JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

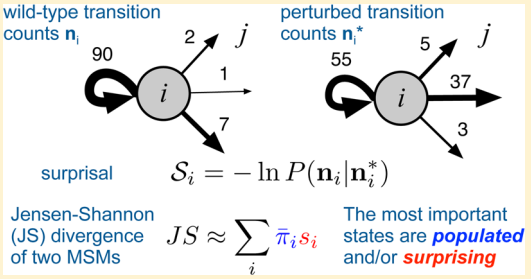
Surprisal Metrics for Quantifying Perturbed Conformational Dynamics in Markov State Models

Vincent A. Voelz,* Brandon Elman, Asghar M. Razavi, and Guangfeng Zhou

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

Supporting Information

ABSTRACT: Markov state models (MSMs), which model conformational dynamics as a network of transitions between metastable states, have been increasingly used to model the thermodynamics and kinetics of biomolecules. In considering perturbations to molecular dynamics induced by sequence mutations, chemical modifications, or changes in external conditions, it is important to assess how transition rates change, independent of changes in metastable state definitions. Here, we present a surprisal metric to quantify the difference in metastable state transitions for two closely related MSMs, taking into account the statistical uncertainty in observed transition counts. We show that the surprisal is a relative entropy metric closely related to the Jensen−Shannon divergence between two MSMs, which can be used to identify conformational states most affected by perturbations. As examples, we apply the surprisal metric to a two-dimensional lattice model of a protein hairpin with mutations to hydrophobic residues, all-atom simulations of the Fs peptide α-helix with a salt-bridge mutation, and a comparison of protein G β-hairpin with its trpzip4 variant. Moreover, we show that surprisal-based adaptive sampling is an efficient strategy to reduce the statistical uncertainty in the Jensen−Shannon divergence, which could be a useful strategy for molecular simulation-based *ab initio* design.



INTRODUCTION

Markov state models (MSMs) are kinetic network models describing conformational dynamics as transitions between metastable states.[1] This kinetic network approach offers an efficient multiscale sampling strategy in which many short, non-equilibrium trajectories can be used to sample local transition rates between states, while the network as a whole provides long-time scale dynamics and equilibrium properties. In recent years, MSM approaches have proved very successful at modeling the conformational dynamics of proteins and other biomolecules on long time scales,[2−8] including the first example of *ab initio* folding simulation on the millisecond time scale.[9]

The requirements to build an MSM are straightforward: (1) a definition of the relevant metastable states—this is usually done using conformational clustering algorithms and/or kinetic-based lumping schemes to choose the appropriate number of states, and (2) an appropriate choice of lag time τ over which the conformational dynamics is approximately Markovian; i.e., the probability $T_{ij}^{(\tau)}$ of transitioning from state i to state j in time τ is independent of preceding transitions.[10,11] Estimation of the transition rates often requires a large amount of conformational sampling, especially when there are large numbers of states. Because of this, some of the most successful strategies for constructing MSMs have involved massively parallel simulation on computing platforms such as Folding@home,[12] GPUGrid,[13] or Google Exacycle.[14]

Once the relevant metastable states have been identified, MSMs offer an extremely useful framework for analyzing biomolecular dynamics. Diagonalization of the transition matrix $\mathbf{T}$ yields a complete picture of thermodynamic and kinetics properties:[15] a spectrum of eigenmodes describing the relaxation dynamics, with the equilibrium populations given by the stationary eigenvector. The statistical uncertainty of these results can be arbitrarily improved using adaptive sampling methods, in which particular conformational transitions are iteratively chosen for biased sampling.[16−18] Moreover, kinetics-based methods for coarse-graining metastable states[19−22] and analysis of pathway fluxes[2,23] can greatly aid human understanding of complicated kinetic mechanisms.

Here, we explore an additional advantage of MSMs: the efficient evaluation of different systems with similar conformational dynamics. The key idea is that the metastable states of MSMs can serve as a conformational roadmap that remains conserved, while the effects of sequence differences, chemical modifications or other perturbations (force field parameters, denaturant, pH, salt concentrations, etc.) alters the transition rates between metastable states. Thus, different systems can be evaluated in an MSM framework by efficient estimation of rate perturbations. Indeed, several authors have already used this idea to model allosteric conformational change[24] and mutational effects in intrinsically unstructured peptides.[25]

We expand on this idea by explicitly considering the statistical uncertainties involved when building MSMs for a perturbed (mutant) system versus unperturbed (wild-type) system. In practice, MSMs are (usually) built from finite

Received: September 11, 2014

   dx.doi.org/10.1021/ct500827g | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

samples of observed transition counts collected from simulation trajectories. So, provided that the metastable states are conserved, the detectable difference between two such systems are the numbers of observed transition counts. We present a statistical metric, called the *surprisal*, which can be used to quantify the difference in observed transition counts. We show that the surprisal metric is closely related to the information-theoretic Jensen−Shannon divergence, which allows us to identify MSM states that contribute most to the statistical difference between wild-type and mutant MSMs. This analysis provides mechanistic insight into how perturbations of local transition rates result in global changes in equilibrium populations and conformational kinetics. Moreover, when combined with adaptive sampling schemes,[16−18] we show that this approach can be used to very efficiently converge statistical estimates of the effects of perturbations, with implications for MSM-based computational design.

In the sections that follow, we first present a derivation of the surprisal metric, and its relation to the Jensen−Shannon divergence. We present an analysis of surprisal properties in a simple, exact 2D lattice model of folding, as a proof of principle. We then present a surprisal analysis of mutations for two MSMs built from all-atom simulation systems: the α-helical Fs peptide and a R9E salt-bridge mutant, and the GB1 and trpzip β-hairpin variants. In both simple lattice and all-atom models, our analysis shows that consideration of non-native states is essential. Finally, we show how surprisal-based adaptive sampling can be used to efficiently converge estimates of the Jensen−Shannon divergence between two MSMs, with implications for MSM-based computational design.

THEORY

Surprisal Metrics. Consider an MSM with M states. The MSM is fully described by an $M\times M$ transition matrix $\mathbf{T}$, where $T_{ij}^{(\tau)}$ is defined as the probability of transitioning from state i to state j in time τ. A valid MSM requires that $\sum_j T_{ij} = 1$ (i.e., transition probability is conserved), and that the detailed balance condition is obeyed, i.e., $\pi_i T_{ij} = \pi_j T_{ji}$, where π_i are the stationary state (equilibrium) probabilities.

Suppose we observe $\sum_j n_{ij} = N_i$ transitions originating from state i, where n_{ij} is the number of transition counts from state i to state j. A central question is this: If we observe some new set of transition counts, $\sum_j n_{ij}^* = N_i^*$, for a system subject to some small perturbation, what is the probability of obtaining these new counts, given our knowledge of the system from the previous counts? To be clear, by "small perturbation" we mean small enough that the perturbed system has a conserved set of metastable states. Henceforth we assume that a perturbation induces changes in transition probabilities $T_{ij} \rightarrow T_{ij}^*$, but not the state definitions.

First, consider the probability of observing a set of transition counts $\mathbf{n}_i = (n_{i1}, n_{i2}, ..., n_{iM})$ given a transition matrix with rows $\mathbf{T}_i = (T_{i1}, T_{i2}, ..., T_{iM})$. We can think of $\mathbf{n}_i$ as being drawn from a multinomial distribution, with probability

$$P(\mathbf{n}_i|\mathbf{T}_i) = \frac{N_i!}{\prod_j^M n_{ij}!}\prod_j^M T_{ij}^{n_{ij}} \tag{1}$$

While we do not know *a priori* the transition probabilities $\mathbf{T}_i$, we can use Bayes's theorem to obtain its posterior distribution, given the observed counts $\mathbf{n}_i$,

$$P(\mathbf{T}_i|\mathbf{n}_i) = \frac{P(\mathbf{n}_i|\mathbf{T}_i)P(\mathbf{T}_i)}{\int P(\mathbf{n}_i|\mathbf{T}_i)P(\mathbf{T}_i)\,\mathrm{d}\mathbf{T}_i} \tag{2}$$

where $P(\mathbf{T}_i)$ is a *prior* distribution of transition probabilities. To model the prior distribution, we use a Dirichlet distribution,

$$\mathrm{Dir}(\mathbf{T}_i|\alpha_i) = \frac{\Gamma(\sum_j \alpha_{ij})}{\prod_j^M \Gamma(\alpha_{ij})}\prod_j^M T_{ij}^{\alpha_{ij}-1} \tag{3}$$

where $\alpha_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iM})$ are a set of *pseudocounts* representing any prior knowledge about the expected number of counts. Because the Dirichlet distribution is the conjugate prior of the multinomial distribution, the posterior distribution in eq 2 is analytically solvable. The well-known result is that the posterior is also a Dirichlet distribution with pseudocounts $\mathbf{n}_i + \alpha_i$,

$$P(\mathbf{T}_i|\mathbf{n}_i) = \mathrm{Dir}(\mathbf{T}_i|\mathbf{n}_i + \alpha_i) \tag{4}$$

Now, consider drawing a new set of counts $\mathbf{n}_i^* = (n_{i1}^*, n_{i2}^*, ..., n_{iM}^*)$ from the distribution $P(\mathbf{n}_i^*|\mathbf{T}_i)$. The marginal distribution of drawing these counts, given the posterior distribution of $P(\mathbf{T}_i||\mathbf{n}_i)$ from our previous counts $\mathbf{n}_i$, is

$$\begin{aligned}
P(\mathbf{n}_i^*|\mathbf{n}_i) &= \int P(\mathbf{n}_i^*|\mathbf{T}_i)P(\mathbf{T}_i|\mathbf{n}_i)\,\mathrm{d}\mathbf{T}_i \\
&= \int\left[\frac{N_i^*!}{\prod_j^M n_{ij}^*!}\prod_j^M T_{ij}^{n_{ij}^*}\right]\left[\frac{\Gamma(N_i + \sum_j \alpha_{ij})}{\prod_j^M \Gamma(n_{ij} + \alpha_{ij})}\prod_j^M T_{ij}^{n_{ij}+\alpha_{ij}-1}\right]\mathrm{d}\mathbf{T}_i \\
&= \int\left[\frac{N_i^*!}{\prod_j^M n_{ij}^*!}\right]\left[\frac{\Gamma(N_i + \sum_j \alpha_{ij})}{\prod_j^M \Gamma(n_{ij} + \alpha_{ij})}\prod_j^M T_{ij}^{n_{ij}^*+n_{ij}+\alpha_{ij}-1}\right]\mathrm{d}\mathbf{T}_i \\
&= \left[\frac{N_i^*!}{\prod_j^M n_{ij}^*!}\right]\left[\frac{\Gamma(N_i + \sum_j \alpha_{ij})}{\prod_j^M \Gamma(n_{ij} + \alpha_{ij})}\right]\left[\frac{\prod_j^M \Gamma(n_{ij}^* + n_{ij} + \alpha_{ij})}{\Gamma(N_i^* + N_i + \sum_j \alpha_{ij})}\right] \\
&\quad \int \mathrm{Dir}(\mathbf{T}_i|\mathbf{n}_i^* + \mathbf{n}_i + \alpha_i)\,\mathrm{d}\mathbf{T}_i
\end{aligned}$$

The Dirichlet distribution at the end integrates to unity, leaving

$$P(\mathbf{n}_i^*|\mathbf{n}_i) = \left[\frac{N_i^*!}{\prod_j^M n_{ij}^*!}\right]\left[\frac{\Gamma(N_i + \sum_j \alpha_{ij})}{\prod_j^M \Gamma(n_{ij} + \alpha_{ij})}\right]\left[\frac{\prod_j^M \Gamma(n_{ij}^* + n_{ij} + \alpha_{ij})}{\Gamma(N_i^* + N_i + \sum_j \alpha_{ij})}\right] \tag{5}$$

Surprisal. For a random outcome X observed with probability $P(X)$, the *surprisal* is defined as $-\ln P(X)$. This term was originally coined by Myron Tribus,[26] and is also known as self-information.[27] Surprisal-based analysis methods

dx.doi.org/10.1021/ct500827g | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

have been previously applied in many different contexts,[28] although the term does not enjoy common parlance. The surprisal is a positive number that quantifies the "order of magnitude" of how unlikely it is to observe a particular outcome. Here, the term "surprisal" is very apt, because our central question addresses how *surprising* it is to obtain a set of transition counts $\mathbf{n}_i^*$, given that some counts $\mathbf{n}_i$ have already been observed.

Let us define the surprisal for state i to be $\mathcal{S}_i = -\ln P(\mathbf{n}_i^*|\mathbf{n}_i)$. The factorial and Γ function terms in eq 5 can be simplified using Stirling's approximation, $\ln n! = \ln \Gamma(n+1) \approx n \ln n - n$, to get (see Appendix):

$$\mathcal{S}_i = -\ln P(\mathbf{n}_i^*|\mathbf{n}_i) = (N_i^* + \hat{N}_i)\tilde{H}_i^{\text{comb}} - N_i^*\tilde{H}_i^* - \hat{N}_i\tilde{H}_i \tag{6}$$

where

$$\tilde{H}_i^* = \sum_j^M -\frac{n_{ij}^*}{N_i^*} \ln \frac{n_{ij}^*}{N_i^*}$$

$$\tilde{H}_i = \sum_j^M -\frac{\hat{n}_{ij}}{\hat{N}_i} \ln \frac{\hat{n}_{ij}}{\hat{N}_i}$$

$$\tilde{H}_i^{\text{comb}} = \sum_j^M -\frac{n_{ij}^* + \hat{n}_{ij}}{N_i^* + \hat{N}_i} \ln \frac{n_{ij}^* + \hat{n}_{ij}}{N_i^* + \hat{N}_i}$$

To simplify the above expression, we have defined $\hat{N}_i = (N_i + \sum_j^M \alpha_{ij} - 1)$ and $\hat{n}_{ij} = (n_{ij} + \alpha_{ij} - 1)$. Note that, for the common (and sensible) choice of pseudocount parameters $\alpha_{ij} = 1/M$, $\hat{N}_i = N_i$. Henceforth, we will assume this to be the case, and drop the hat notation on N_i and n_{ij}.

The terms $\tilde{H}_i$, $\tilde{H}_i^*$, and $\tilde{H}_i^{\text{comb}}$ in eq 6 are estimates of entropy rates for three different stochastic processes: $\tilde{H}_i$ estimates the entropy rate of transition counts from state i based on the observed counts n_{ij}, $\tilde{H}_i^*$ estimates the entropy rate based on counts n_{ij}^*, and $\tilde{H}_i^{\text{comb}}$ estimates the entropy rate based on the combined counts $n_{ij} + n_{ij}^*$. Because $P(\mathbf{n}_i^*|\mathbf{n}_i) \leq 1$, $\mathcal{S}_i \geq 0$. $\mathcal{S}_i$ is exactly zero only in the limiting case when the distribution of observed transition counts n_{ij}^*/N_i^* is identical to the distribution of observed transition counts n_{ij}/N_i, in which case $\tilde{H}_i^* = \tilde{H}_i = \tilde{H}_i^{\text{comb}}$ and $P(\mathbf{n}_i^*|\mathbf{n}_i) = 1$.

We note that an identical expression has been derived by Bowman for use in the Bayesian Agglomerative Clustering Engine (BACE) algorithm.[21] In that context, the surprisal is used to approximate the Jensen–Shannon divergence (discussed below) incurred as a result of lumping two microstates into a single macrostate.

Since $\mathcal{S}_i$ is an extensive quantity that grows linearly with the number of observed counts, a more useful quantity for comparing two MSMs is $\mathcal{S}_i$ divided by the total number of transitions counts observed starting from state i:

$$s_i = \frac{\mathcal{S}_i}{N_i + N_i^*} = \tilde{H}_i^{\text{comb}} - \frac{N_i^*}{N_i^* + N_i}\tilde{H}_i^* - \frac{N_i}{N_i^* + N_i}\tilde{H}_i \tag{7}$$

The normalized surprisal s_i is an estimate of the excess entropy rate[27] of Markov transitions from a model $T_{ij}^{\text{comb}} \sim (n_{ij}^* + n_{ij})/(N_i^* + N_i)$, versus a model where $N_i^*/(N_i^* + N_i)$ of the Markov transitions come from a model $T_i^* \sim n_{ij}^*/N_i^*$, and $N_i/(N_i^* + N_i)$ of the Markov transitions come from a model $T_{ij} \sim n_{ij}/N_i$. For the remainder of this article, when we refer to surprisal, we mean the *normalized* surprisal, s_i.

We define the *total surprisal* as the (normalized) negative log-probability of observing an entire set of $N = \sum_i N_i$ old counts and $N^* = \sum_i N_i^*$ new counts across all states i:

$$s = \frac{\sum_i \mathcal{S}_i}{N + N^*} = \frac{-\sum_i \ln P(\mathbf{n}_i^*|\mathbf{n}_i)}{N + N^*} \tag{8}$$

$$= \sum_i \frac{(N_i + N_i^*)}{N + N^*}\left[\tilde{H}_i^{\text{comb}} - \frac{N_i^*}{N_i + N_i^*}\tilde{H}_i^* - \frac{N_i}{N_i + N_i^*}\tilde{H}_i\right] \tag{9}$$

$$= \sum_i \frac{(N_i + N_i^*)}{N + N^*} s_i \tag{10}$$

Thus, the total surprisal is the count-weighted average surprisal across all states i.

Estimation of Surprisal Uncertainty. Non-parametric methods such as the bootstrap method can easily be used for estimating uncertainty in the surprisal (for instance, resampling transition counts with replacement from the observed sample). Because this can be expensive, we can use the following approach to obtain an analytical estimate of the uncertainty in the expected surprisal values for each state i.

We utilize the fact that in the limit of large numbers of counts, the distribution of observed counts n_{ij} approach a multivariate normal distribution with some event probabilities p_{ij}. With a non-informative prior enforced, the estimate that maximizes the posterior distribution of p_{ij} is $\hat{p}_{ij} = (n_{ij} + 1/2)/\sum_j(n_{ij} + 1/2)$. The covariance matrix for observed counts n_{ij} can then be estimated as $\mathbf{V}_i$, where $(\mathbf{V}_i)_{jk} = (\sum_j n_{ij})\hat{p}_{ij}(1 - \hat{p}_{ij})$ for $j = k$ and $(\mathbf{V}_i)_{jk} = -N_{ij}\hat{p}_{ij}\hat{p}_{ik}$, where $N_i = \sum_j n_{ij}$.

A first-order Taylor series approximation yields an estimate of the variances σ_i^2 of expected surprisal values s_i, which can be computed as

$$\sigma_i^2 = \mathbf{q}^{\mathbf{T}}\begin{bmatrix}\mathbf{V}_i & 0\\ 0 & \mathbf{V}_i^*\end{bmatrix}\mathbf{q}$$

where $\mathbf{V}_i$ and $\mathbf{V}_i^*$ are the covariance matrices for counts $\mathbf{n}_i$ and $\mathbf{n}_i^*$, respectively, and $\mathbf{q}$ is a vector of *sensitivities*, defined as

$$\mathbf{q} = \left[\frac{\partial s_i}{\partial n_{i1}}, \frac{\partial s_i}{\partial n_{i2}}, ..., \frac{\partial s_i}{\partial n_{iM}}, \frac{\partial s_i}{\partial n_{i1}^*}, \frac{\partial s_i}{\partial n_{i2}^*}, ..., \frac{\partial s_i}{\partial n_{iM}^*}\right]$$

where the partial derivatives work out to be

$$\frac{\partial s_i}{\partial n_{ij}} = -\frac{1}{N_i + N_i^*}\left(s_i + \ln \frac{n_{ij} + n_{ij}^*}{N_i + N_i^*} - \ln \frac{n_{ij}}{N_i}\right)$$

$$\frac{\partial s_i}{\partial n_{ij}^*} = -\frac{1}{N_i + N_i^*}\left(s_i + \ln \frac{n_{ij} + n_{ij}^*}{N_i + N_i^*} - \ln \frac{n_{ij}^*}{N_i^*}\right)$$

Because the transition counts from each state i are statistically independent, the variance of the total surprisal is $\sigma^2 = \sum_i \sigma_i^2 (N_i + N_i^*)/(N + N^*)$.

To validate our analytical estimate of surprisal, we compared against bootstrap estimators (Figure 1), using a wide range different states M (between 2 and 200), various numbers of sampled transition counts (between 1 and 10^6), and

dx.doi.org/10.1021/ct500827g | J. Chem. Theory Comput. XXXX, XXX, XXX–XXX

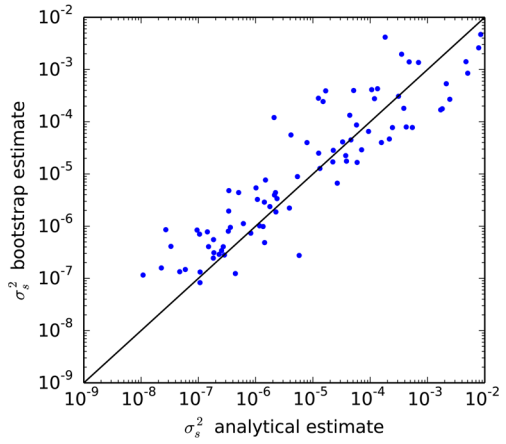


Figure 1. A comparison of surprisal variance estimated analytically versus by a bootstrap estimator using 10 000 samples. 100 calculations are compared for cases where transition probabilities are randomly selected from a uniform distribution, using between $M = 2$ and 200 states, and numbers of samples are selected between 1 and 10^6.

distributions of transition probabilities (drawn from a random uniform distribution, then normalized). Analytical and bootstrap estimates using 10000 samples agree with a correlation coefficient of $R^2 = 0.95$, and standard deviations of $\log_{10}(\sigma_i^2)$ are 0.75.

Relation to the Jensen−Shannon Divergence. The surprisal is directly related to an information-theoretical quantity called the generalized Jensen−Shannon divergence. Namely, if the transition counts are observed at equilibrium, then the surprisal is an *estimate* of the generalized Jensen−Shannon divergence, JS($\mathbf{T}^*$,$\mathbf{T}$), between two models $\mathbf{T}^*$ and $\mathbf{T}$.

Previous studies have considered the Kullback−Leibler divergence (otherwise known as the relative entropy) between a "test" MSM and "reference" MSM to quantify the convergence of MSM adaptive sampling methods.[18] The Kullback−Leibler divergence between transition matrices $\mathbf{T}^{\text{test}}$ and $\mathbf{T}^{\text{ref}}$ is given by

$$D(\mathbf{T}^{\text{test}} \| \mathbf{T}^{\text{ref}}) = \sum_i^M \pi_i \sum_j^M T_{ij}^{\text{test}} \ln \frac{T_{ij}^{\text{test}}}{T_{ij}^{\text{ref}}} \tag{11}$$

where π_i is the equilibrium probability of state i for the "test" MSM.

One drawback of the relative entropy is that, in most cases, it is not symmetric, i.e., $D(\mathbf{T}^{\text{test}}\|\mathbf{T}^{\text{ref}}) \neq D(\mathbf{T}^{\text{ref}}\|\mathbf{T}^{\text{test}})$, and thus it cannot be used as a proper distance metric. A symmetric metric is provided by the generalized JS divergence, which is defined for a collection of transition matrices $\mathbf{T}_k$ as

$$\text{JS}(\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_K) = \sum_{K=1}^K p_k D(\mathbf{T}_k \| \mathbf{T}^{\text{comb}}) \tag{12}$$

where $\mathbf{T}^{\text{comb}} = \sum_k p_k \mathbf{T}_k$, Here, p_k is the weight of the contribution of each $\mathbf{T}_k$ to the JS divergence, where $\sum p_k = 1$.

The JS divergence can also be written

$$\text{JS}(\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_K) = H^{\text{comb}} - \sum_k p_k H_k \tag{13}$$

where $H_k = -\sum_i^M (\pi_k)_i \sum_j^M (T_k)_{ij} \ln(T_k)$ and $H^{\text{comb}} = -\sum_i^M \pi_i^{\text{comb}} \sum_j^M T_{ij}^{\text{comb}} \ln T_{ij}^{\text{comb}}$ are MSM entropy rates. With this definition of the JS divergence, we can easily see that if the counts n_{ij} and n_{ij}^* are observed under equilibrium conditions, i.e., if $\pi_i \approx N_i/N$ and $\pi_{ij}^* \approx N_i^*/N^*$, then the total surprisal in eq 10 is equivalent to a count-based estimate of the JS divergence JS($\mathbf{T}$,$\mathbf{T}^*$) with contribution weights $N/(N + N^*)$ and $N^*/(N + N^*)$ for $\mathbf{T}$ and $\mathbf{T}^*$ respectively.

In practice, however, the observed transition counts used to estimate MSM transition matrices are usually not collected under equilibrium conditions, and depending on the situation, disproportionate numbers of transition counts may be sampled from the unperturbed system versus the perturbed system. Moreover, the use of MSM adaptive sampling schemes (discussed below) can bias the sampling toward specific states. It is thus important to accurately estimate the equilibrium probabilities π_i and π_i^*, given the available sampling. This can be challenging because naive estimates of transition probabilities from observed counts lead to π_i values that do not obey detailed balance. There are many methods available to do this,[10,29−31] not all of which are useful in this context. For example, estimating the transition matrix as a row-normalized matrix of symmetrized counts $(n_{ij} + n_{ji})/2$ gives poor estimates because the π_i estimates in this case are proportional to the number of samples obtained from each state i. We find that the maximum likelihood estimation (MLE) method implemented in MSMBuilder2[31] is time-consuming and can give poor results for sparsely sampled count matrices. As an alternative to these methods, we use a very simple scheme of computing modified transition matrix elements $T_{ij} \leftarrow \min(1, T_{ji}\pi_j/T_{ij}\pi_j)T_{ij}$ to achieve a transition matrix and equilibrium populations consistent with detailed balance.

To facilitate a state-wise analysis of the total JSD, the following equation can be used as a close approximation for JS($\mathbf{T}$,$\mathbf{T}^*$):

$$\text{JS}(\mathbf{T}, \mathbf{T}^*) \approx \sum_i \frac{\pi_i + \pi_i^*}{2}\left[H_i^{\text{comb}} - \frac{\pi_i}{\pi_i + \pi_i^*}H_i - \frac{\pi_i^*}{\pi_i + \pi_i^*}H_i^*\right] \tag{14}$$

where π and π^* are estimates of state probabilities at equilibrium. The above expression stems from a few simple assumptions: (1) that the unperturbed and perturbed systems represented by $\mathbf{T}$ and $\mathbf{T}^*$ receive equal weight (i.e., $p_k = 1/2$), and (2) that $(\pi + \pi^*)/2$ approximates the equilibrium state probabilities π^{comb} corresponding to $\mathbf{T}^{\text{comb}} = (\mathbf{T} + \mathbf{T}^*)/2$. This is a good assumption when perturbations are small (see Figure 4a), specifically when the elements of $(\mathbf{T} - \mathbf{T}^*)(\pi - \pi^*)$ are small.

In eq 14, the term in brackets can be thought of as a surprisal quantity, corrected for the relative numbers of counts from state i that we expect to observe for the perturbed versus unperturbed systems at equilibrium. Assuming that perturbations are small (i.e., $\pi_i \approx \pi_i^*$), we can further approximate eq 14 as

$$\text{JS}(\mathbf{T}, \mathbf{T}^*) \approx \sum_i \overline{\pi}_i s_i \tag{15}$$

where $\overline{\pi}_i = (\tilde{\pi}_i + \tilde{\pi}_i^*)/2$, and s_i is computed using equal numbers of counts N_i and N_i^*. This approximation is particularly useful in interpreting the JS divergence. According to the above equation, the largest contributions to the overall divergence will come from those states i which are either highly populated at equilibrium (large $\overline{\pi}_i$), highly surprising (large s_i), or both. The quantity $\overline{\pi}_i s_i$ provides a convenient way to identify states i that

dx.doi.org/10.1021/ct500827g | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

contribute most to the difference between an unperturbed and perturbed MSM.

METHODS

2D Lattice Model of Protein Folding. As a model system to demonstrate the utility of surprisal metrics, we use a modified version of the HP model of Lau and Dill,[32] in which a protein is modeled as a self-avoiding chain of monomers living on a square lattice. Toy models of proteins such as this are useful because they are computationally tractable, and provide exact solutions to which we can compare finite-sample estimates.

There is an enumerable set of unique conformations of the chain (conformations related by rotation and reflection are counted as the same conformation) that comprise the complete set of microstates available to the system. Each monomer can either be interacting or neutral. Only the interacting monomers are assigned a particular sequence of amino acids. The energy of each microstate i is computed as

$$E_i = \sum_k \sum_{l>(k+2)} \varepsilon(a_k, a_l)\delta_{kl}(i) \tag{16}$$

where $\delta_{kl}(i) = 1$ if monomers k and l are both interacting and adjacent on the lattice (i.e., in contact), and 0 otherwise; a_k and a_l are the amino acid identities of monomers k and l, respectively, and $\varepsilon(a_k,a_l)$ is the pairwise contact energy between residues a_k and a_l, given by the statistical potential of Miyazawa and Jernigan.[33]

To model conformational dynamics, transition probabilities from microstate i to microstate j are calculated from the Monte Carlo move set Ω of all possible transitions for (i) three-bead flips, (ii) end flips, (iii) crankshaft moves, and (iv) rigid rotations, as described in ref 34. Transition matrix elements T_{ij} for $i \neq j$ are computed as

$$T_{ij} = \frac{\sum_{(i\to j)\in\Omega} \nu(i \to j) P_{\text{accept}}(i \to j)}{\sum_{(i\to j)\in\Omega} 1} \tag{17}$$

and

$$T_{ii} = 1 - \sum_{j\neq i} T_{ij} \tag{18}$$

where $\nu(i \to j) = 1$ if the move is viable (i.e., the chain is still a self-avoiding walk), $\nu = 0$ otherwise, and $P_{\text{accept}}(i \to j) = \min(1, \exp[-(E_j - E_i)/k_BT])$ is the Metropolis acceptance probability for Boltzmann's constant k_B and temperature T.

Molecular Simulation. Molecular dynamics simulations of Fs peptide at 300 K were performed using MPI-enabled GROMACS 4.5.4[35] on the Owlsnest high-performance computing cluster. The AMBER ff03 potential[36] and TIP3P water model were used, with ∼9000 atoms in a $(45\ \text{Å})^3$ periodic box. Na^+ and Cl^- counterions were added at 100 mM to neutralize change. An integration time of 2 fs was used, with constrained hydrogen bonds and PME electrostatics. NVT ensemble temperatures were controlled by a Berendsen thermostat. One hundred trajectories were simulated for a total of 65 μs of simulation time for wild-type Fs, and 30 μs for mutant R9E. Initial configurations for trajectories came from conformational clustering of prior REMD simulations.

Molecular simulations of the GB1 hairpin and trpzip4 variant at 300 K were performed using GROMACS 4.5.3 on the Folding@home distributed computing platform.[12] The simulation system was composed of protein (using the AMBER ff99sb-ildn potential), 1491 TIP3P waters, and 6 Na^+ and 3 Cl^- ions. A total of 1815 trajectories were produced for the GB1 hairpin, totaling 557 μs of simulation data, with an average trajectory length of 0.50 μs. A total of 1189 trajectories were produced for the trpzip4 hairpin, totaling 609 μs of simulation data, with an average trajectory length of 0.55 μs. Twenty different initial configurations were taken from a conformational clustering of simulations in AMBER ff96 using the OBC GBSA implicit solvent model[37] at 350 K.

MSMBuilder2 was used for all construction and analysis of MSMs.[31] Detailed balance was enforced for estimated transition matrices using the default maximum-likelihood estimation method available in MSMBuilder2.[31] Implied relaxation time scales τ_k were estimated as $\tau_k = -\tau/\log(\mu_k)$, where μ_k are the eigenvalues of $\mathbf{T}^{(\tau)}$, and τ is the chosen lag time of the model. The BACE algorithm was used to construct macrostate models.[21] Transition path theory (TPT) calculations[2,23] were performed as previously described.[4,9]

RESULTS

Mutation Effects in a 2D Lattice Model of Protein Folding. We consider two similar 12-mer sequences, A-A-A-A-A-A and A-A-F-A-A-A, where A and F represent the interacting amino acids alanine and phenylalanine, and "-" represents a non-interacting residue (white beads in Figure 2). For ease of notation, we will henceforth omit the non-interacting residues and simply describe the sequences as AAAAAA and AAFAAA.

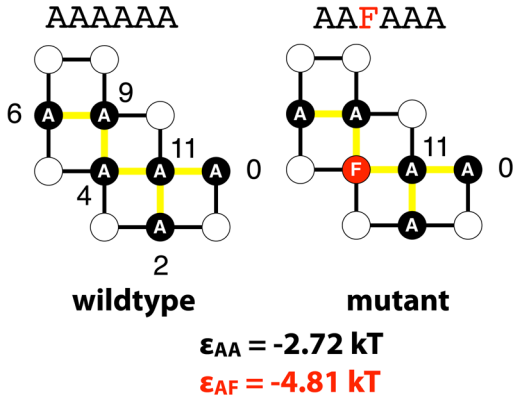


Figure 2. Folded conformations of a 12-residue 2D lattice protein, shown for a wild-type and mutant sequence. In this model, only the colored residues interact. Contacts between interacting residues (shown in yellow) are sequence-dependent, with energies of $\varepsilon_{AA} = -2.72k_BT$ for alanine−alanine contacts, and $\varepsilon_{AF} = -4.81k_BT$ for alanine−phenylalanine contacts (taken from Miyazawa and Jernigan[33]).

The sequence-dependent contact energies of interacting residues are $\varepsilon_{AA} = -2.72k_BT$ and $\varepsilon_{AF} = -4.81k_BT$. Because of the pattern of favorably interacting residues, both the wild-type and mutant sequences are *foldable*, meaning there is a single microstate conformation having the lowest energy, that we call the *native state*. The native state for these sequences is a hairpin-like conformation, shown in Figure 2.

For a 12-mer chain, there are 15 037 microstates, requiring a large, sparse transition matrix. As described in Methods, we compute a transition matrix for each sequence based on a particular model of Monte Carlo conformational sampling, with a lag time of one Monte Carlo step.

dx.doi.org/10.1021/ct500827g | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

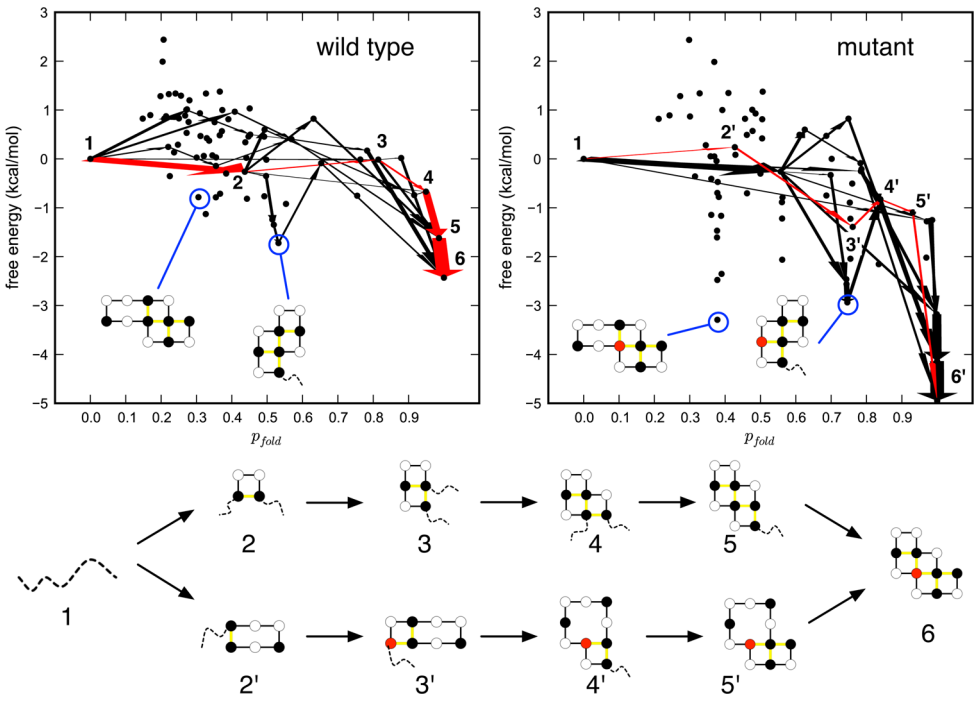


Figure 3. Transition path theory (TPT) analysis of steady-state folding pathway fluxes. Shown for wild type (AAAAAA, left) and mutant (AAFAAA, right) are the 10 macrostate pathways with the largest folding fluxes, with the dominant (largest-flux) pathway shown in red. The highest-flux folding pathway of the wild-type model exhibits a zipping mechanism, whereas the mutant pathway begins with long-range non-native contacts forming first (bottom).

To simplify the analysis, we lump microstates into a coarse-grained set of *macrostates*, where each macrostate is a set of microstates all sharing a unique set of contacts between interacting monomers. For the 12-mer hairpin, there are 72 possible macrostates, providing a convenient number of states to study convergence and the performance of adaptive sampling methods. The number of microstates (15 037) and macrostates (72) is similar to many MSMs constructed to date from all-atom simulations of protein folding, typically utilizing thousands to tens of thousands of microstates, and tens to hundreds of macrostates.

The sequence mutation from AAAAAA to AAFAAA perturbs the thermodynamics and folding kinetics of the toy lattice protein in several ways. Overall, the native-state hairpin population is stabilized, increasing from a population of 0.325 to 0.620 at 300 K. Non-native states are also stabilized, resulting in deeper non-native free energy minima, as well as more pronounced kinetic traps (Figure 3). The slowest relaxation time scale in each model corresponds to the overall folding transition, and the mutation induces a 16-fold increase in the folding time scale, from τ = 6760 steps to τ = 112 800 steps.

A macrostate TPT analysis of steady-state folding fluxes from the unfolded (no-contact) macrostate to the folded native macrostate reveals changes in preferred folding pathways upon mutation. The highest-flux folding pathway in the wild-type sequence is a "zipping" pathway starting from formation of the turn, while the highest-flux pathway for the mutant sequence is dominated by long-range contacts and early non-native interactions.

Surprisal-Based Estimates of the Jensen−Shannon Divergence. An advantage of lattice models is the ability to calculate exact values of the JS divergence, with which we can compare surprisal-based estimates. The total JS divergence of the wild-type and mutant lattice models is 1.688×10^{-5}. Our estimate of the JS divergence is $\mathrm{JS}(\mathbf{T},\mathbf{T}^*) \approx \sum_i \bar{\pi}_i s_i$, where $\bar{\pi}_i = (\tilde{\pi}_i + \tilde{\pi}_i^*)/2$ is an approximation of the equilibrium state populations π_i^{comb} derived from the combined transition matrix $\mathbf{T}^{\mathrm{comb}} = (\mathbf{T} + \mathbf{T}^*)/2$. A comparison of $\bar{\pi}_i$ and π_i^{comb} shows very good agreement (Figure 4a).

Only a small number of macrostates contribute significantly to the total JS divergence, with only 12 states contributing ~90% of the divergence (Figure 4b). The macrostates with the largest contributions to the JS divergence include the highly populated native state, but also a number of low-population, *highly surprising* states, all which have contacts involving the mutated residue. While some of these states contain only native contacts, others contain non-native contacts. The importance of these non-native states implies, as several studies have suggested,[38−40] that consideration of non-native states in mutational analysis of folding is critical. Many of the macrostates contributing most to the JS divergence are also found in the highest-flux pathways connecting the unfolded and native macrostates: for example, state (iii) identified in our surprisal analysis (Figure 4) is the same as state 4 in the highest-flux wild-type pathway (see Figure 3).

Salt Bridge Mutation Effects in All-Atom Simulations of α-Helix Folding. Salt bridges between ionizable residues have long been known to play a key role in protein stability.[41,42] Recent structural bioinformatics work and mutagenesis studies suggest that salt bridges are designable, pH-tunable features that can be used to stabilize protein native structures.[43,44] To elucidate how a salt bridge mutation affects the folding of a simple alpha helical protein, we performed molecular simulations of the Fs peptide (Ace-A_5(AAARA)$_3$A-Nme) and mutant sequence (R9E), as described in Methods. The wild-type sequence is polyalanine, but with three regularly spaced arginines to make the helix water-soluble. Substitution of a glutamic acid residue at position 9 creates opportunities for salt

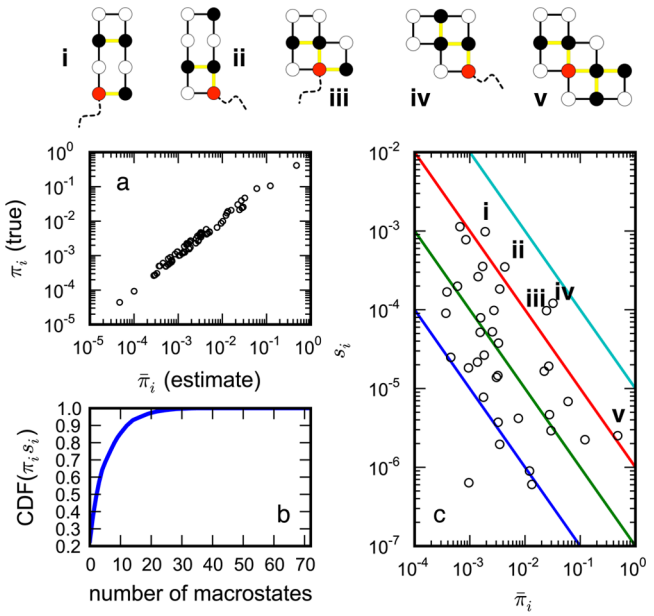


Figure 4. Estimation of the Jensen−Shannon divergence in a toy lattice model. (a) Estimates of equilibrium populations $\bar{\pi}_i = (\pi_i^* + \pi_i)/2$ compare well to the actual values of π_i^{comb} for combined transition matrix $\mathbf{T}^{comb} = (\mathbf{T} + \mathbf{T}^*)/2$. (b) Cumulative distribution of $\bar{\pi}_i s_i$ values across all 72 macrostates. (c) Scatter plot showing s_i versus $\bar{\pi}_i$ for each macro state i. Colored lines are constant-JS contours. Only 12 states contribute 90% of the JS divergence, with the most important contributions coming from the native state (v), as well as a number of low-population, but highly surprising non-native states (i−iv).

bridge formation. Our goal here is to use surprisal metrics to investigate the consequences of this mutation on folding stability and kinetics, at least as predicted by MSMs built from molecular simulation data.

To construct an MSM for both the wild-type Fs peptide and R9E mutant, we performed hybrid k-medoid clustering on the combined trajectory data using backbone + C_β atom positions, resulting in 115 microstates. MSMs were then built for each system by assigning transitions separately based on the combined cluster generators, using a lag time of 10 ns. Validation studies building separate versus combined models confirm the excellent overlap of metastable states for the Fs peptide system (Figures S1 and S2).

Surprisal analysis was performed on the two MSMs (Figure 5), revealing several states with the highest contributions to the JS divergence. Like the lattice model results, we find that some of these states are important because they have high equilibrium populations (e.g., state 0, the native-state helix), while others are important because the differences in outgoing transition probabilities are highly *surprising*. The eight states with the highest JS divergence all lie along transition paths between the helical native state and a two-helix bundle conformation (Figure 6). Structural examination of these states reveals that all have conformations that make possible salt-bridge interactions between R9E and the other arginines. State 48, for example, has a backbone kink allowing significant interaction between E9 and R14; state 10 has a turn allowing interaction between E9 and R19.

These findings confirm our intuition that salt-bridge interactions should be perturbed, and also highlights the importance of sampling non-native states when predicting changes in stability and kinetics, even in a short helical peptide.

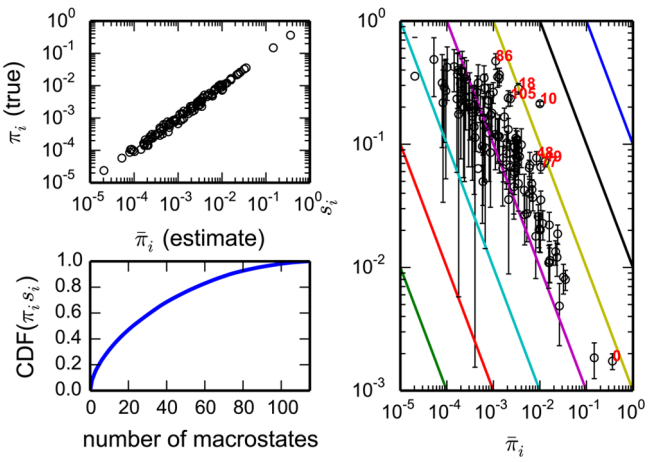


Figure 5. Surprisal analysis of MSMs built from all-atom molecular simulations of Fs peptide and a R9E variant. (A) Estimates of equilibrium microstate populations $\bar{\pi}_i = (\pi_i^* + \pi_i)/2$ compare well to the actual values of π_i^{comb} for combined transition matrix $\mathbf{T}^{comb} = (\mathbf{T} + \mathbf{T}^*)/2$. (B) Cumulative distribution of $\bar{\pi}_i s_i$ values across all 115 microstates. (C) Scatter plot showing s_i versus $\bar{\pi}_i$, with error bars denoting estimated uncertainties in s_i. Colored lines denote constant-JS divergence contours. Labeled are the eight microstates with the largest contributions to the total JS divergence.

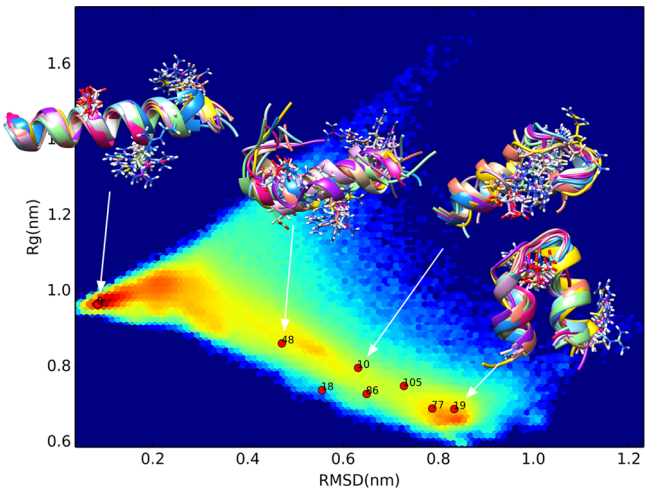


Figure 6. Landscape of sampled conformations in combined molecular simulations of Fs peptide and a R9E variant, projected onto backbone rmsd and radius of gyration reaction coordinates. Denoted as red circles are cluster generator conformations corresponding to the eight microstates with the largest contributions to the JS divergence. These states all lie along a reaction path going from helix to two-helix bundle, and all show significant salt-bridging between residue R9E and the other arginines. Selected states are annotated with a rendering of 10 random conformations from each microstate. Only conformations from the R9E ensemble are shown to better illustrate salt-bridge opportunities. State 48, for example, has a backbone kink allowing significant interaction between E9 and R14; state 10 has a turn allowing interaction between E9 and R19.

The slowest implied time scales from the MSMs built for Fs peptide and R9E are ∼80 and ∼150 ns, respectively, suggesting that salt-bridge interactions indeed create kinetic traps that can increase folding times.

Mutational Effects in All-Atom Simulations of β-Hairpin Folding. Since its discovery as a stable hairpin in solution,[45] the C-terminal hairpin of protein G B1 (GEW-

dx.doi.org/10.1021/ct500827g | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

TYDDATKTFTVTE) has been an important model system for protein folding, allowing direct comparisons between theory and experiment.[46−59] Experimental characterization of folding stability and kinetics of GB1 hairpin variants have been used to probe sequence-dependent folding mechanisms. Using the GB1 hairpin as a template, Cochran et al.[60] designed a series of so-called "tryptophan zipper" proteins—hyperstable hairpin folds made possible by the substitution of several tryptophan residues—whose folding kinetics were found to be more complex.[48,50] In particular, the trpzip4 (TZ4) variant (GEWTWDDATKTWTWTE) was found to have a more stable hairpin fold, but slower folding kinetics. Several studies have suggested that residual structures and/or non-native kinetic traps account for the slower folding times.[49,54]

To probe sequence-dependent folding mechanisms in these hairpins, we compared MSMs constructed for wild-type GB1 hairpin (WT) and TZ4, using the simulated trajectory data described in Methods. The trajectory data sets are quite large; using a lag time of 5 ns resulted in 4.8 million and 5.2 million transition counts for WT and TZ4, respectively. We performed a hybrid *k*-medoid clustering of the combined trajectory data using an rmsd metric on combined backbone + C_β atom positions, resulting in 3000 microstates. The BACE algorithm[21] was used to lump microstates into a 20-macrostate model for TPT[2,23] and surprisal analysis. Because GB and TZ4 differ by the mutation of three bulky residues, we chose a relatively coarse-grained representation to ensure good overlap of metastable states. Implied time scales for separate versus combined data sets (Figure S3) and structural analysis validate that metatstable states are well-conserved (Figure S4−S6).

Surprisal analysis of GB1 versus TZ4 identifies a set of six macrostates (indices 1, 2, 9, 10, 12, and 17) contributing most to the JS divergence (Figure 7). Due to the large number of observed transitions from each macrostate, the statistical uncertainty of the computed surprisal values is very low. According to the surprisal analysis, the two most important states are state 10, which corresponds to the native state, and state 1, corresponding to what we will call the "unfolded state", containing coil and helical conformations. All six macrostates show hydrophobic residues in close proximity, as may be expected for conformational states that are perturbed when hydrophobic residues are mutated to tryptophan (Figure 8).

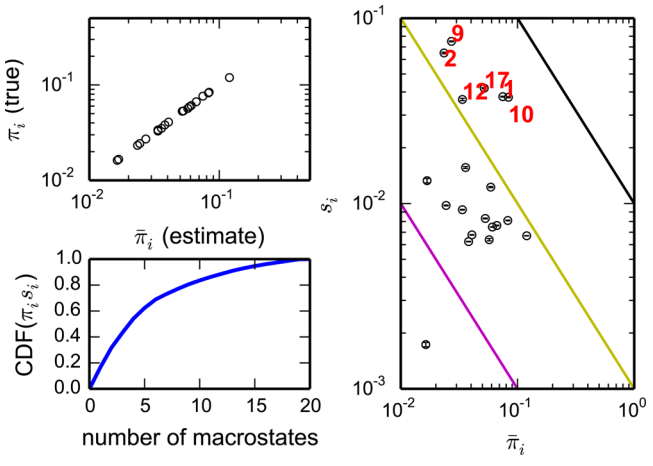


Figure 7. Surprisal analysis of MSMs built from all-atom molecular simulations of GB1 hairpin and trpzip4 variant. (A) Estimates of equilibrium macrostate populations $\bar{\pi}_i = (\pi_i^* + \pi_i)/2$ compare well to the actual values of π_i^{comb} for combined transition matrix $\mathbf{T}^{comb} = (\mathbf{T} + \mathbf{T}^*)/2$. (B) Cumulative distribution of $\bar{\pi}_i s_i$ values across all 20 macrostates. (C) Scatter plot showing s_i versus $\bar{\pi}_i$, with error bars denoting estimated uncertainties in s_i. Colored lines denote constant-JS divergence contours. Labeled are the six macrostates with the largest contributions to the total JS divergence.

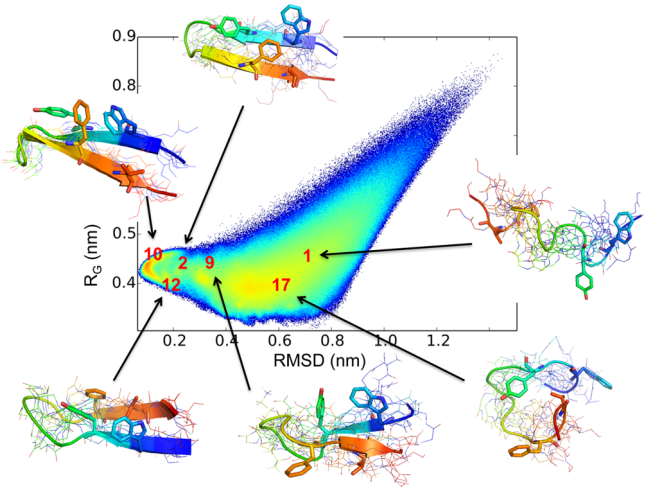


Figure 8. Landscape of sampled conformations in combined molecular simulations of GB1 hairpin and trpzip (TZ4) variant, projected onto backbone rmsd and radius of gyration reaction coordinates. Shown are the six macrostates with the largest contributions to the JS divergence, rendered using 10 GB1 trajectory frames randomly selected from each macrostate. All six macrostates show hydrophobic residues in close proximity, as may be expected for conformational states that are perturbed when hydrophobic residues are mutated to tryptophan. State 10 is the native state, and state 1 corresponds to what we will call the "unfolded state", containing coil and helical conformations,. The remaining states are well-structured non-native traps expected to be greatly affected by tryptophan mutations. State 2 is a "near-native" state with the same topology as the native state, but with much more structural heterogeneity and poorly packed side chains. States 9 and 12 are misregistered hairpins with one or more strands "flipped", while state 17 has end-to-end hydrophobic interactions between W3 and V14/W14.

State 2 is a "near-native" state with the same topology as the native state, but with much more structural heterogeneity and poorly packed side chains. States 9 and 12 are misregistered hairpins; state 9 has the C-terminal strand "flipped", while state 12 has both strands flipped. The main feature of state 17 is an end-to-end hydrophobic interaction between W3 and V14/W14.

TPT analysis reveals changes in the predominant folding pathways (Figure 9), not unlike changes observed in the 2D toy lattice model (Figure 3). Steady-state folding fluxes and committor (p_{fold}) values were computed from state 1 (unfolded) to state 20 (native). States 2, 12, and 9 are found to be near the transition state (with p_{fold} values near 0.5) for both models, with a large flux coming into the native state from state 2. Not coincidentally, these three states are among highest contributors to the JS divergence, indicating their importance in controlling the kinetics and distribution of folding pathways. From the TPT analysis, we can see that tryptophan mutations in TZ4 produce changes reminiscent of those seen in the lattice model: they stabilize a number of unfolded/misfolded states, and slow down folding kinetics. Microstate MSM models constructed for WT and TZ4 predict folding relaxation time scales of ∼170 and ∼280 ns, respectively (Figure S3). These

dx.doi.org/10.1021/ct500827g | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

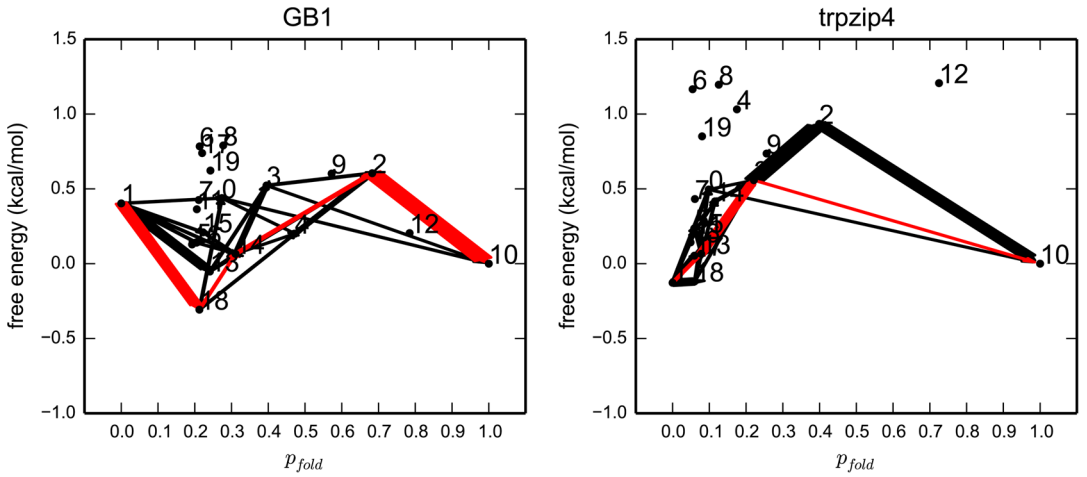


Figure 9. Transition path theory (TPT) analysis of steady-state folding pathway fluxes for the protein GB1 hairpin and its trpzip4 variant. Shown are the 10 macrostate pathways with the largest folding fluxes from the unfolded state (state 1) to the folded state (state 10). Line thicknesses are proportional to the inter-macrostate flux values, and the single largest-flux pathway is shown in red.

values are consistent with the slower folding times for TZ4 measured in experiment,[49] but are accelerated more than an order of magnitude over the experimental values. Predicted time scales agree much better with experiment when using the tICA[61,62] approach to define metastable states (unpublished results). The TPT analysis also reveals a shift in the transition state ensemble upon mutation to TZ4 toward more native-like structures. The "near-native" state 2, for example, is pushed from the product side of the reaction to the reactant side.

In both our lattice model example and in all-atom molecular simulations of folding, we see unfolded and misfolded states as key contributors to the JS divergence, underscoring their importance in controlling the kinetics and thermodynamics of folding. We believe the elucidation of these states via surprisal analysis can be useful for understanding folding mechanisms in detail, and for identifying states requiring more focused conformational sampling.

Surprisal-Based Adaptive Sampling. In adaptive sampling schemes, multiple rounds of simulation trajectories are sampled iteratively in a biased fashion so as to efficiently reduce the uncertainty of some quantity of interest.[16−18] For example, the adaptive sampling scheme of Hinrichs et al.[17] (which we will call SA sampling, for "sensitivity analysis") iteratively identifies the state i that contributes most to the uncertainty of the largest eigenvalue of the transition matrix (corresponding to the slowest relaxation rate). For MSMs of protein folding, successive rounds of SA adaptive sampling result in converged estimates of folding rates. Bowman et al.[18] used SA sampling along with adaptive state discovery to show that multiple short trajectories chosen adaptively and simulated in parallel can greatly reduce the time required to construct statistically accurate MSMs.

Here, we examine the use of surprisal-based adaptive sampling to produce converged estimates of the JS divergence of two MSMs, as described in the following algorithm:

1. Collect equal numbers of transition counts n_{ij} and n_{ij}^* from simulations of the unperturbed and perturbed systems, respectively.
2. Compute estimates of the variance $\sigma^2_{\overline{\pi}_i s_i}$ for each state i.
3. For both unperturbed and perturbed systems, start an equal number of new simulations from the state i with the largest value of $\sigma^2_{\overline{\pi}_i s_i}$, adding new observed transition counts to the previous counts.
4. Repeat from step 2, until the estimate of $\sigma^2_{\overline{\pi}_i s_i}$ is below some desired tolerance.

To estimate the statistical uncertainty of the quantity $\overline{\pi}_i s_i$, we use the approximation $\sigma^2_{\overline{\pi}_i s_i} \approx \overline{\pi}_i^2 \sigma^2_{s_i}$, at the risk of underestimating the true variance. (This shortcoming could be remedied in future work using the method by Hinrichs et al. to estimate eigenvector uncertainties.[17])

Surprisal-Based Adaptive Sampling Is Efficient. We used the toy hairpin model described above to investigate the efficiency of surprisal-based adaptive sampling. We explored two different adaptive sampling schemes differing in the way states are chosen, and the way transition counts are collected. We define *focused* adaptive sampling of state i to mean collecting a set of transition counts, all starting from state i, at each iteration of adaptive sampling. By contrast, we define *trajectory* sampling as generating a series of transition counts from short trajectories starting from state i, at each iteration. Unlike focused sampling, trajectory sampling can generate a chain of transitions through neighboring states. Finally, we define *unbiased* sampling to mean generating a Markov chain with no adaptive bias; this is akin to generating a single, long molecular dynamics trajectory.

Adaptive sampling experiments were performed by first sampling 10 000 initial transition counts in proportion to equilibrium populations π_i for AAAAAA and π_i^* for AAFAAA models. Then, 1000 iterations of adaptive sampling were performed, with each iteration consisting of 1000 sampled transitions. To determine how efficient surprisal-based sampling is compared to other adaptive sampling methods, we compared our results to the SA sampling method of Hinrichs et al.,[17] where the matrix of combined counts was used as an input.

Figure 10 shows the total JS divergence (as estimated by $\sum_i \overline{\pi}_i s_i$) of the two MSMs over the course of adaptive sampling. The results presented are averaged over five independent sampling runs. We find that surprisal-based sampling is the most efficient way to decrease the uncertainty in the JS divergence. For focused adaptive sampling (Figure 10a), surprisal-based is the most efficient, while SA sampling is

dx.doi.org/10.1021/ct500827g | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

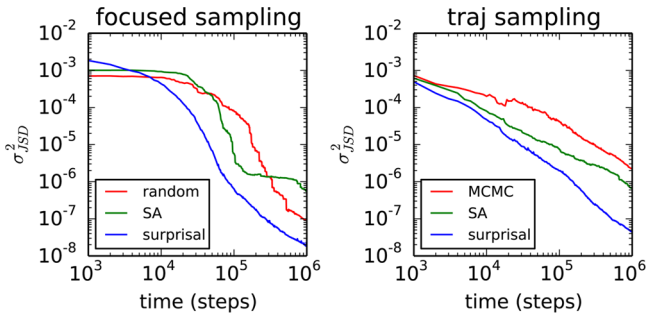


Figure 10. Surprisal-based adaptive sampling efficiently converges estimates of the JS divergence between a wild-type AAAAAA and mutant AAFAAA model. The statistical uncertainty in the total Jensen−Shannon divergence (calculated as $\sum_i \overline{\pi}_i s_i$)was estimated over multiple 1000 rounds of adaptive sampling, with each round consisting of 1000 samples. (a, Left) Results of focused adaptive sampling (transition counts all drawn starting from the same state) performed for random adaptive choices (red), states chosen by sensitivity analysis (green) and surprisal-based sampling (blue). (b, Right) Results of trajectory adaptive sampling (a Markov chain of transitions generated each round, starting from the adaptively chosen state) performed for a single continuous trajectory (red), states chosen by sensitivity analysis (green) and surprisal-based sampling (blue).

comparable to choosing a random state to perform focused sampling at each iteration. In general, trajectory sampling (Figure 10b) is less efficient than focused sampling, with SA sampling showing efficiencies worse than suprisal-based sampling, but better than Markov Chain Monte Carlo sampling.

We also ran adaptive sampling tests using identical MSM models (both with sequence AAAAAA). In this case, the JS divergence should approach zero in the limit of large numbers of sampled transition counts, and the only source of uncertainty is finite sampling error. The results are similar to the results found for different MSM models (Figure 11); surprisal-based

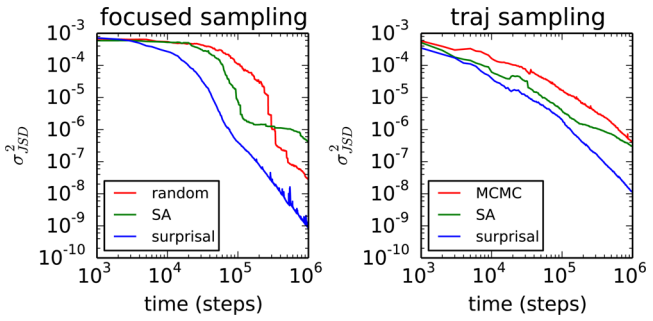


Figure 11. Surprisal-based adaptive sampling efficiently converges estimates of the JS divergence between two wild-type models (AAAAAA). In this case, the total JS divergence converges to zero, with the uncertainty solely a product of finite sampling error. The description is the same as in Figure 10.

sampling is the most efficient, followed by SA sampling. These results suggest that surprisal-based sampling could be useful for choosing states to bias adaptive sampling so as to converge a single MSM model.

We note that, since the toy lattice model dynamics are exactly Markovian by definition, the adaptive sampling results we present are equivalent to sampling from an MSM. We hope to publish similar results using all-atom molecular simulation trajectories in the near future.

DISCUSSION

Implications for Folding Mechanisms. In both lattice models and all-atom molecular simulation models, we find that surprisal metrics identify key non-native states. Many of these states are on- or off-pathway "traps" that can significantly alter folding mechanisms and kinetics. In the case of the lattice model hairpin, we see that the stabilization of such states leads to changes from the "zipping" mechanism of turn-first folding, toward an "end-to-end" mechanism. For our simulations of the protein G hairpin and its trpzip4 variant, we find similar perturbations to hairpin stability, kinetics and folding pathways, with the largest contributors to the JS divergence being states consisting of non-native hairpin conformations. Surprisal analysis can automatically identify these states and determine their significance.

For our simulations of the Fs peptide and R9E variant, we find that intermediate states that enable salt-bridge interactions are important determinants of folding mechanisms. The ubiquitous nature of salt-bridging interactions in folding, binding and aggregation[25,63,64] suggests that surprisal analysis may be very useful for analyzing biologically relevant mutations and their impact on human health.

Implications for Simulation-Based Design. We believe surprisal-based sampling is an attractive strategy for simulation-based first-principle molecular design. The main obstacle in using molecular simulation to design proteins and peptidomimetics is the inability to explore sequence space efficiently. There are recent exciting examples of molecular simulations being used to design proteins,[65] but they are limited by the requirement of performing separate independent simulations for each candidate sequence. Protein design algorithms such as Rosetta[66] overcome this limitation by decoupling the protein conformational search from the sequence search using fast side-chain rotamer libraries.[67] Similarly, MSMs offer an efficient way to decouple conformational space (the MSM metastable state definitions) from sequence space (sequence-dependent rate perturbations).

While here we present results comparing two MSMs, the general approach of estimating the JS divergence should extend to multiple MSM models, so that many different perturbation effects can be screened in parallel. Future work will focus on this, as well as automating surprisal-based sampling algorithms for use with all-atom simulation models.

One of the caveats of the surprisal-based analysis presented here is the assumption that metastable states are conserved. Although for the Fs-peptide system mutation of a single residue does not incur significant changes in metastable states, that may not be true for larger systems. In a recent proof-of-principle study aimed toward the computational design of cyclic peptide β-hairpin mimetics,[68] it was found that a small chemical perturbation (methylation) did not perturb the relevant metastable states of MSMs built using the tICA method,[61,62] while a mutation from L- to D-valine induced structural changes significant enough that separate MSMs needed to be built. This was in part due to the fact that the valine coordinates comprised a large component of the slowest degrees of freedom identified by tICA, making the clustering results sensitively dependent on this residue.

Future work needs to be done to more systematically identify situations where poor metastable state conservation exists, and to guide the use of more coarse-grained representations where appropriate. One strategy for doing this would be to use a

dx.doi.org/10.1021/ct500827g | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

combined surprisal metric that takes into account differences incurred by coarse-graining of metastable states (as the BACE algorithm does[21]) along with differences incurred by rate perturbations (mutation and/or chemical modification). Such an algorithm may be able to prevent truly distinct metastable states from getting lumped together, according to rigorous and quantitative criteria.

Another scenario that our approach does not deal with explicitly is the discovery of new metastable states throughout the course of adaptive sampling. We believe that surprisal-based sampling could be useful in this case, provided that metastable states are periodically recomputed. In this case, newly discovered states would be expected to have a large uncertainty in the JS divergence, and would receive high priority in surprisal-based adaptive sampling schemes, once discovered.

To be clear, the surprisal-based adaptive sampling algorithm presented here is—like previously published adaptive sampling approaches for MSMs[16−18]—a heuristic procedure that is only empirically shown to give efficient convergence. In this sense, surprisal-based sampling is an instance of the "multi-arm bandit" problem,[69,70] in which the optimal sampling of multiple unknown stochastic processes are inexorably tied to their exploration. Such problems currently have no formal solution in computer science, but certain strategies can be shown to be optimal in specific cases.[71] Although we have empirically shown that surprisal-based adaptive sampling can efficiently decrease the uncertainty in the JS divergence between two MSMs, we must be careful to point out that the discovery of new metastable states may impact this efficiency in a non-trivial way.

CONCLUSION

In this work, we have presented a surprisal metric for quantifying the difference between an unperturbed and perturbed MSM. As we have shown, the surprisal is closely related to the Jensen−Shannon divergence, which can be used to identify metastable states exhibiting the most important differences. These states are found to be those with high equilibrium populations, as well as those with highly *surprising* differences in outgoing transition probabilities. In both lattice models and all-atom simulations, such states reveal much about the mechanistic differences in folding induced by mutations. Moreover, surprisal-based adaptive sampling efficiently decreases the uncertainty of the Jensen−Shannon divergence, to achieve statistically converged MSMs. We believe strategies like this will be useful for simulation-based computational design, as it allows the effects of sequence mutations and/or chemical modifications to be assessed separately from the conformational search.

APPENDIX

Equation 5 can be simplified using Stirling's approximation, $\ln n! = \ln \Gamma(n+1) \approx n \ln n - n$, to get

$$
\begin{aligned}
\mathcal{S}_i &= -\ln P(\mathbf{n}_i^*|\mathbf{n}_i) \\
&= -N_i^* \ln N_i^* + N_i^* - \sum_j^M n_{ij}^* \ln n_{ij}^* - \sum_j^M n_{ij}^* \\
&\quad - (N_i + \sum_j^M \alpha_{ij} - 1) \ln(N_i + \sum_j^M \alpha_{ij} - 1) \\
&\quad + (N_i + \sum_j^M \alpha_{ij} - 1) + \sum_j^M (n_{ij} + \alpha_{ij} - 1) \\
&\quad \ln(n_{ij} + \alpha_{ij} - 1) - \sum_j^M (n_{ij} + \alpha_{ij} - 1) \\
&\quad + (N_i^* + N_i + \sum_j^M \alpha_{ij} - 1) \ln(N_i^* + N_i \\
&\quad + \sum_j^M \alpha_{ij} - 1) - (N_i^* + N_i + \sum_j^M \alpha_{ij} - 1) \\
&\quad - \sum_j^M (n_{ij}^* + n_{ij} + \alpha_{ij} - 1) \ln(n_{ij}^* + n_{ij} + \alpha_{ij} \\
&\quad - 1) + \sum_j^M (n_{ij}^* + n_{ij} + \alpha_{ij} - 1)
\end{aligned}
$$

After cancelation, combination, and rearrangement of terms, we get

$$
\begin{aligned}
\mathcal{S}_i &= [(N_i^* + N_i + \sum_j^M \alpha_{ij} - 1) \ln(N_i^* + N_i \\
&\quad + \sum_j^M \alpha_{ij} - 1) - \sum_j^M (n_{ij}^* + n_{ij} + \alpha_{ij} - 1) \\
&\quad \ln(n_{ij}^* + n_{ij} + \alpha_{ij} - 1)] \\
&\quad - [N_i^* \ln N_i^* - \sum_j^M n_{ij}^* \ln n_{ij}^*] \\
&\quad - [(N_i + \sum_j^M \alpha_{ij} - 1) \ln(N_i + \sum_j^M \alpha_{ij} - 1) \\
&\quad - \sum_j^M (n_{ij} + \alpha_{ij} - 1) \ln(n_{ij} + \alpha_{ij} - 1)] \\
&= (N_i^* + \hat{N}_i) H_i^{\mathrm{comb}} - N_i^* H_i^* - \hat{N}_i H_i
\end{aligned}
$$

where $\hat{N}_i = (N_i + \sum_j^M \alpha_{ij} - 1)$ and H_i^*, H_i, and H_i^{comb} are the entropy terms defined in the main text.

ASSOCIATED CONTENT

Supporting Information

Supporting Figures S1−S6. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: voelz@temple.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation through grant NSF-MCB-1412508, start-up

funding through Temple University, and major research instrumentation grant no. NSF-CNS-09-58854. We thank the participants of Folding@home for their continued contributions to this research.

REFERENCES

(1) Chodera, J. D.; Noé, F. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135−144.
(2) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011−19016.
(3) Zheng, W.; Gallicchio, E.; Deng, N.; Andrec, M.; Levy, R. M. *J. Phys. Chem. B* **2011**, *115*, 1512−1523.
(4) 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*, 12565−12577.
(5) De Sancho, D.; Mittal, J.; Best, R. B. *J. Chem. Theory Comput.* **2013**, *9*, 1743−1753.
(6) Beauchamp, K. A.; McGibbon, R.; Lin, Y.-S.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 17807−17813.
(7) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2011**, *133*, 18413−18419.
(8) Sadiq, S. K.; Noé, F.; De Fabritiis, G. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 20449−20454.
(9) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526−1528.
(10) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, No. 174105.
(11) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Dill, K. A. *Multiscale Model. Sim.* **2006**, *5*, 1214−1226.
(12) Shirts, M.; Pande, V. S. *Science* **2000**, *290*, 1903−1904.
(13) Harvey, M. J.; De Fabritiis, G. *Drug Discovery Today* **2012**, *17*, 1059−1062.
(14) Kohlhoff, K. J.; Shukla, D.; Lawrenz, M.; Bowman, G. R.; Konerding, D. E.; Belov, D.; Altman, R. B.; Pande, V. S. *Nat. Chem.* **2014**, *6*, 15−21.
(15) Buchete, N.-V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.
(16) Singhal, N.; Pande, V. S. *J. Chem. Phys.* **2005**, *123*, No. 204909.
(17) Hinrichs, N. S.; Pande, V. S. *J. Chem. Phys.* **2007**, *126*, No. 244101.
(18) Bowman, G. R.; Ensign, D. L.; Pande, V. S. *J. Chem. Theory Comput.* **2010**, *6*, 787−794.
(19) Deuflhard, P.; Weber, M. *Linear Algebra Appl.* **2005**, *398*, 161−184.
(20) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, No. 155101.
(21) Bowman, G. R. *J. Chem. Phys.* **2012**, *137*, No. 134111.
(22) Yao, Y.; Cui, R. Z.; Bowman, G. R.; Silva, D.-A.; Sun, J.; Huang, X. *J. Chem. Phys.* **2013**, *138*, No. 174106.
(23) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. *Multiscale Model. Sim.* **2009**, *7*, 1192−1219.
(24) Long, D.; Brüschweiler, R. *J. Am. Chem. Soc.* **2011**, *133*, 18999−19005.
(25) Lin, Y.-S.; Pande, V. S. *Biophys. J.* **2012**, *103*, L47−L49.
(26) Tribus, M. *Thermostatics and thermodynamics: an introduction to energy, information and states of matter, with engineering applications*; van Nostrand: New York, 1961.
(27) Cover, T. M.; Thomas, J. A. *Elements of information theory*; John Wiley & Sons: New York, 2012.
(28) Levine, R.; Bernstein, R.; Kahana, P.; Procaccia, I.; Upchurch, E. *J. Chem. Phys.* **1976**, *64*, 796−807.
(29) Bacallado, S.; Chodera, J. D.; Pande, V. *J. Chem. Phys.* **2009**, *131*, No. 045106.
(30) Trendelkamp-Schroer, B.; Noé, F. *J. Chem. Phys.* **2013**, *138*, No. 164113.
(31) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. *J. Chem. Theory Comput.* **2011**, *7*, 3412−3419.
(32) Lau, K. F.; Dill, K. A. *Macromoleules* **1989**, *22*, 3986−3997.
(33) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623−644.
(34) Dill, K. A.; Bromberg, S.; Yue, K.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D.; Chan, H. S. *Protein Sci.* **1995**, *4*, 561−602.
(35) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D. *Bioinformatics* **2013**, No. btt055.
(36) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T. *J. Comput. Chem.* **2003**, *24*, 1999−2012.
(37) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins: Struct., Funct. Bioinf.* **2004**, *55*, 383−394.
(38) Cho, J.-H.; Meng, W.; Sato, S.; Kim, E. Y.; Schindelin, H.; Raleigh, D. P. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 12079−12084.
(39) Cho, J.-H.; Raleigh, D. P. *J. Am. Chem. Soc.* **2006**, *128*, 16492−16493.
(40) Weinkam, P.; Pletneva, E. V.; Gray, H. B.; Winkler, J. R.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 1796−1801.
(41) Marqusee, S.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 8898−8902.
(42) Marqusee, S.; Robbins, V. H.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 5286.
(43) Donald, J. E.; Kulp, D. W.; DeGrado, W. F. *Proteins: Struct., Funct. Bioinf.* **2011**, *79*, 898−915.
(44) Lau, W. L.; DeGrado, W. F.; Roder, H. *Biophys. J.* **2010**, *99*, 2299−2308.
(45) Blanco, F. J.; Rivas, G.; Serrano, L. *Nat. Struct. Mol. Biol.* **1994**, *1*, 584−590.
(46) Munoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196−199.
(47) Dinner, A. R.; Lazaridis, T.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9068−9073.
(48) Snow, C. D.; Qiu, L.; Du, D.; Gai, F.; Hagen, S. J.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4077−4082.
(49) Du, D.; Zhu, Y.; Huang, C.-Y.; Gai, F. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15915−15920.
(50) Du, D.; Tucker, M. J.; Gai, F. *Biochemistry* **2006**, *45*, 2668−2678.
(51) Fesinmeyer, R. M.; Hudson, F. M.; Andersen, N. H. *J. Am. Chem. Soc.* **2004**, *126*, 7238−7243.
(52) Olsen, K. A.; Fesinmeyer, R. M.; Stewart, J. M.; Andersen, N. H. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15483−15487.
(53) Zhou, R.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12777−12782.
(54) Yang, W. Y.; Pitera, J. W.; Swope, W. C.; Gruebele, M. *J. Mol. Biol.* **2004**, *336*, 241−251.
(55) Lin, E.; Shell, M. S. *J. Chem. Theory Comput.* **2009**, *5*, 2062−2073.
(56) Zhuang, W.; Cui, R. Z.; Silva, D.-A.; Huang, X. *J. Phys. Chem. B* **2011**, *115*, 5415−5424.
(57) Juraszek, J.; Bolhuis, P. G. *J. Phys. Chem. B* **2009**, *113*, 16184−16196.
(58) Best, R. B.; Mittal, J. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108* (27), 11087−11092.
(59) Best, R. B.; Mittal, J. *Proteins: Struct., Funct. Bioinf.* **2011**, *79*, 1318−1328.
(60) Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5578−5583.
(61) Schwantes, C. R.; Pande, V. S. *J. Chem. Theory Comput.* **2013**, *9*, 2000−2009.
(62) Perez-Hernandez, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. *J. Chem. Phys.* **2013**, *139*, 015102.
(63) Ciani, B.; Jourdan, M.; Searle, M. S. *J. Am. Chem. Soc.* **2003**, *125*, 9038−9047.
(64) Tarus, B.; Straub, J. E.; Thirumalai, D. *J. Am. Chem. Soc.* **2006**, *128*, 16159−16168.

(65) Piana, S.; Sarkar, K.; Lindorff-Larsen, K.; Guo, M.; Gruebele, M.; Shaw, D. E. *J. Mol. Biol.* **2011**, *405*, 43−48.
(66) Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D. *Methods Enzymol.* **2004**, *383*, 66−93.
(67) Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L. *Proteins: Struct., Funct. Bioinf.* **2009**, *77*, 778−795.
(68) Razavi, A. M.; Wuest, W. M.; Voelz, V. A. *J. Chem. Inf. Model.* **2014**, *54*, 1425−1432.
(69) Auer, P.; Cesa-Bianchi, N.; Fischer, P. *Mach. Learn.* **2002**, *47*, 235−256.
(70) Scott, S. L. *Appl. Stoch. Model. Bus. Ind.* **2010**, *26*, 639−658.
(71) Burnetas, A. N.; Katehakis, M. N. *Adv. Appl. Math.* **1996**, *17*, 122−142.