Supporting Information

Unfolded state dynamics and structure of protein L characterized by simulation and experiment

Vincent A. Voelz, Vijay R. Singh, William J. Wedemeyer, Lisa J. Lapidus, Vijay S. Pande

This supplement contains:

Supporting Figures

Figure S1. Average values of the RMSD (C_α) over time for simulated ensembles starting from native (black), extended (red) and coil (green) conformations. Non-native ensembles simulated at 300K, 330K and 370K (thin lines) converge to a compact globule state on the microsecond time scale, while the 450K (thick lines) ensembles converge to a random-coil state. The native state is very stable in the simulations, deviating only ~2.5Å RMSD from the starting conformation. At 450K, the native state simulation (tick black line) begins to unfold by 1 μs.

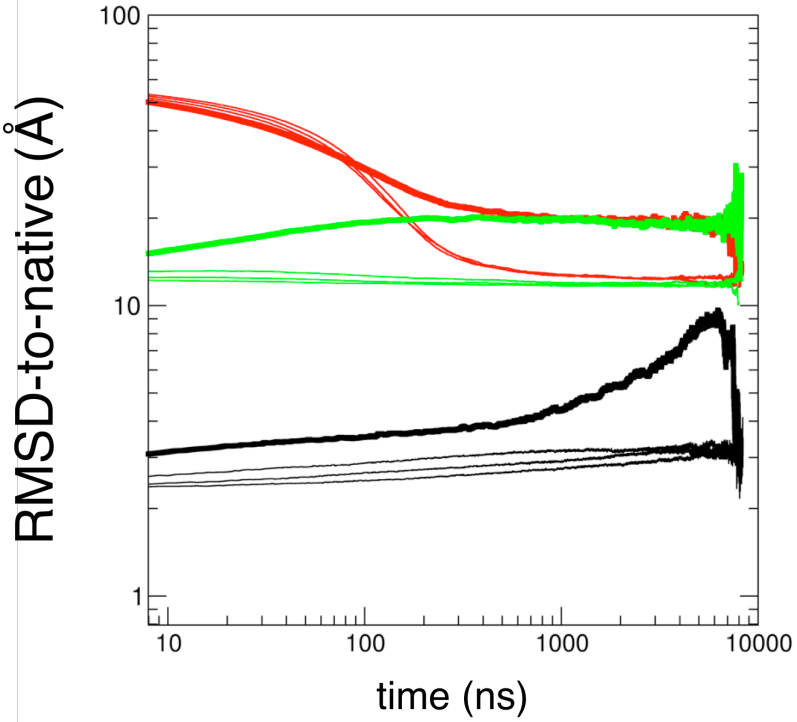


Figure S2. (a) Simulated distributions of K23-W47 distances over time, for various starting states, at 300K and 450K. The vertical axis is on a logarithmic scale to enhance detail. (b) The relative entropy of $P_{ext}(r)$ (dashed line) and $P_{native}(r)$ (solid line) with respect to $P_0(r)=P_{coil}(r)$ over time, at various temperatures. Smaller values of relative entropy, computed as $\int dr P(r)\log[P(r)/P_0(r)]$, reflect more similar distributions. The increase in the relative entropy between $P_{ext}(r)$ and $P_{coil}(r)$ at ~100 ns is an artifact of using a single reaction coordinate: the two distributions briefly share similar modes.

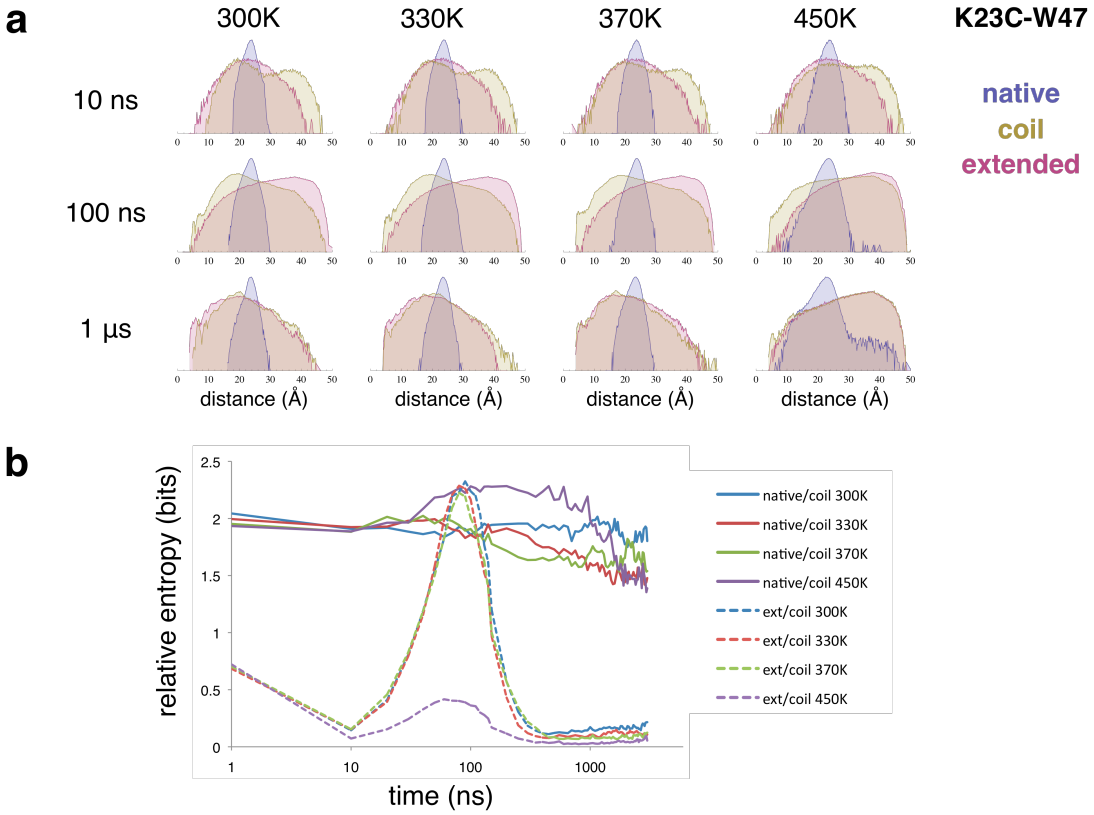


Figure S3. Polymer scaling coefficients for unfolded ensembles. Each plot is of the average radius of gyration for all pairs (i,j) of C_α atoms, calculated from a 1.5 – 2.0 μs window of trajectory data, versus of the sequence separation $|i-j|$, shown on a log-log scale. The data from ensembles started from an extended state at 300K and 450K were least-squared fit to a curve $R_g = R_0 N^\nu$ to extract the scaling coefficients ν. The 300K ensemble has a scaling coefficient of 0.372, which is close to the polymer theoretical expectation of ν=1/3 for a compact globule, while the 450K ensemble has a scaling coefficient of 0.592, which is very close to the expected ν=0.588 for a random-coil chain with excluded volume [1]. Native contacts in the compact ensemble (red) are more compact on average than non-native contacts.

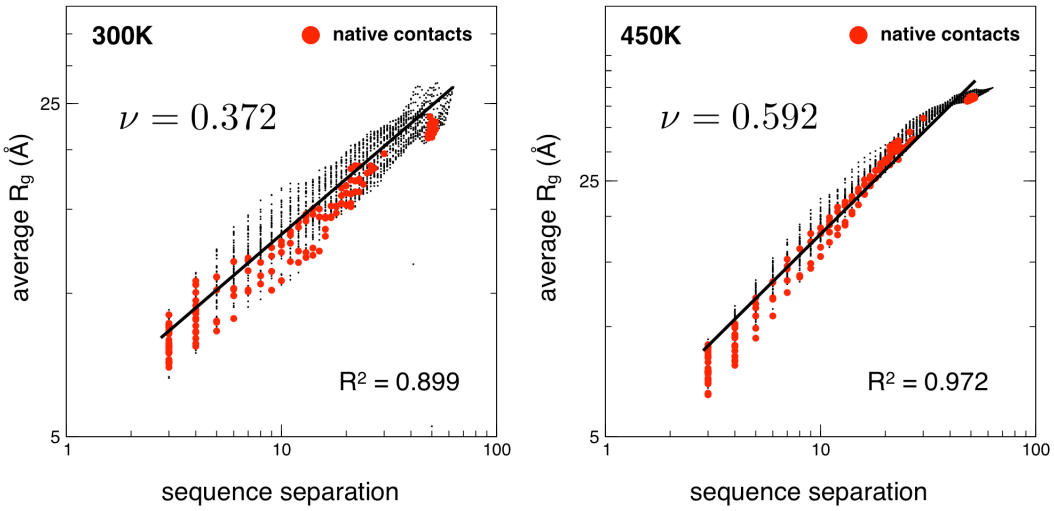
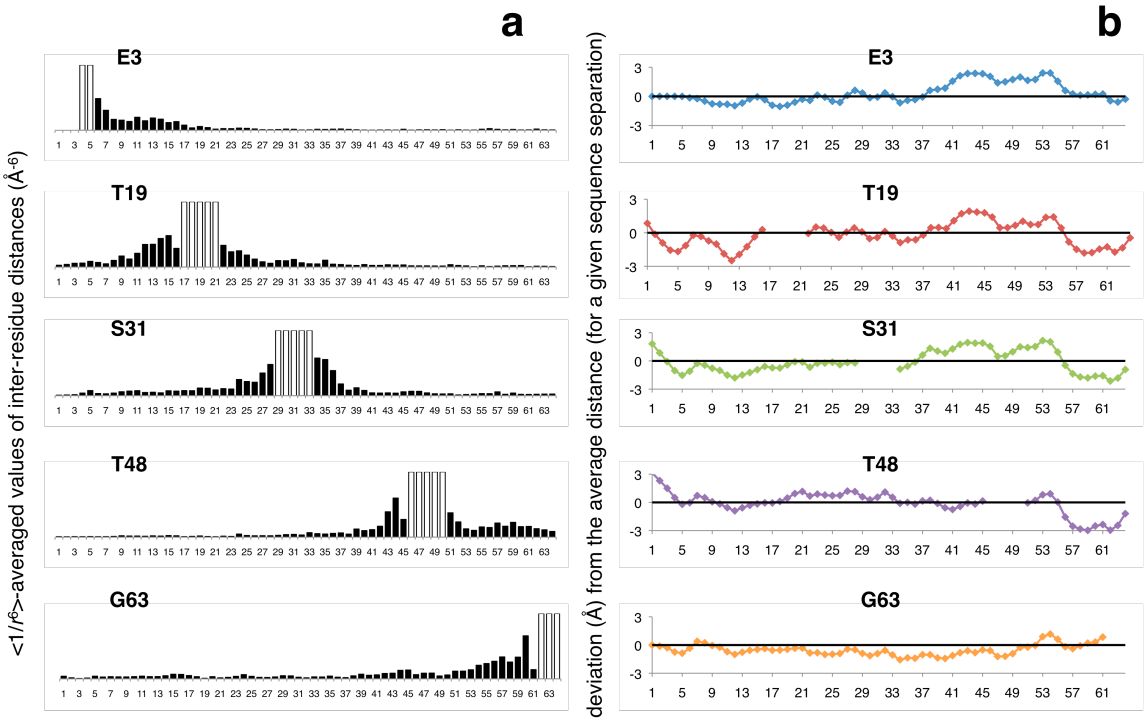


Figure S4. Distance profiles for selected residues E3, T19, S31, T48 and G63, corresponding to nitroxide-labeled positions used in paramagnetic relaxation enhancement studies by Yi et al. (2000). Panels (a) and (c) show average values of $\langle 1/r^6 \rangle$ plotted for inter-residue distances along the chain, for unfolded ensembles started from extended (a) and coil (c) states, after 1 μs. The maximum value on the vertical axis is 4 x 10^{-5} ($Å^{-6}$). The white bars, scaled to this maximum, are placeholders for residue pairs whose sequence separation is less than 4 residues, which we ignore. Panels (b) and (d) the deviations of the average inter-residue distance from the ensemble average over all sequence separations, for unfolded ensembles started from extended (a) and coil (c) states, after 1μs. In both unfolded ensembles, we see indications of native-like structure, including lower-than-average distances between adjacent strand regions in the hairpin regions, and an oscillating pattern of values corresponding to the central helix region. In particular, the ensemble started from the coil state shows more structuring than the ensemble started from the extended state, with both hairpins showing closer-than-average distances with the central helix.



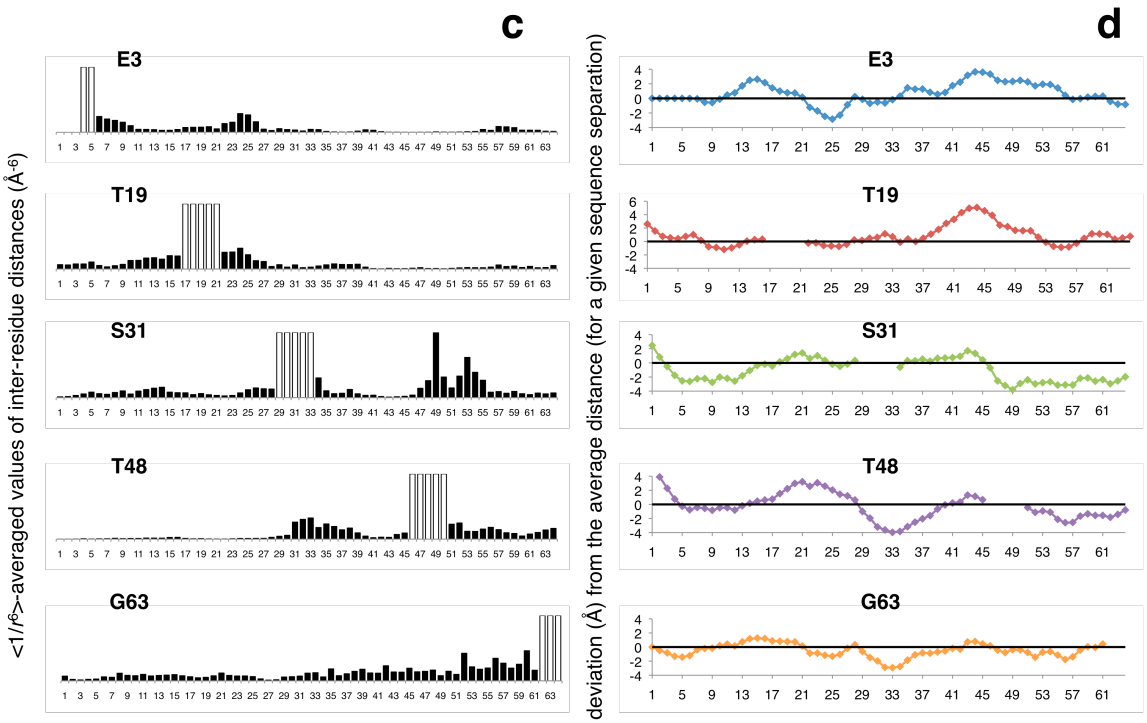

c
d
E3
T19
S31
T48
G63
<1/r^6>-averaged values of inter-residue distances (Å^-6)
deviation (Å) from the average distance (for a given sequence separation)

Figure S5. Secondary structure content after 1 μs for unfolded-state ensembles started from the extended state. Simulated ensembles of the wild-type (WT) sequence at 450K contain 18% more helical secondary structure than the 300K ensemble, as calculated by the DSSP algorithm. The F22A sequence has slightly more helicity than WT, most likely due to local sequence effects.

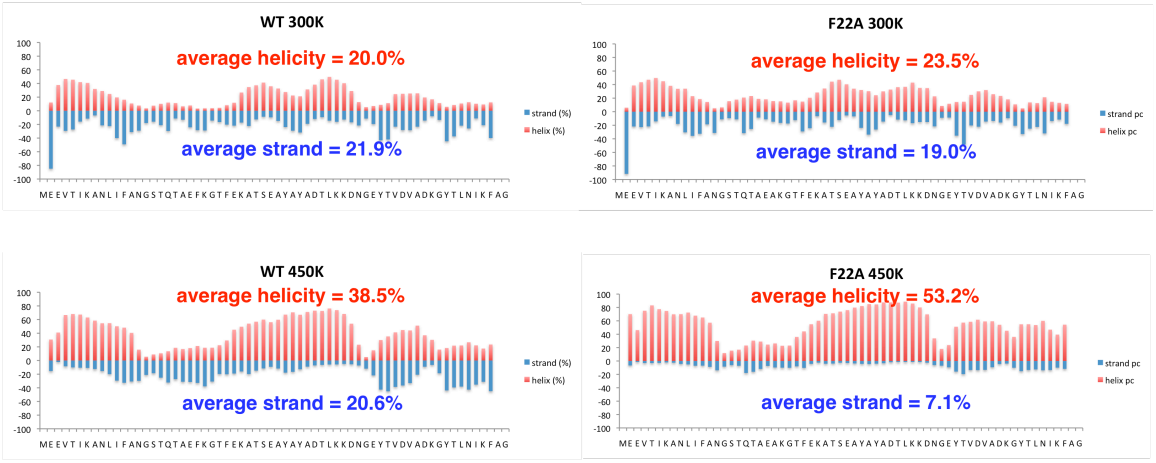
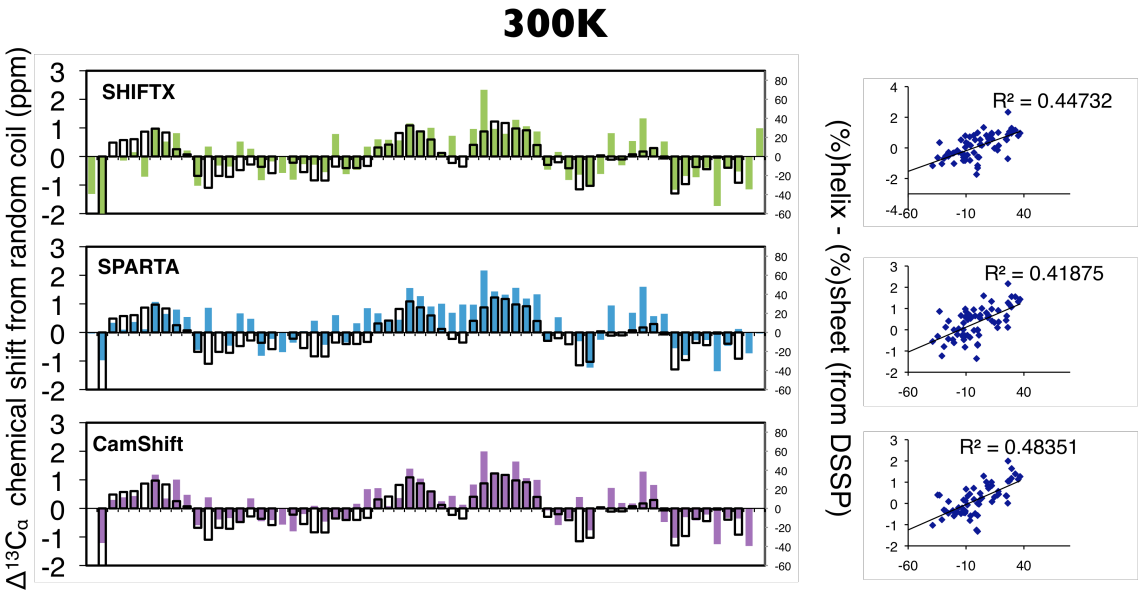


Figure S6. Comparison of chemical shift prediction algorithms applied to protein unfolded-state data at 300K (a) and 450K (b). Secondary structure propensities after 1 μs for the 300K (top) ensemble and 450K (middle) ensemble started from the extended state (black outline), plotted against predicted chemical shifts from three different algorithms: ShiftX [2] (green bars), SPARTA [3] (blue bars), and CamSHIFT (http://www-vendruscolo.ch.cam.ac.uk/camshift/camshift.php) (purple bars). Propensities are calculated as percent α-helix minus percent β-sheet, using the DSSP algorithm. The correlation between simulated propensities and predicted chemical shifts is shown to the right. Random coil values in all figures are from Wishart and Sykes (1994). We also ran all predictions (not shown) using random coil values from Spera and Bax (*JACS* 1991), with consistently poorer results. In all cases CamSHIFT gave the highest correlation with the actual secondary structures as computed by DSSP.

a



b

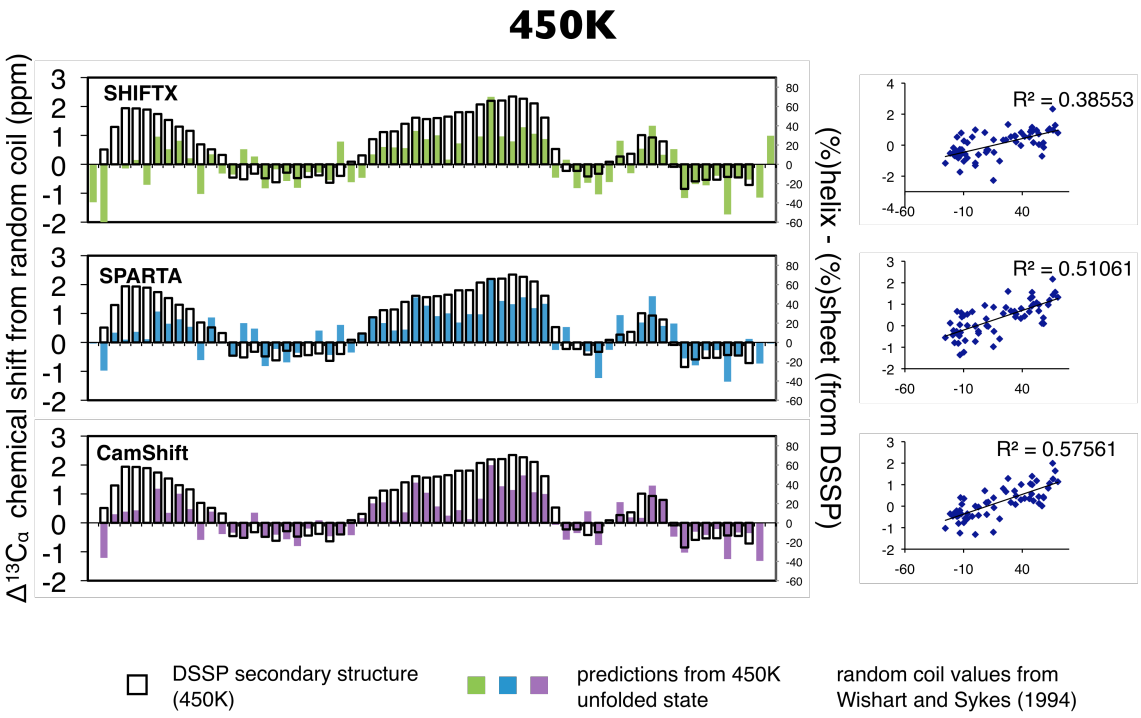

450K
SHIFTX
SPARTA
CamShift
Δ13Cα chemical shift from random coil (ppm)
(%)helix - (%)sheet (from DSSP)
R² = 0.38553
R² = 0.51061
R² = 0.57561
DSSP secondary structure (450K)
predictions from 450K unfolded state
random coil values from Wishart and Sykes (1994)

Figure S7. Comparison of unfolded state structure in wildtype (WT) versus F22A protein L. (a) Secondary structure profiles for unfolded state ensembles for WT and F22A simulated at 300K (starting from an extended state, for data after 1 μs) show an increase in helicity for the mutant, particularly near residue F22A. (b) The R_g values across all simulation temperatures increase only slightly (~5%) for F22A compared to WT. (c) Differences in average inter-residue distances for F22A compared to WT are plotted on a contact map using a red-to-green color scale, with the native contact map shown transparently in blue. Above (below) the diagonal are shown results from simulations started from an extended (coil) state. Changes in non-local interactions from WT include an increase in helical propensity beyond the N-terminal end of the central helix, and an increase between residues in central helix and C-terminal hairpin.

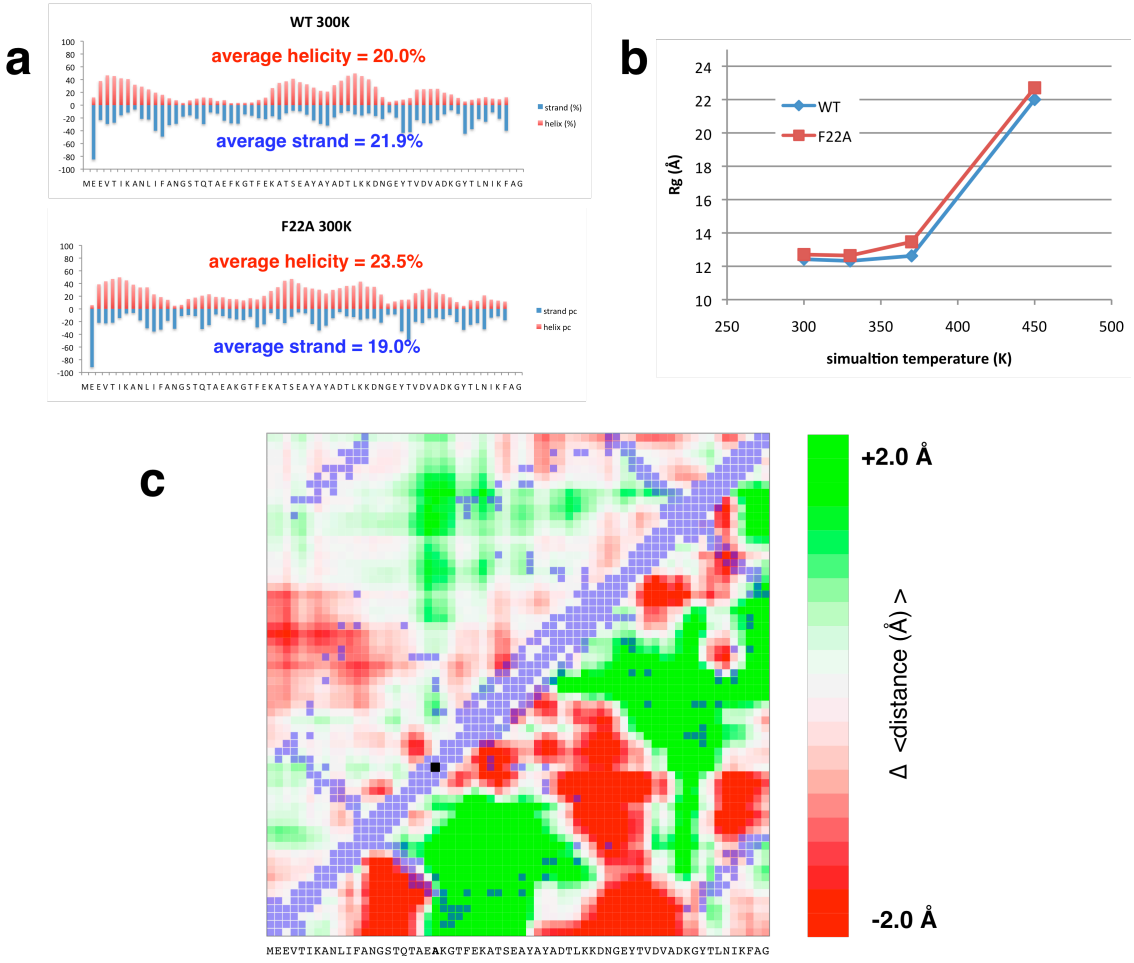


Figure S8. Simulated intramolecular diffusion coefficients over time. Diffusion coefficients *D* versus time were calculated by linearly fitting mean squared displacements of K23-W47 (solid line) and W47-T57 (dashed line) distances over time in windows of 50 ns. Shown are *D* versus time for WT simulations at 300K (a) and 450K (b), and simulations of the hydrophobic deletion mutant F22A at 300K (c) and 450K (d).

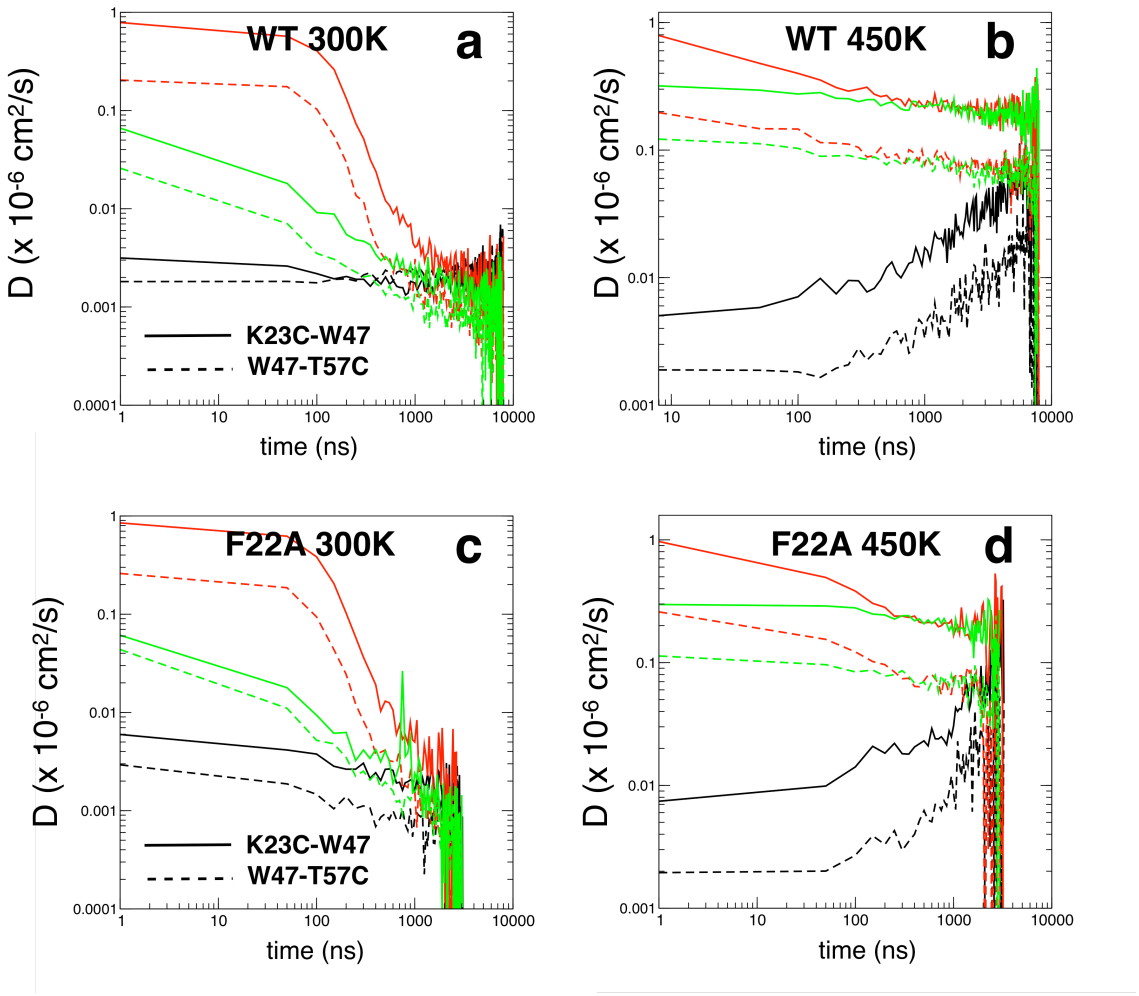


Table S1. Autocorrelation times for simulated Trp-Cys distances after 1 μs.

Temperature (K)	Distance	Ensemble starting from	τ_1 (ns)	τ_2 (ns)
300	K23-W47	extended	1.06	**276**
330	K23-W47	extended	1.98	**216**
370	K23-W47	extended	1.41	**212**
450	K23-W47	extended	12	**116**
300	K23-W47	coil	0.8	**206**
330	K23-W47	coil	1.05	**232**
370	K23-W47	coil	1.61	**230**
450	K23-W47	coil	12.5	**107**
300	W47-T57	extended	1.1	**114**
330	W47-T57	extended	1.95	**98**
370	W47-T57	extended	5.5	**100**
450	W47-T57	extended	2.5	**460**
300	W47-T57	coil	9.2	**142**
330	W47-T57	coil	8.9	**105**
370	W47-T57	coil	5.11	**112**
450	W47-T57	coil	7.95	**70**

Supporting Text

Materials and Methods

Distributed molecular dynamics simulation on GPUs

Simulations of the WT protein L sequence (with W47, but no cysteine mutations) were performed using a version of the GROMACS molecular dynamics program [4] optimized for graphics processing units [5]. Up to 10,000 parallel simulations (each with randomized Boltzmann-distributed initial velocities) were distributed using the Folding@Home platform [6]. The AMBER96 and AMBER03 forcefields were used, with the Generalized Born/Surface Area (GBSA) solvation model of Onufriev, Bashford and Case [7]. Stochastic integration was performed at 300K, 330K, 370K and 450K, using a 2 fs timestep and Berendsen temperature coupling. A water-like solvent (shear) viscosity of 91 ps^{-1} was used, with full $O(N^2)$ electrostatic and vdW interactions. Hydrogen bonds were constrained using the SHAKE algorithm. Trajectory snapshots were recorded every 1 ns. Diffusion coefficients were calculated by linear fits of mean squared displacement over time in 100-ns windows using snapshots between 10 and 90 ns. Error bars for simulated k_R were computed from a bootstrap of 5 subsets of the $P(r)$ data.

Starting conformations for the native state of protein L were taken from the crystal structure 2PTL, and steepest-descent minimized for 5000 steps. (Minimization was done using the GBSA model of Still et al.) Five starting conformations for the random coil ensemble were taken from snapshots of Monte Carlo trajectories in which dihedral angles were randomized under a potential rewarding compact R_g. The dihedral probabilities came from the TOP500 database [8]. Starting conformations for extended structures were constructed by setting dihedral angles to their canonical values. Trp-Cys distances were computed as the distance between the 'CD2' atom of W47 (closest to the center of mass of the side chain), and the C_γ atoms of K23 and T57 (corresponding to the cysteine sulfur).

Standard MD simulation

Standard MD simulations using the AMBER9 molecular dynamics package were performed at 9 different temperatures: 250 K, 300 K, 350 K, 400 K, 500 K, 600 K, 700 K, 800 K and 1000 K. The AMBER ff99 forcefield was used with the GBSA model Tsui and Case [9] (igb=1). For all temperatures except 300K, a single 40-ns trajectory was generated started from a fully extended state. Seventy 10-ns trajectories were generated at 300 K, each starting from a different random-coil configuration taken from the 600 K, 700 K, and 800 K simulations and energy minimized for 500 steps. Initial velocities for dynamics were assigned randomly from a Maxwell-Boltzmann distribution. A Langevin thermostat with collision frequency of 1 ps^{-1} was used with a 1 fs time step. Coordinates were saved every 1 ps. The Trp-Cys distance was measured as the distance between the sulfur atom of cysteine side chain and the center of mass of tryptophan.

Trp-Cys quenching experiments

The kinetics of intramolecular contact formation in the unfolded state can be modeled using a two-step model [10]. Intramolecular diffusion brings the excited tryptophan and the

cysteine residues together with a forward rate, k_{D+}, and the two residues diffuse apart at a backward diffusion rate, k_{D-}. When the two residues are in close contact, irreversible quenching can occur with a rate k_q. Therefore, the observed rate can be expressed as

$$k_{obs} = \frac{k_{D+}k_q}{k_q + k_{D-}} \qquad (1)$$

If $k_q >> k_{D-}$ then $k_{obs} = k_{D+}$. Since previous studies on the Trp-Cys quenching system have shown that diffusion and quenching rates are comparable [10], k_{obs} can be rewritten as

$$\frac{1}{k_{obs}} = \frac{1}{k_q K} + \frac{1}{k_{D+}} = \frac{1}{k_R(T)} + \frac{1}{k_{D+}(T,\eta)} \qquad (2)$$

where $K \equiv k_{D+}/k_{D-}$ is the equilibrium constant for forming the Trp-Cys contact and $k_R \equiv k_q K$.

To derive the reaction-limited quenching rate, k_R, and the forward intramolecular contact rate k_{D+}, a set of k_{obs} measurements at a particular denaturant concentration, 5 temperatures and 4 concentrations of sucrose was modeled as

$$k_R(T) = k_{R0}\exp\left(\frac{E_0(T-T_0)}{RTT_0}\right) \qquad (3)$$

$$k_{D+}(T) = \frac{k_{D+0}T}{\eta T_0}\exp\left(\gamma(T-T_0)\right) \qquad (4)$$

where T is the absolute temperature, T_o = 273 K, η is the solution viscosity and k_{R0}, k_{D+0}, E_0, and γ are fitting parameters. For the lowest concentrations of GdnHCl in which the unfolded state was observed, the temperature dependence of k_R disappeared so E_0 was constrained to be zero for K23C and F22A K23C.

Calibrating simulated unfolded ensembles with experiment

Comparing simulated and experimental unfolded ensembles is a challenging task. Unfolding experiments typically rely on chemically denaturants such as GdnHCl or urea, due to the technical difficulties of making measurements at high temperatures. Molecular dynamics simulations, on the other hand, must rely primarily on high temperature to produce unfolded ensembles, as well-parameterized implicit models of chemical denaturants do not exist, and explicit models are too computational costly and untested for the purpose of generating converged unfolded states.

Although the molecular mechanism by which chemical denaturants destabilize proteins may be very different from the effects of increased temperature, we can compare the two

by examining the coil-globule transition, which can be compared using polymer theory approaches. Free energies of transferring a monomer into solvent can be computed based on the extent of chain compaction, and compared across different GdnHCl concentrations and temperatures. This is the approach we take here.

Recently, Ziv and Haran [11] have had success in using a Flory-type mean-field polymer theory to model the coil-globule transition as a function of denaturant concentration. In these models, the Flory χ parameter [12], which quantifies the free energy of solute transfer per monomer, was calculated for several proteins as a function of denaturant concentration by fitting to conformational distributions derived from single-molecule FRET experiments. For a large set of proteins, including protein L, they found that the transfer free energy has a linear dependence on denaturant concentration.

One of the key features of the polymer theory is that the mean-field transfer free energy is model-independent. This means that although temperature and denaturants may destabilize proteins by different molecular mechanisms, their effect as order parameters in the coil-globule transition can be compared. Thus, we can compare the simulated $\chi_{sim}(T)$ value at a given temperature T to the calculated $\chi(D)$ value at a given denaturant concentration D to obtain the relationship between simulated temperature and experimental denaturant concentration, with no other fitting parameters required.

Polymer theory models for the coil-globule transition

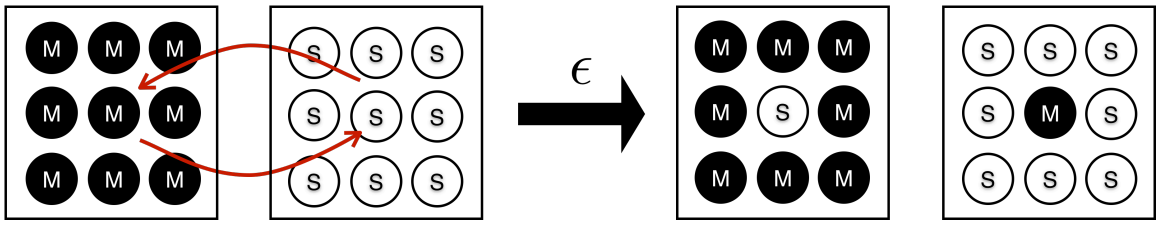
In Flory-type polymer models of the coil-globule transition, the expansion/contraction of chain molecules is described as a competition between two driving forces: mixing entropy and interactions between monomers [13]. The free energy can be described in the form:

$$\frac{\Delta F}{nk_BT} = \left(\frac{1-\rho}{\rho}\right)\ln(1-\rho) + 1 - \chi\rho$$

where ΔF is the free energy of the chain, n is the chain length, k_B is the Boltzmann constant, T is temperature, ρ is the volume density of monomers, and χ is the Flory chi-parameter, $\chi(T) = \varepsilon/2k_BT$, which describes a mean-field free energy of transferring a monomer to solution[12]. Specifically, ε is the free energy of the following combined transfer process: a monomer in a pure solution of monomers is transferred to a solution of solvent molecules, and a solvent molecule in pure solution of solvent molecules is transferred to a solution of monomers (**Figure S9**). The volume density ρ is defined such that $\rho = 1$ in the maximally compact state (usually assumed to the native state).

Figure S9. An illustration of the mean-field transfer energy, ε. The Flory χ parameter is $\chi(T) = \varepsilon/2k_BT$, where ε is the mean-field free energy of the following combined transfer process: a monomer in a pure solution of monomers is

transferred to a solution of solvent molecules, and a solvent molecule in pure solution of solvent molecules is transferred to a solution of monomers.



For this study, we examined similar two polymer theory models in the literature, both of which draw heavily from the theory of Sanchez[14]. The predictions from each model were compared against each other to check the robustness of the results.

The first of these is from Dill [15,16], and we will refer to it as the "Dill" model. In this model, the free energy difference ΔF_{Dill} with respect to the maximally compact state is modeled as

$$\frac{\Delta F_{Dill}}{nk_BT} = -\frac{7}{2}(\rho_0^{2/3} - (\rho_0/\rho)^{2/3}) + \frac{2}{n}\ln\rho + \left(\frac{1-\rho}{\rho}\right)\ln\rho + \frac{\varepsilon}{2k_BT}(1-\rho)$$

where ρ_0 is the volume density at the coil-globule transition. This value is fixed to be $\rho_0=[19/27n]^{1/2}$, which is the second-order transition critical point in the limit of chain length $n\to\infty$, according to the Landau theory of phase transitions [14]. Also note that, whereas the original Dill model [15] accounted for the fraction of solvophobic residues and fraction of buried and surface exposed residues in the native state, we have ignored these parameters so that ε reflects a mean-field transfer energy across all residues.

The second of the models we examined is the modified-Sanchez model of Ziv and Haran [11] which we will refer to as the "Ziv" model. This model starts with the empirical ideal-chain Flory-Fisk distribution [17] of the radius of gyration $R = \langle R^2\rangle^{1/2}$ and weights it by a Boltzmann factor capturing the transfer free energy of the chain:

$$P(R) \sim R^6 \exp(-7R^2/2R_0^2)\exp(-ng(\rho,\varepsilon)/k_BT)$$

where

$$g(\rho,\varepsilon) = -\frac{1}{2}\rho\varepsilon + k_BT[(1-\rho)/\rho]\ln(1-\rho)$$

Here, R_0 is the radius of gyration of the ideal chain, which is fixed by the relation

$$\rho_0 = \frac{R_{native}^3}{R_0^3}$$

Fitting polymer theory models to simulated radius of gyration data.

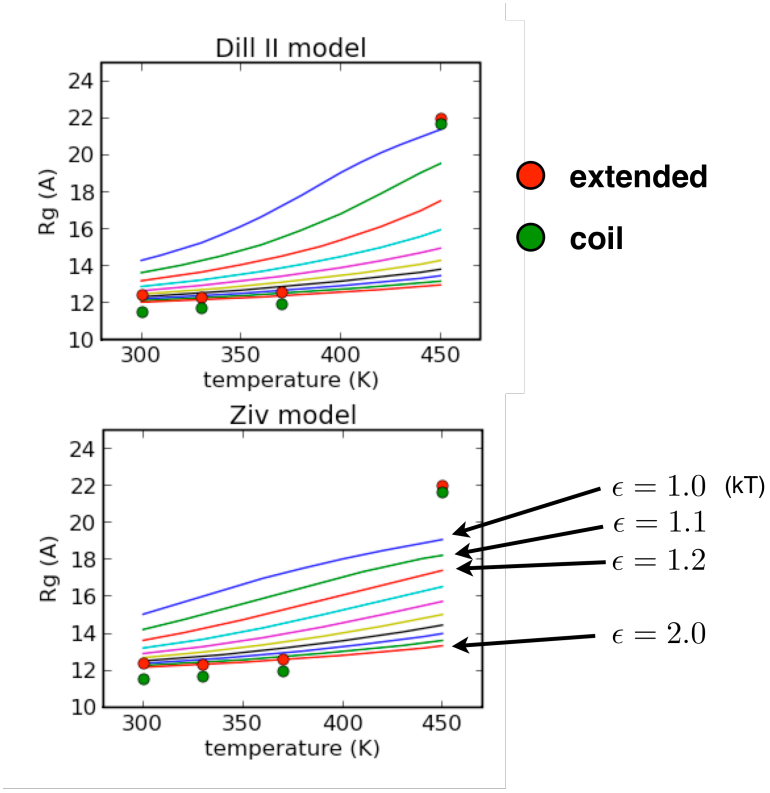
The average radius of gyration after 1 microsecond, for the ensemble started from the extended state and simulated at temperatures T=300K, 330K, 370K and 400K was used for fitting. R_{native}, the radius of gyration of the maximally compact state, was set to 11.5Å, the lowest value of mean R_g observed in our simulations. A computer program was written to find the best ε parameters using least-squares fitting to the R_g vs. T data (**Table S2**), by numerically calculating the volume density ρ that minimizes the free energy for the Dill and Ziv models.

Table S2. Simulated temperature and mean radius of gyration for ensembles started from the extended state.

Temperature (K)	average R_g (after 1 μs) (Å)
300	12.42
330	12.32
370	12.63
450	21.99

Because there is no explicit temperature-dependence in the GBSA implicit solvent model used [7], the simplest assumption would be to hold ε constant. However, this model produces very poor fits to the polymer-theory models (**Figure S10)**. Whereas the polymer theory predicts a slow continuous chain expansion as temperature increases, our simulations show a very sharp transition: compact R_g for 300K, 330K and 370K, and an expanded state at 450K.

Figure S10. Various constant values of $\varepsilon(T) = \varepsilon_0$ yield poor polymer theory fits to the R_g versus temperature data.



To produce a satisfactory fit, we found it was necessary to introduce a temperature-dependent $\varepsilon(T)$. We used a simple linear model of the transfer energy $\varepsilon(T) = \varepsilon_0 - (T-T_0)\Delta S$, where T_0= 300K. The values of ε_0 and ΔS were then found by least-squares fitting of the polymer theory prediction to the data, for both the Dill and Ziv models. We also fit using two variations of how the chain length n is defined. Polymer theory models typically assume a freely-jointed chain of monomers, with the persistence length as the bond length. The Dill model uses $n=N/1.4$, where N is the number of residues in the protein, accounting for the fact that the persistence length of polypeptides is about 1.4 amino acids [15] whereas the Ziv model uses $n=N$. In practice, we find that the results are not sensitively dependent on the value of n, an additional test of the robustness of the model.

The fits using all four kinds of models (Dill and Ziv models, each tested with $n=N/1.4$ and $n=N/1.4$) are all very similar, showing good agreement between the Dill and Ziv models (**Figure S11a**).

Interestingly, the best-fit models show large free energies (~4-6 kcal/mol) of transfer at low temperatures compared to the range of values computed by Ziv and Haran for

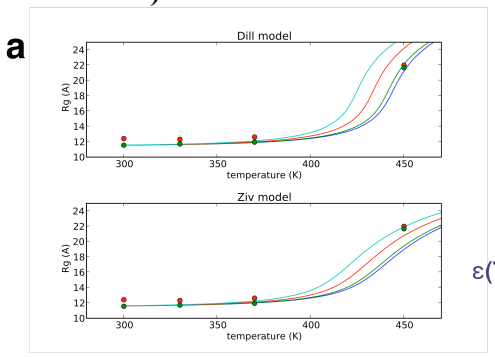
various proteins in low concentrations [GdnHCl] (~1-2 kcal/mol) (**Figure S11b**). This suggests that GBSA implicit solvent models overstabilize compact states, perhaps more significantly than currently appreciated. We discuss the implications for GB models in a latedr next section.

As further validation of the Ziv model, we took the additional step of fitting for the parameters ε_0 and ΔS that yield a model $P_{\mathrm{Ziv}}(R_{\mathrm{g}})$ that best empirically reproduces the simulated distribution, $P_{\mathrm{sim}}(R_{\mathrm{g}})$ (compiled for trajectory data after 1 μs). The fit was performed by minimizing the relative entropy, D, between the model and simulated distributions equally weighting all temperatures (T = 300, 330, 370, 450K) and both ensembles (*ens* = coil, extended) simultaneously.

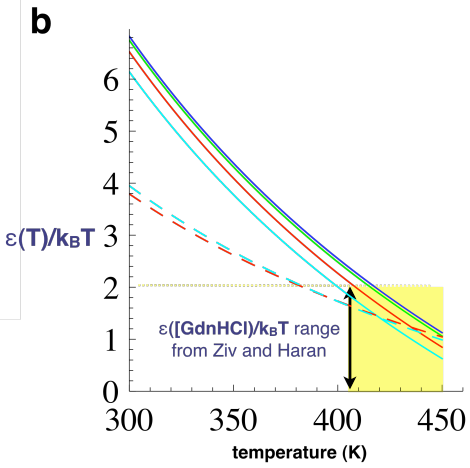
$$D = \sum_{T,ens} \sum_{R_g} P_{Ziv}(R_g|T, ens) \log \frac{P_{Ziv}(R_g|T, ens)}{P_{sim}(R_g|T, ens)}$$

The result of the fitting procedure is shown in **Figure S11b and S11c**. Overall, the results are very similar to the fits obtained when only considering the mean radius of gyration at each temperature. The Ziv model distributions are slightly narrower at low temperatures, a consequence of the fit, in which the more smoothly-expanding $P_{\mathrm{Ziv}}(R_{\mathrm{g}})$ is forced to accommodate the more abruptly expanding $P_{\mathrm{sim}}(R_{\mathrm{g}})$. We also tried fitting R_{native} as a free parameter in these fits, and obtained poorer fits, especially at high temperature. Both of these findings suggest that the simulation data may be better fit by a higher-order expression for $\varepsilon(T)$ describing a more abrupt, non-linear dependence on temperature. Given the limited set of radius of gyration values considered, and lack of further theoretical justification, we avoid fitting any higher-order expression for $\varepsilon(T)$, which would lead to over-fitting.

Figure S11. (a) Four different temperature-dependent polymer theory models with $\varepsilon(T) = \varepsilon_0 - (T-T_0)\Delta S$, yield robust fits to the average R_g versus temperature. The best parameters found by least-squared fitting for each model is shown below. (b) A plot of $\varepsilon(T)/k_BT$ for the four models shows large free energies of transfer at low temperatures compared to the range of values (yellow) computed by Ziv and Haran for various proteins in increasing concentrations of GdnHCl denaturant (see Figure 1 of reference[11]). The dashed lines are the $\varepsilon(T))/k_BT$ found in a best-fit of the Ziv model $P(R_g)$ to the distributions observed in our simulations. (c) Best parameters found for a fit of the Ziv model $P(R_g)$ to the simulated distributions. Simulated (blue) and best-fit (black) Ziv model $P(R_g)$ distributions are shown for $n=N$ ($n=N/1.4$ gives similar results).

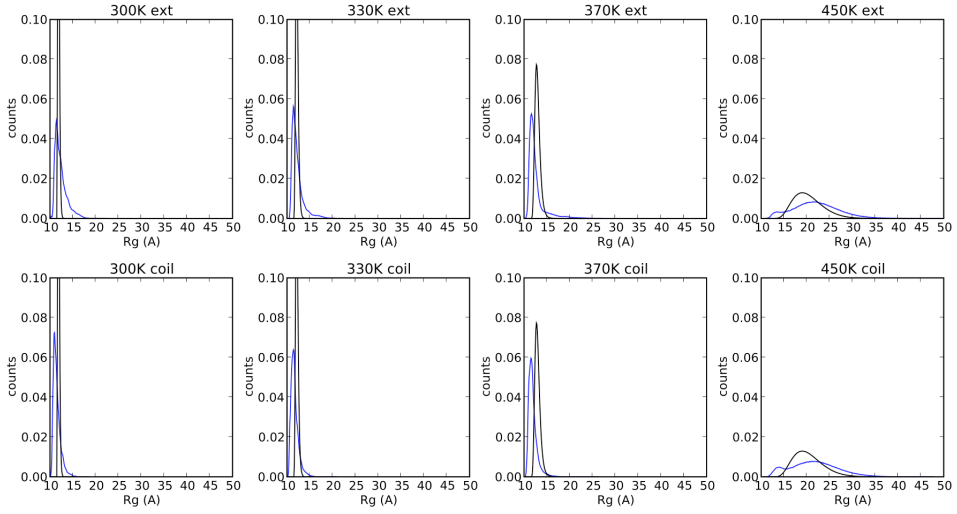


model	ℓ (a.a.)	ε_0 (kcal/mol)	ΔS (cal/mol/K)
Dill	1.0	4.067	20.403
Dill	1.4	4.022	20.489
Ziv	1.0	3.890	20.888
Ziv	1.4	3.659	20.663



c

model	ℓ (a.a.)	ε_0 (kcal/mol)	ΔS (cal/mol/K)
Ziv	1.0	2.26	8.85
Ziv	1.4	2.35	9.81



Effective [GdnHCl] for comparing simulated and experimental unfolded ensembles.

The 450K ensemble can be compared objectively to experimental data to derive an effective concentration of GdnHCl at which the experimental ensemble might be compared to the simulated ensemble. Ziv and Haran [11] computed free energies of transfer ε([GdnHCl]) and expansion factors α^2([GdnHCl]) from histograms of FRET efficiencies published in two different single-molecule FRET studies of protein L by Sherman et al. [18] and Merchant et al. [19]. The expansion factor is equivalent to $\alpha^2=R_g^2/R_0^2$, where R_g is the radius of gyration of the ensemble, and R_0 is the radius of gyration for the ideal chain at the coil-globule transition, which is fixed phase transition theory by $R_{native}^3/R_0^3=[19/27n]^{1/2}$.

Using R_{native}=11.5Å, R_g(450K) = 21.99 and n=64, we compute the expansion factor for our 450K ensemble to be α^2 ~0.81. The value of [GdnHCl] with an equivalent expansion factor is ~2.45M for the Merchant et al. data, and ~3.8M for the Sherman et al. data (see Figure 1 in reference [11]).

Thus for purposes of comparing the simulated and experimental unfolded ensembles, we conclude that the 450K ensemble is comparable to an experimental GdnHCl concentration of ~3.2M ± 1M.

The simulated unfolded states at low temperatures (300K, 330K and 370K) exhibit a high degree of compaction, becoming even more compact than the native ensembles beyond 1 μs (see Figure 1 in the main text). This observation, combined with the large free energies of transfer predicted by the polymer theory and small values of the expansion factors for the observed R_g, indicate that the low-temperature simulations can best be compared to experimental conditions in the absence of denaturant.

Implications for generalzed Born (GB) implicit solvation models

Our results suggest that GBSA implicit solvent models overstabilize compact states, perhaps more significantly than currently appreciated. Our microsecond distributed simulations of protein L unfolded states present a unique opportunity to sensitively probe this effect in a different context than the small-molecule hydration free energies and short MD trajectories often used to parameterize such models [7].

It is well known that melting temperatures in implicit solvent models are unphysically high[20]. This can be been attributed to several factors, including 1) the lack of temperature-dependence in the solvation model, 2) the overstabilization of salt-bridge interactions [21,22] and 3) forcefield torsion parameters. With our simulations of protein L unfolded states, we find that the coil-globule transition may additionally be unrealistically sharp. A similar observation has been made with simulated-tempering simulations of the N-terminal domain (residues 1-39) of ribosomal protein L9 (Xuhui Huang, personal communication).

GBSA models are commonly available in molecular simulation packages such as AMBER, and are widely used for simulating proteins. Efforts to improve GB models, by better accounting for the lack of compensating van der Waals interactions with solvent, and better parameterization of polarization and cavitation (surface area or volume) terms, have already been undertaken by several groups [23-26]. We hope that the unfolded-state simulation data generated in this work can help to validate and improve these models, ultimately bringing them into wider use. Improved GB models would not only enable better predictions of unfolded states, but would improve sampling in expanded Hamiltonian simulations (such as replica exchange or simulated tempering) that use implicit solvent models.

References

(1) Le Guillou, J. C.; Zinn-Justin, J. *Phys. Rev. Lett.* **1977**, *39*, 95-98.
(2) Neal, S.; Nip, A. M.; Zhang, H.; Wishart, D. S. *J. Biomol. NMR* **2003**, *26*, 215-240.
(3) Shen, Y.; Bax, A. *J. Biomol. NMR* **2007**, *38*, 289-302.
(4) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen], H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701-1718.
(5) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. *J. Comput. Chem.* **2009**, *30*, 864-872.
(6) Shirts, M.; Pande, V. *Science* **2000**, *290*, 1903-1904.
(7) Onufriev, A.; Bashford, D.; Case, D. *Proteins* **2004**, *55*, 383-394.
(8) Lovell, S. C.; Davis, I. W.; W. B. Arendall, I.; Bakker, P. I. W. d.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. *Proteins* **2003**, *50*, 437-450.
(9) Tsui, V.; Case, D. A. *Biopolymers* **2001**, *56*, 275-291.
(10) Lapidus, L. J.; Eaton, W. A.; Hofrichter, J. *PNAS* **2000**, *97*, 7220-7225.
(11) Ziv, G.; Thirumalai, D.; Haran, G. *PNAS* **2009**, *11*, 83-93.
(12) Flory, P. J. *Principles of Polymer Chemistry*; Cornell University Press: Ithaca, New York, 1953.
(13) Chan, H. S.; Dill, K. A. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 447-490.
(14) Sanchez, I. C. *Macromolecules* **1979**, *12*, 980-988.
(15) Dill, K. A. *Biochemistry* **1985**, *24*, 1501-1509.
(16) Dill, K. A.; Alonso, D. O. V.; Hutchinson, K. *Biochemistry* **1989**, *28*, 5439-5449.
(17) Flory, P. J.; Fisk, S. *J. Chem. Phys.* **1966**, *44*, 2243-2248.
(18) Sherman, E.; Haran, G. *PNAS* **2006**, *103* 11539-11543.
(19) Merchant, K. A.; Best, R. B.; Louis, J. M.; Gopich, I. V.; Eaton, W. A. *PNAS* **2007**, *104*, 1528-1533.
(20) Pitera, J. W.; Swope, W. *PNAS* **2003**, *100*, 7587-7592
(21) Zhou, R.; Berne, B. J. *PNAS* **2002**, *99*, 12777-12782.
(22) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.n* **2006**, *2*, 115-127.

(23) Leontyev, I. V.; Stuchebrukhov, A. A. *J. Chem. Phys.* **2009**, *130*.
(24) Labute, P. *J. Comput. Chem.* **2008**, *29*, 1693-1698.
(25) Chen, J.; III, C. L. B. *PCCP* **2008**, *10*, 471-481.
(26) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479-499.