JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

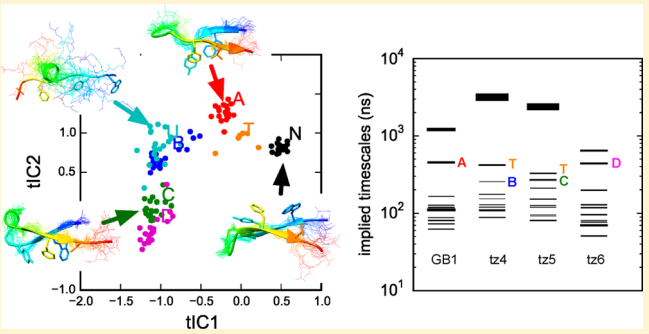
Kinetic Network Models of Tryptophan Mutations in β-Hairpins Reveal the Importance of Non-Native Interactions

Asghar M. Razavi and Vincent A. Voelz*

Department of Chemistry, Temple University, Philadelphia, Pennsylvania 19122, United States

Supporting Information

ABSTRACT: We present an analysis of the most extensive explicit-solvent simulations of β-hairpins to date (9.4 ms in aggregate), with the aim of probing the effects of tryptophan mutations on folding. From molecular simulations of GB1 hairpin, trpzip4, trpzip5, and trpzip6 performed on Folding@home, Markov State Models (MSMs) were constructed using a unified set of metastable states, enabling objective comparison of folding mechanisms. MSM models display quantitative agreement with experimental structural observables and folding kinetics, and predict multimodal kinetics due to specific non-native kinetic traps, which be identified as on- or off-pathway from the network topology. We quantify kinetic frustration by several means, including the *s*-ensemble method to evaluate glasslike behavior. Free-energy profiles and transition state movement clearly show stabilization of non-native states as Trp mutations are introduced. Remarkably, we find that "β-capped" sequences (trpzip4 and trpzip5) are able to overcome this frustration and remain cooperative two-state folders with a large time-scale gap. These results suggest that, while β-capping motifs are robust, fold stabilization by tryptophan generally may require overcoming significant non-native kinetic traps, perhaps explaining their under-representation in natural proteins.



INTRODUCTION

In the standard model of two-state protein folding, the folding free-energy landscape is minimally frustrated, with native-state interactions playing a dominant role.[1,2] However, there are many circumstances where we would like to add detail to this standard model—systems for which understanding non-native interactions may be crucial, such as intrinsically disordered proteins, those prone to misfolding and/or aggregation, synthetic foldamers, or even globular proteins.[3] Recent biophysical studies showing significant denatured-state interactions in NTL9 suggests that the importance of non-native interactions may be understated, and that a truly detailed understanding of mutational effects may require consideration of non-native states.[4]

Here, as a model system for understanding how mutations perturb the folding reaction, we consider a series of small beta-hairpin mini-proteins based on the C-terminal hairpin of the B1 domain of *Streptococcal* protein G,[5] otherwise known as the GB1 peptide. Based on GB1 peptide and other hairpins, Cochran et al. designed several remarkably stable "tryptophan zipper" (trpzip) sequences, by mutating hydrophobic core residues to tryptophan (see Table 1).[6]

Folding mechanisms of trpzip peptides have been extensively studied both experimentally[7−15] and computationally,[8,9,16−18] mostly for trpzip2 (SWTWENGKWTWK), because of its small size of only 12 residues. Compared to GB1 hairpin, the 16-residue trpzip variants are more stable and have slower observed folding relaxation times[7,8,19−21] (~14 μs for trpzip4, versus ~3.7 μs for GB1 hairpin) with slower folding and unfolding rates for trpzip4.[7] Using a combination of replica exchange molecular dynamics (REMD) simulations (0.7 μs in aggregate) and thermal unfolding experiments, Yang et al. concluded that the folding/unfolding of trpzip2 was extremely heterogeneous, with many kinetic traps contributing to "glassy" folding behavior and slow kinetics.[9] Indeed, recent single-molecule optical force microscopy studies of trpzip2 and trpzip4 suggest distinct unfolding intermediates stabilized by tryptophan pairing.[22]

Time-resolved spectroscopic studies, coupled with mutational analysis, generally support a zipping mechanism for hairpin folding, the formation of which, in turn, nucleates subsequent cross-strand pairing.[11,19,21,23−26] However, more recent experimental studies that achieve increased resolution with isotopic labeling[27,28] and/or multiple spectroscopic probes,[29] suggest a more-complicated mixture of zipping and hydrophobic collapse, depending greatly on peptide sequence. A recent FTIR study of hydrophobic mutations in trpzip2 showed a highly stabilizing role for edge-to-face Trp-Trp pairs compared to other hydrophobic pairs (e.g., Val-Val), which display faster and more-complex kinetics.[30] These folding kinetics are consistent with work from the Andersen laboratory, showing that a β-hairpin capping motif composed of a distal

Received: January 30, 2015

DOI: 10.1021/acs.jctc.5b00088
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

cross-strand pair of Trp residues confers remarkable stability to hairpin folds.[31−35]

Computational studies of hairpin folding mechanisms have provided useful mechanistic detail[36,37] and have continued to increase in resolution, with molecular simulations of β-hairpins in explicit solvent now able to routinely access the microsecond time scale.[38−40] Most recently, Best et al. has performed a large-scale explicit-solvent simulation of GB1 peptide (0.7 ms of aggregate trajectory data), and used Markov State Model (MSM) analysis to model the folding kinetics.[41−43] Notably, the MSM approach was able to more consistently estimate folding rates; a previous study based on first-passage times gave results that are sensitively dependent on the initial starting conformations.[43]

Here, we utilize the Folding@home distributed computing platform to perform the most extensive explicit-solvent simulations of β-hairpins to date (9.4 ms in aggregate), with the specific aim of examining the effects of tryptophan mutations. We examine GB1 hairpin as well as trpzip4, trpzip5, and trpzip6 variants (Table 1), to investigate how sequence mutations perturb folding. Using a tICA approach,[3,44] we build MSMs for all four sequences, and quantitatively compare folding mechanisms using a 150-macrostate MSM constructed from the combined set of data. We find that MSMs show quantitative agreement with experimental structural observables and folding kinetics, with multimodal kinetics dominated by misfolded kinetic trap states consisting of non-native conformations stabilized by hydrophobic interactions. In agreement with recent experimental studies,[30,33,34] our MSMs predict that the sequences with distal cross-strand tryptophan pairs (trpzip4 and trpzip5) are the most stable, and fold in a more two-state manner,[45] despite the fact that increased numbers of tryptophans introduce significant non-native interactions. As the number of tryptophan mutations grows, our MSMs exhibit slower folding relaxations, increased stabilization of unfolded states due to non-native interactions, and a shift in transition-state structure.[46] We quantify non-native interactions using computed kinetic frustration scores,[47] MSM network topologies,[48] and computed glass activities,[49] which indicate more inactive kinetics than mini-proteins of similar size. Overall, it appears tryptophans must be introduced judiciously to hairpin sequences in order to overcome kinetic traps, which may explain the absence of β-capping motifs in nature.

Table 1. Sequences for GB1 Hairpin and Selected Tryptophan Mutants

name	sequence
GB1	GEWTYDDATKTFTVTE
trpzip4	GEWTWDDATKTWTWTE
trpzip5	GEWTYDDATKTFTWTE
trpzip6	GEWTWDDATKTWTVTE

METHODS

Molecular Dynamics Simulation. Molecular coordinates for trpzip4 were obtained from the Protein Data Bank (PDB ID 1LE3), and threaded using UCSF Chimera[50] to generate structures for GB1, trpzip5, and trpzip6. Before performing explicit-solvent simulations, we first performed extensive implicit-solvent simulations (>160 μs for each sequence) at 300 K on the Owlsnest high-performance computing cluster, using the AMBER ff96 force field[51] with the OBC GBSA implicit solvation model,[52] a combination previously shown to be most accurate for *ab initio* implicit solvent folding studies.[53−55] MSMBuilder2[56] was used to cluster the resulting trajectory data into 2000 microstates using a tICA approach with backbone + C_β atom pairs. Microstates were then coarse-grained into 20 macrostates using the BACE algorithm.[57] The generator of the most populated microstate in each macrostate was used to initiate explicit-solvent simulations on the Folding@home distributed computing platform.[58]

Explicit-solvent simulations for all sequences used the AMBER ff99sb-ildn force field[59] with 1507 TIP3P water molecules, 6 Na^+ and 3 Cl^- ions in a $(36.4\ Å)^3$ cubic box. Covalent hydrogen bond lengths were constrained using LINCS, and PME electrostatics were used with a nonbonded cutoff of 9 Å. NVT simulations were performed using stochastic (Langevin) integration at 300 K, using a 2 fs time step and snapshots written every 100 ps. Protein and solvent atom groups were coupled separately using an inverse friction constant of 1.0 ps. More than 2.3 ms of trajectory data was generated for each sequence (see Table 2), with average trajectory lengths of ∼400 ns, and maximum trajectory lengths more than 2 μs (see Figure S1 in the Supporting Information).

Table 2. Total Simulation Time for Implicit and Explicit Simulations at 300 K

sequence	implicit (μs)	explicit (ms)	number of trajectories	mean trajectory length (ns)	maximum trajectory length (ns)
GB1	170	2.38	7618	312.2	2140
trpzip4	245	2.44	5278	462.5	2140
trpzip5	162	2.33	5185	449.4	2270
trpzip6	160	2.34	5533	423.1	2090

All simulations used unsymmetrized[60] force fields, since the computed observables considered here did not require symmetrization.

Markov State Model (MSM) Construction. MSMBuilder2 was used to construct MSMs, which entails defining a set of discrete conformational microstates, and estimating transition rates from the simulation trajectory data. Microstate conformational states were computed separately for each sequence, using the recently developed tICA (time-lagged independent component analysis) method.[3,44] The tICA method finds a low-dimensional principal component subspace that captures the slowest conformational motions, by diagonalizing a time-lagged covariance matrix (TLCM) of interatomic distances, with respect to a corresponding covariance matrix (CM).[61] The resulting coordinate system is more suitable for conformational clustering as a means to identify kinetically metastable states. The TLCM and CM were constructed using backbone + C_β atom pair distances for all residues (in total 3003 distances). A lag time of 5 ns was chosen to construct TLCM, although we found that the TLCM is not very sensitive to lag time: using shorter or longer lag times attains similar results (see Figure S2 in the Supporting Information). As others have shown previously,[3,44] we find that tICA gives much better results, compared to other distance-based conformational clustering metrics, including root-mean-square deviation (rmsd), radius of gyration, or dihedral angle clustering, which consistently yield much shorter implied relaxation time scales (see Figure S3 in

DOI: 10.1021/acs.jctc.5b00088
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

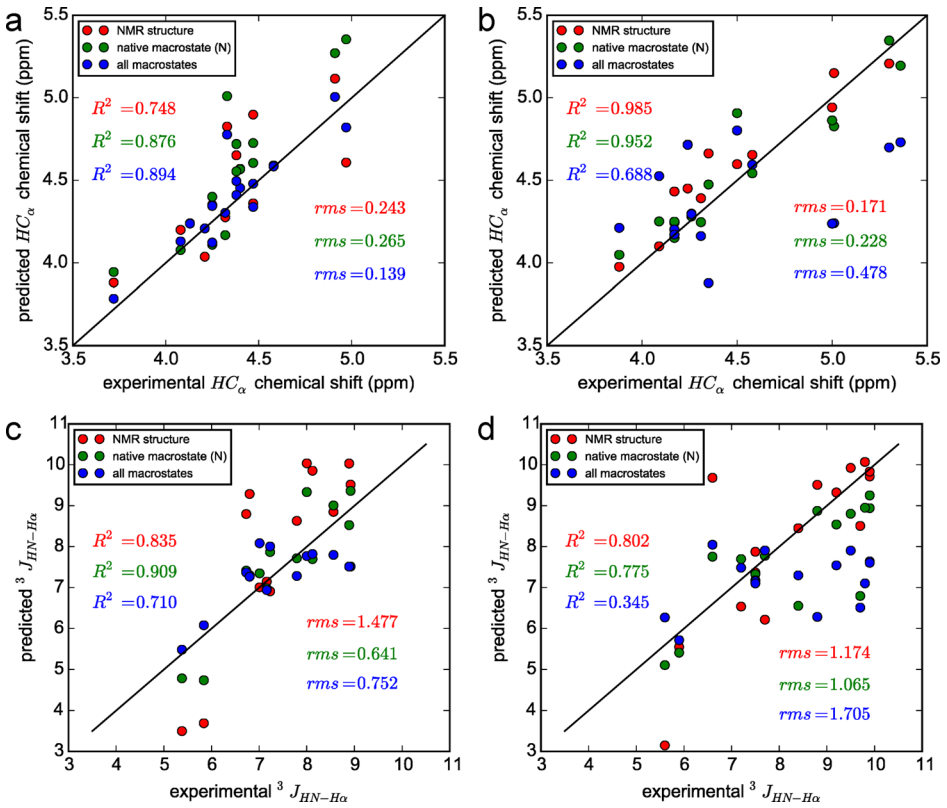


Figure 1. Comparison of experimental chemical shifts and *J*-couplings with simulation predictions. Experimental versus predicted HC_α chemical shifts for (a) GB1 hairpin and (b) trpzip4, and experimental versus predicted $J_{HN\text{-}HA}$ values for (c) GB1 hairpin and (b) trpzip4. Predicted values were computed using the published NMR structure for trpzip4 (red, 1LE3), the native macrostate ensemble predicted by the MSM (green), and the entire MSM ensemble (blue). Correlation coefficients (R^2) and rms devations from experiment are reported on each panel.

the Supporting Information), indicating more discretization error in defining metastable states.[62]

Selecting an appropriate dimensionality to represent the tICA subspace is important: using too few tICA components reduces spatial resolution, while using too many components reduces connectivity. To find the most practical tICA subspace for our explicit-solvent trajectory data, we systematically constructed over 800 different MSMs to explore the effects of varying the number of tICA eigenvectors and number of microstates. We also examined different tICA lag times and the MSM lag times for estimating transition rates. From a detailed analysis of the resulting MSMs, we determined that optimal microstate models should use four tICA components and 2000 microstates (see Figures S4 and S5 in the Supporting Information).

After hybrid *k*-medoids clustering in the tICA subspace, all trajectory snapshots were assigned to microstate indices, and sliding-window transition counts were used to estimate a transition probability matrix $\mathbf{T}$, whose elements T_{ji} contain the probabilities of transitioning from state i to state j in lag time τ. The maximum likelihood estimator (MLE) implemented in MSMBuilder2[56] was used to obtain estimates of $\mathbf{T}$. The eigenvalues μ_n of $\mathbf{T}$ yield the implied molecular relaxation time scales,

$$\tau_n = -\frac{\tau}{\ln \mu_n}$$

while the associated eigenvectors describe the relaxation modes, with the equilibrium probability distribution given by the (first) stationary eigenvector, $\pi = \mathbf{T}\pi$. All MSM models reported here use a lag time of τ = 10 ns, chosen based on a study of the implied time scales computed as a function of lag time, which begin to plateau after 10 ns. We additionally report folding time scales at τ = 40 ns.

In the Results section, we present projections of (right) eigenvectors φ_n of $\mathbf{T}$ onto rmsd-to-native reaction coordinates using the eigenvector sign convention that $\langle\psi_n|\mathbf{p}\rangle$ is positive, where ψ_n is n^{th} left eigenvector of $\mathbf{T}$, and $\mathbf{p}$ is a uniform distribution of initial populations. This way, we can think of each eigenvector as contributing to the time evolution of $\mathbf{p}(t)$ with positive amplitude, according to the solution of the master equation $d\mathbf{p}/dt = \mathbf{Kp}$, where $\mathbf{T} = \exp(\tau\mathbf{K})$, given by $\mathbf{p}(t) = \sum_n \langle\psi_n|\mathbf{p}(0)\rangle\ \varphi_n\ \exp(-t/\tau_n)$, where $\mathbf{p}(0)$ are the initial (uniform) state populations.[63,64]

Macrostate MSMs for Comparing Hairpin Sequences. To objectively compare differences in folding mechanisms across different hairpin sequences, we performed conformational clustering of the combined trajectory data of all four hairpin sequences, using only the subset of backbone + C_β atoms, which are the same for all sequences. The tICA method was used with a lag time of 5 ns, and the first four tICA components were used as a subspace for conformational clustering into 1000 microstates. This microstate clustering is fairly fine-grained, and not all microstates were visited by all four sequences, because of finite sampling effects. To achieve a macrostate MSM sampled by all representative hairpin sequences, we used the BACE algorithm[57] to obtain a coarse-grained MSM with 150 macrostates (see Figure S6 in the Supporting Information). The macrostate assignments for each

DOI: 10.1021/acs.jctc.5b00088
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 3. Previously Published Experimental and Simulated Relaxation Times for GB1 and trpzip Hairpins

		τ_{obs} (μs)[a]	τ_f (μs)	τ_u (μs)	T (K)
GB1	experiment (ref 19)[b]	3.5	6	6	297
trpzip4	experiment (ref 7)[c]	14.0	14.9 ± 0.6	234.0 ± 10.7	297
GB1	experiment (ref 21)[d]	4.2	18.2	5.5	298
GB1	simulation (ref 41)[e]	13 ± 6			300
GB1	simulation (ref 36)[f]		0.08−0.12		298
GB1	simulation (ref 37)[f]		30−60		298
GB1	simulation (ref 77)[g]		1−5		300

[a]The observed folding relaxation time is given as $\tau_{obs} = (\tau_f^{-1} + \tau_u^{-1})^{-1}$, where τ_f and τ_u are the folding and unfolding times, respectively; the τ_{obs} values can be compared to MSM implied time scales. [b]As measured by T-jump time-resolved fluorescence. [c]As measured by T-jump infrared spectroscopy. [d]Extracted from NMR exchange broadening data, using a revised estimate of unfolded-state stability (ca. 30%). [e]Using explicit-solvent simulations in AMBER ff03*/TIP3P. [f]From discrete path sampling (DPS), using an implicit solvent model. [g]Implicit-solvent simulations using parallel distributed computing.

sequence were then used to construct four separate MSMs, each utilizing the same macrostate definitions.

We find that the macrostate representation yields accurate models for all sequences, because of the high sequence similarly of GB1 hairpin and trpzip. The macrostate model recapitulates the (individually constructed) microstate MSM implied time scales and their associated fluxes (see Figure S11 in the Supporting Information), suffering only marginally from the increased discretization error (see Figure S7 in the Supporting Information). Structural comparison of molecular conformations from each macrostate for different sequences show striking similarities (see Figure 5, presented later in this work). Projections of trajectory data onto the first two tICs, and projections onto radius of gyration and rmsd observables (Figure S8 in the Supporting Information) confirm the high degree of conformational similarity of macrostate conformations across all hairpin sequences.

Transition Path Theory (TPT) calculations[65] were performed as described previously[54,66,67] to compute committor (p_{fold}) values and steady-state folding fluxes along macrostate pathways. Gephi[68] was used to visualize macrostate MSM networks, using the Force Atlas algorithm for graph minimization and layout. Kinetic frustration scores were computed as described by Savol and Chennubhotla,[47] with the native basin being the folded native-state macrostate. Mean glass activities were computed using s-ensemble techniques previously described by Weber et al.[49]

Uncertainties in predicted macrostate free energies were estimated using a cross-validation procedure, leaving out one-fourth of the trajectories over four different trials, with nonsliding-window estimates of transition counts.

RESULTS

Comparison of Simulated and Experimental Structural Observables. The chosen force field (AMBER ff99sb-ildn) correctly predicts the native structure of GB1 and trpzip variants. The distribution of rmsd-to-native (backbone + C_β) seen in the trajectory data—calculated using the NMR solution structure of trpzip4 as a reference[6]—has a folded-state maximum at ∼1.2 Å for all sequences (see Figure S9 in the Supporting Information, as well as Figure 3, presented later in this work). Experimentally measured NMR observables such as chemical shifts and vicinal J-couplings provide information about conformationally averaged structural observables that can be used to validate simulation predictions.[69] For this purpose, we compared predicted conformational ensembles of GB1 hairpin and trpzip4 to experimentally measure data.[5,6] Predicted HC_α (Figure 1) and HN chemical shifts (Figure S10 in the Supporting Information) were computed using SHIFTX2,[70] and J-couplings were calculated for HN-HA using coefficients from refs 71−73.

For GB1 hairpin, HC_α chemical shifts predicted using the full ensemble of MSM macrostates (weighted by predicted equilibrium probabilities) have rmsd values of 0.139 from experimental values, consistent with previous simulation studies. For example, an REMD simulation study of GB1 peptide in implicit solvent (OPLS-AA/AGBNP) found rmsd values for HCα chemical shifts of ∼0.2 at 300 K, with the best agreement of ∼0.1 occurring at 400 K.[69] REMD studies of GB1 and designed hairpin GB1m3 using CHARMM36, a force field optimized and validated extensively against NMR data, yielded rmsd values of HCα chemical shifts of 0.132 and 0.22, respectively.

Predicted $J_{HN\text{-}HA}$ coupling constants for GB1 hairpin have an rmsd of 0.75 Hz from experimental values. This is comparable to the recent results of Beauchamp et al., who tested the accuracy of several force fields and water models in predicting NMR observables (chemical shifts and coupling constants) for short peptides (di-, tri-, and tetra-peptides) and ubiquitin, and found that for $^3J_{HN\text{-}HA}$, the best-in-class force field and water model (ff99sb-ildn-NMR + TIP4P-EW) gave a root-mean-square (rms) error of 0.55 Hz.[74] Our results are also similar to the accuracy of predictions for capped amino acid residues (rms errors of ∼0.5 Hz) calculated from the ABSINTH implicit solvent model when coupled with several popular atomic force fields (e.g., OPLS-AA, GROMOS, and AMBER ff99).[75]

Moreover, we find that both chemical shifts and J-coupling constants calculated from the entire simulated ensemble of GB1 peptide have lower rms deviations from experimental values than those calculated using only the native structure or native macrostate, which is consistent with the findings of Li and Brüschweiler, that conformational averaging present in microsecond trajectories significantly improves chemical shift predictions.[76]

For trpzip4, predicted HC_α chemical shifts agree less well with experiment, but are still comparable to the accuracy of SHIFTX2 predictions (0.4412).[70] Predicted HN chemical shifts (see Figure S10 in the Supporting Information) are poorer than the average HN accuracy of SHIFTX2 for all ensembles. Interestingly, much better agreement is obtained if we compare to the native structure alone. This result, and other findings discussed in more detail below, suggest that simulations underestimate the native population of trpzip4.

DOI: 10.1021/acs.jctc.5b00088
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Predicted Folding Kinetics for GB1 and trpzip Mutants Agree with Experimental Measurements. The experimental folding relaxation times for GB1 and trpzip4 are ~3.5−4.2 μs and 14.9 ± 0.6 μs, respectively[7,21] (see Table 3). The 2.8-fold slower folding kinetics of trpzip4 is recapitulated by our MSMs for GB1 and trpzip4, whose slowest (i.e., folding) implied time scales are 2.0 and 4.7 μs, respectively, using a lag time of 10 ns (2.35-fold slower). Predicted time scales using a lag time of 40 ns are 3.5 and 10.3 μs for GB1 and trpzip4, respectively (trpzip4 is 2.9-fold slower, see Figure 2). While implied time scales using the 40 ns lag time more closely match the experiment, we have conservatively chosen a lag time of 10 ns for all subsequent MSM analyses. In either case, our folding rate predictions match experimental values more closely than previous simulation-based estimates.[36,37,41,77]

Predictions of the folding time scales for trpzip5 and trpzip6 variants are similar to that of GB1 and trpzip4, with folding times ordered from shortest to longest, according to the ranking trpzip6 < GB1 < trpzip5 < trpzip4. To our knowledge, the folding kinetics of trpzip5 and trpzip6 have not been measured, making our calculations a blind prediction to be validated experimentally.

Multimodal Folding Kinetics Arise from Misfolded Kinetic Traps. A hallmark of two-state folding is a well-separated gap between the slowest and next-slowest implied time scale.[54,78] Even in the absence of time scale gap, folding may appear two-state if only one of several relaxation eigenmodes dominates the folding flux, or if spectroscopic probes are preferentially sensitive to particular structural observables.[79,80] Following Beauchamp et al.,[78] we assess the importance of each relaxation eigenmode, computing the equilibrium folding flux through the nth eigenmode as $\|\varphi_n\|^2$, where φ_n is the properly π-normalized nth eigenvector of $\mathbf{T}$. Results for the 10 slowest implied time scales are presented for the 2000-microstate MSM (see Figure S11 in the Supporting Information) and the 150-macrostate MSM for all hairpin sequences (Figure 2). For all sequences except trpzip6, the slowest relaxation mode has the greatest folding flux, with a time scale gap indicative of two-state folding. trpzip4 and trpzip5, the two sequences with distal cross-strand Trp-Trp pairs, show the least deviation from two-state folding, while GB1 hairpin and trpzip6 show a greater degree of complexity, with smaller time scale gaps and more-disperse folding fluxes. Time scales and fluxes for microstate versus macrostate models compare quite favorably, with macrostate MSMs preserving the salient features and suffering only modestly from lumping artifacts (slight decreases in implied time scales and flux distributions).

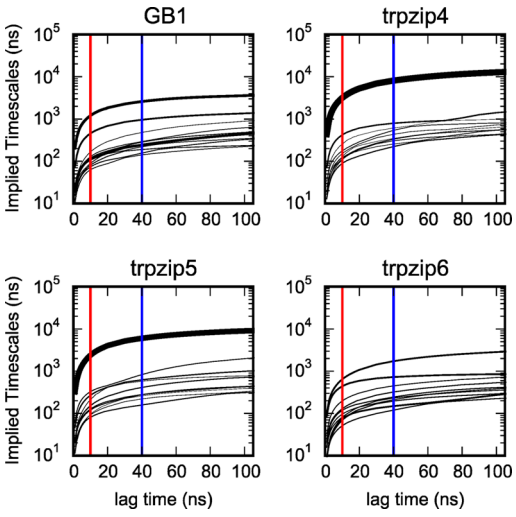


Figure 2. Ten slowest implied time scales versus lag time for 150-macrostate MSM models of all hairpin sequences. Line widths are proportional to equilibrium fluxes $\|\varphi_n\|^2$ through each eigenmode φ_n, computed using a lag time of 10 ns, and normalized so that the sum of the top 10 fluxes is constant in all sequences. Red and blue lines mark the 10 and 40 ns lag times, respectively, at which we report predicted folding relaxation time scales (see Table 4).

To identify structural changes corresponding to the relaxation eigenmodes for each sequence, we projected the first four eigenvectors of the microstate MSM onto a single rmsd-to-native observable (backbone + C_β atoms; see Figure 3) using a kernel density estimation approach.[81] The projection is a sum of Gaussian kernels with mean rmsd r_i and standard deviation σ_i for each microstate i, with each kernel weighted by the value of each eigenvector element. For all sequences, the first (stationary) eigenvector (corresponding to the equilibrium populations) and second eigenvector (corresponding to unfolded-to-folded population flux) are similar: the stationary eigenvector shows two main peaks, corresponding to the native and unfolded ensembles, while the second eigenvector shows negative and positive peaks for folded and unfolded populations, respectively. However, the third and fourth eigenvectors are unique for each sequence and show peaks at non-native rmsd values. Inspection of the molecular conformations comprising these peaks reveals unique misfolded kinetic trap states, stabilized by hydrophobic interactions and alternative backbone hydrogen bonding.

To get a clearer picture of these states and their physical significance, we hereafter will utilize the 150-macrostate MSMs, in which the key microstates are mapped to one or a handful of particular macrostates that dominate the third and fourth eigenvectors (see Figure S12 in the Supporting Information). A summary of these macrostates, and their role in the third and fourth eigenmodes, is provided in Figure 4. Note that these states are chosen to help illustrate differences in folding mechanisms across the four hairpin sequences, and that many other macrostates contribute to conformational dynamics.

Molecular visualizations of selected macrostates for all hairpin sequences are shown in Figure 5. Visually, macrostate conformations are extremely similar across the different hairpin sequences, attesting to the high degree of sequence similarly, and the success of tICA clustering and lumping using the combined trajectory data. The native macrostate (N) is well-defined and easily discerned as the global free-energy minima. The third eigenmode of trpzip4 and trpzip5 is dominated by a near-native macrostate (T) with native-like strand-pairing and stacking of hydrophobic groups, but a more open turn region in which the carboxylate group of residue Asp6 coordinates with backbone amide hydrogens in the turn (see Figure S13 in the Supporting Information). This state is somewhat reminiscent of a "bulged" folding intermediate of trpzip2 recently characterized by Jones et al., using temperature-jump two-dimensional (2D) infrared (IR) spectroscopy.[15] Macrostate A has a native-like turn region, similarly stabilizing by the Asp7-Lys10 salt bridge, but is otherwise completely inverted, with all four of the key hydrophobic residues interacting on the opposite plane of the hairpin. Macrostate B has a non-native turn, and a flipped N-terminal strand compared to the native state. Macrostates C and D consist of two groups of conformations with very similar

DOI: 10.1021/acs.jctc.5b00088
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 4. Numbers of Folding/Unfolding Trajectories, and Predicted Folding Relaxation Rates τ_{obs} (μs) at 300 K for GB1 and trpzip Hairpins

			MSM type: 2000 microstates		MSM type: 150 macrostates	
	n_{fold} [a]	n_{unfold} [b]	lag time: τ = 10 ns	lag time: τ = 40 ns	lag time: τ = 10 ns	lag time: τ = 40 ns
GB1	99	94	2.0 ± 0.3	3.5 ± 2.5	1.2 ± 0.1	2.6 ± 0.3
trpzip4	73	49	4.7 ± 1.3	10.3 ± 5.5	3.2 ± 0.3	8.1 ± 1.1
trpzip5	120	64	2.6 ± 0.6	6.3 ± 2.5	2.4 ± 0.3	6.0 ± 0.6
trpzip6	137	64	1.7 ± 0.5	5.9 ± 1.8	0.64 ± 0.11	1.7 ± 0.3

[a] n_{fold} is the number of trajectories that start from unfolded states (macrostate p_{fold} < 0.45) and later reach native macrostate N. [b] n_{unfold} is the number of trajectories that sample the native macrostate N, and later reach an unfolded state.

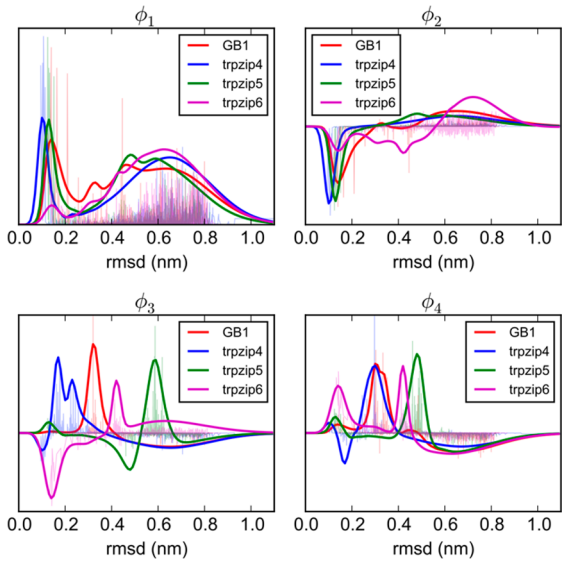


Figure 3. Transition matrix eigenvectors φ_n, projected on rmsd space using a kernel density method, for all hairpin sequences. The eigenvectors φ_1 and φ_2 reflect equilibrium populations and the slowest folding relaxation mode, respectively, while φ_3 and φ_4 are dominated by positive peaks indicative of non-native kinetic traps.

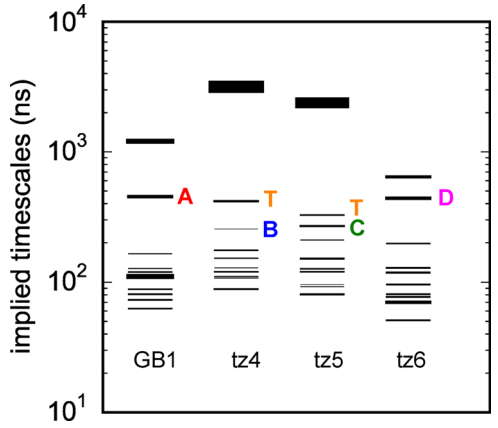


Figure 4. Implied time scale spectra from 150-macrostate MSMs (lag time = 10 ns) of all hairpin sequences, annotated with "trap" macrostates T, A, B, C, and D that dominate particular eigenmodes (see Figure 5).

α-helical turn regions and misregistered strands. One group of conformations has both strands inverted, while the other has only the N-terminal strand inverted (see Figure S13 in the Supporting Information). There is some variability in the assignment of microstates to macrostates C and D, because of the proximity of these two conformational groups in the tICA subspace (Figure 5b), with the presence of non-planar Val14 versus planar Trp14 as a probable influence. Macrostates C and D play very similar roles in folding and would likely be lumped together by the BACE algorithm in more-coarse-grained models. Macrostate U corresponds to the "unfolded state", which we describe as such because a large number of microstates map to it, and it has the lowest non-native free-energy minimum.

The predicted equilibrium probabilities of the trap macrostates are small, but conceivably detectable by experiment: trpzip4 and trpzip5 are predicted to populate macrostate T with populations 1.8% and 1.0%, respectively; GB1 hairpin populates macrostate A at 2.4%, trpzip5 populations macrostate C at 1.7%, and trpzip6 populates macrostate D at 1.9% (see Table S1 in the Supporting Information for a full list).

Tryptophan Mutations Perturb Folding via Significant Non-Native Interactions. To compare the folding mechanisms of the four different sequences, we used Transition Path Theory[65,66,82−85] (TPT) to calculate equilibrium folding fluxes from the unfolded state U to the native state N, along with p_{fold} values (i.e., committors) for each macrostate, which reflect the probability that trajectories starting from each macrostate reach N before U. Macrostates with p_{fold} values near 0.5 correspond to the transition-state ensemble. Figure 6 shows the conformational free energy of each macrostate as a function of p_{fold}, along with the 10 pathways with the largest folding fluxes, for all four sequences. A large number of competing non-native macrostates compose the unfolded-state ensemble near zero p_{fold} values, with "trap" macrostates A, T, C, and D being low-free energy unfolded states for GB1, trpzip4, trpzip5, and trpzip6, respectively. Superimposed on each panel is a qualitative free-energy profile, calculated from a kernel density estimate of the partition function $Z(p_{fold})$ using a Gaussian kernel of width σ = 0.1.

To illustrate how the folding landscape is perturbed by sequence mutations, we select three intermediate macrostates, I1, I2 and I3, near the transition-state ensemble, i.e., with p_{fold} values near 0.5 for all sequences. Free-energy profiles near these intermediates show that folding barriers are largest for trpzip4, and smallest for trpzip6, in expected correspondence with the predicted folding rates across the four sequences. As the number of tryptophan residues increases, the MSM predicts that unfolded states are stabilized, with respect to the transition state. The stabilization mechanism is almost certainly dependent on interactions of hydrophobic residues, which are ubiquitous in non-native macrostates (see Figure S14 in the Supporting Information).

The predicted unfolded-state stabilization can also be seen through the movement of transition-state macrostates I1, I2, and I3 toward more-structured conformations as more tryptophans are introduced into the hairpin sequence. Since

DOI: 10.1021/acs.jctc.5b00088
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

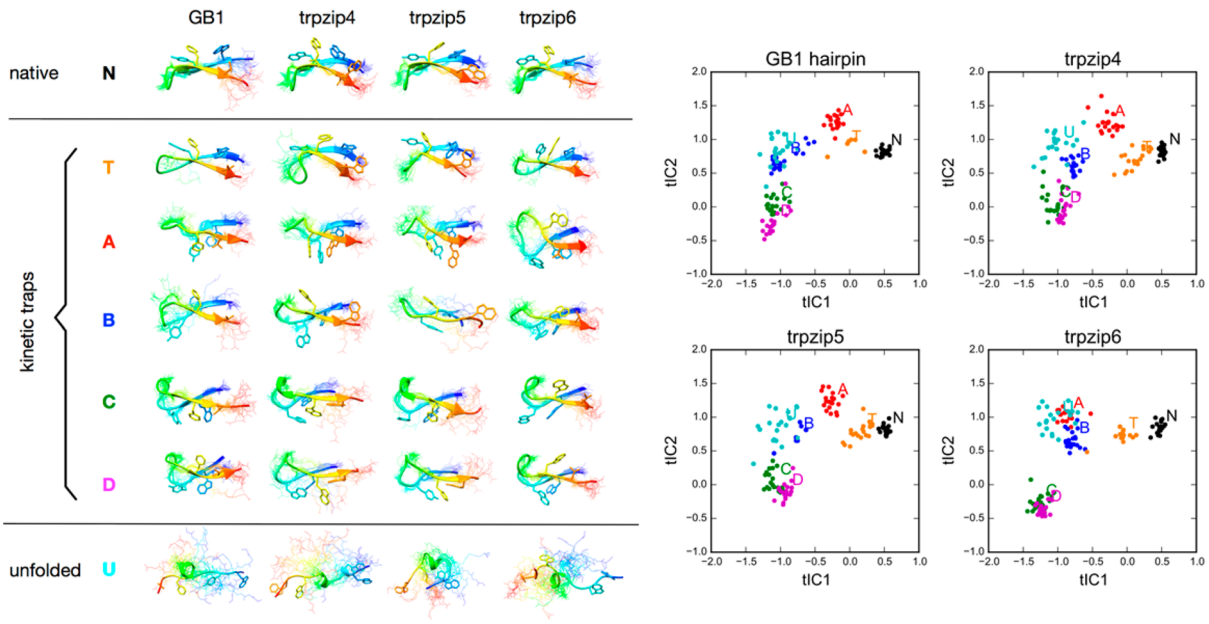


Figure 5. Molecular visualization of relevant macrostates from MSMs of GB1 hairpin, trpzip4, trpzip5, and trpzip6 using a unified definition of metastable states. (Left) Twenty conformations (wireframe) were randomly selected from each macrostate, with one exemplar rendered as a ribbon backbone with hydrophobic side chains. (Right) The same 20 conformations from each macrostate, projected onto the two largest time-lagged independent components (tIC1 and tIC2). All four sequences populate highly similar metastable conformational states, enabling comparisons by MSMs.

p_{fold} is a *kinetic* reaction coordinate, we alternatively order macrostates I1, I2, and I3 along a *structural* reaction coordinate (see Figure 7). We see that, for the GB1 hairpin, macrostate I1 has a p_{fold} near 0.5, whereas for trpzip5 and trpzip6, macrostate I2 has a p_{fold} near 0.5, and, for trpzip4, macrostate I3 has a p_{fold} near 0.5. This movement is consistent with the Hammond postulate,[46] which holds that as reactants are stabilized, transition states more closely resemble product structures.

Inspection of transition-state structures confirms the conclusions of previous studies, that hairpin folding proceeds via a so-called "zipping" mechanism, whereby turn formation precedes strand pairing. The progression of intermediates I1, I2, and I3 show an intact hairpin topology, with hydrophobic interactions "zipping" from the turn, although the pathways of greatest net flux show concerted assembly as well, especially for trpzip4. This heterogeneity of folding routes is consistent with computational[38] and experimental studies,[29] suggesting the existence of both zipping and "hydrophobic collapse" folding mechanisms.

Kinetic Frustration Metrics and Network Topologies. To further elucidate the mechanistic roles of kinetic traps, we computed kinetic frustration scores for each macrostate following the recent work of Savol and Chennubhotla.[47] The kinetic frustration score $f_{nat}(i)$ quantifies the percent change in the mean first passage time from macrostate U to macrostate N, upon the removal of macrostate i from the network. Macrostates with negative frustration scores are said to *inhibit* folding (i.e., a trap), while positive frustration scores are said to *facilitate* folding (i.e., an intermediate). Generally, we find that frustration scores are correlated with both p_{fold} (see Figure S15 in the Supporting Information) as well as macrostate conformational free energies (Figure S16 in the Supporting Information), with the majority of macrostates in the unfolded state ensemble acting as significant inhibitors to folding. We also find that our frustration scores are more polarized (ranging from −8% to +80%) than typical values obtained by Savol and Chennubhotla (ranging from −2% to +3%). We attribute this discrepancy to their more expansive definition of the native ensemble, whereas we define the native ensemble as a single macrostate (N). Macrostate networks for all four sequences were visualized using Gephi[68] using force-minimization (with unweighted edges) for graphical layout. Graph nodes representing each macrostate are presented with sizes proportional to the equilibrium populations, and are colored according to their kinetic frustration score (Figure 8). This macrostate network representation reveals several interesting mechanistic similarities and differences across the four sequences.

The overall topology of the folding network is highly conserved. Automatic graphical layout with force minimization consistently produced highly funnel-like graphs for the four hairpin sequences that could be effectively superimposed. The unfolded-state ensemble consists of a densely connected region of mostly disordered macrostates, as well as more sparsely populated regions, one of which bridges the native state (N) through a narrow channel of transition state intermediates, including macrostates I1, I2, and I3. Despite the similarities in network topology, distinct difference can be seen across the four hairpin sequences, the most prominent being the relatively high population of near-native intermediates for GB1 hairpin versus the trpzip sequences. These intermediates are comprised largely of states lacking distal backbone hydrogen bonds and distal hydrophobic contacts. Such states are not as populated for the trpzip4 and trpzip5 sequences, likely due to the strong folding cooperativity conferred by distal cross-strand Trp pairs..

The subregion of the unfolded state ensemble farthest from the native state (in terms of edge distances) contains the states with the highest degree of kinetic frustration. The extent of frustration varies modestly for each hairpin sequence, roughly as expected by the number of Trp residues in the sequence, with GB1 hairpin being the least frustrated, and trpzip4 and

DOI: 10.1021/acs.jctc.5b00088
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

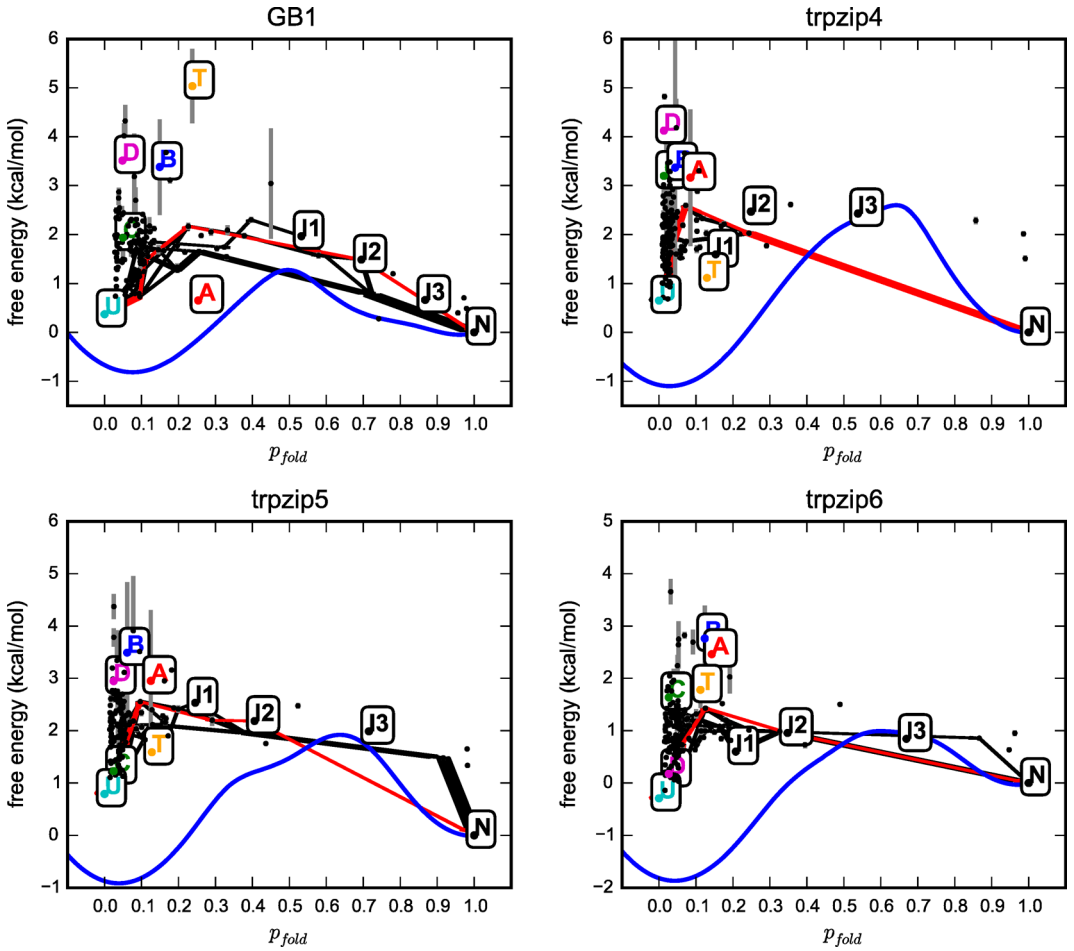


Figure 6. Macrostate free energies versus p_{fold} (committor) values for the U → N reaction, for all hairpin sequences, calculated by Transition Path Theory (TPT).[65,66] Gray vertical lines denote uncertainties in free energies, estimated from leave-one-out cross-validation. Key macrostates (U, N, T, A, B, C, and D) are labeled. Lines denote the top 10 pathways of reactive flux from U → N, with the top path in red. Free-energy profiles calculated from a kernel density estimator are shown as blue lines.

trpzip6 being the most frustrated. The average kinetic frustration scores for all macrostates with negative scores (i.e., *inhibitors* of folding) are −1.11%, −1.15%, −1.14%, and −1.16% for GB hairpin, trpzip4, trpzip5, and trpzip6, respectively.

A key issue in the experimental hairpin folding literature is the determination of on-pathway versus off-pathway intermediates.[15,86] Insight, in this respect, is given by the location of macrostate trap states (T, A, B, C, and D) within the larger network topology. The most distant of these traps are macrostates B, C, and D. As mentioned above, states C and D have similar conformational properties, but are not lumped together by the BACE algorithm, most likely due to sequence-dependent differences in the C-terminal. Macrostates C and D have a completely non-native turn with mis-registered strand pairs; Macrostate B is similar to C and D, but with a flipped N-terminal strand. All three correspond to conformations that would need to completely unfold before folding, in correspondence with their distant location in the network. Macrostates A and T, in contrast, reside more closely to the folded state. Macrostate A retains a native-like turn, while macrostate T retains native-like strand pairing and hydrophobic packing. In each case, complete unfolding is not required to reach the folding state. The prominence of macrostate A in the folding of GB1 hairpin, and its relative unimportance for trpzip4, trpzip5, and trpzip6, suggests that hydrophobic mutations to Trp may help preclude native-like turn formation in inverted forms.

Quantification of Glasslike Folding Behavior. Given the high degree of unfolded-state kinetic frustration, and the presence of misfolded kinetic traps, can we describe the folding kinetics of GB1 and trpzip hairpins as "glassy"? Recent work by Weber et al. has applied the *s*-ensemble method to a set of MSMs constructed from molecular simulations of small mini-proteins.[49] In this method, activities (numbers of nonself-transitions over some observed time t_{obs}) are computed for an ensemble of biased transition matrices parametrized by s, $\mathbf{T}(s) = \mathbf{U}e^{-s} + \mathbf{D}$, where $\mathbf{U}$ and $\mathbf{D}$ are the nondiagonal and diagonal elements of an unbiased MSM transition matrix, respectively. In the limit of s much greater than zero, self-transitions dominate in the so-called *inactive* dynamical phase, and the mean activity per time step (i.e., normalized by t_{obs}) $K(s) \to 0$. In the limit of s much less than zero, nonself-transitions dominate in the so-called *active* phase, and $K(s) \to 1$. The unbiased transition matrix is recovered when $s = 0$. The coexistence value s^*, where $K(s^*) = 0.5$ and the active and inactive phases coexist, is a measure of whether dynamics is active ($s^* > 0$) or glassy ($s^* < 0$).

Based on their survey of 16 protein folding MSMs, Weber et al. proposed a relationship between the folding time of a

DOI: 10.1021/acs.jctc.5b00088
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

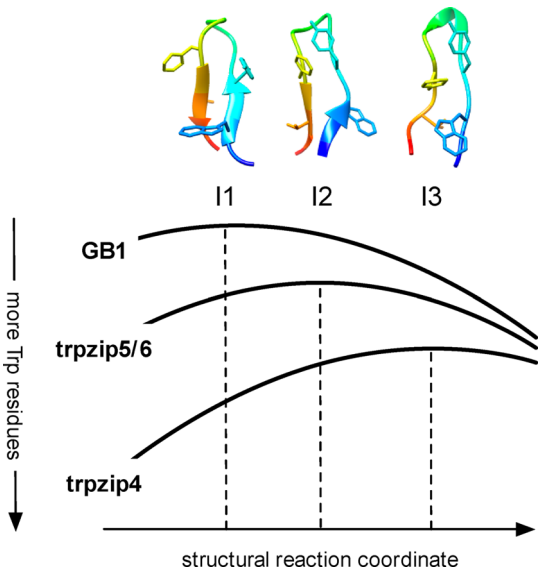


Figure 7. A schematic illustration showing the movement of transition-state intermediates I1, I2, and I3 along a structural reaction coordinate as tryptophans are introduced into the hairpin sequence. Macrostates I1, I2, and I3 share a preformed hairpin topology, and display a progression of hydrophobic side-chains "zipping" up into native-like conformations from the turn region. Black lines denote changes in the shape of free-energy landscape (the vertical placement is arbitrary).

protein and active vs "glassy" dynamics, with proteins with folding times longer than ∼10 μs exhibiting glassy dynamics. To be clear, the energy landscape for a truly glassy system is quite different from that of natural proteins.[87] Nevertheless, Weber et al. find that some protein folding MSMs predict active dynamics, while others predict inactive dynamics, reminiscent of a glass transition. To compare our results with these findings, we computed mean activities $K(s)$ from the 150-macrostate MSMs of the four GB1 and trpzip hairpins, setting t_{obs} to the slowest relaxation folding time for each of the models. For all hairpin sequences, we find that coexistence values s^* are nearly zero, indicating dynamical behavior straddling active vs glassy behavior (Figure 9a). The s^* values for GB hairpin, trpzip4, trpzip5, and trpzip6 are 0.0525, 0.0125, 0.0175, and 0.0875, respectively.

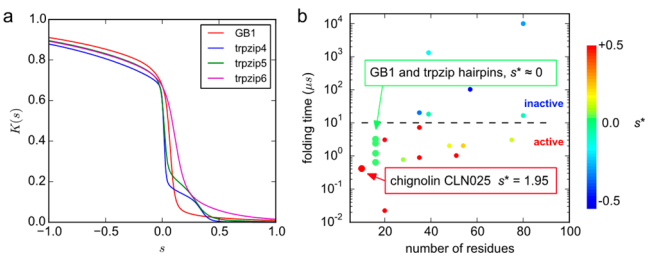


Figure 9. (a) Mean activities $K(s)$ vs s computed from 150-macrostate MSMs of GB1 and trpzip hairpins. (b) Coexistence values s^* near zero suggests dynamics intermediate between active and glassy phases, in accordance with similar findings from Weber et al.[49] Activity results for GB1 and trpzip hairpins, plotted as a function of folding time and number of residues, superimposed upon the plot of 16 MSMs studied by Weber et al., with dashed line denoting their proposed ∼10 μs boundary between active and inactive dynamics.

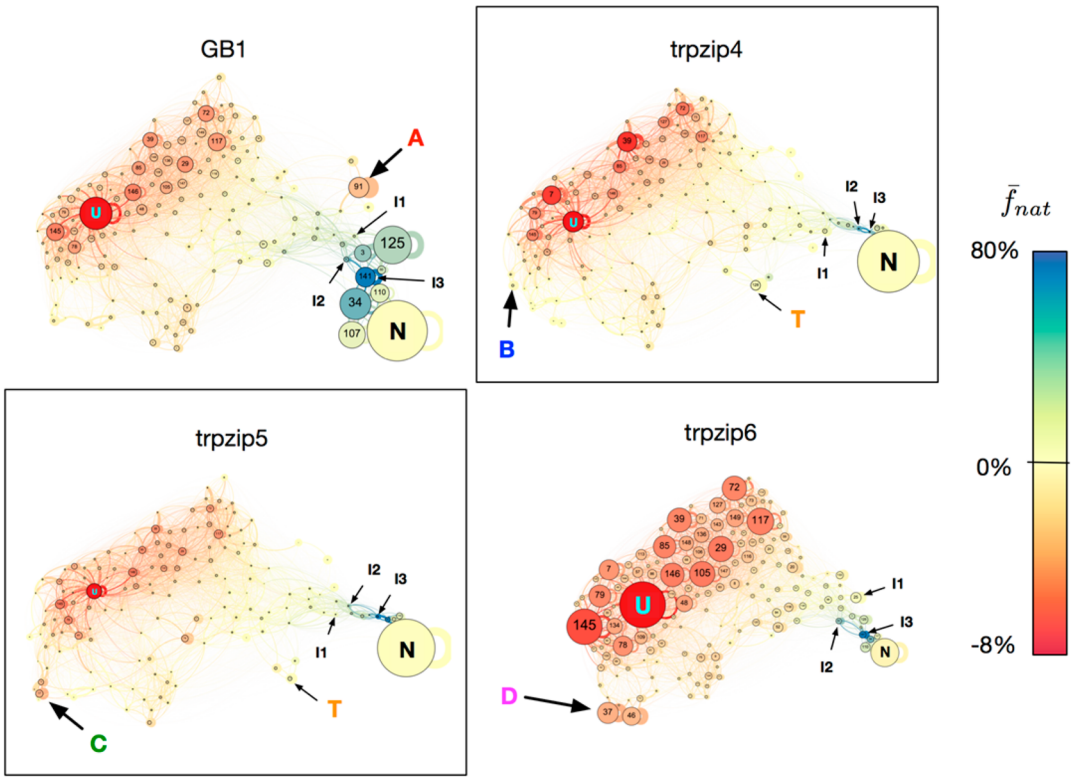


Figure 8. Network representations of 150-macrostate MSMs for GB1 hairpin, trpzip4, trpzip5, and trpzip6, each built using the same metastable definitions, with key macrostates (N, U, A, T, B, C, D) labeled by letter (the rest labeled by index). Nodes are sized in proportion to macrostate equilibrium populations, and colored according to computed kinetic frustration scores, $f_{nat}(i)$. Graphical layout was performed using Gephi.

DOI: 10.1021/acs.jctc.5b00088
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

The folding times of our hairpin MSMs are all faster than the proposed ∼10 μs boundary, and much more inactive than proteins of similar size (Figure 9b). These results suggest that GB1 and trpzip hairpins may be "glassier" than typical mini-proteins. It is particularly interesting to compare our results with those computed by Weber et al. for CLN025 (a 10-residue designed hairpin based on chignolin) from a comparable 150-state MSM (t_{obs} = 400 ns). CLN025 is predicted to be very much in an active dynamical phase (s^* = 1.95), whereas GB1 and trpzip hairpins are not. Since our computed s^* values do not change significantly with the number of introduced Trp residues, the near-glassy dynamical behavior may be a consequence of more general structural properties, such as β-strand formation, as conjectured by Weber et al.[49]

DISCUSSION

Our simulation and analysis of GB1 and trpzip hairpin folding is significant, both in terms of methodology and application. This study represents the most extensive explicit-solvent simulation trajectory data yet presented for GB1 hairpin and trpzip variants, and the best agreement to date with experimental folding kinetics. This is a significant achievement, although we note that implicit-solvent simulations have generated many times this amount of data,[8] and the longest time scales accessed in master equation models (i.e., kinetic transition network, or MSM models) have been obtained by discrete path sampling with implicit solvent.[36,37] To gain insight into the effects of sequence mutations on folding mechanisms, we have pursued a unique strategy of defining a coarse-grained set of MSM macrostates able to capture the dynamics of multiple sequences. To do this, we have made very careful use of the tICA approach in identifying low-dimensional coordinate systems to best capture slow conformational dynamics. As the quality of our results attest, particularly in the robust prediction of folding time scales in macrostate modes, this approach can be extremely useful for the objective comparison of multiple sequences, provided adequate sampling. We expect that MSM approaches like the type presented here will be used in the future to objectively compare conformational dynamics for mutations important in biological processes, as well as for protein design problems. Major advantages of the MSM approach include the ease with which adaptive sampling can be performed,[53,88] and decreased reliance on preconceived reaction coordinates.

A key question raised in this study and others[4,42,89] is the importance of non-native interactions on folding. Even at the level of microscopic detail obtained by simulations, the overall mechanistic picture of GB1 and trpzip hairpins as two-state folders remains intact. That said, non-native interactions clearly play a large role in determining the folding mechanisms of trpzip hairpins. We can learn a great deal by comparing the results for difference sequences in this respect: As the number of tryptophan residues in the sequence increases, MSMs predict that unfolded states are stabilized with respect to the transition state, impeding productive folding, and increasing kinetic frustration.

A striking difference among the four sequences studied here is between trpzip4 and trpzip5, which display strongly cooperative two-state folding behavior with a well-defined time scale gap, versus GB hairpin and trpzip6, whose folding is predicted to be more heterogeneous. The distinguishing factor separating these two groups is the presence of distal cross-strand Trp-Trp pairing in trpzip4 and trpzip5, demonstrated experimentally to be highly stabilizing.[30,33,34] The cooperative reward for folding, made possible by this β-capping motif, is likely to be significant enough to overpower the increased propensity for kinetic traps that come with more tryptophan residues. An apparent drawback of the capping motif is that it may not be very robust to sequence mutations. Indeed, our simulations predict that a conservative W14 V mutation (i.e., trpzip4 to trpzip6) results in a loss of two-state folding and a destabilized native state. A recent experimental study of folding kinetics for trpzip2 variants shows a similar effect (lowering the melting temperature by 30 K) for distal Trp → Val mutations that induce faster folding kinetics and less stability (fraying), which the authors attribute to increased populations of intermediates.[30] The lack of robustness to sequence mutations may partly explain why β-capping motifs are under-represented in natural protein sequences,[6,31] although this may be due to its propensity to aggregate,[90] or the high metabolic cost of tryptophan use.[91] Peptides with β-capping motifs are currently being used as a scaffold for designing functional peptidomimetics,[33] and our results suggest the need to carefully consider the effects of mutations on folding properties, especially as hairpins are functionalized for other properties such as binding. As shown here, we can use molecular simulation to elucidate important kinetic traps, so that we may design sequences that robustly fold.[92]

A related question raised by our work is the extent to which experimental studies can discern kinetic traps and other deviations from two-state folding of small hairpins. We believe the future is ripe for continued investigations in this direction, fueled in part by high-resolution modeling studies like the type presented here. As we have shown above, our MSMs have the advantage of being able to discriminate on-pathway versus off-pathway intermediates. This, and the ability of MSMs to make detailed predictions of time-resolved observables,[67,93] suggests that MSMs should play a key role alongside spectroscopic studies to unravel the details of folding mechanisms of hairpins.[94] Evidence for multiple folding pathways and relaxation time scales involving intermediates have been found recently for CLN025[29] and a disulfide-cross-linked variant of trpzip4.[86] Jones et al. have recently used 2D IR spectroscopy of isotopologues to detect hairpin folding intermediates with equilibrium populations as low as 1%.[15] Our MSMs predict kinetic trap intermediates populated to 1%−2% at equilibrium, involved in relaxations with significant folding flux.

An area for improvement identified in this work is the underestimation of native-state populations in our molecular simulations (see Table S1 in the Supporting Information), as evidenced by disagreement of simulated trpzip4 ensembles with NMR structural observables. Our simulations of trpzip6 display the most obvious example of this, with the equilibrium population of native macrostate N predicted to be smaller than that of the unfolded macrostate U. While molecular simulation studies have long been used to understand mechanistic differences in folding despite other shortcomings (e.g., poor predictions of melting temperatures), absolute prediction of fold stability is vitally important for making quantitative connections with experimental observations. Here, we have several reasons to believe that discrepancies in folded populations may be due to force field bias, rather than finite sampling errors. First, our computed equilibrium macrostate populations and folding relaxation rates are robust with respect to the leave-one-out cross-validation used to calculate

DOI: 10.1021/acs.jctc.5b00088
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

uncertainties, suggesting that observed transition counts have adequate statistics. Furthermore, for all sequences, we not only observe many folding trajectories, but also a significant number of unfolding trajectories (see Table 4). For example, we observe 99 folding trajectories for trpzip4 and 94 unfolding trajectories. Given the experimental unfolding time of ∼230 μs, however, we should expect to observe many fewer unfolding events in the <2.2 μs trajectories studied here. While the exact reason for force field inaccuracies remains unclear, we speculate that the nonpolarizable force field used here may inaccurately estimate face-to-edge tryptophan interaction energies.

Despite these inaccuracies in predicted absolute stabilities, we think there is strong evidence that our simulations provide accurate relative information about folding mechanisms, and the key features of folding that control kinetics and folding pathways:

(1) our excellent agreement with absolute folding rates,
(2) the correct rank ordering of predicted rates for the different sequences, and
(3) predictions of high folding cooperativity for β-cap sequences trpzip4 and trpzip5.

CONCLUSION

We have presented a Markov State Model (MSM) analysis of the most extensive explicit-solvent simulations of β-hairpins to date (9.4 ms in aggregate), for GB1 hairpin and sequence variants trpzip4, trpzip5, and trpzip6. We make quantitative comparisons against experimental structural observables and folding kinetics, and use a combined macrostate MSM approach to objectively compare hairpin folding mechanisms. Our results reveal multimodal kinetics arising from misfolded kinetic trap states, which are increasingly stabilized as the number of Trp residues increases, resulting in clear transition-state movement. Computed kinetic frustration scores,[47] network topologies, and glass activities[49] quantify the significance of non-native interactions in folding. Importantly, our simulations predict that sequences with distal cross-strand tryptophan pairs (trpzip4 and trpzip5) are more stable and fold in a more two-state manner, consistent with the remarkable stability of β-capping motifs identified in experimental protein design studies. Taken as a whole, this work shows MSMs to be a highly effective tool for the quantitative comparison of sequence-dependent conformational dynamics, suggesting new avenues for computational design, and a better understanding of how mutations perturb folding mechanisms.

ASSOCIATED CONTENT

Supporting Information

Table S1 and Figures S1−S16 are provided as Supporting Information. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00088.

AUTHOR INFORMATION

Corresponding Author

*E-mail: voelz@temple.edu.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation, through Major Research Instrumentation Grant No. CNS-09-58854, and through NSF MCB-1412508. We thank Brandon Kier for useful feedback, and the participants of Folding@home, without whom this work would not be possible. We thank Ted Delikatny for initial simulation work, which helped to motivate the current study.

ABBREVIATIONS

MSM, Markov State Model; REMD, replica exchange molecular dynamics; NMR, nuclear magnetic resonance; MFPT, mean first passage time; TPT, Transition Path Theory; GBSA, generalized Born/surface area

REFERENCES

(1) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins: Struct., Funct., Bioinf.* **1995**, *21*, 167.
(2) Best, R. B.; Hummer, G.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 17874.
(3) Schwantes, C. R.; Pande, V. S. *J. Chem. Theory Comput.* **2013**, *9*, 2000.
(4) Cho, J. H.; Meng, W.; Sato, S.; Kim, E. Y.; Schindelin, H.; Raleigh, D. P. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 12079.
(5) Blanco, F. J.; Rivas, G.; Serrano, L. *Nat. Struct. Biol.* **1994**, *1*, 584.
(6) Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5578.
(7) Du, D.; Zhu, Y.; Huang, C. Y.; Gai, F. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15915.
(8) Snow, C. D.; Qiu, L.; Du, D.; Gai, F.; Hagen, S. J.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4077.
(9) Yang, W. Y.; Pitera, J. W.; Swope, W. C.; Gruebele, M. *J. Mol. Biol.* **2004**, *336*, 241.
(10) Smith, A. W.; Tokmakoff, A. *J. Chem. Phys.* **2007**, *126*, 045109.
(11) Du, D.; Tucker, M. J.; Gai, F. *Biochemistry* **2006**, *45*, 2668.
(12) Streicher, W. W.; Makhatadze, G. I. *J. Am. Chem. Soc.* **2006**, *128*, 30.
(13) Narayanan, R.; Pelakh, L.; Hagen, S. J. *J. Mol. Biol.* **2009**, *390*, 538.
(14) Hwang, S.; Hilty, C. *J. Phys. Chem. B* **2011**, *115*, 15355.
(15) Jones, K. C.; Peng, C. S.; Tokmakoff, A. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 2828.
(16) Radford, I. H.; Fersht, A. R.; Settanni, G. *J. Phys. Chem. B* **2011**, *115*, 7459.
(17) Settanni, G.; Fersht, A. R. *Biophys. J.* **2008**, *94*, 4444.
(18) Zhuang, W.; Cui, R. Z.; Silva, D.-A.; Huang, X. *J. Phys. Chem. B* **2011**, *115*, 5415.
(19) Munoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196.
(20) Dyer, R. B.; Maness, S. J.; Peterson, E. S.; Franzen, S.; Fesinmeyer, R. M.; Andersen, N. H. *Biochemistry* **2004**, *43*, 11560.
(21) Olsen, K. A.; Fesinmeyer, R. M.; Stewart, J. M.; Andersen, N. H. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15483.
(22) Yu, Z.; Selvam, S.; Mao, H. *Biochemistry* **2014**, *53*, 5978.
(23) Xu, Y.; Du, D.; Oyola, R. *J. Phys. Chem. B* **2011**, *115*, 15332.
(24) Xu, Y.; Oyola, R.; Gai, F. *J. Am. Chem. Soc.* **2003**, *125*, 15388.
(25) Dyer, R. B.; Maness, S. J.; Peterson, E. S.; Franzen, S.; Fesinmeyer, R. M.; Andersen, N. H. *Biochemistry* **2004**, *43*, 11560.
(26) Jas, G. S.; Hegefeld, W. A.; Middaugh, C. R.; Johnson, C. K.; Kuczera, K. *J. Phys. Chem. B* **2014**, *118*, 7233.
(27) Hauser, K.; Ridderbusch, O.; Roy, A.; Hellerbach, A.; Huang, R.; Keiderling, T. A. *J. Phys. Chem. B* **2010**, *114*, 11628.
(28) Hauser, K.; Krejtschi, C.; Huang, R.; Wu, L.; Keiderling, T. A. *J. Am. Chem. Soc.* **2008**, *130*, 2984.
(29) Davis, C. M.; Xiao, S.; Raleigh, D. P.; Dyer, R. B. *J. Am. Chem. Soc.* **2012**, *134*, 14476.

DOI: 10.1021/acs.jctc.5b00088
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(30) Popp, A.; Wu, L.; Keiderling, T. A.; Hauser, K. *J. Phys. Chem. B* **2014**, *118*, 14234.
(31) Santiveri, C. M.; Jimenez, M. A. *Biopolymers* **2010**, *94*, 779.
(32) Andersen, N. H.; Olsen, K. A.; Fesinmeyer, R. M.; Tan, X.; Hudson, F. M.; Eidenschink, L. A.; Farazi, S. R. *J. Am. Chem. Soc.* **2006**, *128*, 6101.
(33) Kier, B. L.; Andersen, N. H. *J. Pept. Sci.* **2014**, *20*, 704−715.
(34) Kier, B. L.; Shu, I.; Eidenschink, L. A.; Andersen, N. H. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 10466.
(35) Wu, L.; McElheny, D.; Huang, R.; Keiderling, T. A. *Biochemistry* **2009**, *48*, 10362.
(36) Carr, J. M.; Wales, D. J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 3341.
(37) Evans, D. A.; Wales, D. J. *J. Chem. Phys.* **2004**, *121*, 1080.
(38) Juraszek, J.; Bolhuis, P. G. *J. Phys. Chem. B* **2009**, *113*, 16184.
(39) Yang, C.; Jang, S.; Pak, Y. *Nat. Commun.* **2014**, *5*, 5773.
(40) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517.
(41) De Sancho, D.; Mittal, J.; Best, R. B. *J. Chem. Theory Comput.* **2013**, *9*, 1743.
(42) Best, R. B.; Mittal, J. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 11087.
(43) Best, R. B.; Mittal, J. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 1318.
(44) Perez-Hernandez, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. *J. Chem. Phys.* **2013**, *139*, 015102.
(45) Streicher, W. W.; Makhatadze, G. I. *J. Am. Chem. Soc.* **2006**, *128*, 30.
(46) Matouschek, A.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 7814.
(47) Savol, A. J.; Chennubhotla, C. S. *J. Chem. Theory Comput.* **2014**, *10*, 2964.
(48) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299.
(49) Weber, J. K.; Jack, R. L.; Pande, V. S. *J. Am. Chem. Soc.* **2013**, *135*, 5501.
(50) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004**, *25*, 1605.
(51) Kollman, P. A. *Acc. Chem. Res.* **1996**, *29*, 461.
(52) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297.
(53) Voelz, V. A.; Elman, B.; Razavi, A. M.; Zhou, G. *J. Chem. Theory Comput.* **2014**, *10*, 5716.
(54) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526.
(55) Hayre, N. R.; Singh, R. R. P.; Cox, D. L. *J. Chem. Phys.* **2011**, *134*, 035103.
(56) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. *J. Chem. Theory Comput.* **2011**, *7*, 3412.
(57) Bowman, G. R. *J. Chem. Phys.* **2012**, *137*, 134111.
(58) Shirts, M.; Pande, V. S. *Science* **2000**, *290*, 1903.
(59) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950.
(60) Małolepsza, E.; Strodel, B.; Khalili, M.; Trygubenko, S.; Fejer, S. N.; Wales, D. J. *J. Comput. Chem.* **2010**, *31*, 1402.
(61) Schwantes, C. R.; Pande, V. S. *J. Chem. Theory Comput.* **2013**, *9*, 2000.
(62) Djurdjevac, N.; Sarich, M.; Schütte, C. *Multiscale Model. Simul.* **2012**, *10*, 61.
(63) Buchete, N.-V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057.
(64) Prinz, J. H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schutte, C.; Noe, F. *J. Chem. Phys.* **2011**, *134*, 174105.
(65) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. *Multiscale Model. Simul.* **2009**, *7*, 1192.
(66) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011.
(67) Voelz, V. A.; Jäger, M.; Yao, S.; Chen, Y.; Zhu, L.; Waldauer, S. A.; Bowman, G. R.; Friedrichs, M.; Bakajin, O.; Lapidus, L. J.; Weiss, S.; Pande, V. S. *J. Am. Chem. Soc.* **2012**, *134*, 12565.
(68) Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. Presented at *2009 International AAAI Conference on Weblogs and Social Media (ICWSM 09)*, San Jose, CA, 2009.
(69) Weinstock, D. S.; Narayanan, C.; Felts, A. K.; Andrec, M.; Levy, R. M.; Wu, K.-P.; Baum, J. *J. Am. Chem. Soc.* **2007**, *129*, 4858.
(70) Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S. *J. Biomol. NMR* **2011**, *50*, 43.
(71) Vögeli, B.; Ying, J.; Grishaev, A.; Bax, A. *J. Am. Chem. Soc.* **2007**, *129*, 9377.
(72) Schmidt, J. M.; Blümel, M.; Löhr, F.; Rüterjans, H. *J. Biomol. NMR* **1999**, *14*, 1.
(73) Hu, J.-S.; Bax, A. *J. Am. Chem. Soc.* **1997**, *119*, 6360.
(74) Beauchamp, K. A.; Lin, Y.-S.; Das, R.; Pande, V. S. *J. Chem. Theory Comput.* **2012**, *8*, 1409.
(75) Vitalis, A.; Pappu, R. V. *J. Comput. Chem.* **2009**, *30*, 673.
(76) Li, D.-W.; Brüschweiler, R. *J. Phys. Chem. Lett.* **2010**, *1*, 246.
(77) Zagrovic, B.; Sorin, E. J.; Pande, V. *J. Mol. Biol.* **2001**, *313*, 151.
(78) Beauchamp, K. A.; McGibbon, R.; Lin, Y.-S.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 17807.
(79) Noé, F.; Doose, S.; Daidone, I.; Löllmann, M.; Sauer, M.; Chodera, J. D.; Smith, J. C. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 4822.
(80) Keller, B. G.; Prinz, J.-H.; Noé, F. *Chem. Phys.* **2012**, *396*, 92.
(81) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2011**, *133*, 18413.
(82) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415.
(83) Wales, D. J. *Int. Rev. Phys. Chem.* **2006**, *25*, 237.
(84) Wales, D. J. *J. Chem. Phys.* **2009**, *130*, 204111.
(85) Berezhkovskii, A.; Hummer, G.; Szabo, A. *J. Am. Chem. Soc.* **2009**, *130*, 205102.
(86) Markiewicz, B. N.; Yang, L.; Culik, R. M.; Gao, Y. Q.; Gai, F. *J. Phys. Chem. B* **2014**, *118*, 3317.
(87) Heyes, D. M.; Okumura, H. *J. Chem. Phys.* **2006**, *124*, 164507.
(88) Du, W.; Bolhuis, P. G. *J. Am. Chem. Soc.* **2014**, *140*, 195102.
(89) Deng, N.-J.; Dai, W.; Levy, R. M. *J. Phys. Chem. B* **2013**, *117*, 12787.
(90) Markiewicz, B. N.; Oyola, R.; Du, D.; Gai, F. *Biochemistry* **2014**, *53*, 1146.
(91) Wagner, A. *Mol. Biol. Evol.* **2005**, *22*, 1365.
(92) Razavi, A. M.; Wuest, W. M.; Voelz, V. A. *J. Chem. Inf. Model.* **2014**, *54*, 1425.
(93) Voelz, V. A.; Pande, V. S. *Proteins: Struct., Funct., Bioinf.* **2012**, *80*, 342.
(94) Baiz, C. R.; Lin, Y.-S.; Peng, C. S.; Beauchamp, K. A.; Voelz, V. A.; Pande, V. S.; Tokmakoff, A. *Biophys. J.* **2014**, 1.

DOI: 10.1021/acs.jctc.5b00088
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX