Biophysical Journal, Volume 96

Supporting Material

Force field bias in protein folding simulations

Peter L. Freddolino, Sanghyun Park, Benoit Roux, and Klaus Schulten

Supplementary material for "Force field bias in protein folding simulations"

Peter L. Freddolino, Sanghyun Park, Benoit Roux, and Klaus Schulten

April 8, 2009

1 Cluster analysis of folding trajectories

Cluster analysis of the folding trajectories was performed using the `g_cluster` module of GROMACS 3.3 [1] with the gromos clustering method [2]. For general analysis of the folding trajectories, RMSDs were calculated using all heavy atoms except those which are chemically equivalent to another atom in the same residue (as in [3]), with a cutoff chosen based on the distribution of pairwise frame-frame RMSDs for each selection. Frames for clustering were taken once every 300 ps for each folding trajectory. The selected cutoffs were 3.5 Å, 3.5 Å, 2.8 Å, and 3.5 Å for SimFold_1, SimFold_2, SimFold_3, and SimFold_4, respectively (*n.b.* the cutoffs described here are not the same as the cutoffs used to define the DM ensembles, which are described separately in Section 3). Distributions of the sizes of the first several clusters are shown in Fig. 1, and the cluster present over time in each trajectory is shown in Fig. 2; the first 20 clusters represent 36.8%, 27.4%, 39.7%, and 65.4% of the total set of clusters for the three folding simulations. In all four cases new clusters do appear throughout the entire duration of the simulation (data not shown). Representative conformations (those with the lowest average pairwise RMSD to other members of the same cluster) for each cluster are shown in Figs. 3 and 4. While the use of a less strict RMSD cutoff for identification of clusters might lead to a smaller number of clusters and reduced appearance of further conformations throughout the simulations, the qualitative conclusion that the WW domain explores a variety of distinct misfolded conformations throughout the length of each folding trajectory is unaffected by the exact cutoff used. For example, over the last two microseconds of SimFold_1 the protein fluctuates between clusters 2, 5, and 7, which are representative of the HELIXU, HELIXV, and HELIXL states, respectively.

During identification of representative conformations for use in deactivated morphing, all-protein-atom RMSDs were used in place of the heavy atom metric described above, and a uniform cutoff of 4.0 Å was applied. As described in Section 3, DM ensembles were considered using 2.0 Å, 3.0 Å, and 4.0 Å cutoffs around the selected reference states.

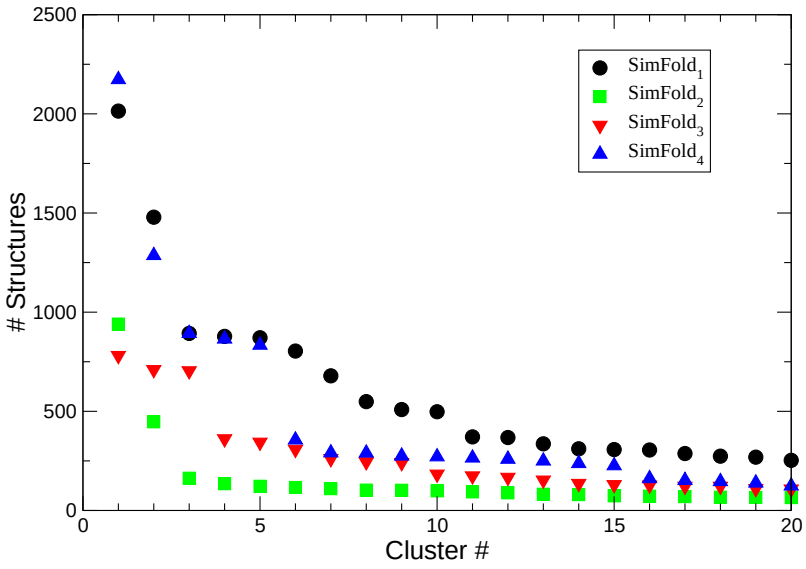

Figure 1: Number of conformations contained in the top 20 clusters of each folding simulation.

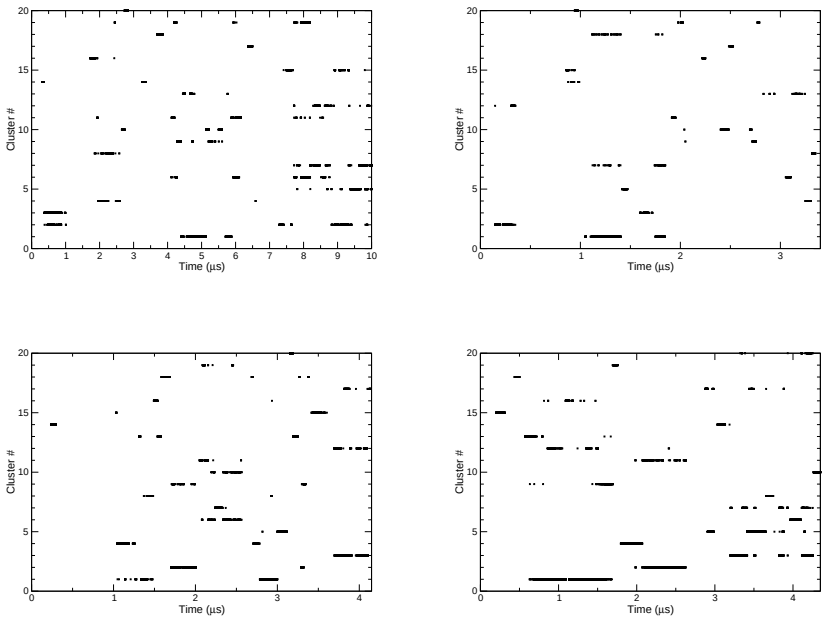

Figure 2: Cluster occupied as a function of time through each folding trajectory. Only the 20 most highly occupied clusters are shown. Clockwise from top left: SimFold_1, SimFold_2, SimFold_4, SimFold_3.

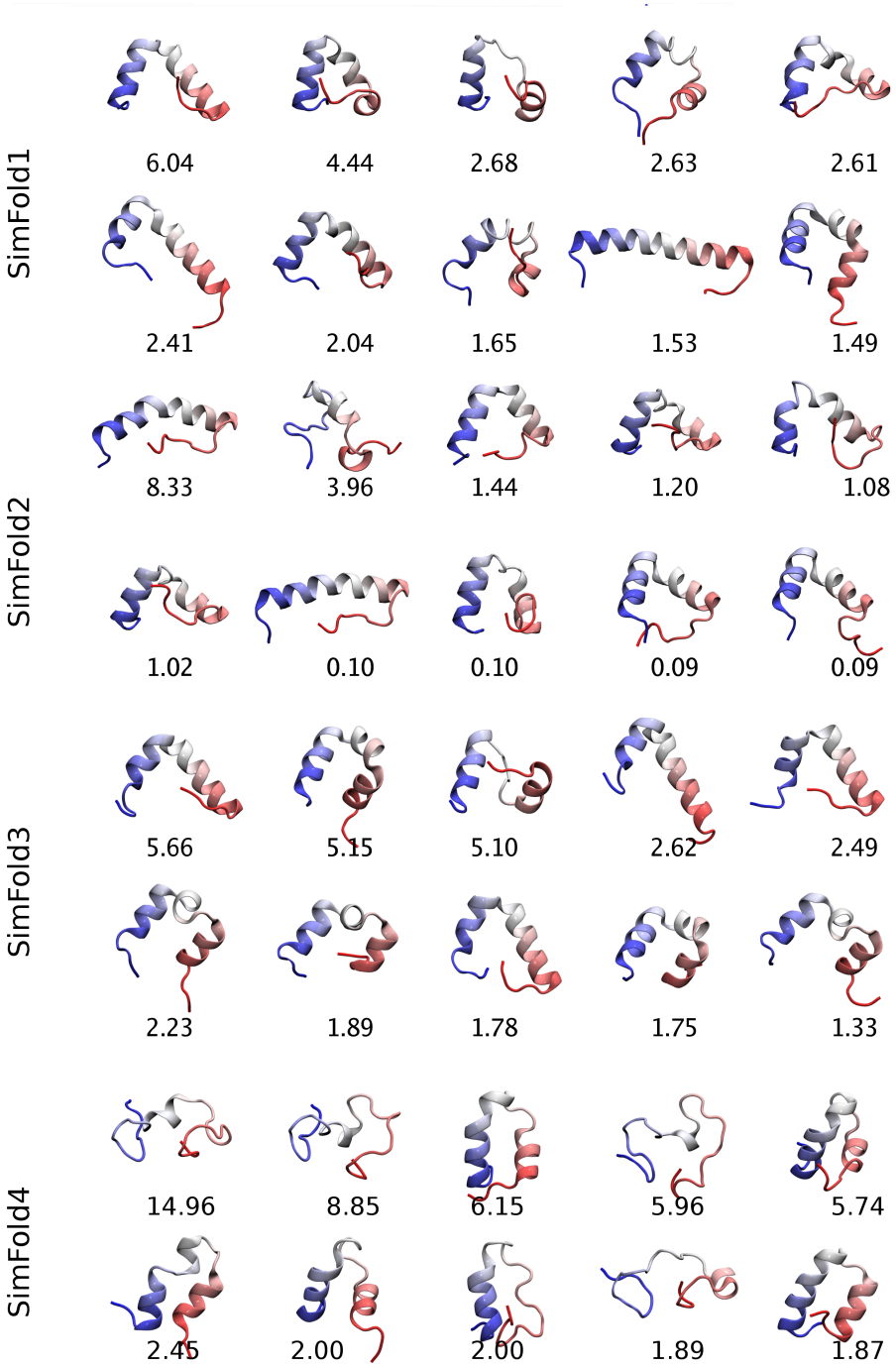

Figure 3: Cartoon representations of the 10 most highly occupied clusters from folding simulations. Coloring runs blue to red from N-terminus to C-terminus. The percentage of timesteps from a given trajectory falling in each cluster is also shown.

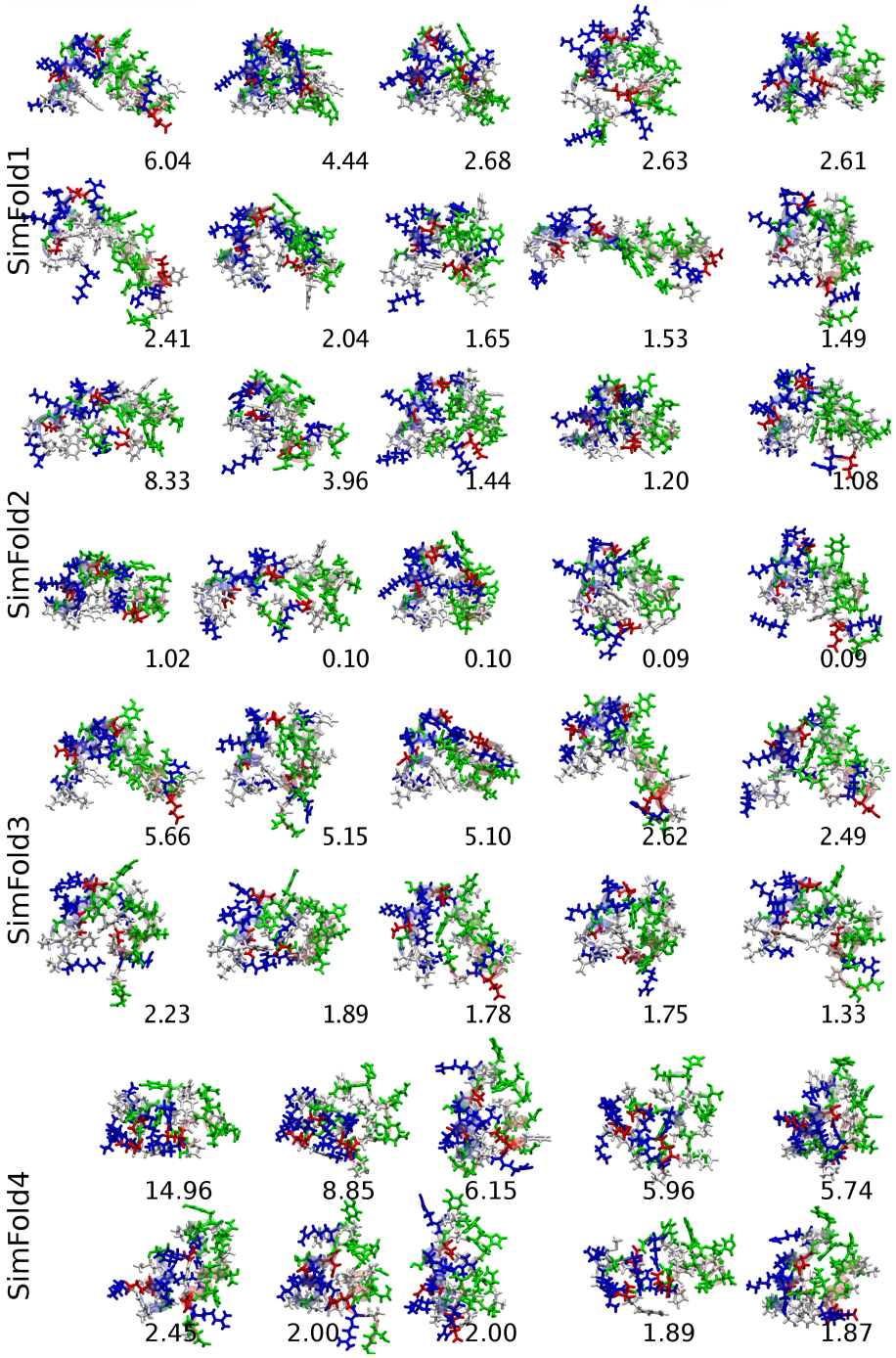

Figure 4: Licorice representations of the 10 most highly occupied clusters from folding simulations. Residues are colored by type: white for hydrophobic, green for polar, red for negatively charged, and blue for positively charged. Coloring of the cartoon backbone runs blue to red from N-terminus to C-terminus. The percentage of timesteps from a given trajectory falling in each cluster is also shown.

2 Hydrogen bonding analysis

To analyze the distribution of hydrogen bonding conformations for comparison with data from structural databases [4] and QM calculations [5], the distribution of the hydrogen-acceptor distance δ_{HA}, donor-hydrogen-acceptor angle Θ, hydrogen-acceptor-acceptor antecedent angle Ψ, and dihedral about the acceptor-acceptor antecedent bond X were calculated; the notation and definitions used here are those of Kortemme and coworkers [4]. For each of the four reference conformations, 2000 evenly spaced frames were taken from the 20 ns trajectory for the intermediate restraint state κ_{70} (k=0.001 kcal/(mol Å^2)) and hydrogen bonds identified based on a donor-acceptor distance cutoff of 3.5 Å and Θ angle cutoff of no more than 80 degrees from collinearity. Backbone-backbone, backbone-sidechain, and sidechain-sidechain interactions were binned separately; the resulting distributions, after a correction for bin volumes, are shown in Figs. 5, 6, 7, and 8. When compared to the results from a survey of crystallographic structures presented in Fig. 2 of Kortemme *et al.* [4], the distributions observed in our simulations show an overpopulation of values of δ_{HA} greater than 2.2 Å, values of Ψ below 110° and above 160°, and values of Θ less than 120°.

Because we noted that the inclusion of hydrogen bonds with Θ values deviating more than 40° from 180° introduced a significant population of hydrogen bonds with large (>2.2 Å) values of δ_{HA} and values of $\Psi < 100°$, data are plotted separately for a reduced set of hydrogen bonds with Θ no more than 35 degrees from linearity in Figs. 9, 10, 11, and 12. In this reduced set the distributions of δ_{HA} and Θ are much closer to those extracted from crystal structures, but the backbone Ψ angle distribution is still shifted closer to linearity for all cases, particularly SHEET, and an overpopulation of Ψ values greater than 140° is observed for sidechain-sidechain interactions. Examining the subset of sidechain-sidechain hydrogen bonds with $\delta_{\mathrm{HA}} < 2.1$ Å, unlike the results of Kortemme and coworkers, does not significantly alter the Ψ angle distribution (data not shown). The overpopulation of Ψ angles closer to 180°, rather than the ideal geometry of 110° expected from *ab initio* calculations on formamide dimer [5], would be expected for a pure dipole treatment of hydrogen bonding [4].

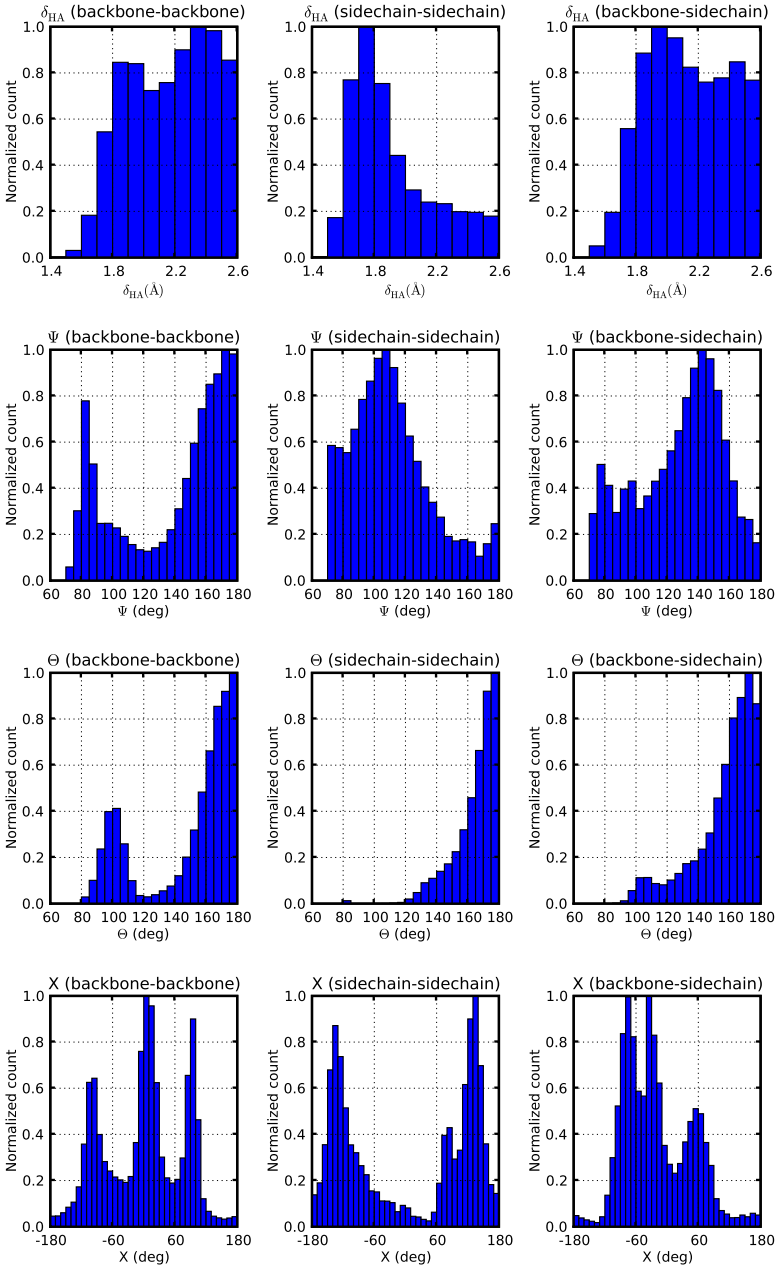

Figure 5: Distribution of hydrogen bond parameters observed during simulation of SHEET.

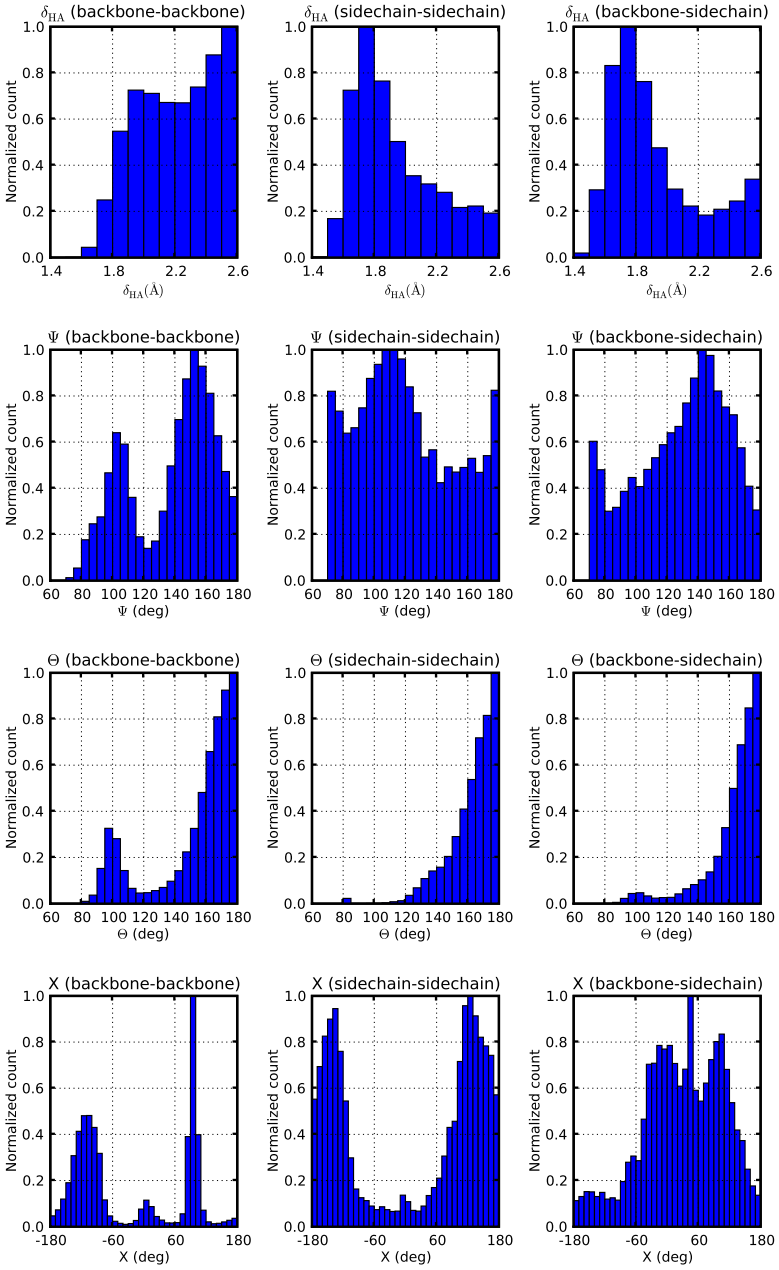

Figure 6: Distribution of hydrogen bond parameters observed during simulation of HELIXL.

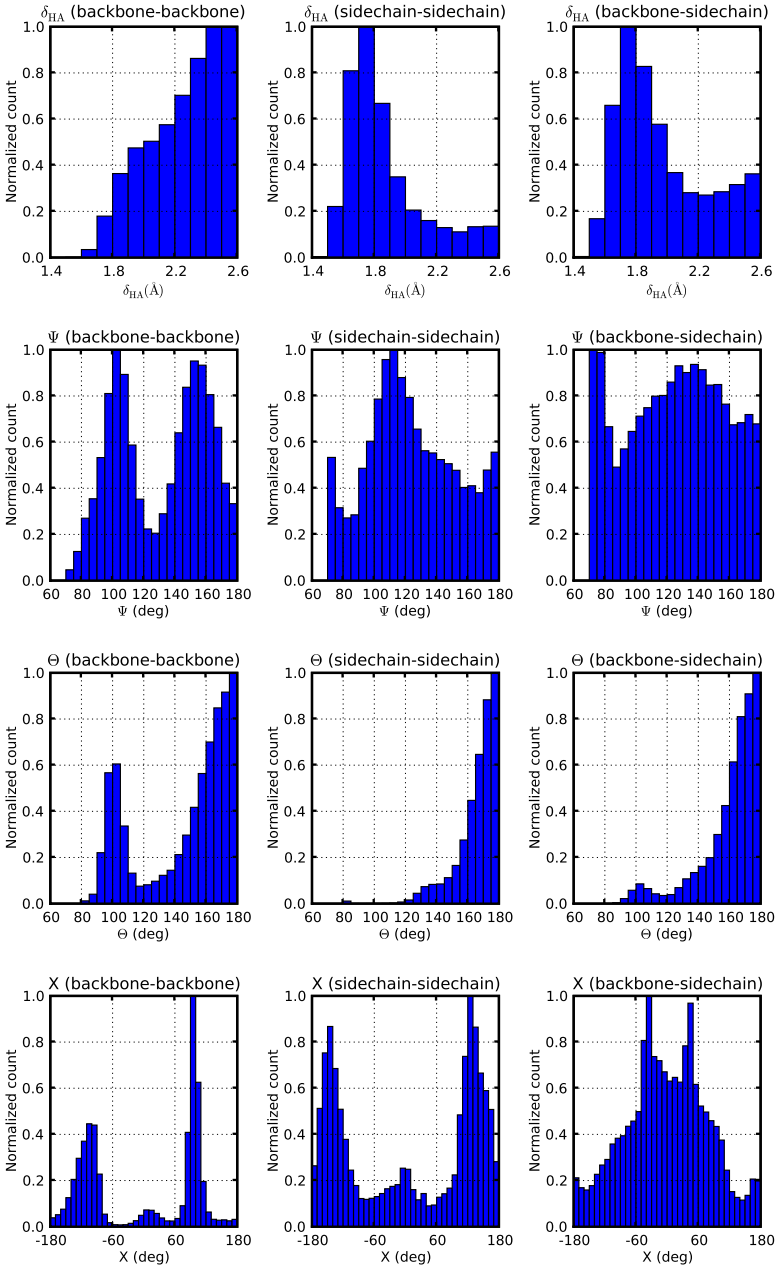

Figure 7: Distribution of hydrogen bond parameters observed during simulation of HELIXU.

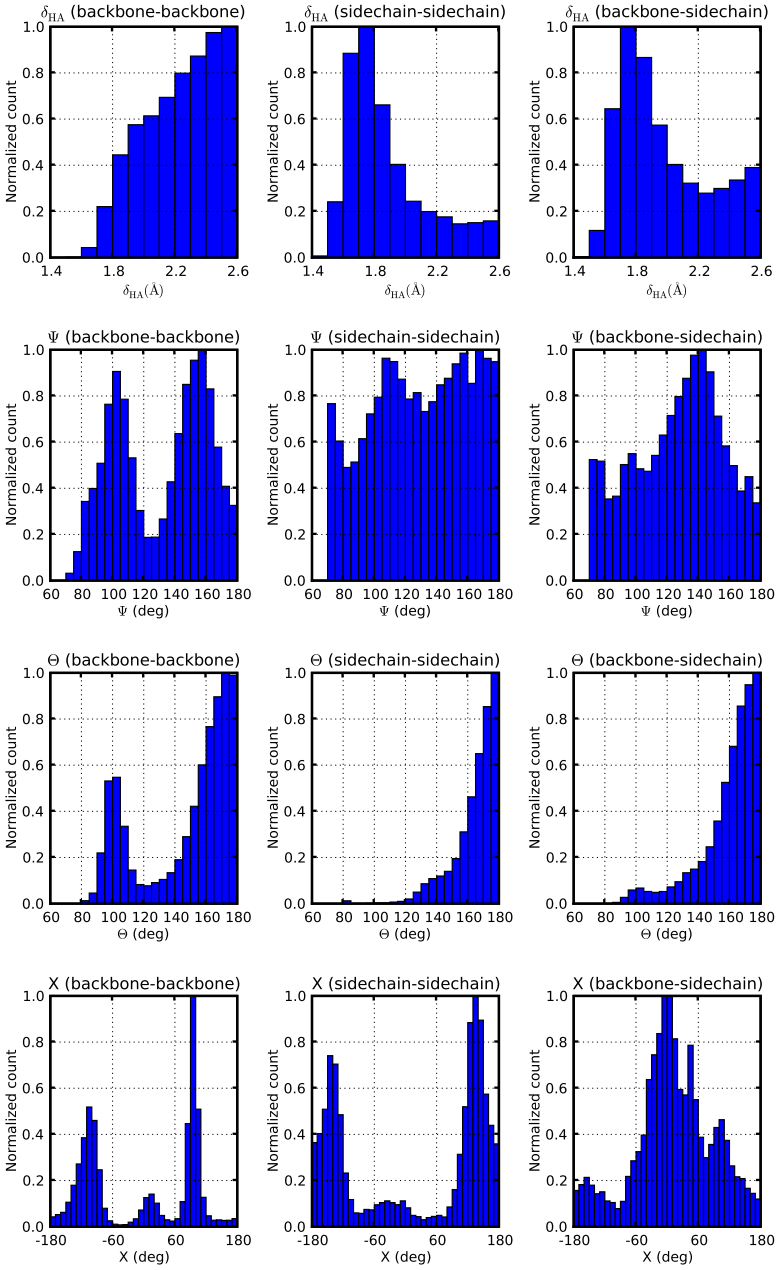

Figure 8: Distribution of hydrogen bond parameters observed during simulation of HELIXV.

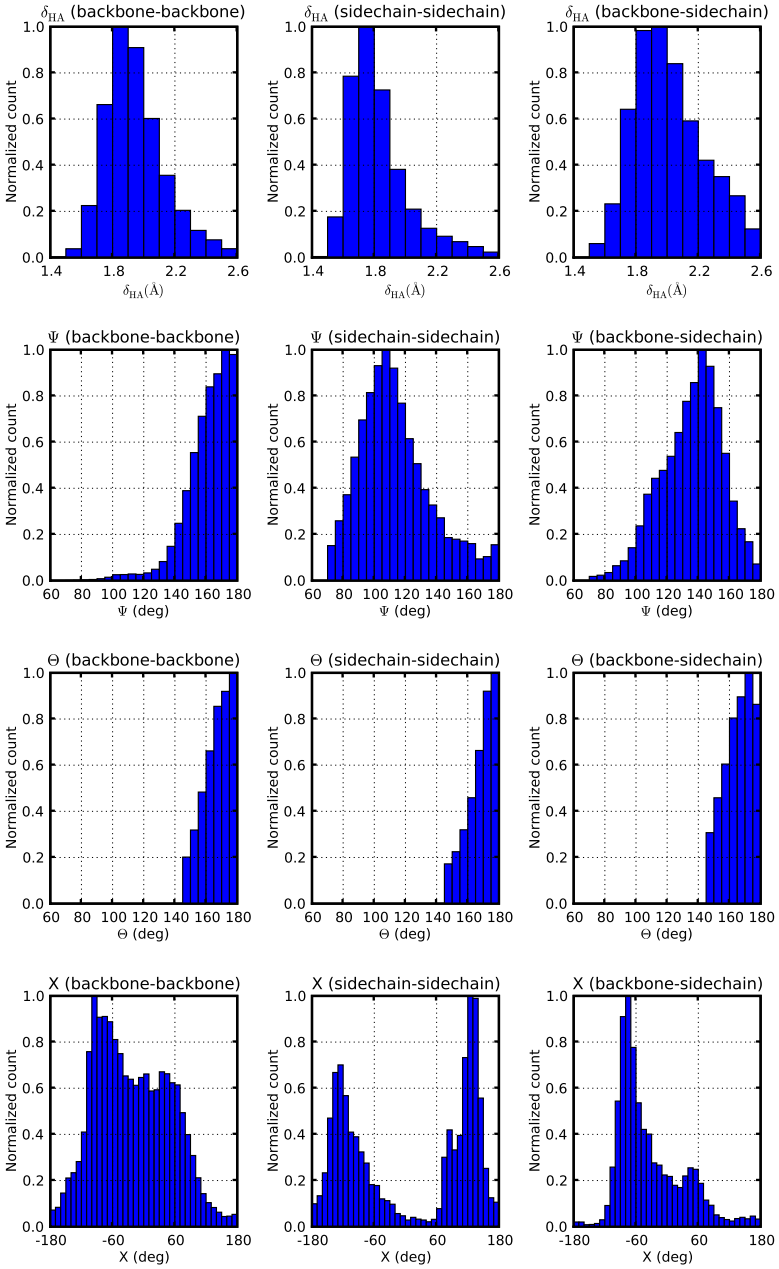

Figure 9: Distribution of hydrogen bond parameters observed during simulation of SHEET, taking only hydrogen bonds with $\Theta > 145.0°$.

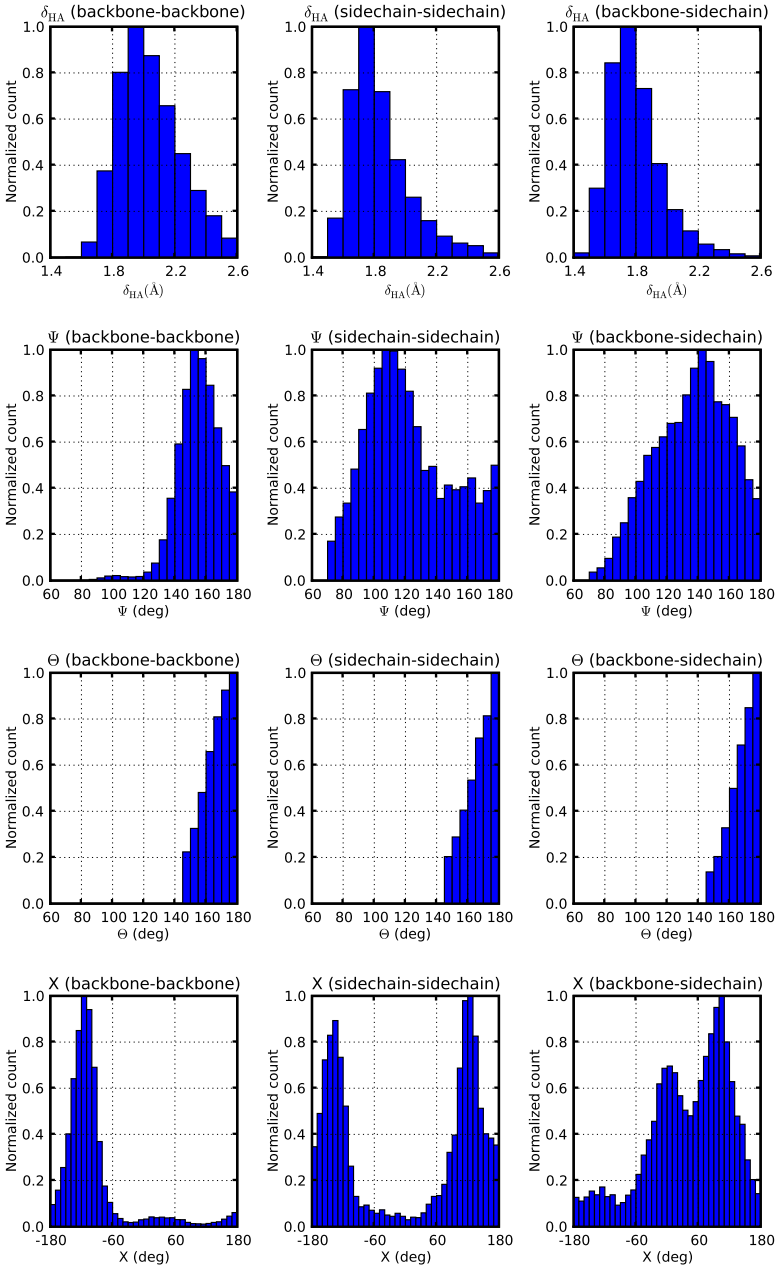

Figure 10: Distribution of hydrogen bond parameters observed during simulation of HELIXL, taking only hydrogen bonds with $\Theta > 145.0°$.

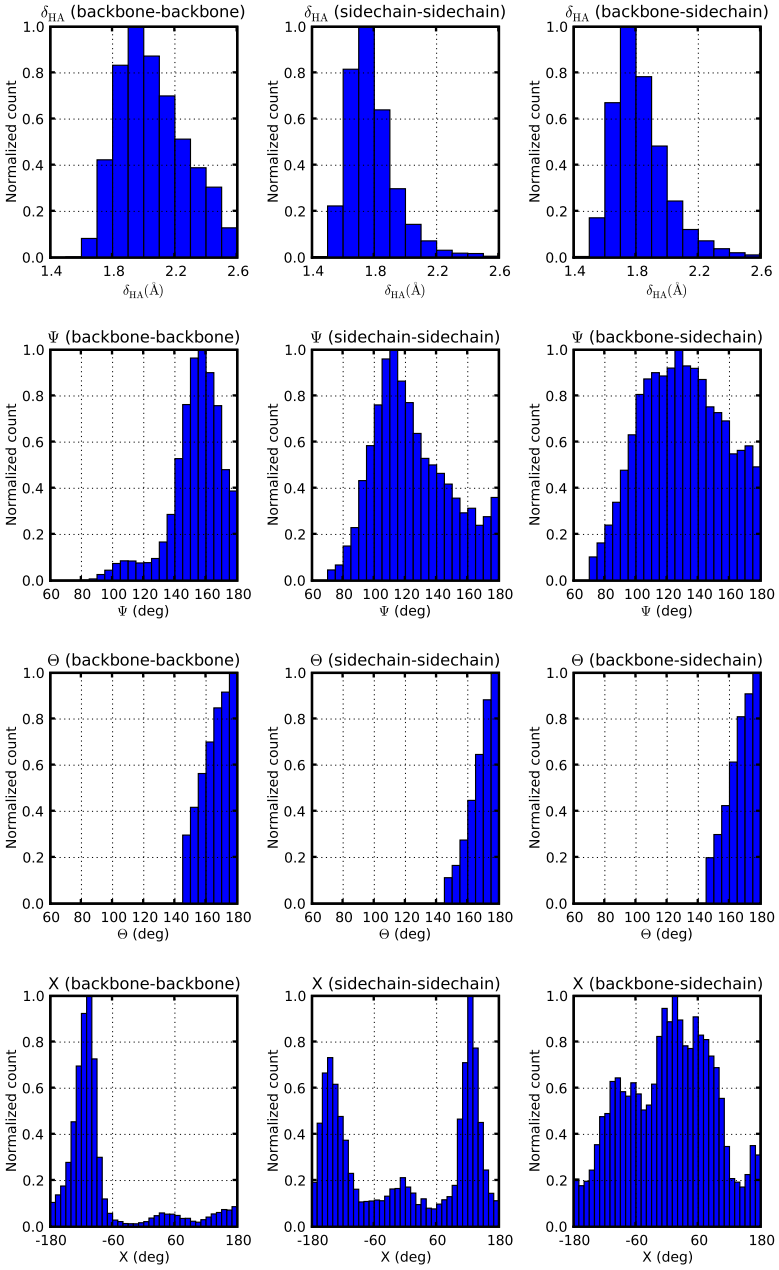

Figure 11: Distribution of hydrogen bond parameters observed during simulation of HELIXU, taking only hydrogen bonds with $\Theta > 145.0°$.

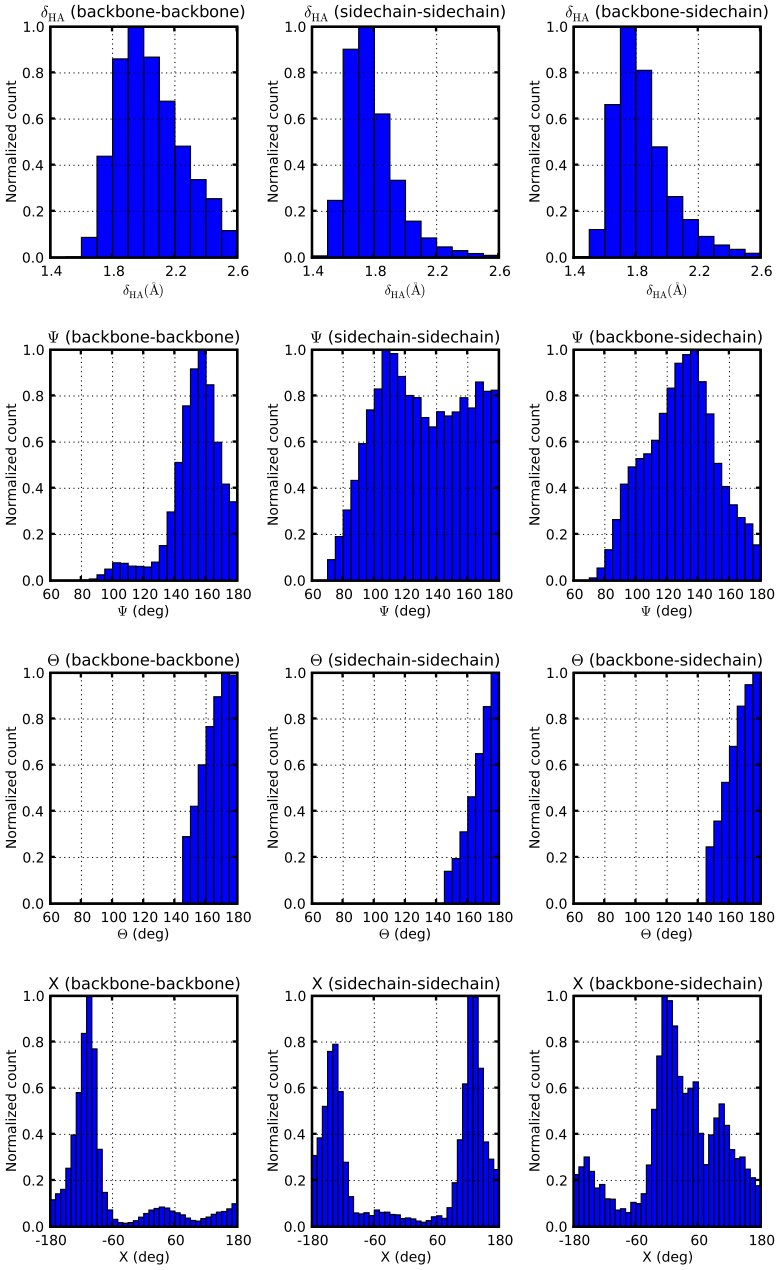

Figure 12: Distribution of hydrogen bond parameters observed during simulation of HELIXV, taking only hydrogen bonds with $\Theta > 145.0°$.

3 The deactivated morphing procedure

The deactivated morphing procedure used here was identical to that presented by Park *et al.* in [6] except as noted below. This procedure circumvents the usual difficulty in calculation of free energy differences for nontrivial conformational changes of a protein by utilizing a series of unphysical intermediate states in which the protein is restrained to a reference conformation representative of the conformational ensemble of interest, all internal interactions removed, and the protein is then "morphed" to a set of coordinates corresponding to a different reference conformation of interest, and the first two steps then reversed. The stages involved in the DM calculations performed in this study are illustrated in Fig. 14; relevant details of each stage of the calculation are given below. As noted in the main text, for a given reference conformation A the intermediates represent the ensemble of structures within a specified protein RMSD cutoff of the reference (E(A)), the state with harmonic restraints applied to all protein atoms restraining it to conformation A with κ=1000 kcal/mol Å^2 (K$_1$(A)), the "deactivated" state with all protein atoms restrained to their coordinates in reference state A (Q(A)), and a "dummy" state with a uniform set of van der Waals parameters and charges applied (D(A)).

Candidate reference structures were initially identified by clustering analysis as described in the main text, and then subjected to 50 ns of unrestrained MD. The single conformation from simulation of each candidate reference state with the lowest RMSD to all other frames was then minimized for 6000 steps, and the resulting structure used as a DM reference conformation.

All free energy calculations were performed using the Bennett Acceptance Ratio (BAR) [7], in which the free energy difference between two related systems with slightly different potential energy functions U_1 and U_2 is calculated by sampling L_1 and L_2 conformations using U_1 and U_2, respectively, and then self-consistently solving

$$e^{\beta\Delta G} = \sum_{l=1}^{L_1+L_2} \left[L_1 e^{-\beta\Delta G} + L_2 e^{-\beta\Delta U(\mathbf{R}_l)} \right]^{-1} \tag{1}$$

for ΔG, with $\Delta U \equiv U_2(\mathbf{R}) - U_1(\mathbf{R})$ and $\beta \equiv \frac{1}{\mathrm{k_B T}}$. Here $\mathbf{R}$ includes all protein, water, and ion coordinates. Equation 1 is used for all steps of deactivated morphing and analysis of the effects of different potentials on the relative free energies on conformations, except for calculation of the E(A) $\rightarrow$ K$_1$(A) free energy difference, which was estimated as in [6]:

$$e^{-\beta(F_E - F_{K_S})} = \sum_{l=1}^{L_1+\ldots+L_S} \Theta(\mathbf{X}_l) \left[\sum_{i=1}^{S} L_i \frac{e^{-\beta\frac{\kappa_i}{2}(\mathbf{X}_l - \hat{\mathbf{X}})^2}}{e^{-\beta(F_{\kappa_i} - F_{\kappa_S})}} \right]^{-1} \tag{2}$$

with L_i the lengths of the intermediate restraint simulations with spring constants κ_i, $[\mathbf{X}_1, ..., \mathbf{X}_{L_1}]$ the set of conformations from state K$_1$, and $\Theta(\mathbf{X}) \equiv 1$ if the RMSD between $\mathbf{X}$ and the reference conformation $\hat{\mathbf{X}}$ is less than a specified RMSD cutoff, and 0 otherwise. For the DM calculations all of the restraining

states K_i were used in calculation of $G_E - G_{K_S}$, with RMSD cutoffs of 2.0, 3.0, and 4.0 Å.

The effects of modifying nonbonded interactions and the force field's CMAP terms were also calculated. In the case of nonbonded interactions, 20 ns simulations were performed with cutoffs of 8.0, 9.0, 10.0, 12.0 Å. Likewise, to assess the effects of CMAP on different conformations, 20 ns simulations were performed with the CMAP grid contribution scaled by 1.0, 0.8, 0.5, 0.3, 0.2, 0.1 and 0.0. The full set of states used for these calculations is shown in Fig. 13. For all of these calculations a spring constant of 0.001 kcal / (mol Å^2) was applied to restrain the simulation to the reference conformation; in these cases only one state was used to calculate the transition to the unrestrained ensemble, and thus Eq. 2 simplifies to

$$e^{-\beta(F_E - F_K)} = \sum_{l=1}^{L} \Theta(\mathbf{X}_l) L^{-1} e^{\beta \frac{\kappa}{2} (\mathbf{X}_l - \hat{\mathbf{X}})^2} \tag{3}$$

Only a 4.0 Å cutoff for the unrestrained ensemble was used for the modified potential cases; in all cases the RMSD distribution for the the restrained ensemble significantly populated the region around 4.0 Å. Simulation parameters were the same as those for the DM calculations unless otherwise noted.

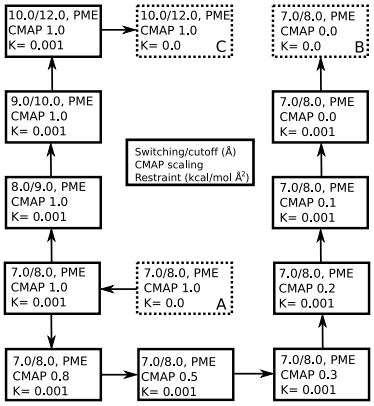

Figure 13: Steps in free energy calculations using alternate potentials. Boxes with solid lines represent intermediate states that were simulated; those with dashed lines represent unrestrained ensembles with free energies calculated based on the neighboring simulated state (see Methods). "CMAP scaling" refers to the scaling of the gridded CMAP correction to the backbone potential.

Poisson-Boltzmann electrostatics calculations were carried out on the reference states using APBS 1.0.0 [8] including apolar contributions [9]. CHARMM22 atomic charges and vdW parameters were used except for the polar contribution to the solvation free energies, where optimized atomic Born radii [10] were used instead. Calculations were carried out on a cubic lattice with an edge length of 60 Å using 129 points along each dimension, with a dielectric constant of 78.54

in solvent and 1.0 in the protein interior and in vacuum, and monovalent mobile ions present at 30 mM. As with the MD simulations a temperature of 337 K was used. For the apolar portion of the calculation, a solvent radius of 1.4 Å, solvent pressure P of 36.0 cal / (mol Å^3), and solvent surface tension γ of 3.0 cal / (mol Å^2) [9] were used.

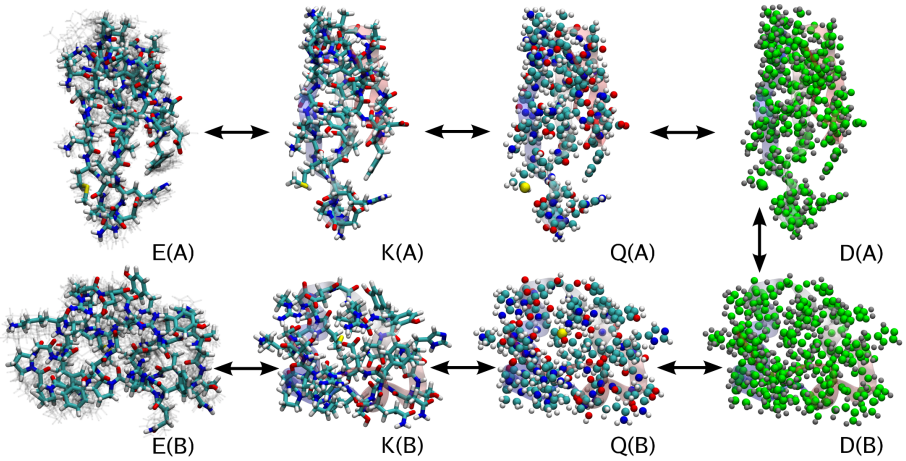

Figure 14: Schematic of the intermediate states involved in a deactivated morphing calculation. The transition between each of these states involves calculation of the free energy changes along a path of one or more intermediates. In this case states A and B correspond to SHEET and HELIXU, respectively, from the main text.

Restraining states The transition between E(A) and K$_1$(A) was for all cases calculated through a series of 70 intermediate states ranging from $\kappa_1 = 1000.0$ kcal/(mol Å^2) to $\kappa_{70} = 0.001$ kcal/(mol Å^2). The full set of restraining states for each simulation, along with the durations of the calculations, are shown in Table 1. The timestep used for different spring constants was varied to ensure stability of the simulation; SHEET, HELIXU, and HELIXV used identical timesteps in all cases, but slightly shorter timesteps were needed in some cases for HELIXL. Plots of the RMSD distributions for different spring constants showing the degree of overlap of adjacent ensembles are shown in Fig. 15. In all cases data for free energy calculations were taken once per 100 fs. An additional 1.0 ns equilibration was performed for the κ_1 state in each case to allow proper equilibration of the heavily restrained system with the thermal bath.

State	κ (kcal / mol Å)	Duration (ns)	Timestep (fs)
1	1000.00000	2.0	0.5 (0.4)
2	975.00000	2.0	0.5 (0.4)
3	950.00000	2.0	0.5 (0.4)
4	900.00000	2.0	0.5 (0.4)

5	850.00000	2.0	0.5 (0.4)
6	800.00000	2.0	0.5 (0.4)
7	760.00000	2.0	0.5 (0.4)
8	720.00000	2.0	0.5
9	640.00000	2.0	0.5
10	608.00000	2.0	0.5
11	576.00000	2.0	0.5
12	544.00000	2.0	0.5
13	512.00000	2.0	0.5
14	460.80000	2.0	0.5
15	409.60000	2.0	0.5
16	368.64000	2.0	0.5
17	327.68000	2.0	0.5
18	294.91200	2.0	0.5
19	262.14400	2.0	0.5
20	235.92960	2.0	1.0 (0.5)
21	209.71520	2.0	1.0 (0.5)
22	188.74368	2.0	1.0 (0.5)
23	167.77216	2.0	1.0
24	134.21773	2.0	1.0
25	107.37418	2.0	1.0
26	85.89935	2.0	1.0
27	68.71948	2.0	1.0
28	54.97558	2.0	1.0
29	43.98047	2.0	1.0
30	35.18437	2.0	1.0
31	28.14750	2.0	1.0
32	22.51800	2.0	1.0
33	18.01440	2.0	1.0
34	14.41152	2.0	1.0
35	11.52922	2.0	1.0
36	9.22337	2.0	1.0
37	7.37870	2.0	1.0
38	5.90296	2.0	1.0
39	4.72237	2.0	1.0
40	3.77789	2.0	1.0
41	3.02231	2.0	1.0
42	2.41785	2.0	1.0
43	1.93428	2.0	1.0
44	1.54743	2.0	1.0
45	1.23794	2.0	1.0
46	0.99035	2.0	1.0
47	0.79228	2.0	1.0
48	0.63383	2.0	1.0
49	0.50706	2.0	1.0
50	0.40565	2.0	1.0

51	0.32452	2.0	1.0
52	0.25961	2.0	1.0
53	0.20769	2.0	1.0
54	0.16615	2.0	1.0
55	0.13292	2.0	1.0
56	0.10634	2.0	1.0
57	0.08507	2.0	1.0
58	0.06806	2.0	1.0
59	0.05445	2.0	1.0
60	0.04356	2.0	1.0
61	0.03484	4.0	1.0
62	0.02788	5.0	1.0
63	0.01742	5.0	1.0
64	0.01394	5.0	1.0
65	0.01000	20.0	2.0 (1.0)
66	0.00500	20.0	2.0 (1.0)
67	0.00375	20.0	2.0
68	0.00250	20.0	2.0
69	0.00175	20.0	2.0
70	0.00100	20.0	2.0

Table 1: List of restraint states used in deactivated morphing calculations. Timesteps shown in parenthesis are for HELIXL, if different from the other states.

Deactivating states The deactivation step of the DM procedure involved simulation of 35 intermediate states, each with one additional residue of the protein fixed, starting from the N-terminus. When a residue is "fixed", the coordinates of all of its atoms are constrained to those in the reference state; in addition, no bonded or short-range interactions are calculated between pairs of fixed atoms. For all deactivating calculations a harmonic potential with a spring constant of $\kappa_1 = 1000.0$ kcal/(mol Å^2) was applied to restrain all unconstrained atoms to their reference coordinates. A timestep of 0.5 fs was used for all cases except HELIXL, which required a 0.4 fs timestep. Data were taken once per 100 fs, with a total of 1 ns of simulation in each deactivation intermediate state. Each intermediate was equilibrated for 1 ns prior to data collection due to the strong restraints used; in the future this additional equilibration could be avoided by beginning each deactivation run from the coordinates and velocities of the previous state, as was done for the restraining states.

Dummying The most significant difference between the DM procedure used here and that originally used in [6] is that instead of performing a morphing step in which each atom in the initial state A was translated to its position in state B, an additional intermediate referred to as the "dummy" state was introduced. In this state all protein heavy atoms have identical van der Waals parameters (those of a TIP3P oxygen) and all protein hydrogens have a separate set of

vdW parameters (those of a TIP3P hydrogen in the CHARMM22 forcefield). Similarly, the dummies were given uniform charges by evenly spreading the total charge of the protein over the dummies. Note that because the dummies were only used in states with all protein atoms fixed, no bonded interactions were needed.

The free energy change for the transition from a fully deactivated to dummied state was calculated using the FEP module of NAMD [11]; simulations were performed at $\lambda = 0.0$, 0.001, 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.08, 0.85, 0.9, 0.95, 0.99, 0.999, and 1.0. For each λ value the system was equilibrated for 100 ps, and then simulated with data gathered for 1 ns. A timestep of 1.0 fs was used for all dummying calculations.

Morphing As noted above, the use of dummied protein atoms makes the morphing process significantly more efficient; instead of forcing all protein atoms to travel the complete distance between their conformations in the initial and final states, a distance-minimizing mapping [12] was instead calculated for all heavy atom and hydrogen positions in state A moving to positions in state B. Prior to calculation of the distance-minimizing map the principal axes of the final state were aligned to the initial state. The average distance moved by atoms during the morphing step was 3.4 Å, compared with 11.3 Å if a distance-minimizing mapping was not used. Fifty intermediate states were used for each morphing calculation, with 1 ns of data collected in each state using a timestep of 1.0 fs. Diagrams of the free energy profile over the course of each morph are shown in Fig. 16. Separate morphs were carried out for all pairwise combinations of conformations, although only morphs involving SHEET are shown in Fig. 3 of the main text.

Site-bound waters In two cases, HELIXU and HELIXV, we noted that a single water molecule appeared to be in an inaccessible binding site inside of the protein, such that it did not exchange with bulk solvent after the constraint step. The free energy differences calculated from any other state to HELIXU or HELIXV using the procedure outlined above would thus not be correct, because the morphing endpoint does not contain the corresponding site-bound water. Beginning the morphs from structures with the bound water can lead to convergence difficulties due to the trapped water being very close to moving dummies (and cannot solve the problem of morphing from a structure with one bound water to another structure with a bound water at a different location, without adding further constraints to the water and including it in the morph). Instead, for morphs involving HELIXU and/or HELIXV a small additional step was added where the free energy change for transferring a water from bulk solution to the binding site in the fixed, dummied reference structure D(X) was calculated. Thus, for morphs involving HELIXU or HELIXV, we considered an additional intermediate state in which the dummied protein was in the HELIXU/HELIXVconformation but the site bound water was absent; the

morphing method described above was used to calculate the free energy difference between this intermediate and the other dummied reference state, and the method of Roux and coworkers [13] (described below) was used to calculate the difference between that intermediate and the true HELIXU/HELIXV dummied state.

In the scheme used to calculate the free energy of water binding to an internal site, the site-bound water is harmonically restrained, decoupled from the environment, the effects of the restraints removed analytically, and the cycle completed using the free energy of coupling the water to a bulk water box [13], allowing calculation of the ratio of protein with and without water bound at the identified site, R_1 (Eq. 14 of [13]):

$$R_1 = \rho_{\text{bulk}} \left(\frac{2\pi k_B T}{k_{\text{harm}}} \right) e^{-\left[\Delta G^{0\to 1}_{\text{cavity}} - \Delta G^{0\to 1}_{\text{bulk}}\right]/k_B T} \tag{4}$$

ρ_{bulk} represents the bulk density of water (in waters / Å^3), $\Delta G^{0\to 1}_{\text{cavity}}$ the free energy change between a state with an unrestrained water in the cavity of the dummied protein and with the same water restrained to a position in the cavity but decoupled from the rest of the system, and $\Delta G^{0\to 1}_{\text{bulk}}$ the free energy of decoupling a single TIP3P water molecule from a box of water. The overall free energy change for water binding to the internal site was then taken to be $\Delta G_{\text{site}} = -k_B T \log(R_1)$.

The optimal spring constants k_{harm} for restraining the water were calculated using Eq. 24 of [13] to be 15.0 kcal/(mol Å^2) for HELIXU and 4.0 kcal/(mol Å^2) for HELIXV. Restraints were applied to the bound water in 5 steps (for HELIXU) or 3 steps (for HELIXV), with 0.5 ns of simulation in each state for HELIXU and 1.0 ns for HELIXV. Decoupling was performed using a single-topology FEP approach on the restrained water using 31 intermediate values of λ with 50 ps of equilibration and 500 ps of production data at each state. Decoupling of a single water from a water box was performed using a box of 10,917 TIP3P water molecules under conditions identical to the DM calculations using 14 intermediate values of λ (0.0, 0.01, 0.05, 0.15, 0.25,..., 0.95, 0.99, 1.0), using the softcore potential of Zacharias and coworkers [14] to improve convergence. At each value of λ, 25 ps of equilibration and 250 ps of sampling were used. The calculation yielded a free energy for decoupling a single TIP3P water from bulk solvent ($\Delta G^{0\to 1}_{\text{bulk}}$) of 5.84 ± 0.13 kcal/mol under the conditions used (*n.b.* at a temperature of 337 K). The overall free energies for water binding to the buried sites in HELIXU and HELIXV, ΔG_{site}, were calculated to be 2.050 ± 0.098 kcal/mol and 2.200 ± 0.191 kcal/mol, respectively. In the main text and all other discussion, the additional contribution of the site bound water calculations described here are in all cases included as part of the morphing step; for reference, the free energy changes for morphing between SHEET and the water-free structures of HELIXU and HELIXV were calculated to be -2.601 ± 0.490 kcal/mol and -4.862 ± 0.595 kcal/mol, respectively.

Approximations to deactivated morphing While deactivated morphing provides a computationally tractable pathway for calculating free energy differences between radically different conformations of biomolecules, the required sampling to obtain precise results is still expensive, and it is thus attractive to consider potential shortcuts for the steps involved in DM. For the case of the deactivation step the correlation with internal interactions of the protein provides a simple solution; in development of the deactivated morphing method [6] it was observed that the deactivation free energy (G_Q - G_{K_1}) could be approximated by -U_P, the internal interaction energy of the protein in the associated reference state. In the case of the WW domain, because PME was used for long range electrostatics G_Q corresponds to a state with all short-range protein-protein interactions removed, but with long range electrostatics still present; however, the effective change in free energy due to internal interactions between two conformations, $\Delta G_{\mathrm{int},A\to B}$ = ($G_{K_1,B}$ - $G_{Q,B}$) - ($G_{Q,A}$ - $G_{K_1,A}$), would still be expected to be approximately equal to the difference in short-range potential energies for the two reference states, $\Delta U_{P,short}$. This relationship also appears to hold for the WW domain conformations studied here; the values of $\Delta U_{P,short}$ relative to SHEET are -31.60, -8.10, and -49.45 kcal/mol for HELIXL, HELIXU, and HELIXV, respectively. These values are in good agreement with the differences in deactivation energy from DM of -31.44, -7.16, and -48.57 kcal/mol. We define $U_{P,short}$ as the internal van der Waals and real space electrostatic energy of the protein using the same nonbonded switching and cutoffs used in the simulations.

Park *et al.* also noted that calculation of solvation free energies using Poisson-Boltzmann methods might be usable as an approximation for the morphing free energies, since the morphing step is expected to primarily reflect the effects of interaction with solute. From the complete deactivated morphing calculations, the morphing step (defined as the $Q(A) \to Q(B)$ transition) favors SHEET by 15.32, 9.01, and 49.44 kcal/mol relative to HELIXL, HELIXU, and HELIXV, respectively. Poisson-Boltzmann solvation free energies for all four conformations are shown in Table 3 of the main text; while these calculations are qualitatively correct in their ranking of HELIXU and HELIXV as having less favorable solvation energies than SHEET, HELIXL is incorrectly stabilized relative to SHEET; in addition, the magnitude of the differences from PB calculations is significantly larger than that calculated from DM. Thus, unlike the approximation to the deactivation step discussed above, PB calculations do not appear to provide an effective approximation to the morphing step of DM.

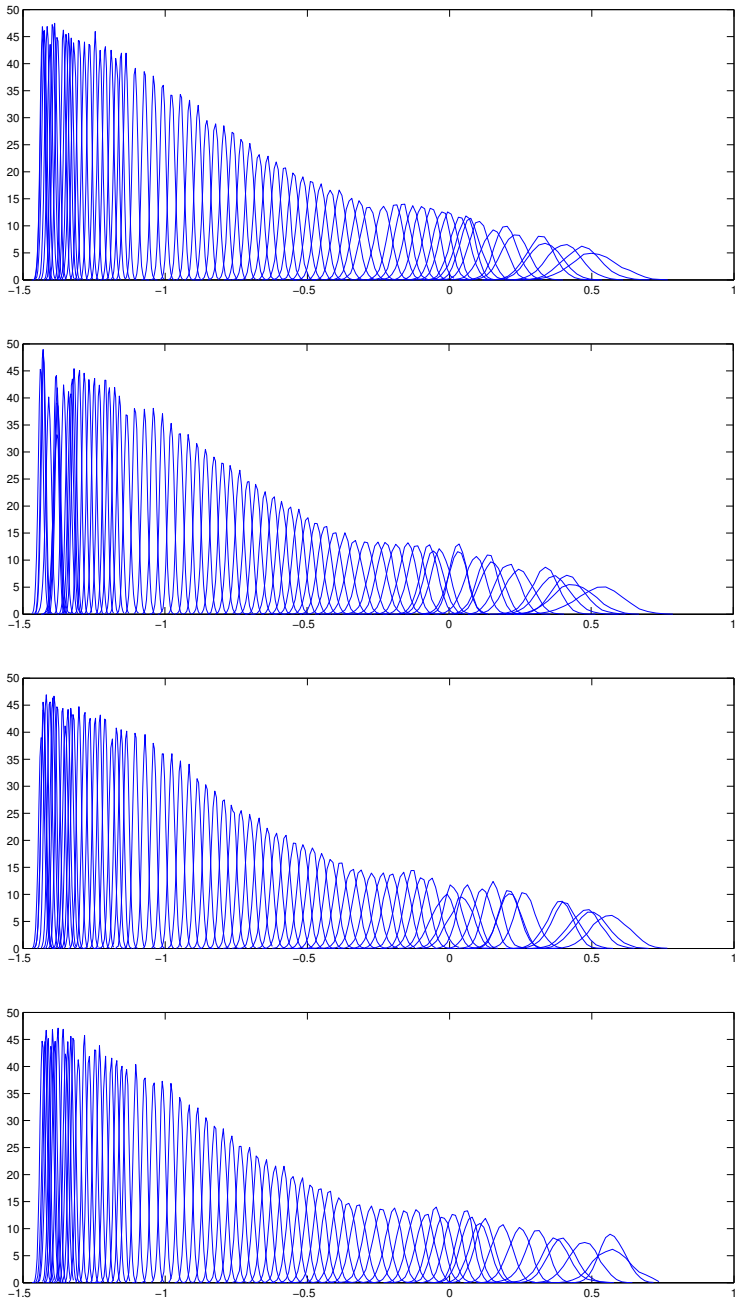

Figure 15: Overlap of adjacent ensembles used in restraining the reference conformations. From top, SHEET, HELIXL, HELIXU, and HELIXV are shown. Spring constants decrease going from left to right.

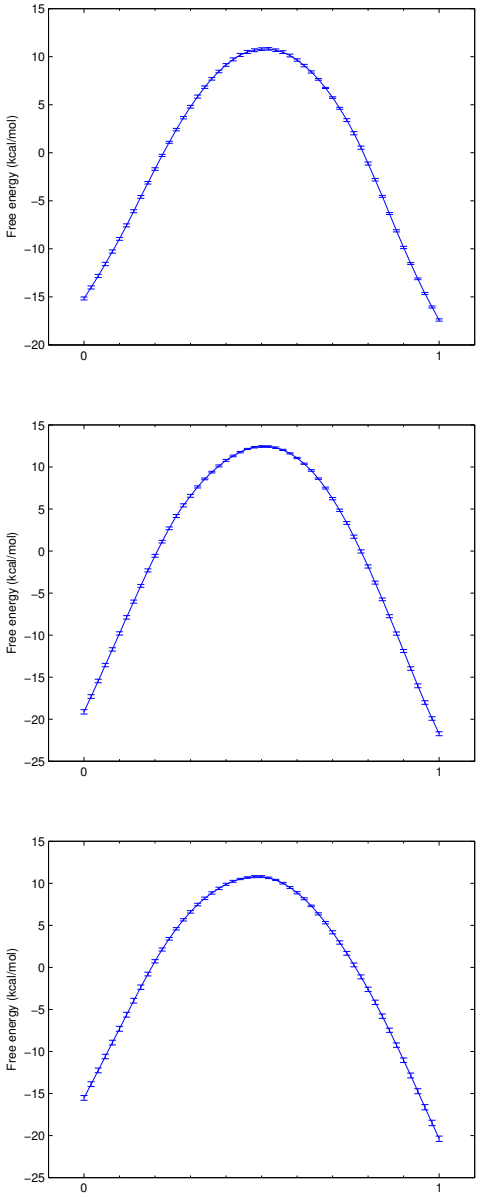

Figure 16: Free energy profiles associated with morphing. Profiles are shown for morphing SHEET (λ=0) to one of the helical states (λ=1); from top, HELIXL, HELIXU, or HELIXV.

4 Corrective perturbations to the force field

Given the clear structural differences observed between SHEET and the helical DM reference states (most notably, their significantly different secondary structure), it should be possible to identify simple perturbations to the force field that would cause SHEET to be lower in free energy than the other states. If we assume that some small perturbation δ is applied to the potential energy function to yield a new potential

$$U_{\text{new}}(\mathbf{X}) = U(\mathbf{X}) + \delta(\mathbf{X}), \tag{5}$$

and that the effective change in the free energy of a given reference state is well-approximated by the first-order perturbative formula

$$\Delta G_{\delta}(\text{state}) \simeq \langle \delta(\mathbf{X}) \rangle_{\text{state}}\,, \tag{6}$$

then we must simply identify a sufficient perturbation δ such that the inequality

$$\Delta G_{\delta}(\text{helix}) - \Delta G_{\delta}(\text{sheet}) > \Delta G_{\text{DM}}(\text{helix} \rightarrow \text{sheet}) \tag{7}$$

is satisfied for all helical DM reference states; that is, the net change in free energy difference between SHEET and each helical state is greater than the free energy difference from deactivated morphing between SHEET and the helical conformation being considered.

Identification of a sufficient δ is particularly simple if one considers perturbations to the CMAP correction. As in the parameterization of CMAP [15], we define a given (ϕ, ψ) angle combination as being in the α region for ($-150.0 \leq \phi \leq -30.0$, $-90.0 \leq \psi \leq 30.0$), and in the β region for ($-180.0 \leq \phi \leq -30.$, $60.0 \leq \psi \leq 180.0$); using these definitions, the average occupancies of α and β regions of the Ramachandran plot are shown in Table 2.

We can then seek some quantity k such that, if δ were given by altering the CMAP corrections by $+k$ for all α conformations and by $-k$ for all β conformations, the inequality in Eq. 7 is satisfied for all three helical reference states. We thus seek k such that $k\left((N_{\alpha} - N_{\beta})_{\text{helix}} - (N_{\alpha} - N_{\beta})_{\text{sheet}}\right) > \Delta G_{\text{DM}}(\text{helix} \rightarrow \text{sheet})$ for each helical state, with N_{α} and N_{β} the average number of residues in the α and β portions of the Ramachandran map, respectively, in a given state. Using the data in Table 2 it is straightforward to find that for $k \geq 0.23$ kcal/mol, SHEET will be more stable than any of the three helical reference states; this stability should, furthermore, increase for larger values of k. Thus, altering the CMAP correction to stabilize all β conformations by 0.23 kcal/mol and destabilize all α conformations by 0.23 kcal/mol would be expected to make SHEET lower in free energy than the helical DM reference states.

As an alternative to changing the CMAP correction, given that all three helical reference states have more interactions with water than SHEET (see Table 2), it should also be possible to stabilize SHEET relative to the helical states by making the Lennard-Jones interactions between water and protein less attractive. Using the average numbers of contacts from Table 2, reduction of the strength

State	N_α	N_β	Water Contacts	ΔG_{DM}
SHEET	6.10 ± 0.09	21.75 ± 0.04	1230.87 ± 25.42	–
HELIXL	24.34 ± 0.20	4.88 ± 0.06	1291.95 ± 57.55	8.07 kcal/mol
HELIXU	23.72 ± 0.33	5.98 ± 0.27	1288.78 ± 34.07	4.59 kcal/mol
HELIXV	21.92 ± 0.07	6.94 ± 0.01	1243.36 ± 19.16	4.37 kcal/mol

Table 2: Statistics from the κ_{70} simulations from the deactivated morphing restraint step. N_α and N_β refer to the average numbers of residues in the α and β region of (ϕ,ψ) space, respectively. "Water contacts" refers to the average number of pairwise contacts (using a 4.0 Å cutoff) between protein atoms and water oxygens. Error bars are calculated using block averaging as described in the Methods section of the main text. ΔG_{DM} refers to the free energy difference between a given state and SHEET from deactivated morphing calculations.

of all Lennard-Jones interactions between water oxygens and protein atoms by 0.35 kcal/mol would be sufficient to make SHEET more stable than any of the helical conformations. We note, however, that application of this perturbation would amount to average changes in hydration free energy of ∼13 kcal/mol for each residue, significantly larger that the accepted margin of error in hydration free energies for the CHARMM force field [16, 17]. In addition, given the large uncertainties in the numbers of water contacts, this method would be less certain to correct the helix/sheet balance in the WW domain than the CMAP alterations described above, but does illustrate an alternative (and unrelated) path toward possible corrections.

The analysis presented above, while providing several perturbations to the potential that should make SHEET lower in free energy than the helical decoys used in deactivated morphing, primarily illustrates the danger in attempting to "fix" a force field based only on simulations of one protein. Given the variety of data used in the parameterization of CHARMM Lennard-Jones parameters [18] and the CMAP correction [15], it is extremely unlikely that the "corrections" proposed here would in fact improve the performance of the CHARMM force-field in any test except that of stabilizing the SHEET conformation of the WW domain. Indeed, even in the realm of changes to the CMAP correction the perturbation proposed above is not unique; either raising the energy of all α conformations by 0.45 kcal/mol or lowering the energy of all β conformations by 0.45 kcal/mol would also be sufficient to satisfy Eq. 7. The correction of force field inaccuracies such as those observed in the present study must instead be solved by a combination of basic physical principles and testing on a wide variety of proteins; *ad hoc* corrections based on a single protein are likely to be non-unique and unphysical, and not to generalize well to other systems.

References

[1] van der Spoel, D., E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. 2005. Gromacs: Fast, flexible, and free. *J. Comp. Chem.* 26:1701–1718.

[2] Daura, X., K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark. 1999. Peptide folding: When simulation meets experiment. *Angewandte Chemie International Edition.* 38:236–240.

[3] Chodera, J. D., N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope. 2007. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J Chem Phys.* 126:155101.

[4] Kortemme, T., A. V. Morozov, and D. Baker. 2003. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol.* 326:1239–1259.

[5] Morozov, A. V., T. Kortemme, K. Tsemekhman, and D. Baker. 2004. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc Natl Acad Sci U S A.* 101:6946–6951.

[6] Park, S., A. Y. Lau, and B. Roux. 2008. Computing conformational free energy by deactivated morphing. *J Chem Phys.* 129:134102.

[7] Bennett, C. H. 1976. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics.* 22:245–268.

[8] Baker, N. A., D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. 2001. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA.* 98:10037–10041.

[9] Wagoner, J. A. and N. A. Baker. 2006. Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms. *Proc Natl Acad Sci U S A.* 103:8331–8336.

[10] Nina, M., D. Beglov, and B. Roux. 1997. Atomc radii for continuum electrostatics calculations based on molecular dynamics free energy simulations. *J. Phys. Chem. B.* 101:5239–5248.

[11] Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. 2005. Scalable molecular dynamics with NAMD. *J. Comp. Chem.* 26:1781–1802.

[12] Anitescu, M. and S. Park. 2009. A linear programming approach for the least-squares protein morphing problem. *Mathematical Programming.* In press.

[13] Roux, B., M. Nina, R. Poms, and J. C. Smith. 1996. Thermodynamic stability of water molecules in the bacteriorhodopsin proton channel: a molecular dynamics free energy perturbation study. *Biophys J.* 71:670–681.

[14] Zacharias, M., T. P. Straatsma, and J. A. McCammon. 1994. Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration. *J. Chem. Phys.* 100:9025–9031.

[15] MacKerell, A. D., Jr., M. Feig, and C. L. Brooks III. 2004. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comp. Chem.* 25:1400–1415.

[16] Shirts, M. R. and V. S. Pande. 2005. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *J Chem Phys.* 122:134508.

[17] Deng, Y. and B. Roux. 2009. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *J Phys Chem B*:in press.

[18] MacKerell, A., Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, I. W. E. Reiher, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* 102:3586–3616.