Supporting Information for: A Strategy for Proline and Glycine Mutations to Proteins with Alchemical Free Energy Calculations

Ryan L. Hayes[†] and Charles L. Brooks III[*,†,‡]

†*Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109, United States*

‡*Biophysics Program, University of Michigan, Ann Arbor, Michigan 48109, United States*

E-mail: brookscl@umich.edu

Phone: (734) 647-6682

The Soft Bond Potential

In free energy perturbation and other similar approaches, the soft bond need only be sufficiently smooth to ensure sufficient phase space overlap between the alchemical endpoint at $\lambda = 0$ and the neighboring λ window. In contrast, MSλD requires a smoother alchemical free energy because sharp changes near the endpoint can lead to trapping and because they can lead to errors due to the MSλD histogram based free energy estimator which includes points near the endpoint.

Consequently, we explored parameters of the soft bond potential, including using r_α of 1 Å or ∞ (which corresponds to a scaled hard bond), scaling the soft bond by λ^{n_α}, and scaling the angles and other bonded terms by λ^{n_θ}. The tests in the table below were performed on the unfolded ensemble of the Y25P proline mutation. The soft bond does not change the

endpoints, only the pathway between them, so ΔG should be the same for all simulations, but is systematically lower for simulations with linear bond or angle scaling ($n_\alpha = 1$ or $n_\theta = 1$). The systems with $n_\alpha = n_\theta = 2$ are more likely to be correct because their free energy profiles are smoother, as quantified by the smaller magnitude of the endpoint bias determined from ALF (ω). Simulations using soft bonds ($r_\alpha = 1$ Å) rather than hard bonds ($r_\alpha = \infty$) also appear to have improved transition rates and more accurate ΔG, as judged by their agreement with $n_\alpha = n_\theta = 2$ simulations. Because we wanted to used soft bonds to scale angles linearly with $n_\theta = 1$ as described in the next section, and because $n_\theta = 2$ lowers transition rates, $r_\alpha = 1$ Å, $n_\alpha = 2$, and $n_\theta = 1$ was chosen as a compromise between accuracy and efficiency.

Table S1: Soft Bond Optimization

r_α (Å)	Bond exp (n_α)	Angle exp (n_θ)	Transitions (1/ns)	ΔG (kcal/mol)	ω (kcal/mol)
1	1	1	7.895	33.94 ± 0.08	-7.80
1	2	1	8.090	34.13 ± 0.07	-5.84
1	2	2	7.265	34.20 ± 0.10	-1.42
∞	1	1	7.995	33.89 ± 0.05	-7.74
∞	2	1	7.740	34.04 ± 0.07	-5.80
∞	2	2	7.145	34.25 ± 0.09	-1.90

Restraints and Scaled Angles

The whole residue approach tightly restrains analogous atoms together. This is implemented in CHARMM with NOE restraints in the cons module. For a site with N_s interconverting residues, each atom is harmonically restrained to each of the other $N_s - 1$ analogous atoms with a harmonic force constant of $59.2/(N_s - 1)$ kcal/mol/Å^2, (where the harmonic potential has the standard prefactor of one half.)

Three angles through the Cα atom were unscaled, which is two more than allowed. Furthermore, in the whole residue strategy, two angles through N are also unscaled, which is one more than allowed. We responded to this problem in three ways, first by ignoring it

(which is not rigorous, but gives satisfactory results), second by scaling all but one of the offending angles by λ, and finally by scaling all of the offending angles by λ and adding an angle restraint between Cα-Cβ bonds.

Scaling all but one of the angles was done as follows. For the old side chain perturbation strategy, the Cβ-Cα-N and Cβ-Cα-Hα angles were scaled and the Cβ-Cα-C angle remained unscaled. For the new whole residue perturbation strategy, the Cβ-Cα-Hα and Cβ-Cα-C angles were scaled and the Cβ-Cα-N angle was unscaled, and the HN-N-Cα angle was also scaled, leaving the HN-N-C angle unscaled. This strategy was unsatisfactory because the Cα-Cβ bond sampled many unphysical orientations, which degraded the quality of the results.

In the third strategy, to ensure the Cα-Cβ bond remained in a physical orientation, the angle between each Cα-Cβ bond vector, and every other Cα-Cβ bond vector was harmonically restrained with GEO ANGLE (or a newly implemented GEO AANGLE) in the mmfp module of CHARMM, with a harmonic force constant of $59.2/(N_s - 1)$ kcal/mol/radian^2. This restraint took the role of the one allowed unscaled angle interaction, so all angles through Cα (and N in the whole residue strategy) were scaled. This maintained full physical rigor, but without the sampling of nonphysical Cα-Cβ bond orientations in the previous approach. This strategy gave indistinguishable results from the strategy of keeping all Cα angles unscaled, indicating that leaving these angles unscaled has a negligible effect on accuracy (Figure S1).

Scaling an angle A-B-C was implemented in CHARMM by declaring a soft bond between atoms A and C. While there is no bond between A and C, except possibly a Urey-Bradley interaction, which is treated with a soft bond, this ensures the A-B-C angle including both these atoms is scaled by λ.

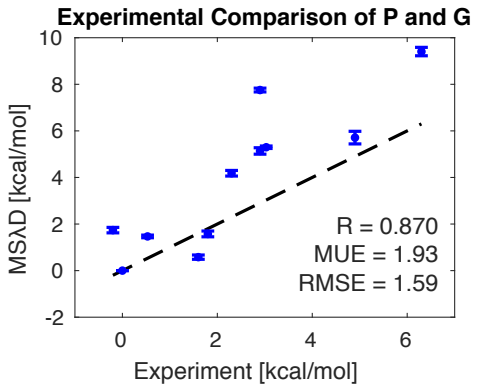

Figure S1: Comparison of MSλD with experiment for proline and glycine mutations, but with all three Cα angles unscaled. (See Main Text Figure 3 for results with the three Cα angles scaled.) While calculations with the angles scaled were run with $r_\alpha = 1$ Å, $n_\alpha = 2$, and $n_\theta = 1$ and scaled improper torsions, these calculations with the angles unscaled were run with $r_\alpha = \infty$, $n_\alpha = 2$, and $n_\theta = 2$ and unscaled improper torsions. The dashed line is $y = x$.

Table S2: Proline and Glycine Predictions (kcal/mol)

Mutation	Experiment	MSλD Unscaled Angles	MSλD Scaled Angles
Native	0.00	0.00 ± 0.00	0.00 ± 0.00
Y25P	2.30	4.18 ± 0.12	3.92 ± 0.11
L33G	2.90	7.75 ± 0.08	7.80 ± 0.11
P37G	-0.20	1.74 ± 0.12	1.68 ± 0.15
S44G	0.53	1.47 ± 0.06	1.38 ± 0.04
S44P	3.03	5.30 ± 0.05	5.91 ± 0.10
G56M	1.80	1.58 ± 0.12	1.36 ± 0.10
T59G	1.60	0.58 ± 0.09	0.40 ± 0.14
Q69P	2.90	5.14 ± 0.14	5.07 ± 0.16
L99G	6.30	9.41 ± 0.18	9.30 ± 0.06
V149G	4.90	5.71 ± 0.27	6.43 ± 0.27

Sources of Error

The errors relative to experiment were larger for the proline and glycine dataset than for the dataset lacking proline and glycine. While this is disappointing, it is not surprising. As mentioned in the main text, glycine mutations have been previously observed to give larger errors relative to experiment than other mutation types, and it is not surprising that prolines do as well. A further difficulty is that the $\Delta\Delta G$ values for this data were computed using at least three different equations from thermal unfolding at much higher temperatures, and reported for temperatures between 51.68 and 63.2 Celsius, whereas our simulations were performed at 300 K, or 26.85 C (Table S3). Using $\Delta G = \Delta H - T\Delta S$ to linearly extrapolate back to 300 K for the eight mutations for which the necessary data was reported gives markedly different $\Delta\Delta G$ values, especially in the case of the worst outlier, L33G. Thus we expect at least some of our error comes from using $\Delta\Delta G$ measured at a high temperature, which may not reflect $\Delta\Delta G$ at a lower temperature. This may also partially explain why the results were better for the dataset lacking proline and glycine: those mutations were mostly made at a lower pH with a lower unfolding temperature, so the $\Delta\Delta G$ values were not measured as far from 300 K; for example, sites 42 and 98 were measured at 40 Celsius. For this set of eight mutations and calculations with unscaled angles, the RMSE falls from 1.50 kcal/mol with the reported values to 1.41 kcal/mol with the extrapolated values, and more notably, the mean signed error, which captures systematic overprediction, falls from 1.73 kcal/mol to -0.47 kcal/mol.

The sites where large errors occur mostly make sense. Y25 and L33 are paired in a β sheet. S44 and Q69 are both fully solvent exposed in α helices. P37, G56, and T59 are surface exposed in a loop. L33, L99, and V149 are all deeply buried, and their mutation to glycine leaves a void that must be filled by water, structural rearrangement, or simply a buried cavity. Unsurprisingly, mutations in the surface loops have small effects and agree most closely with experiment. More surprisingly, for surface helix mutations, simulations seem to overestimate the effect of the mutation by 2 kcal/mol. Finally, our worst outliers are

for deeply buried mutations to glycine, and notably the two leucine to glycine mutations that remove four carbons are worse than the valine to glycine mutation, which only removes three carbons. Either experiments underestimate the large destabilizing effect of these mutations, or other slow degrees of freedom that aren't observable in the time scale of our simulations relax to accommodate the void.

Table S3: Extrapolating Experimental $\Delta\Delta G$ back to 300 K

Mutation	Reported $\Delta\Delta G$	Reported T (C)	Extrapolated $\Delta\Delta G$	pH
Y25P	2.3	60	4.61	5.4
L33G	2.9	60	8.13	5.4
P37G	-0.2	60	-0.32	5.4
S44G	0.53	51.68	0.79	3.0
S44P	3.03	51.68	5.30	3.0
G56M	1.8	60	3.88	5.4
T59G	1.6	63	-	6.5
Q69P	2.9	63.2	-	6.5
L99G	6.3	60	10.31	5.4
V149G	4.9	59	8.68	5.4

Controls

In order to demonstrate the theoretical validity of this strategy, Ramachandran distributions of the endpoint ensembles were examined, and closed thermodynamic cycles were constructed. These controls were performed with the three extra angles unscaled because this provides an upper bound on the error; simulations with the extra angles all scaled have one less source for possible artifacts.

The Ramachandran distributions for the pentapeptide modeling the P37G unfolded ensemble were examined (Figure S2) along with one dimensional free energy profiles along ϕ and ψ (Figure S3). In addition to proline and glycine, a third alanine substituent was added at the same site. Standard molecular dynamics (MD) simulations containing only one of the three residues were performed for 40 or 400 ns with five independent trials, the first quarters of the simulations were discarded for equilibration, and snapshots were saved every

10 ps. Analogously, simulations with the MSλD perturbation strategy were run with proline, glycine, and alanine all present, but λ variables were fixed to one of the three thermodynamic endpoints. Ramachandran distributions run for 40 ns showed deviation between MD and MSλD results due to noise caused by the small number of transitions between basins. Consequently, longer 400 ns simulations were run to allow the Ramachandran distributions to equilibrate fully. Due to the length of these simulations, they were run in the unpublished BLaDE software package, which can perform MSλD simulations more than five times faster than CHARMM on a GTX 1080 TI GPU. For these longer simulations, the Ramachandran distributions agree quite well.

Closed thermodynamic cycles are not important for error correction in MSλD as they are in other alchemical methods like free energy perturbation, because all perturbations can be compared in the same MSλD simulation without the need for a network of pairwise comparisons. Still, thermodynamic cycles are useful for highlighting possible theoretical artifacts or coding errors, and have been previously used with MSλD to highlight the need for soft core interactions (Hayes et al, JPC B 2017). Since the free energy change around a thermodynamic cycle should be zero to within statistical noise, deviations highlight potential artifacts. Consequently, closed thermodynamic cycles, from proline to glycine to alanine and back to proline were evaluated in the unfolded ensemble for the P37G mutation. Since the corresponding folded cycle is not considered, artifacts which might cancel out between the folded and unfolded cycles are still present, and this represents a stricter test of correctness. Convergence is also likely slower in the folded ensemble due to slow backbone relaxation, so focusing on the unfolded ensemble minimizes noise due to sampling issues. Even within the unfolded ensemble, convergence is still nontrivial.

After flattening, five independent trials were run for 40 ns, and after reoptimization of biases, another five trials of 40 ns were run. Results from both are reported in Table S4. It is readily apparent that bootstrapping significantly underestimates the uncertainty in these simulations because the two results differ from each other by substantially more than the

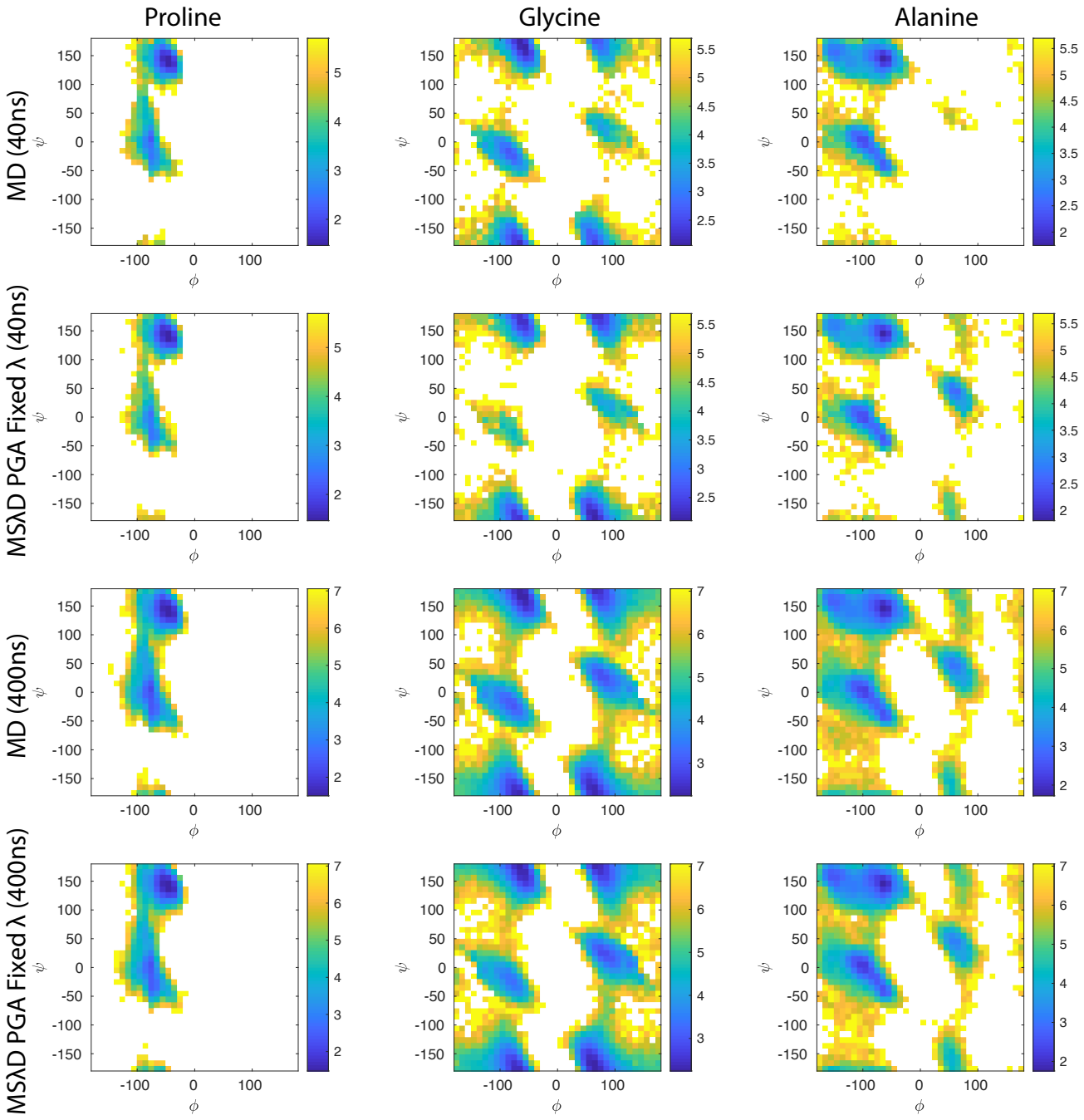

Figure S2: Ramachandran distributions for standard molecular dynamics (MD), and MSλD for proline, glycine, and alanine with a fixed λ state at one of the three thermodynamic endpoints (MSλD PGA fixed λ). Distributions show some deviation for 40 ns simulations due to small numbers of transitions between basins, but agree well after 400 ns of sampling. Color axis is in units of kcal/mol.

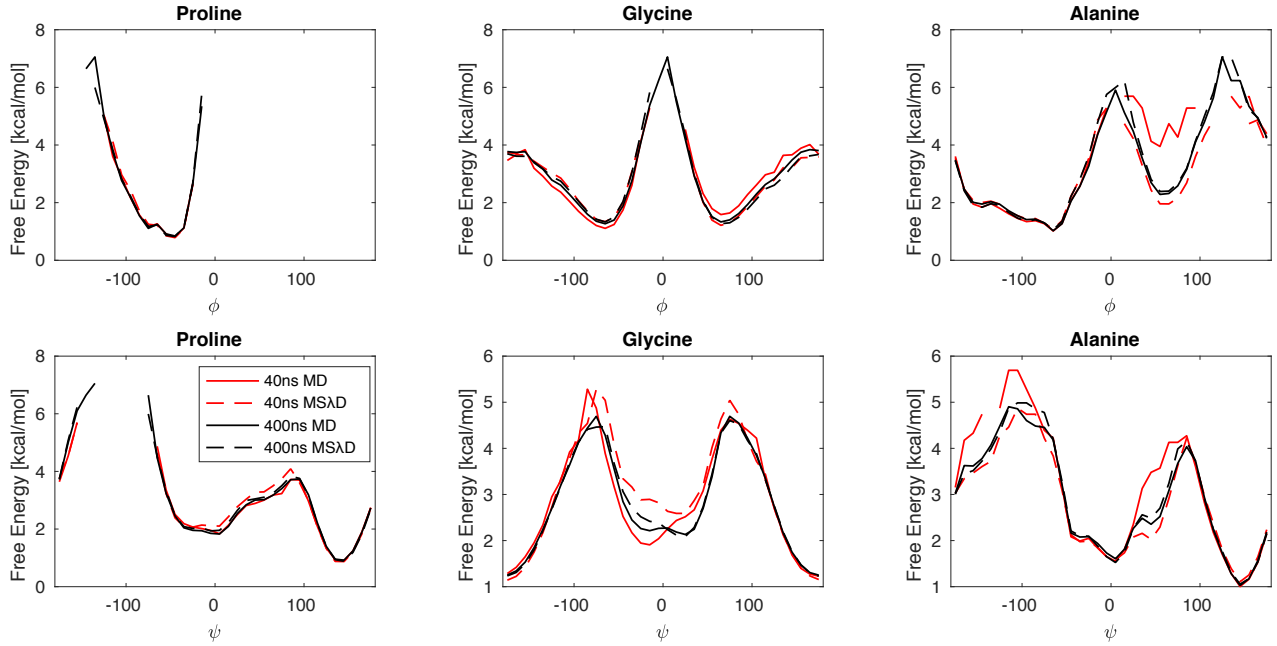

Figure S3: One dimensional ϕ and ψ backbone dihedral distributions for standard molecular dynamics (MD), and MSλD for proline, glycine, and alanine with a fixed λ state at one of the three thermodynamic endpoints (MSλD). Distributions show some deviation for 40 ns simulations due to small numbers of transitions between basins, but agree well after 400 ns of sampling.

estimated uncertainty. The majority of the inconsistency between the two runs can be traced to the proline to glycine perturbation. Inspection of the λ trajectories reveals two states: a state that is trapped in glycine and a state that rapidly transitions between proline and glycine (corresponding to positive and negative ϕ in the Ramachandran plots, respectively). Within the 150 ns used for free energy estimation (5 trials times the final 75% of a 40 ns simulation), only four transitions from one state to the other were observed, which explains the poor convergence. Consequently, the BLaDE software package was used to access longer time scales. After flattening, five trials of 40 ns were used to reoptimize the bias for five trials of 400 ns. By random chance, the BLaDE 40 ns runs were quite close to zero. In the 400 ns runs, slightly over 50 transitions between the states were observed, implying substantially improved convergence. While the result differs from zero by more than the bootstrapped uncertainty, it is closer, and the uncertainty is likely underestimated as it was in the 40 ns simulations. While the cycle closure error of 0.19 kcal/mol is quite small, we expect longer simulations would bring it even closer to zero. If they do not, the two best candidate explanations are the three unscaled angles, which could mildly perturb the ensemble, or the free energy estimator, which approximates the free energy of the endpoint by the population of states within a small λ distance from the true endpoint. Both of these effects might be mitigated by subtracting the corresponding folded cycle.

Table S4: Closed Thermodynamic Cycles (kcal/mol)

	ΔG_{PG}	ΔG_{GA}	ΔG_{AP}	$\Delta G_{\mathrm{cycle}}$
First 40 ns run	-55.514 ± 0.409	12.011 ± 0.072	42.501 ± 0.035	-1.002 ± 0.417
Second 40 ns run	-54.345 ± 0.078	12.147 ± 0.017	42.491 ± 0.051	0.293 ± 0.095
BLaDE 40 ns run	-54.740 ± 0.160	12.012 ± 0.052	42.548 ± 0.056	-0.180 ± 0.177
BLaDE 400 ns run	-54.720 ± 0.063	12.033 ± 0.025	42.495 ± 0.020	-0.192 ± 0.070

Given that longer 400 ns runs were required to obtain well converged results for Ramachandran distributions and closed cycles in the unfolded ensemble, it is likely longer simulations would improve upon the results presented in the paper. Indeed, preliminary 400 ns simulations of the ten proline and glycine mutations run in CHARMM and BLaDE

obtained RMSE values of 1.59 kcal/mol and 1.50 kcal/mol, respectively. While this represented a modest improvement of 0.1 to 0.2 kcal/mol over the results obtained with 40 ns simulations, we chose to present results from the 40 ns simulations both for consistency with the previous study of T4 lysozyme and because the modest improvement required a tenfold increase in computational effort.

Patching

In order to implement whole residue scaling, an elaborate set of CHARMM patches were designed. Patches are used in CHARMM to make changes to the topology of one or more residues, including adding a disulfide bond, protonating a titratable residue, or phosphorylating a residue. Patches were generated with a python script to avoid introduction of errors due to typos. Each patch included all atoms from a residue with new unique names, and bonds, impropers, and CMAP interactions between them and with the previous and next residue. Angles and dihedrals are automatically generated by CHARMM and need not be included in the patch, though nonphysical angles and dihedrals generated between different patches must be removed. Additional patches are employed to duplicate the CMAP and C improper from the the previous residue and the CMAP and N improper from the next residue, which should interact in a scaled fashion with each copy of the perturbed residue.

While only single mutations were considered in this work, this approach has already been used within our group for applications with multiple mutations, which requires additional modifications. If two consecutive residues are mutating, additional CMAP, improper, and bond patches are required to link each copy of the neighboring residues, and if two residues separated by a third residue are mutating, or if three consecutive residues are mutating, additional CMAP only patches are required. This ensures that even though the CMAP terms are scaled by the product of all three λ values, they always add up to the equivalent of one and only one effective CMAP interaction. These patches are all applied appropriately

and automatically by a CHARMM script posted online.

Adaptive Landscape Flattening

In a previous study of T4 lysozyme, adaptive landscape flattening (ALF) was run for 50 iterations of 100 ps, 10 iterations of 1 ns, 3 iterations of 5 ns, and then production simulations of 40 ns (or 20 ns when using variable bias replica exchange) were run until a converged result was obtained. A result was considered converged if the range of changes in the fixed bias ($\Delta\phi$) was less than $3kT$, the minimum $\Delta\phi$ was greater than $-2kT$, and the uncertainties of the native and non-native ΔG were less than 0.3 and 0.5 kcal/mol respectively. If these criteria were not met, the biasing potential was reoptimized based on the production run, and another production run was launched.

In this study the same approach was used for the mutations excluding proline and glycine, except instead of 3 sequential 5 ns simulations, a short production of 5 independent trials of 5 ns simulations run in parallel was used. This slightly increased computational cost in GPU-hours, but decreased the wait time to obtain results by 10 ns of sampling. Following this the full production simulation was run. In some cases the A98 site was stopped before production simulations had reached the above convergence criteria because it was judged unlikely results would improve by running it yet again.

The proline and glycine mutations were only considered in pairs, which reduced the level of noise substantially, but still had large noise in some cases. The flattening strategy consisted of 100 iterations of 100 ps, 10 iterations of 1 ns, and a production run of 5 independent trials of 5ns. To reduce confusion about the number of production runs, two production runs of 5 independent trials of 40 ns were performed, though only one or two sites needed the second run for convergence, and results from the second run were reported. In all cases, results were not compared to experiment until after they were judged to have converged to avoid biasing the results artificially.