$ has a steep change, the second stage has λ equal to 0.00922, 0.04794, 0.115, 0.20634, 0.316, 0.43738, 0.56262, 0.68392, 0.79366, 0.88495, 0.95206, 0.99078. For the folded state, one set of the TI calculations began with the C-capping

glycine in place and used the crystal structures. Dummy atoms were added to the experimental structures to give the starting structures for the second stage of the calculations. Starting structures for the third stage were obtained by changing the Gly in the experimental structures to D-Ala. The alternate set of TI calculations was derived from the last frames of a 50 ns standard MD simulations of the D-Ala mutants. The structures resulting from these simulations were converted back to the Gly containing variants to provide starting structures for the first stage of the calculations.

For the folded state, MD simulations used the same set up as the standard MD simulations described above except that the length of the simulation was set to 12ns for each window. The blocked peptides, which model local interactions in the unfolded state, were converted from Gly to D-Ala in three stages using the same λ values that were used for the folded states. The same sampling enhancement strategy described above was used for all stages and λ windows. Only data from 298K was collected. Numerical integration was performed using trapezoidal integration. Three ΔΔG° values were obtained by dividing simulations of each λ window for the folded states and unfolded states into three blocks. Error bars for the calculated ΔΔG° were the standard deviation of the three ΔΔG° values.

Energy Decomposition and Analysis of First Shell Water Molecules

The van der Waals potential energy between Gly or D-Ala and the rest of protein was calculated by post processing MD simulation trajectories. 1-4 van der Waals interactions were considered as van der Waals interactions with a scaling factor of 0.5. $\Delta\Delta E_{vdw}$ is defined as:

$$\Delta\Delta E(VDW) = [E^{u}_{D-ala}(VDW) - E^{u}_{Gly}(VDW)] - [E^{f}_{D-ala}(VDW) - E^{f}_{Gly}(VDW)] \quad (5)$$

where "u" and "f" indicate unfolded and folded states respectively. For example, $E^{u}_{D-ala}(VDW)$ is the van der Waals interaction between D-Ala residue and the rest of the protein in the unfolded state.

The first shell water molecules were counted by using Cpptraj [22] in Amber, with a cutoff of 3.4 Å. For the folded states, the first shell water molecules around the amide nitrogen, amide proton, carbonyl carbon

and carbonyl oxygen of residues i-4 to i+1 (i=Gly/D-Ala) were counted because these atoms are structurally close to the C-capping residues. For the unfolded states, the water molecules around amide nitrogen, amide proton, carbonyl carbon and carbonyl oxygen of residues i-1 to i+1 (i=Gly/D-Ala) were counted.

$$\text{Number of water molecules (unfolded} - \text{folded)} = (n^{u}_{D-ala} - n^{u}_{Gly}) - (n^{f}_{D-ala} - n^{f}_{Gly}) \quad (6)$$

Where n is the number of first shell water molecules. The error bars of $\Delta\Delta E_{vdw}$ and number of water molecules (unfolded-folded) are the standard deviation of the 3 sets of simulations.

The desolvation effect on the backbone was also quantified by using Poisson Boltzmann (PB) equation solved by DelPhi[23]. The Amber ff14SB partial charges[13] and Yamagishi, J's radii set[24] were used.

$$\Delta\Delta G(bb\_solvation) = [G^{u}_{D-ala}(bb\_solvation) - G^{u}_{Gly}(bb_{solvation})] - [G^{f}_{D-ala}(bb\_solvation) - G^{f}_{Gly}(bb\_solvation)] \quad (7)$$

Since PB equation is non-linear, the solvation energy of each term on the right side of equation 7 was calculated in two steps. In the first step, we calculated the solvation energy of the whole protein with partial charges on the amide nitrogen, amide proton, carbonyl carbon and carbonyl oxygen of residues i-4 to i+1 (i-1 to i+1 for the unfolded state; i=Gly/D-Ala). In the second step, the partial charges on the amide nitrogen, amide proton, carbonyl carbon and carbonyl oxygen of residues i-4 to i+1 (i-1 to i+1 for the unfolded state; i=Gly/D-Ala) were set to 0 and the solvation energy of the whole protein was calculated again. The difference in the solvation energy obtained from these two step was considered as the solvation energy of the backbone around the Gly/D-Ala.

Calculation of $\Delta\Delta E_{vdw\text{-}gb}$ Using an Implicit-solvent Model

The length of the simulations were 5 ns with stepsize set to 1fs. Amber ff14SBonlysc [25] was used and igb was set to 8 which corresponds to GBneck2 implicit solvent model [26]. Mbondi3 radii set was used [26]. Simulations were conducted under 200K due to low thermostability of proteins in the implicit-solvent

model used here[25]. Langevin dynamics was employed with the collision frequency set to 1 ps^{-1}. No cutoff of non-bond interactions was used. The salt concentration was set to 0.0 M.

For the experimentally tested proteins (EH, GA, HP35, NTL9, PSBD, Trp-cage, UBA and ubiquitin), the starting structures were prepared in the same way as for the simulations in explicit solvent except no solvent was added. For the 120 proteins and their D-Ala variants listed in **Table S3**, any selenomethionines were converted to methionines and all acidic residues were deprotonated and all basic residues except histidines were protonated. The protonation states of histidines depends on whether the hydrogen on δ or ε nitrogen is resolved by X-ray. If neither of the hydrogens is resolved, the ε nitrogen was protonated. Disulphide bonds were added as indicated by the authors of the structures. All non-protein molecules and ions were deleted. Local effects in the unfolded states of proteins were modeled as blocked tetrapeptides. The tetrapeptides were simulated at 400K for 0.4ns, followed by cooling from 400K to 200K in 0.4ns and 0.4ns at 200K. This annealing cycle was repeated 160 times. The van der Waals potential energy between Gly or D-Ala and the rest of protein was calculated by post processing MD simulation trajectories. 1-4 van der Waals interactions were considered as van der Waals interactions instead of bonded interactions. $\Delta\Delta E_{vdw\_gb}$ is defined as:

$$\Delta\Delta E(VDW\_gb) = [E^{u}_{D-ala}(VDW\_gb) - E^{u}_{Gly}(VDW\_gb)] - [E^{f}_{D-ala}(VDW\_gb) - E^{f}_{Gly}(VDW\_gb)] \quad (8)$$

where "u" and "f" indicate unfolded and folded states respectively. For example, $E^{u}_{D-ala}(VDW\_gb)$ is the van der Waals interaction between the D-Ala residue and the rest of the protein in the unfolded state calculated using the implicit-solvent model.

For the 8 experimentally tested proteins, each $E^{f}_{D-ala}(VDW\_gb)$ value and each $E^{f}_{Gly}(VDW\_gb)$ value is the average over 100,000 frames from 10 independent simulations with different random number seeds

for Langevin dynamics. For the 120 target proteins and their variants, $E^{f}_{D-ala}(VDW\_gb)$ values and $E^{f}_{Gly}(VDW\_gb)$ values were averaged over 30,000 frames from 3 independent simulations. For all of the proteins, $E^{u}_{D-ala}(VDW\_gb)$ values and $E^{u}_{Gly}(VDW\_gb)$ values were averaged over 40,000 frames collected from the simulations at 200K.

Protein Chains Dataset and $\Delta\Delta E_{vdw\_gb}$

All protein chains listed here are non-redundant protein chains with BLAST [27] pvalue less than 10e-7. According to the authors of the structures, all of the protein chains are monomeric. All proteins have at least one α-helical C-capping Gly. The criteria for defining a helix was at least 5 sequential residues with $-140^{\circ} \leqslant \varphi \leqslant -30^{\circ}$ and $-90^{\circ} \leqslant \psi \leqslant 45^{\circ}$. A C-capping Gly is the first non-helical residue at the C-terminus of a helix with $20^{\circ} \leq \varphi \leq 125^{\circ}$ and $-45^{\circ} \leq \psi \leq 90^{\circ}$ [28]. $\Delta\Delta E_{vdw\_gb}$ values were only calculated for proteins with high sequence diversity. In order to do so, a table of sequence redundancy in protein data bank was obtained from Molecular Modelling Database [29]. A representative of each non-redundant sequence was chosen according to the ranking provided by this table.

Table S3. Calculated values of $\Delta\Delta E_{vdw\_gb}$ for 160 C-capping sites from 120 non-redundant proteins taken from the pdb bank. Positive $\Delta\Delta E_{vdw\_gb}$ values indicate a stabilizing effect.

pdb code	chain ID	Short description of protein	Organism	Site No.	Calculated $\Delta\Delta E_{vdw\_gb}$ (kcal/mol)
1ABA	A	T4 glutaredoxin	Enterobacteria phage T4 sensu lato	56	0.61
1C44	A	Sterol carrier protein 2	Oryctolagus cuniculus	32	0.15
				86	1.20
				97	0.95
1KAF	A	The DNA Binding Domain Of Phage T4 Transcription Factor MotA	Enterobacteria phage T4 sensu lato	125	0.59
				179	0.36
1KP6	A	Killer toxin kp6 alpha-subunit	Ustilago maydis	9	-0.35
1L8R	A	Dachshund protein	Homo sapiens	255	0.76
1L9L	A	Granulysin from cytolytic T lymphocytes	Homo sapiens	63	0.64
1LWB	A	Phospholipase A2 protein	Streptomyces violaceoruber	75	0.35
1MC2	A	Phospholipase A2 protein	Deinagkistrodon acutus	14	0.34
1MK0	A	The catalytic domain of intron endonuclease I-TevI	Enterobacteria phage T4 sensu lato	38	0.27
1MOL	A	Monellin	Dioscoreophyllum cumminsii	27	0.87
1NWZ	A	Light receptor photoactive yellow protein	Halorhodospira halophila	51	0.31
				86	0.54
1OOH	A	An odorant binding protein LUSH	Drosophila melanogaster	34	1.40
				56	1.08
1ORG	A	A pheromone-binding protein	Rhyparobia maderae	53	1.03
1OSD	A	A mercury-binding protein	Cupriavidus metallidurans	65	0.24
1PBJ	A	A hypothetical protein	Methanothermobacter thermautotrophicus	59	0.40
1Q6V	A	Phospholipase A2 protein	Daboia russelii	14	0.24
1R6J	A	The PDZ2 domain of syntenin	Homo sapiens	231	0.88
1SBX	A	The dachshund-homology domain of Nuclear protooncoprotein SKI	Homo sapiens	165	0.66
1T1J	B	A hypothetical protein	Pseudomonas aeruginosa	43	0.52
				111	0.51
1T8K	A	An apo acyl carrier protein	Escherichia coli	16	0.53
				33	0.37
1TP6	A	A hypothetical protein	Pseudomonas aeruginosa	22	1.21
1TQG	A	CheA phosphotransferase domain	Thermotoga maritima	55	0.20
1U8T	B	CheY protein	Escherichia coli	29	0.37
				102	1.00
1VCD	A	Nudix protein Ndx1	Thermus thermophilus	52	0.33
1VYI	A	The C-terminal domain of a polymerase cofactor	Rabies virus	254	0.70
1WHZ	A	A hypothetical protein	Thermus thermophilus	18	0.60

1WOL	A	An HEPN homologue	Sulfolobus tokodaii	25	0.37
				50	0.67
1WY4	A	A villin headpiece	Gallus gallus	51	-0.31
1XLQ	A	Putidaredoxin	Pseudomonas putida	31	0.44
1XMK	A	The Zβ domain from the RNA editing enzyme ADAR1	Homo sapiens	341	0.78
1YN3	A	An extracellular adherence protein	Staphylococcus aureus	203	0.12
1Z96	A	Mud1 UBA domain	Schizosaccharomyces pombe	307	0.07
1ZMA	A	A bacterocin transport accessory protein	Streptococcus pneumoniae	81	0.38
2ACY	A	An acyl-phosphatase	Bos taurus	34	0.72
2B1L	B	A thiol:disulfide oxidoreductase	Escherichia coli	97	0.37
2B8I	A	A putative bacterial secretion factor	Vibrio vulnificus	56	0.99
2BO1	A	Ribosomal protein L30E	Thermococcus celer	30	0.87
				57	0.65
				75	0.47
2BWF	A	The UBL domain of Dsk2	Saccharomyces cerevisiae	36	-0.05
2CWY	A	A hypothetical protein	Thermus thermophilus	15	0.59
				55	0.48
2CX7	B	Sterol carrier protein 2	Thermus thermophilus	89	0.94
				100	0.65
2D48	A	Interleukin 4	Homo sapiens	95	0.43
2D58	A	An ionized calcium-binding adaptor	Homo sapiens	78	0.55
2FB6	A	A hypothetical protein	Bacteroides thetaiotaomicron	34	0.42
				68	0.59
				82	0.38
				91	0.43
2FC3	A	Ribosomal protein L7Ae	Aeropyrum pernix	46	0.66
				91	0.79
2FE5	A	The second PDZ domain of DLG3	Homo sapiens	270	0.56
2FYG	A	Nsp10	Severe acute respiratory syndrome-related coronavirus	34	0.13
2HC8	A	The actuator domain from Cu^{+}-ATPase	Archaeoglobus fulgidus	277	-0.01
2HL7	A	The periplasmic domain of cytochromes C maturation protein H	Pseudomonas aeruginosa	55	0.72
2HU9	A	A Zn2+ and [2Fe-2S]-containing copper chaperone	Archaeoglobus fulgidus	102	0.78
2I6V	A	Epsc, a crucial component of the type 2 secretion system	Vibrio cholerae	254	0.55
2IAY	A	LP2179, a member of the PF08866 family	Lactobacillus plantarum	31	0.87
2ICT	A	Antitoxin HigA	Escherichia coli	44	0.90
2J5Y	A	An albumin-binding domain	Finegoldia magna	22	0.73
2NT4	A	A response regulator homolog	Myxococcus xanthus	26	0.24
2O0Q	A	A hypothetical protein	Caulobacter vibrioides	20	0.40
				32	0.44
2OGB	A	The C-terminal domain of neuregulin receptor degrading protein 1	Mus musculus	237	0.51

2OY3	A	A macrophage receptor	Mus musculus	463	0.15
2P1H	A	The caspase recruitment domains of apoptotic protease activating factor 1	Homo sapiens	35	0.58
				81	0.22
2P3H	A	The CorC_HlyC domain of a putative hemolysin	Corynebacterium glutamicum	31	0.40
2POS	A	Sylvaticin	Pythium sylvaticum	32	0.30
2PSP	A	A pancreatic spasmolytic polypeptide	Sus scrofa	33	0.39
2PVB	A	Parvalbumin	Esox lucius	34	0.96
2PYQ	C	An uncharacterized protein	uncharacterized protein	20	0.40
				68	0.60
2QJL	A	A ubiquitin-related modifier	Saccharomyces cerevisiae	17	0.30
2RH3	A	The C-terminal domain of VirC2	Agrobacterium tumefaciens	130	0.37
2VB1	A	Triclinic hen egg-white lysozyme	Gallus gallus	16	0.60
				102	0.17
2VSV	A	The PDZ domain of human rhophilin-2	Homo sapiens	55	0.55
2VWR	A	The second PDZ domain of the human numb-binding protein 2	Homo sapiens	379	0.60
2W50	A	The N-terminal domain of human conserved dopamine neurotrophic factor	Homo sapiens	29	0.60
				60	0.71
2WFB	A	The apo Form of the Orange Protein	Desulfovibrio gigas	67	0.54
				88	0.64
2WT8	A	The N-terminal Brct domain of human microcephalin	Homo sapiens	36	0.62
				67	0.59
				83	0.47
2XEV	B	The TPR domain of YbgF	Xanthomonas campestris	15	0.50
				89	0.46
2ZQE	A	The endonuclease domain of an anti-recombination enzyme	Thermus thermophiles	31	0.63
3A0S	A	The PAS domain of histidine kinase ThkA	Thermotoga maritima	448	-0.35
3A0U	A	Response regulator protein TrrA	Thermotoga maritima	25	0.23
3A4R	A	The small ubiquitin-like modifier domain in Nip45	Mus musculus	376	0.66
3B79	A	The N-terminal peptidase C39 like domain of the toxin secretion ATP-binding protein	Vibrio parahaemolyticus	17	0.25
				49	0.73
3BS7	A	The sterile alpha motif domain of hyphen/aveugle	Drosophila melanogaster	71	0.50
3C9P	A	An uncharacterized protein	Streptococcus pneumoniae	25	0.40
				40	0.47
				106	0.41
3CJK	A	Copper transport protein ATOX1	Homo sapiens	59	0.20
3D2Q	B	The tandem zinc finger 3 and 4 domain of muscleblind-like protein 1	Homo sapiens	19	0.57
3E0Z	B	A putative imidazole glycerol phosphate synthase homolog	Agathobacter rectalis	39	0.10

3E11	B	A predicted zincin-like metalloprotease	Acidothermus cellulolyticus	102	0.51
3EZI	B	Histidine kinase NarX sensor domain	Escherichia coli	94	0.75
3FBL	A	An uncharacterized protein	Acidianus filamentous virus 1	66	0.60
3FZ4	A	A possible arsenate reductase	Streptococcus mutans	50	0.26
				68	1.07
3ID4	A	RseP PDZ2 domain	Escherichia coli	239	0.58
3IPJ	A	A domain of the PTS system	Peptoclostridium difficile	83	0.34
3L2A	A	VP35 interferon inhibitory domain	Reston ebolavirus	259	0.31
3LJW	B	The second bromodomain of human polybromo	Homo sapiens	240	1.67
3LLB	A	An uncharacterized protein	Pseudomonas aeruginosa	27	-0.13
				47	-0.06
3M3G	A	An elicitor of plant defense responses	Trichoderma virens	115	0.70
3NIR	A	Crambin	Crambe hispanica	20	0.17
				31	0.27
3NUF	A	A PRD-containing transcription regulator	Lactobacillus paracasei	67	0.59
3O79	B	A Prion protein	Oryctolagus cuniculus	195	0.23
3ODV	A	Kaliotoxin	Androctonus mauritanicus	22	1.28
3PO0	A	A Ubiquitin-like small archaeal modifier proteins	Haloferax volcanii	14	0.27
3QMX	A	Glutaredoxin A	Synechocystis sp. PCC 6803	29	-0.32
3S0A	A	An odorant-binding protein	Apis mellifera	22	0.47
				34	0.90
3SNS	A	The C-terminal domain of lipoprotein BamC	Escherichia coli	263	1.12
				292	0.82
3SVI	A	The Pto-binding domain of HopPmaL	Pseudomonas syringae group genomosp. 3	157	0.48
				173	0.44
3SZS	B	Hellethionin D	Helleborus purpurascens	20	0.33
3T7Z	A	Nop N-terminal domain	Methanocaldococcus jannaschii	60	-0.13
				91	0.56
3UI6	A	Parvulin 14	Homo sapiens	61	0.29
3V1A	A	A Metal interface design	synthetic construct	22	0.61
3W1O	A	A hypothetical protein	Neisseria meningitidis	51	0.86
3WCQ	A	Ferredoxin	Cyanidioschyzon merolae	33	0.31
				73	0.55
3ZR8	X	Rxlr effector AVR3a11	Phytophthora capsici	100	0.65
4CVD	A	A cell wall binding module	Streptococcus phage Cp-1	263	1.08
				279	0.33
4D40	A	Type IV pilin	Shewanella oneidensis	28	0.17
4F55	A	The catalytic Domain of SleB rotein	Bacillus cereus	202	0.56
				222	0.23
4FQN	A	CCM2 C-terminal harmonin homology domain	Homo sapiens	328	0.50
4G9S	A	A goose-type lysozyme	Escherichia coli	60	1.04
4GOQ	A	A hypothetical protein	Caulobacter vibrioides	20	0.41

4HRO	A	Small archaeal modifier protein 1	Haloferax volcanii	14	0.22
4HS5	A	Protein CyaY	Psychromonas ingrahamii	25	0.60
4JIU	A	An uncharacterized protein	Pyrococcus abyssi	14	0.50
4N6X	A	Na(+)/H(+) exchange regulatory cofactor NHE-RF1/Chemokine receptor CXCR2 fusion protein	Homo sapiens	52	0.89
4PXV	A	The LysM domain of chitinase A	Pteris ryukyuensis	32	0.62
4XPX	A	Hemerythrin	Methylococcus capsulatus	69	0.37
				97	0.54

Experimental Thermal Denaturation of EH, HP35, PSBD and Their D-Ala Variants

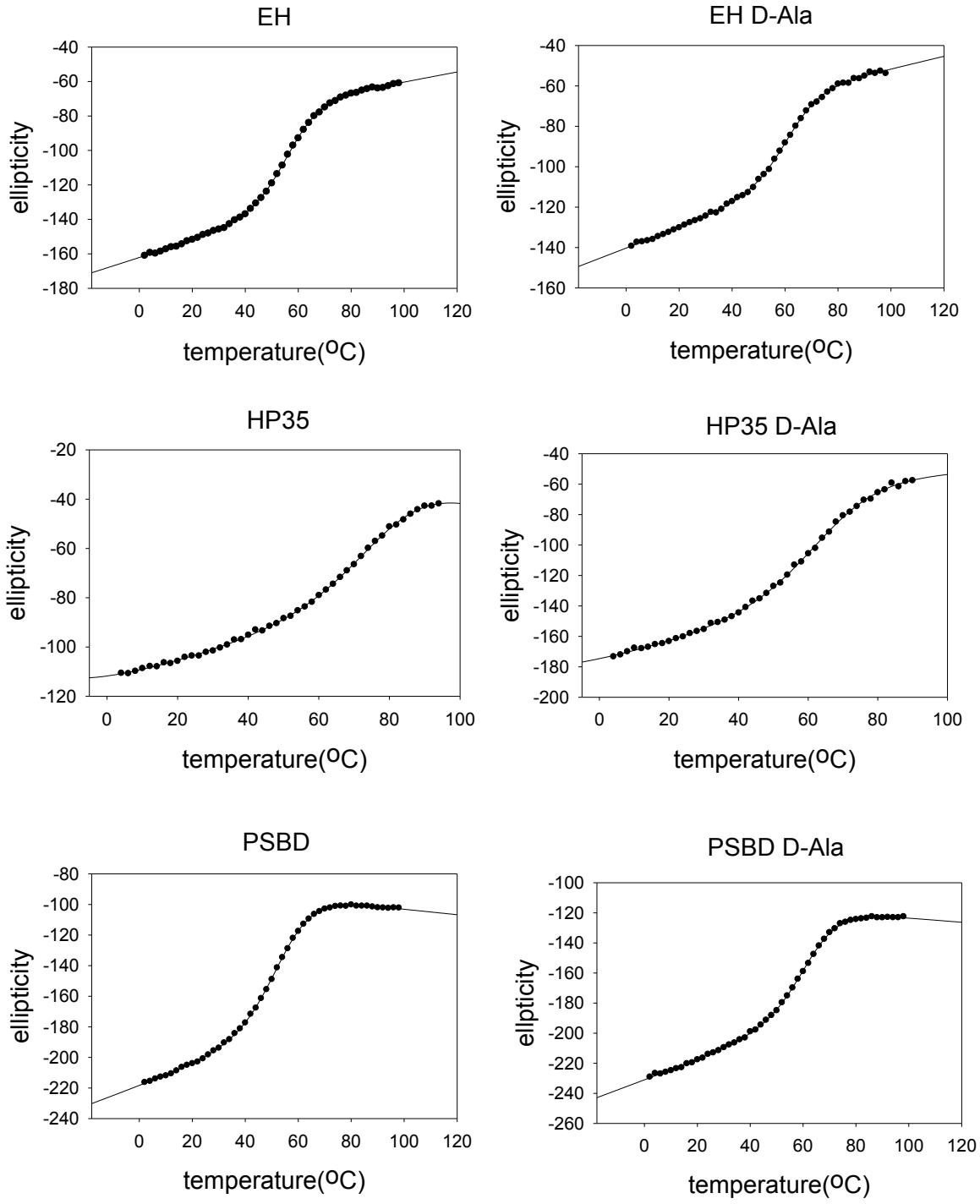
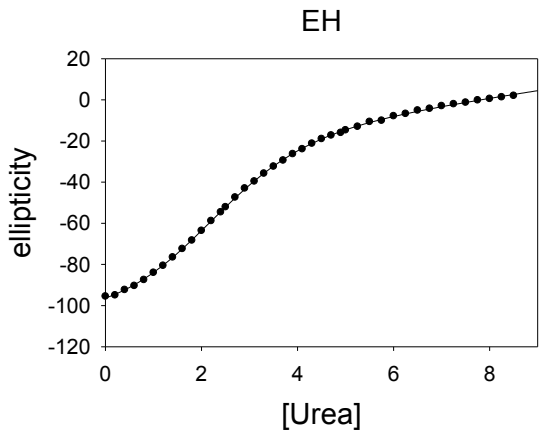
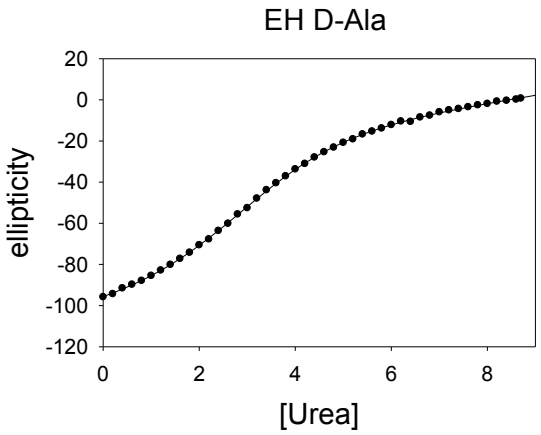
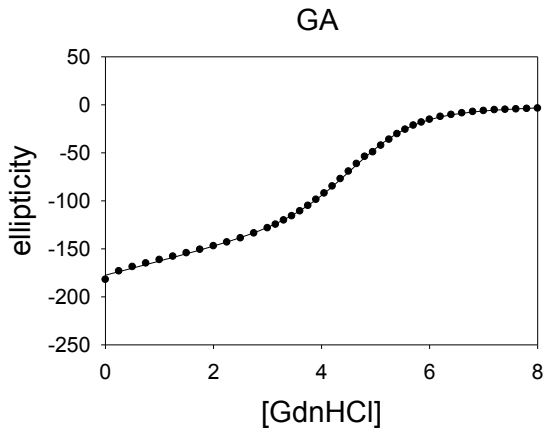


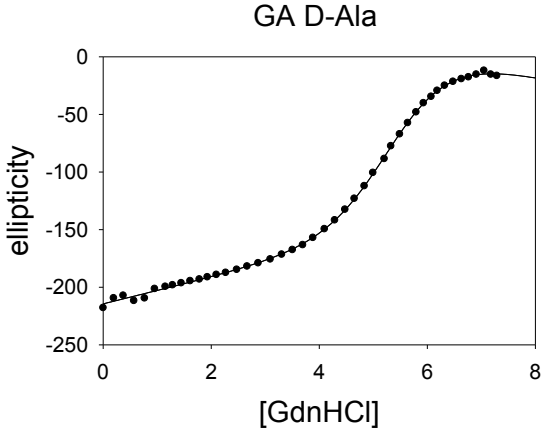
Figure S1. Thermal denaturation of EH, HP35, PSBD and their D-Ala variants. The solid line are the fitted curves.

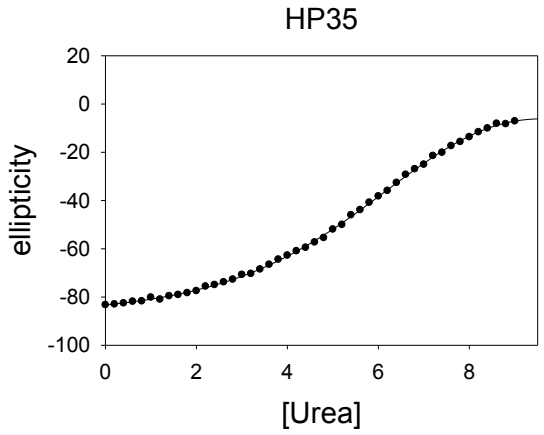
Urea/Guanidine Hydrochloride Denaturation of EH, GA, HP35, PSBD and Their D-Ala Variants

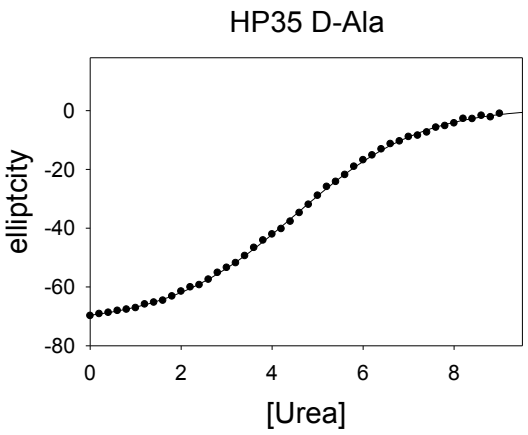

EH
ellipticity
20
0
-20
-40
-60
-80
-100
-120
0
2
4
6
8
[Urea]


EH D-Ala
ellipticity
20
0
-20
-40
-60
-80
-100
-120
0
2
4
6
8
[Urea]


GA
ellipticity
50
0
-50
-100
-150
-200
-250
0
2
4
6
8
[GdnHCl]


GA D-Ala
ellipticity
0
-50
-100
-150
-200
-250
0
2
4
6
8
[GdnHCl]


HP35
ellipticity
20
0
-20
-40
-60
-80
-100
0
2
4
6
8
[Urea]


HP35 D-Ala
elliptcity
0
-20
-40
-60
-80
0
2
4
6
8
[Urea]

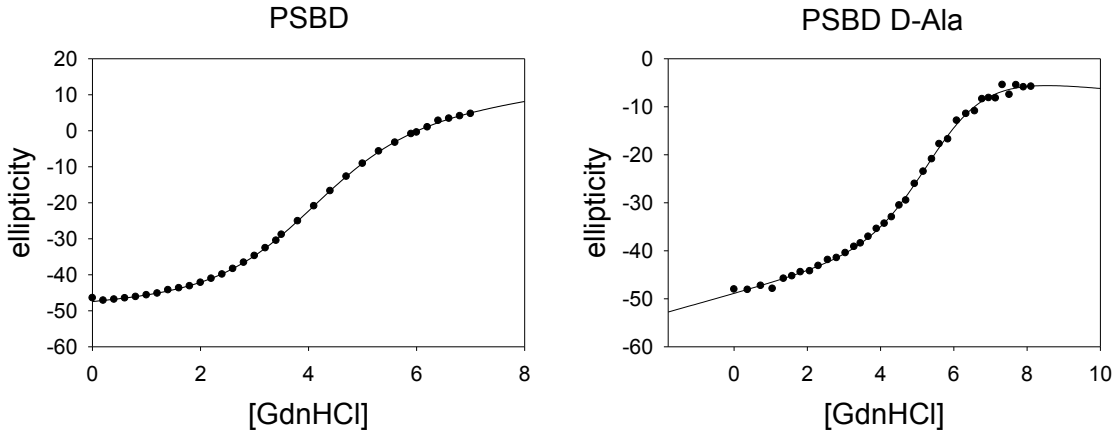


Figure S2. Urea/Guanidine hydrochloride denaturation of EH, GA, HP35, PSBD and their D-Ala variants. The solid lines are the fitted curves.

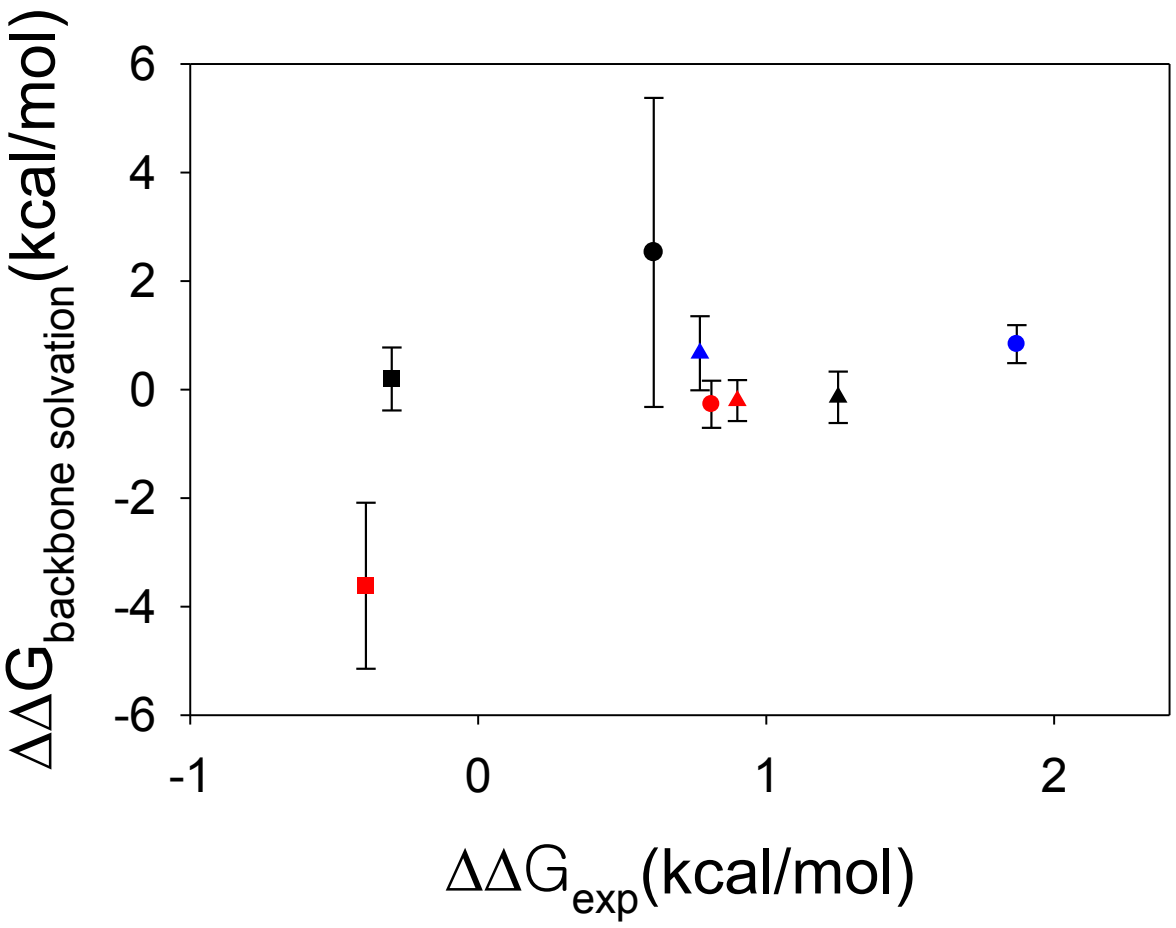


Figure S3. Correlation between $\Delta\Delta G_{backbone\ solvation}$ and $\Delta\Delta G_{exp}$. r=0.52, p=0.19. If only proteins with good convergence are included (GA, NTL9, PSBD, Trp-cage, UBA and ubiquitin), r=0.28, p-value=0.58, slope=0.20. EH ●; GA ●; HP35 ■; NTL9 ●; PSBD ▲; Trp-cage ▲; UBA ▲; Ubiquitin ■;

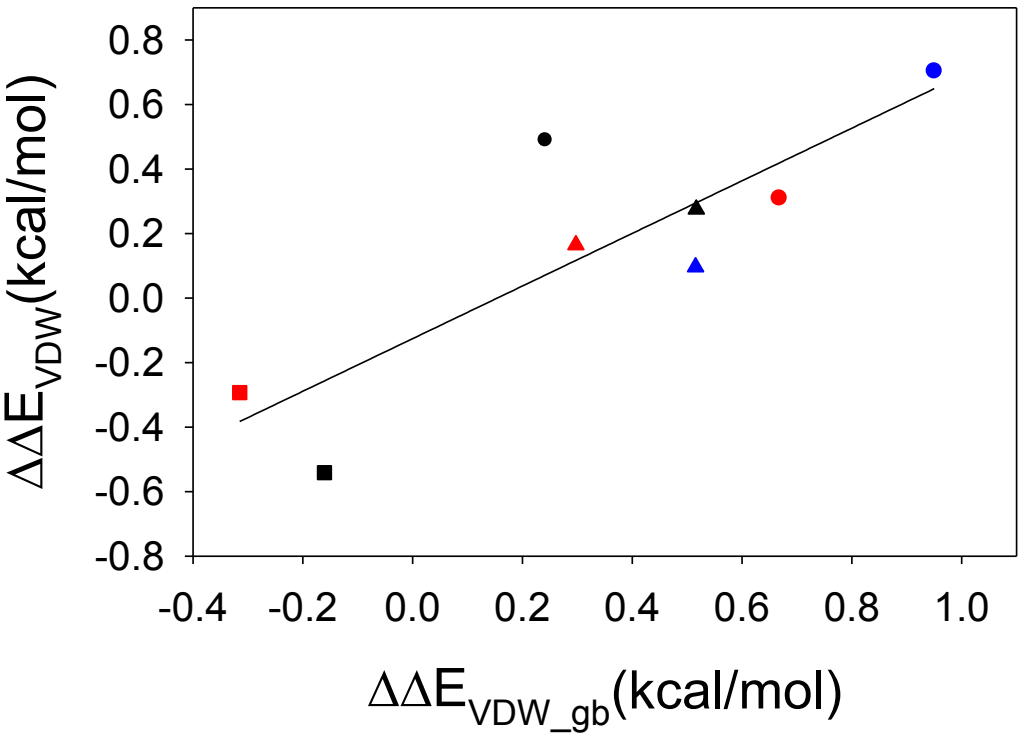


Figure S4. Correlation between $\Delta\Delta E_{vdw}$ and $\Delta\Delta E_{vdw\_gb}$. r=0.84, p=0.0079. EH ●; GA ●; HP35 ■; NTL9 ●; PSBD ▲; Trp-cage ▲; UBA ▲; Ubiquitin ■;

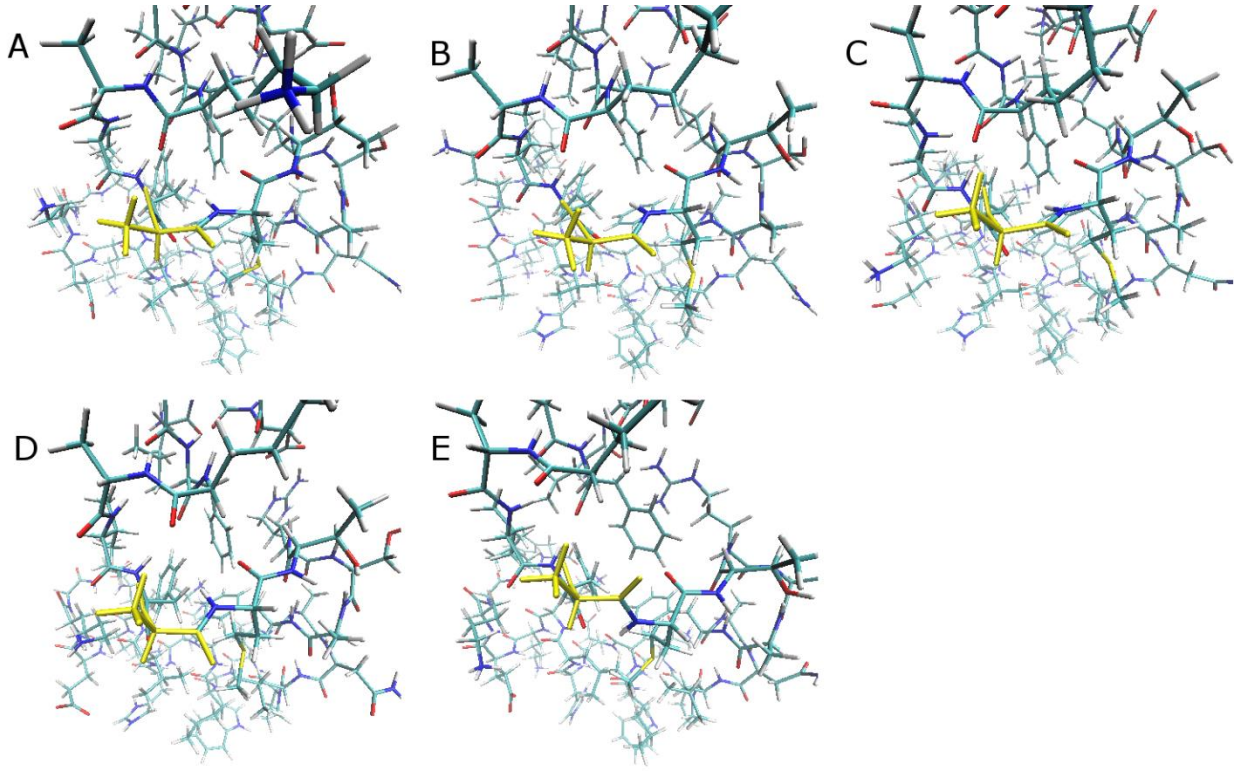


Figure S5. Structure of the HP35 G11D-Ala mutant taken from an MD simulation. 5 snapshots at 40 ns (A), 80 ns (B), 120 ns (C), 160 ns (D) and 200 ns (E) are shown with hydrogen included. The D-Ala residues are colored yellow.

1. Carpino, L. A.; Han, G. Y., *J. Am. Chem. Soc.* **1970,** *92*, 5748-&.
2. Humphrey, W.; Dalke, A.; Schulten, K., *J. Mol. Graph.* **1996,** *14*, 33-38, 27-38.
3. D.A. Case, J. T. B., R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K.M. Merz, G. Monard, P. Needham, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, R. Salomon-Ferrer, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, D.M. York and P.A. Kollman *AMBER 2015*. University of California, San Francisco, 2015.
4. Clarke, N. D.; Kissinger, C. R.; Desjarlais, J.; Gilliland, G. L.; Pabo, C. O., *Protein Sci.* **1994,** *3*, 1779-1787.
5. Johansson, M. U.; de Chateau, M.; Wikstrom, M.; Forsen, S.; Drakenberg, T.; Bjorck, L., *J. Mol. Biol.* **1997,** *266*, 859-865.
6. Chiu, T. K.; Kubelka, J.; Herbst-Irmer, R.; Eaton, W. A.; Hofrichter, J.; Davies, D. R., *Proc. Natl. Acad. Sci. U. S. A.* **2005,** *102*, 7517-7522.
7. Cho, J. H.; Meng, W.; Sato, S.; Kim, E. Y.; Schindelin, H.; Raleigh, D. P., *Proc. Natl. Acad. Sci. U. S. A.* **2014,** *111*, 12079-12084.
8. Kalia, Y. N.; Brocklehurst, S. M.; Hipps, D. S.; Appella, E.; Sakaguchi, K.; Perham, R. N., *J. Mol. Biol.* **1993,** *230*, 323-341.
9. Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H., *Nat. Struct. Biol.* **2002,** *9*, 425-430.
10. Withers-Ward, E. S.; Mueller, T. D.; Chen, I. S. Y.; Feigon, J., *Biochemistry* **2000,** *39*, 14103-14112.
11. Vijaykumar, S.; Bugg, C. E.; Cook, W. J., *J. Mol. Biol.* **1987,** *194*, 531-544.
12. Guex, N.; Peitsch, M. C., *Electrophoresis* **1997,** *18*, 2714-2723.
13. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., *J. Chem. Theory Comput.* **2015,** *11*, 3696-3713.
14. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., *J. Chem. Phys.* **1983,** *79*, 926-935.
15. Khoury, G. A.; Smadbeck, J.; Tamamis, P.; Vandris, A. C.; Kieslich, C. A.; Floudas, C. A., *ACS Synth. Biol.* **2014,** *3*, 855-869.
16. Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R., *J. Chem. Phys.* **1984,** *81*, 3684-3690.
17. Darden, T.; York, D.; Pedersen, L., *J. Chem. Phys.* **1993,** *98*, 10089-10092.
18. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., *J. Comput. Phys.* **1977,** *23*, 327-341.
19. Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A., *Nucleic Acids Res.* **2005,** *33*, W368-371.
20. Bang, D.; Gribenko, A. V.; Tereshko, V.; Kossiakoff, A. A.; Kent, S. B.; Makhatadze, G. I., *Nat. Chem. Biol.* **2006,** *2*, 139-143.
21. Kirkwood, J. G., *J. Chem. Phys.* **1935,** *3*, 300-313.
22. Roe, D. R.; Cheatham, T. E., 3rd, *J. Chem. Theory Comput.* **2013,** *9*, 3084-3095.
23. Li, L.; Li, C.; Sarkar, S.; Zhang, J.; Witham, S.; Zhang, Z.; Wang, L.; Smith, N.; Petukh, M.; Alexov, E., *BMC Biophys.* **2012,** *5*, 9.
24. Yamagishi, J.; Okimoto, N.; Morimoto, G.; Taiji, M., *J. Comput. Chem.* **2014,** *35*, 2132-2139.
25. Nguyen, H.; Maier, J.; Huang, H.; Perrone, V.; Simmerling, C., *J. Am. Chem. Soc.* **2014,** *136*, 13959-13962.
26. Nguyen, H.; Roe, D. R.; Simmerling, C., *J. Chem. Theory Comput.* **2013,** *9*, 2020-2034.
27. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., *J. Mol. Biol.* **1990,** *215*, 403-410.
28. Gunasekaran, K.; Nagarajaram, H. A.; Ramakrishnan, C.; Balaram, P., *J. Mol. Biol.* **1998,** *275*, 917-932.
29. Madej, T.; Lanczycki, C. J.; Zhang, D.; Thiessen, P. A.; Geer, R. C.; Marchler-Bauer, A.; Bryant, S. H., *Nucleic Acids Res.* **2014,** *42*, D297-303.