JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

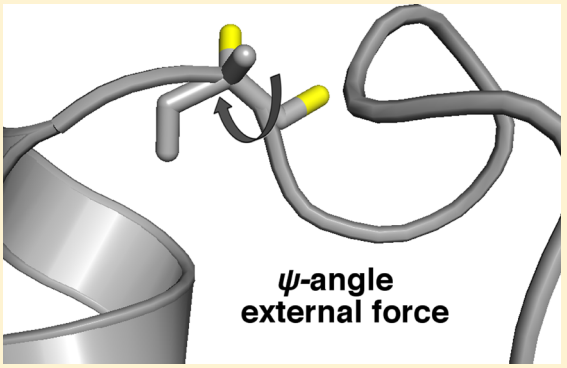
Potential-Based Dynamical Reweighting for Markov State Models of Protein Dynamics

Jeffrey K. Weber and Vijay S. Pande*

Department of Chemistry, Stanford University, Stanford, California 94305, United States

Supporting Information

ABSTRACT: As simulators attempt to replicate the dynamics of large cellular components *in silico*, problems related to sampling slow, glassy degrees of freedom in molecular systems will be amplified manyfold. It is tempting to augment simulation techniques with external biases to overcome such barriers with ease; biased simulations, however, offer little utility unless equilibrium properties of interest (both kinetic and thermodynamic) can be recovered from the data generated. In this Article, we present a general scheme that harnesses the power of Markov state models (MSMs) to extract equilibrium kinetic properties from molecular dynamics trajectories collected on biased potential energy surfaces. We first validate our reweighting protocol on a simple two-well potential, and we proceed to test our method on potential-biased simulations of the Trp-cage miniprotein. In both cases, we find that equilibrium populations, time scales, and dynamical processes are reliably reproduced as compared to gold standard, unbiased data sets. We go on to discuss the limitations of our dynamical reweighting approach, and we suggest auspicious target systems for further application.



INTRODUCTION

At present, small components of the cellular infrastructure can be simulated in isolation with great effort. Scientists ultimately desire, however, to describe proteins, lipids, and sugars in their fully interacting environments and study the biological pathways these molecules dictate. Simulating the entire functional architecture of the cell and its connections to surrounding, living components presents a grand challenge to computational biologists.[1] Burgeoning computational resources will assist researchers in spanning the vast spatial and time scales needed to characterize biophysics at the cellular level. However, physical tools that can be used to understand and augment the dynamics encoded in large data sets will become increasingly important in the pursuit of such extensive simulation outcomes.

In the world of protein folding simulations, Markov state models (MSMs) represent a tool that facilitates the interpretation of large molecular dynamics (MD) data sets.[2,3] In discretizing phase space and parametrizing transition rates among the resulting states, one builds an MSM to define a memoryless dynamical network among a set of mesoscale conformational ensembles. Importantly, an MSM provides a generative model of a protein's long time scale dynamics: by concatenating microscopic transitions between discrete states, one can produce statistically valid trajectories of arbitrary length in a *post hoc* fashion.[4] Though MSMs yield apt descriptions of protein dynamics derived from long MD trajectories, such models can also be parametrized from relatively short trajectories scattered throughout phase space. This fact allows massively parallel computing techniques (like those pioneered by the Folding@home distributed computing project) to be leveraged to great effect. Additionally, the mesoscopic MSM network provides a complete description of a system's kinetic and thermodynamic properties. MSM dynamics, therefore, can be projected onto both structural and kinetic order parameters and allow arbitrary correlation functions to be calculated with ease. Such features simplify comparison with experimental data and simulation results obtained by other means.

In their own right, MSMs offer detailed descriptions of protein free energy landscapes that can be exploited as toy models for further investigation. Efforts throughout the literature have focused on the application of perturbation theories to Markovian networks.[5] In the context of protein dynamics, we have employed random perturbation theories to study dynamical response properties in protein folding MSMs,[6] and we have used so-called "large-deviation" theories to derive insights into far-from-equilibrium phenomena like protein glass transitions and dissipative dynamics in models of molecular engines.[7,8]

One might, however, also consider the converse of this perturbative approach: one could collect perturbed MD data (optimized, for example, to sample a given degree of freedom) and build a reweighted MSM that represents the corresponding equilibrium picture. Methods by which equilibrium populations can be reweighted to better agree with experimental data or

Received: January 13, 2015
Published: April 29, 2015

DOI: 10.1021/acs.jctc.5b00031
J. Chem. Theory Comput. 2015, 11, 2412−2420

other constraints are well developed within the MSM framework.[9] Generating a reweighted model with high fidelity to the desired equilibrium kinetics, though, represents a disparate challenge. Any dynamical reweighting approach must evaluate full trajectories from a probabilistic standpoint, so that proper equilibrium rates for transitions can be parametrized from perturbed (i.e., nonequilibrium) simulation data. How can one build an equilibrium MSM from trajectories sampled from an arbitrary potential energy surface?

Previous work related to parallel tempering methods may provide a hint to the answer. Recently, Chodera et al. leveraged the stochasticity in MD integrators to recover low-temperature autocorrelation functions from high-temperature simulations;[10] in a similar fashion, Prinz et al. developed an ansatz to reweight transition counts from replica exchange simulations and estimate MSM properties as equilibrium expectations.[11] Both approaches rely on an understanding of canonical kinetic energy distributions, and quantitative differences in dependent trajectories, to facilitate the recovery of equilibrium properties.

In this Article, we present an analogous protocol that allows one to build equilibrium MSMs from trajectories sampled from arbitrary potential energy surfaces. We start by outlining the theoretical details for reweighting trajectories between potentials based on path actions, and we next discuss the method by which MSMs can be constructed from the resultant data. We apply these methods to two systems of interest. First, as a proof of principle, we reweight dynamics sampled from a two-well potential. Second, we attempt to accelerate the sampling of folding dynamics in a biopolymeric system by perturbing the dihedral torsion energetics within the Trp-cage miniprotein.[12] In both cases, dynamical reweighting recovers kinetic properties to reasonable agreement with gold-standard equilibrium results. We discuss the qualities of the reweighted MSMs we obtain, and we conclude by enumerating the apparent limitations of the potential surface reweighting approach along with its promising future applications.

THEORY, MODELS, AND METHODS

Derivation. Consider the phase space dynamics of a system on two different potential energy surfaces: one being a "target" potential and the other a "perturbed" potential. To generate target kinetics, we wish to collect samples of trajectories from the perturbed potential and evaluate the likelihood with which these trajectories occur on the target potential. From the perspective of pure Hamiltonian dynamics, of course, the probability of divergent trajectory outcomes is exactly zero: trajectories are uniquely determined by initial velocities, gradients of the potential, any static frictional terms that might be present, etc. Biomolecular simulations, however, are not conducted in isolation: solute and solvent are deemed to be in contact with a heat bath, a phenomenon which is in turn simulated by coupling particles to an external thermostat. In general, thermostatting procedures contain a stochastic term which serves to scatter velocities according to a Maxwell–Boltzmann distribution parametrized by a given temperature. The method of temperature control on which we focus here is Langevin integration, which achieves fidelity to the Maxwell–Boltzmann distribution via a random force, or "noise," term. The prototypical, 1-dimensional Langevin equation can be written as[13−15]

$$m_i \frac{\partial^2 \mathbf{r}_i(t)}{\partial t^2} = \mathbf{F}_i - \lambda \frac{\partial \mathbf{r}_i(t)}{\partial t} + \eta_i(t) \tag{1}$$

where $\mathbf{F}_i$ is the conservative force derived from the system's potential, λ is a static friction coefficient, and $\eta_i(t)$ is a Gaussian random force with zero mean and a standard deviation, σ_{ij}, provided by[13,14]

$$\sigma_{ij} = \sqrt{2kT\gamma m_i \delta_{ij} \delta(t - t')} \tag{2}$$

The delta functions in the above expression enforce an assumption of uncorrelated noise: random forces are uncorrelated in both time ($\delta(t - t')$) and between separate degrees of freedom (δ_{ij}). True collisions from a molecular heat bath, of course, would induce nonzero correlation-times in resultant random forces, but such correlations would be quickly damped by the frequency of integration.

Under the Langevin framework, one can immediately see how a dynamical reweighting scheme might be designed. If a trajectory jumps to a configuration over a large conservative force difference between potentials, one can recast that step in the trajectory as a rarer fluctuation over the less negative gradient. Practically, this procedure amounts to assigning a portion of the conservative force to the random force and evaluating how likely this random force is to occur at the given temperature. Suppose a particle, for single degree of freedom, experiences some conservative force in the target potential, denoted by $\mathbf{F}_i^{\text{target}}$, and a different force, $\mathbf{F}_i^{\text{perturbed}}$, within the perturbed potential. For the same integration step taken in the perturbed potential to occur in the target potential, the random force must cover the deficit in the conservative forces:

$$[\eta_i]_{\text{RW}} = [\mathbf{F}_i^{\text{perturbed}} - \mathbf{F}_i^{\text{target}}] + \eta_i \tag{3}$$

Here, η_i represents the observed random force in the perturbed potential, and $[\eta_i]_{\text{RW}}$ is the reweighted random force needed to reproduce the identical dynamical step in the target potential. The "probability" for each step to occur via random fluctuations is then evaluated using the expression for the Gaussian noise. For the purposes of numerical stability, it is convenient to work with path actions, rather than pure probabilities. The action, S_{step}, corresponding to a single integration time step is simply given by the negative logarithm of that step's probability, P_{step}: $S_{\text{step}} = -\log P_{\text{step}}$.[10] Closed form expressions for the unbiased and reweighted stepwise actions, $[S_i]_0$ and $[S_i]_{\text{RW}}$, are given by

$$[S_i]_0 = \frac{\eta_i^2}{2\sigma_{ij}^2} - \log\left[\frac{1}{\sqrt{2\pi\sigma_{ij}^2}}\right] \tag{4}$$

$$[S_i]_{\text{RW}} = \frac{[\eta_i]_{\text{RW}}^2}{2\sigma_{ij}^2} - \log\left[\frac{1}{\sqrt{2\pi\sigma_{ij}^2}}\right] \tag{5}$$

For a dynamical trajectory (of length t_{obs}) over a full configuration of N particles, one can find the total path action by first summing over the individual action components and next over the configurational actions at each time step:[10]

$$S_0^{\text{traj}} = \sum_{t=1}^{t_{\text{obs}}} \sum_{i=1}^{3N} [S_i]_0 \tag{6}$$

$$S_{\text{RW}}^{\text{traj}} = \sum_{t=1}^{t_{\text{obs}}} \sum_{i=1}^{3N} [S_i]_{\text{RW}} \tag{7}$$

Finally, the action difference, ΔS_{traj}, which we will utilize in the reweighting procedure, is defined as

DOI: 10.1021/acs.jctc.5b00031
J. Chem. Theory Comput. 2015, 11, 2412−2420

$$\Delta S_{\text{traj}} = S_0^{\text{traj}} - S_{\text{RW}}^{\text{traj}} \quad (8)$$

Since we will concern ourselves only with action differences in this work, it is sufficient to compute only the Gaussian exponents (i.e., the first terms) in eqs 4 and 5.

To integrate the path action difference into the MSM framework, we observe that transition counts between states α and β, $C_{\alpha\beta}$, are accumulated over trajectories of fixed lag time, τ_{lag}, starting in state α and terminating in state β. The number of times one should expect to observe one such microscopic trajectory in the target potential, relative to the perturbed potential, is proportional to the exponential of that trajectory's interpotential action difference. The reweighted transition count, $[C_{\alpha\beta}]_{\text{RW}}$, is thus given by[11]

$$[C_{\alpha\beta}]_{\text{RW}} = C_{\alpha\beta} \exp(\Delta S_{\alpha\to\beta}) \quad (9)$$

where the raw count takes a unit value, and $\alpha\to\beta$ represents the specific trajectory of length τ_{lag} in question. With an appropriate choice of lag time (based on heuristics like flattening implied time scales), one can analyze a full set of trajectories to systematically build a reweighted counts matrix, which can in turn be transduced into a equilibrium transition probability matrix via maximum likelihood estimation.

In its essence, this reweighting approach allows one to first sample phase space based on one potential energy surface and then view the resulting dynamics through the lens of another potential in a statistically rigorous fashion. For the procedure to be useful from the standpoint of accelerating sampling, it is clear that one must have some *a priori* intuition for importance sampling in the system of interest. In order to maximize overlap between the sampled and reweighted phase spaces, one needs to limit potential energy perturbations to the greatest extent possible, perhaps focusing on exceptionally slow and local degrees of freedom. A rough protocol for utilizing potential-based reweighting might proceed through the following steps:

1. Define a perturbed potential energy surface to accelerate sampling over a particularly slow degree of freedom. Collect trajectories in this perturbed potential to a satisfactory extent, cataloguing action differences between the target and perturbed potentials at each time step.

2. Cluster collected configurations based on a desired metric to define the state space. Referring to eq 9, construct reweighted count matrices at various lag times.

3. Use a maximum likelihood estimator to construct an equilibrium transition matrix and select a reweighted model based on spectral or likelihood-based properties. Analyze highly populated states, relaxation time scales, and mean first passage times and compare with any available simulation or experimental data.

We attempt to apply this protocol to the two systems described below.

Application 1: Two-Well Potential. As a proof of principle, we first reweight the Langevin dynamics of a single particle traversing a perturbed two-dimensional, two-well potential.[14,16] In a contrived fashion, the system's slowest time scale corresponds to barrier crossing events between the two wells, and auxiliary processes (related to intrawell dynamics) occur at orders-of-magnitude faster time scales. To modulate sampling of interwell dynamics, we systematically (and symmetrically) perturb the well depths until the landscape becomes effectively diffusive. We reweight the dynamics in each case, comparing the resulting slowest time scale to that found with an exhaustive, gold-standard sampling on the target potential. Figure 1 illustrates three examples of landscapes we have sampled: 1) the deep target, or "base," potential, 2) the potential which required the least sampling to attain satisfactory reweighting, and 3) an extreme, nearly diffusive potential for which hallmarks of the wells have been erased. All landscapes were defined using two Gaussian wells (which were perturbed in depth) placed inside a shallow, quadratic bowl; integration was performed with a simple BKK integrator.[14]

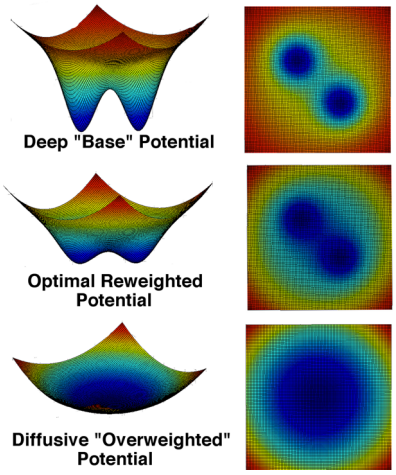


Figure 1. Representative examples of two-well potentials (viewed from the side and from above) sampled under the potential-based reweighting scheme. **At top:** the target, or "base," potential used as a reference for this series of reweighted potentials. Sampling between the two deep wells is slow (requiring approximately 10^8 simulation steps to capture). **Middle:** A potential perturbed such that the two wells are shallow, but the nature of the landscape is still well-defined. From a sampling efficiency perspective, this potential performed the best among the perturbations tested. **Bottom:** A shallow, largely diffusive potential on which hallmarks of the two-well landscape have been washed out. This landscape still offered an enhancement over sampling the deep-welled potential but at a deficit to the potential shown in the middle.

Given the simple nature of the potential energy surface, Markov states were defined in a geometric fashion: an even 20 × 20 grid was overlaid on the landscape, and low population states were trimmed from the network if necessary to ensure ergodicity. Transition counts were determined using a sliding window procedure, and such counts were reweighted according to eq 9. Simple symmetrization was deemed sufficient for the estimation of equilibrium transition probabilities; lag times of approximately 10^3 time steps (in arbitrary time units) proved effective for modeling both raw and reweighted data. Rounds of simulations and model building were repeated until the time scales of interest converged.

For the purposes of comparing sampling efficiencies, all potentials were continuously sampled with trajectories of moderate length (10^4 time steps, initiated from random, already-discovered Markov states) until the slowest reweighted time scale (or raw time scale) reached within a factor of 2 of the gold standard. At this point the sampling efficiency was evaluated through comparison to dynamics on the base potential. The properties of the MSMs built on these simple potentials are detailed in the Results and Discussion section below. Data corresponding to a larger set of potentials (featuring intermediate well depths) are illustrated in Supplementary Figures 1 and 2; specific details and scripts

DOI: 10.1021/acs.jctc.5b00031
J. Chem. Theory Comput. 2015, 11, 2412−2420

used to generate and sample all two-well potentials are included in the Supporting Information (SI).

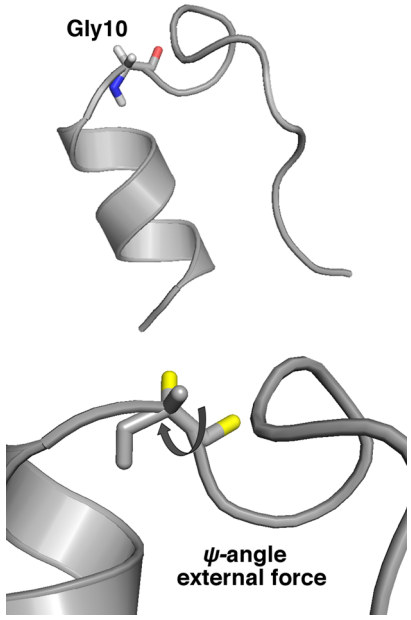


Figure 2. Top: Illustration of the native solution structure for Trp-cage. The highlighted residue (Gly10, near the peak of the hairpin) is targeted for perturbation in our biased dynamics. **Bottom:** A close-up view of Gly10. The external force applied to the system corresponds to a dihedral potential that biases the indicated ψ-angle to an antiparallel configuration (which, in turn, results in the disassembly of the hairpin and exposure of the enclosed tryptophan).

Application 2: Trp-Cage Miniprotein. After validating our protocol on a simple system, we elected to perturb the folding dynamics of the 20-residue miniprotein Trp-cage (PDB: 2JOF).[12] Stabilized by a cage of residues surrounding a central tryptophan, recent simulation studies indicate that Trp-cage folds on time scales of a few microseconds.[17,18] To test our reweighting procedure, we introduce an external dihedral potential on the protein backbone which accelerates the breaking of the native hairpin (and disassembly of the tryptophan cage). We aim to rescale the trajectories obtained from this destabilizing potential in order to recover a reliable description of equilibrium folding behavior.

Figure 2 illustrates the native structure of Trp-cage and demonstrates the nature of the perturbing potential. As indicated in the both panels, the external dihedral force is localized on a glycine near the peak of the hairpin; this modification biases the concomitant backbone ψ-angle toward an antiparallel configuration. The magnitude of the added force was set to a high enough value to make the dihedral perturbation efficacious but not so high that the near-native phase space was seldom revisited. After some manual tuning, the final chosen value for the force constant was 1 $J \cdot mol^{-1} \cdot nm^{-2}$.

The Trp-cage PDB was solvated in a 40 Å cubic box alongside 2082 water molecules, yielding a system containing a total of 6530 particles. Using the GROMACS 4.6.4 suite of software,[19] the system was first energy minimized using a method of steepest descent and next equilibrated with peptide heavy atom restraints at constant temperature and pressure. Starting from an equilibrated frame, production simulations (using the Amber99SB-ILDN force field[20] in concert with the TIP3P water model,[21] under the biased potential) were carried out with the OpenMM 5.2 software package on the OpenCL GPU platform.[22] Fifty separate runs lasting 10 ns each were completed, generating a data set comprised of 500 ns of aggregate trajectories. The external dihedral force was implemented with the CustomTorsionForce class in OpenMM, and action differences were tracked through the use of a custom Verlet-type Langevin integrator.[23] Note that actions are calculated on the fly by the custom integrator at each time step, so that additional configurational and force data need not be stored. Temperatures were constrained to a 298 K target, and the pressure, which was controlled by a Monte Carlo barostat, was set to 1 bar. Representative code written explicitly for these production simulations is included in the SI.

After molecular dynamics data collection was completed, water degrees of freedom were removed, and MSMs were constructed for the raw and reweighted data using the MSMBuilder 2.0 simulation package.[24] Clustering was performed with a simple hybrid k-centers/k-means approach, with cluster radii set to 4 Å. Modified scripts were written to perform the count-based reweighting scheme described by eq 9; transition matrices comprised of 883 states were symmetrized using built-in maximum likelihood estimation functionality. Models were constructed over a range of lag times, and a 7 ns lag time was deemed fit for the state space in question. The modified code used to complete these tasks is once more presented in the SI.

Static and dynamical properties of the MSMs were projected onto RMSDs from the Trp-cage NMR structure and a salt bridge order parameter (the distance between the side chains of Asp9 and Arg16) quoted in the literature.[25] A protein folding MSM provided by Beauchamp et al. was used as a "gold standard" for comparison with reweighted MSM properties.[17,18]

RESULTS AND DISCUSSION

Two-Well Potential. Figure 3 illustrates the relaxation time scales – calculated individually as $\tau_{relax} = -\tau_{lag}/\log(\lambda)$, with λ

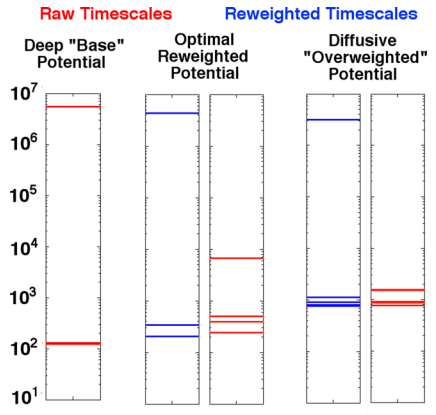


Figure 3. Comparison of MSM relaxation time scales derived from raw data sampled from the target potential (left) and raw and reweighted data taken from the perturbed potentials (right). Raw time scales are shown on the right side for each respective potential and are colored red; reweighted time scales are shown in blue and at left. As the plots indicate, successive biased sampling and reweighting allows for the slowest time scale for the target potential to be recovered with good accuracy. Faster time scales are also recapitulated to reasonable agreement with the unbiased system.

DOI: 10.1021/acs.jctc.5b00031
J. Chem. Theory Comput. 2015, 11, 2412–2420

being a transition matrix eigenvalue − obtained from the deep, best, and diffusive two-well potentials noted previously. The time scales corresponding to the target potential are well behaved: one slow time scale (occurring at $\approx 10^7$ time steps), related to interwell transitions, is distinctly separated from other auxiliary processes. For each of the two perturbed potentials, kinetic properties are presented for MSMs built from both the raw and reweighted trajectory data.

As expected, one observes that equilibration occurs relatively quickly within the raw perturbed potentials: the slowest relaxation time scales fall between 10^3 and 10^4 time steps. This nearly 4 orders of magnitude deficit to the target time scale, however, is easily overcome by the reweighting procedure. Using significantly smaller sets of trajectories, the slowest relaxation time scale of approximately 10^7 time steps was recovered (in the data presented here, within a factor of 2) in both the best-performing and diffuse potentials. While the agreement is not exact between successive faster time scales, the estimates obtained from reweighting are reasonable.

Figure 4 quantifies the gains in sampling efficiency provided by the reweighting approach, where the baseline was obtained

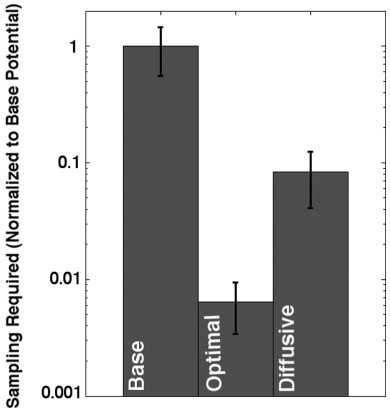


Figure 4. Characterization of relative sampling efficiency observed in dynamics over the potentials illustrated in Figure 1. The target potential was resampled to provide a baseline and minimal estimate of sampling needed to capture dynamics between the deep wells. Error bars are presented as plus or minus one standard deviation from the mean over triplicate simulation runs. The central potential (with shallow, but still-defined wells) provides over 2 orders of magnitude in sampling enhancement; the shallow, nearly diffusive potential offers a less dramatic improvement.

by the alternative sampling of the target potential described in the Methods section. Compared to this alternate sampling, the best-performing potential was able to offer over 2 orders of magnitude of improvement in the amount of aggregate data needed to recover the slowest targeted time scale. The shallow, nearly diffusive potential also yielded a sampling enhancement over the target potential, albeit by only a single order of magnitude.

It is important to reiterate that while phase space, by most measures, would be most rapidly sampled on a (nearly) diffusive landscape, the important regions of phase space for the system of interest − as dictated by statistical mechanics − can simultaneously be undervisited. Provided some meaningful sampling occurs, our reweighting procedure will essentially discard trajectories collected from "irrelevant" regions of phase space, rendering such data uninformative. Notably, the nearly diffusive potential did not outperform other potentials on which the two-well structure is still well-defined. The most drastic potential perturbations will rarely correlate with the best performing modifications, and, to optimize sampling efficiency, one must focus such perturbations on specific slow degrees of freedom. These considerations will be further developed in our analysis of Trp-cage data.

Trp-Cage Miniprotein. Figure 5 provides a illustration of the *ab initio* native state structures of Trp-cage (taken, in

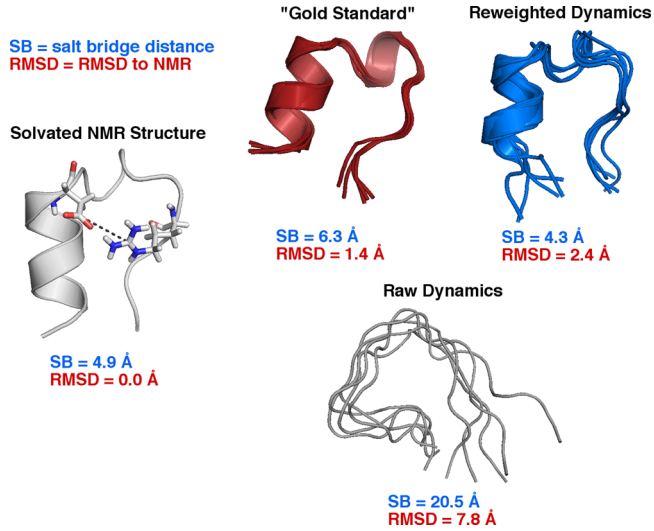


Figure 5. Juxtaposition of *ab initio* native states in raw, reweighted, and gold standard MSMs. Structures from MSMs are compared to the NMR structure at left; the salt bridge distance of interest is indicated with a dashed line on that image. Dynamical reweighting facilitates the recovery of the proper native structure, as the reweighted configurations demonstrate.

respective models, from the Markov state with the highest individual population) derived from MSMs built from raw and reweighted MD data, juxtaposed with the native structure from the gold standard model. The NMR structure (PDB: 2JOF, relaxed for a short while in explicit solvent) to which all native states are compared is also provided for reference, with the Asp9 - Arg16 salt bridge distance highlighted in the image. Examining the raw trajectory data, the external force clearly had an impact on the dynamics: the most populous state in the corresponding "perturbed" MSM is distant from the NMR structure in RMSD and features a widely separated salt-bridge contact. The *ab initio* native state in the reweighted MSM, however, bears a strong resemblance to the NMR reference structure. While the native state derived from reweighting is slightly farther from the NMR structure in RMSD than the gold standard (2.4 Å vs 1.4 Å), the salt bridge contact is actually better aligned with the NMR structure than in the reference MSM (4.3 Å vs 6.3 Å). From the standpoint of recapitulating native state structure, thus, our reweighting procedure appears to have been successful.

We next scrutinize the kinetic properties of the reweighted MSM. Figure 6 displays the five slowest MSM relaxation time scales obtained from the raw and reweighted trajectories, compared with the relaxation time scales reported by Beauchamp et al. The time scales derived from the raw data are far faster than the reference time scales, as the perturbation largely pushes trajectories into the unfolded state. Agreement between the slowest reweighted time scale and that reported by Beauchamp et al., though, is exceptionally good: both estimates

DOI: 10.1021/acs.jctc.5b00031
J. Chem. Theory Comput. 2015, 11, 2412−2420

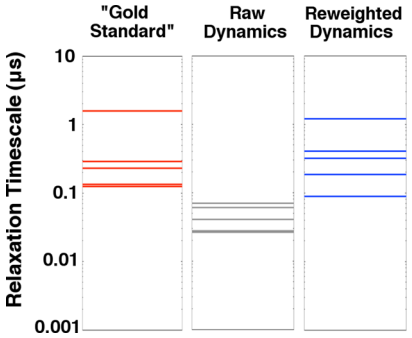


Figure 6. Comparison of slow relaxation time scales obtained from (left) reference, (middle) raw, and (right) reweighted MSMs. The reweighted time scales agree well with those reported by Beauchamp et al.

sit just above one microsecond. As an added bonus, the next four time scales appear in the reweighted data at their expected order of magnitude. Some degree of luck is involved in attaining this extent of congruity after reweighting, as the slowest time scale fluctuates by about 20% when marginally shorter/longer lag times are applied. Nonetheless, the fact that raw time scales are slowed by almost 2 orders of magnitude to come very near to the reference estimate is encouraging.

Though the reweighted time scales appear to be consistent with expectations, one needs to confirm that such time scales correspond to dynamical processes similar to those observed in the reference model. From a simple RMSD-based metric, we only expect to capture the slowest dynamical processes to any accuracy; we thus focus our analysis on the first few eigenvectors of each transition matrix. Figures 7, 8, and 9 show projections of state populations and the models' slowest dynamical eigenvectors onto two order parameters of interest: the RMSD to the NMR structure and the Asp9 - Arg16 salt bridge distance. The dynamical eigenvectors highlight the regions of state space involved in each observed slow process; positive and negative components connote exchange between different basins of population.

The raw population data in Figures 7 and 8 illustrate how the perturbation pushes the system into its unfolded ensemble: equilibrium probability density is largely absent at near-native RMSDs and salt-bridge distances in the raw data, and the raw dynamics occur almost wholly within the unfolded state. The static behavior within the reweighted ensemble, however, is much more in line with that extracted from the reference MSM. The probability peak corresponding to the native state (near 1.5 Å RMSD in the gold standard) is slightly shifted in the reweighted data, as expected from Figure 5; the near-native region is also less populated in the reweighted ensemble by about a factor of 3. From the logarithmic perspective of free energies, such a deviation is not too significant — especially considering the far lesser extent of data used in building the reweighted model (about 500 ns of biased simulation data vs several milliseconds of equilibrium simulation data for the gold standard). In general, the population distribution exhibits more heterogeneity and breadth for the reweighted data; these characteristics are also expected statistically, as sampling of the unfolded basin in Anton simulations was exhaustive. The reweighted data captures the morphology of the distribution over both order parameters quite well and reproduces absolute probabilities to reasonable agreement.

The slowest dynamical eigenvector features a similar amount of unfolded-state heterogeneity in the reweighted case. The analogous mode in the reference model clearly corresponds to a "folding" process, representing density transfer from the broad unfolded basin into near-native states. The reweighted eigenvector encompasses part of a similar folding process, as

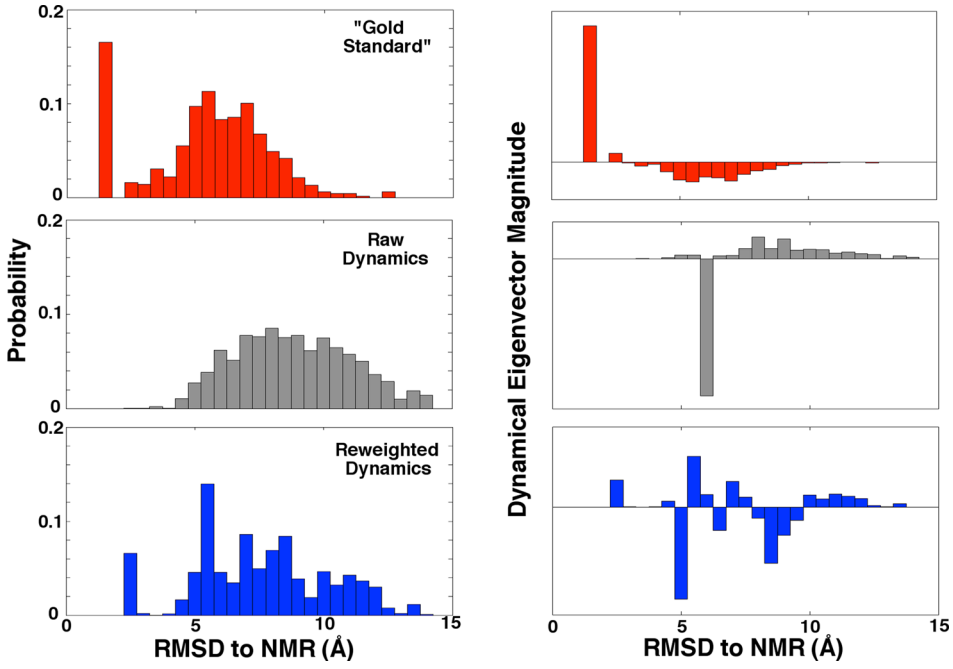


Figure 7. RMSD order parameter projections of (top) gold standard, (middle) raw, and (bottom) reweighted MSM properties corresponding to (left) equilibrium state populations and (right) the slowest dynamical eigenvector observed in each respective model. As all dynamical eigenvectors are scaled according to a standard vector norm, only relative magnitudes of components are shown. Distributions over RMSD agree to a reasonable extent between reweighted and gold standard models.

DOI: 10.1021/acs.jctc.5b00031
J. Chem. Theory Comput. 2015, 11, 2412−2420

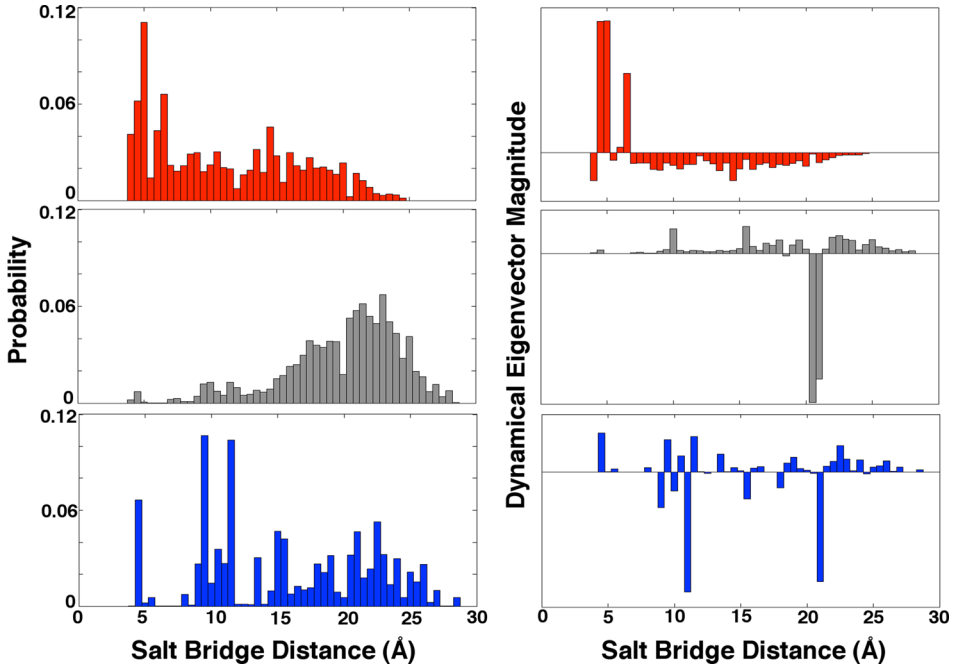


Figure 8. Salt bridge order parameter projections of (top) gold standard, (middle) raw, and (bottom) reweighted MSM properties corresponding to (left) equilibrium state populations and (right) the slowest dynamical eigenvector observed in each respective model. Once more, only relative magnitudes of dynamical eigenvector components are shown. Distributions over the salt bridge distance also demonstrate reasonable agreement between reweighted and gold standard models.

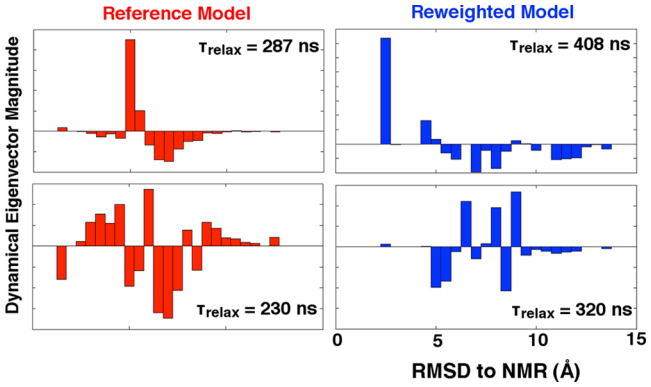


Figure 9. Comparison of the second- and third-slowest eigenvectors for the (left) reference and (right) reweighted models, projected onto RMSD from the NMR structure. Once more, eigenvector inputs are subjected to standard vector normalization.

exemplified by the positive peak that appears in the near-native region, but marked equilibration also occurs within the unfolded state at the same time scale. However, analysis of subsequent eigenvectors – shown in Figure 9 – reveals that the second mode of the reweighted model (with $\tau_{relax} \approx 408$ ns) describes a more canonical folding mechanism and bears a striking resemblance to the slowest reference eigenvector. In a complementary fashion, the second reference eigenvector features unfolded state equilibration similar to that seen in the slowest reweighted mode. Considered in linear combination, thus, the slowest two eigenvectors in each model yield a similar picture of relaxation in the unfolded state and folding dynamics that occur on a near-microsecond time scale. The third-slowest eigenvectors (included at the bottom of Figure 9) share a few common features within the unfolded basin but differ with respect to others. Direct agreement among the faster modes of the reference and reweighted models is not expected, given the associated density increase within and variance between the models' respective eigenvalue spectra. Nonetheless, we can say with some confidence that the slowest reweighted time scales presented in Figure 6 do, in fact, combine to create a dynamical processes similar to that produced by the reference MSM.

It is again important to emphasize that the above reweighting approach does not constitute a "magic bullet" for generating an informative distribution over equilibrium kinetics. Action-based reweighting yields the equilibrium probabilities for trajectories sampled from a biased potential. If such biased trajectories cover little of the phase space relevant to the equilibrium picture, reweighted kinetics will likely not describe the particular equilibrium process of interest to the researcher. As discussed below, selecting a reasonable biasing potential (as the dihedral bias in the present Trp-cage example seems to be) is thus of utmost importance for deriving useful results from reweighted data.

One should also note that evaluation of the custom integrator used here is approximately 5× slower than a standard Langevin integrator under the OpenMM platform. The resulting deficit in integration speed subtracts from estimates of improved sampling efficiency: one could collect almost three microseconds of MD trajectories in the time needed to generate 500 ns of data in this work. However, as one requires much more equilibrium data (say, conservatively, 50 μs) to comprehensively characterize the folding of Trp-cage, the present reweighting method still offers an ≈20× speedup over conventional sampling. The notion that near-equilibrium dynamics can be recovered from such a small, biased data set speaks to the promise of the potential-based reweighting approach. A dedicated integrator code that carries out all reweighting calculations on a GPU would likely enhance the sampling efficiency to an even greater extent.

DOI: 10.1021/acs.jctc.5b00031
J. Chem. Theory Comput. 2015, 11, 2412–2420

Limitations and Future Applications. The above results illuminate two auspicious applications for which potential-based reweighting can provide significant gains in sampling efficiency. However, as we have mentioned previously, this reweighting approach does not come without possible trade-offs in efficiency and efficacy with standard sampling methods. As an illustrative example from the perspective of protein folding, one might consider diminishing native dihedral potentials along a protein's backbone to buffet the rate of conformational sampling. Such a perturbation, no doubt, would allow one to explore a wide swath of the polymer's phase space in an expeditious fashion. However, in applying this modification, one has effectively constructed a realization of Levinthal's paradox:[26,27] without the guidance of a rugged folding free energy landscape that has been meticulously optimized by evolution, the neighborhood of the designed native state will (almost) never be visited. From a reweighting standpoint, data collected at the most distal phase points will be decimated, and little useful information will result.

Even in cases like that described here, wherein only a single dihedral torsion is perturbed, one must take note of the reduced volume of phase space that is sampled due to the external potential's bias. In protein systems that exhibit simple, "two-state"-like folding kinetics (like Trp-cage), it is perhaps sufficient to gently promote unfolding in a generic fashion and reweight the resulting dynamics to recover equilibrium properties. For Trp-cage, this approach yielded good, though imperfect, agreement with folding kinetics obtained from standard means (even though a relatively small fraction of the unfolded basin was actually sampled in the biased ensemble). No guarantee of such easy success, however, can be offered in systems which exhibit many slow and competing dynamical processes. Under such circumstances, one would need to carefully design a perturbing potential which allows the system to sample various unfolded and intermediate basins important to the folding process.

The most promising candidate systems for future applications are those for which one can define a simple and local perturbation that will radically accelerate sampling of interesting processes. The protein BPTI represents an immediate target for this approach, as a particularly slow degree of freedom has been characterized in its folding dynamics.[28] Other proteins with similar properties could no doubt be identified. Even more appropriate applications for reweighting, however, might be found under conditions in which protein dynamics are less dramatic (and for which, consequently, phase space sampling constraints are less severe). In a variety of signaling proteins, for instance, it is known that the unfolding of and movement within short helical motifs is responsible for steps taken toward activation.[29−33] The dynamics of such motifs could easily be biased and reweighted to yield an equilibrium description of related activation processes. A further possible application regards the binding of small molecules to their protein targets: the long times for which small molecules remain tightly fixed to their binding pockets often prohibit accurate determination of dissociation and other binding constants via direct MD simulations.[34,35] Under the framework presented here, one might weaken the interactions of a molecule with its binding site and reweight the observed dissociation kinetics to calculate accurate off-rates for binding. The results of such an analysis could be useful to ventures in drug discovery, not to mention other phenomena in biochemical systems.

Given the inherent difficulties in defining a proper biasing potential from which to reweight, it may be helpful in the future to integrate an action-based reweighting protocol into an existing biased molecular dynamics simulation method. Accelerated molecular dynamics (AMD), for example, implements potential energy biases on the fly and in a manner complementary to the underlying potential energy surface.[36] Computing action differences over the course of AMD's perturbed trajectories should be relatively straightforward (albeit slightly more involved than in the present case, given the dynamic nature of the AMD potential), allowing one to estimate equilibrium kinetics without the direct introduction of custom forces on the part of the user.

CONCLUSION

In this Article, we have derived and evaluated a method by which dynamics can be sampled from a biased potential and rescaled to reflect the properties of the unbiased system. For the systems analyzed here, this potential-based dynamical reweighting allowed for marked gains in sampling efficiency to be attained with little cost to fidelity to equilibrium expectation values. The utility of this approach for its stated applications, as with all enhanced sampling methods, will be tempered by one's savvy directed toward defining a reduced ensemble of states within which external biases are intuitive and resulting trajectories are informative to equilibrium kinetics. After all, the most accurate (though, perhaps, least efficient) means of deriving equilibrium properties will always involve sampling a system's true potential to the exhaustive limit. The increasing impracticality of ergodic sampling in larger and larger systems, however, will serve to nullify hard critiques against biased sampling techniques; in fact, such techniques will likely come to the forefront as we simulate the biological cell and beyond.

ASSOCIATED CONTENT

Supporting Information

Details and code associated with reweighting of 2D potentials; code needed to produce reweighted MSMs of protein dynamics. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00031.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pande@stanford.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the NSF (MCB-0954714) and NIH (R01-GM062868) for their support of this work. J.K.W. was supported by the Fannie and John Hertz Foundation on the endowed Professor Yaser S. Abu-Mostafa Fellowship.

REFERENCES

(1) Tomita, M. *Trends Biotechnol.* **2001**, *19*, 205−210.
(2) 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*, 174105.
(3) Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods* **2010**, *52*, 99−105.
(4) McGibbon, R. T.; Schwantes, C. R.; Pande, V. S. *J. Phys. Chem. B* **2014**, *118*, 6475−6481.

DOI: 10.1021/acs.jctc.5b00031
J. Chem. Theory Comput. 2015, 11, 2412−2420

(5) Lecomte, V.; Appert-Rolland, C.; Van Wijland, F. *J. Stat. Phys.* **2007**, *127*, 51−106.
(6) Weber, J. K.; Pande, V. S. *Biophys. J.* **2012**, *102*, 859−867.
(7) Weber, J. K.; Jack, R. L.; Pande, V. S. *J. Am. Chem. Soc.* **2013**, *135*, 5501−5504.
(8) Weber, J. K.; Pande, V. S. *Phys. Rev. E* **2015**, *91*, 032136.
(9) Beauchamp, K. A.; Pande, V. S.; Das, R. *Biophys. J.* **2014**, *106*, 1381−1390.
(10) Chodera, J. D.; Swope, W. C.; Noé, F.; Prinz, J.-H.; Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2011**, *134*, 244107.
(11) Prinz, J.-H.; Chodera, J. D.; Pande, V. S.; Swope, W. C.; Smith, J. C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 244108.
(12) Barua, B.; Lin, J. C.; Williams, V. D.; Kummler, P.; Neidigh, J. W.; Andersen, N. H. *Protein Eng. Des. Sel.* **2008**, *21*, 171−185.
(13) Coffey, W. T.; Waldron, J.; Kalmykov, Y. P. *The Langevin Equation*; World Scientific:1996; pp 60−66.
(14) Brünger, A.; Brooks, C. L., III; Karplus, M. *Chem. Phys. Lett.* **1984**, *105*, 495−500.
(15) Zwanzig, R. *J. Stat. Phys.* **1973**, *9*, 215−220.
(16) Izaguirre, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D. *J. Chem. Phys.* **2001**, *114*, 2090−2098.
(17) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517−520.
(18) Beauchamp, K. A.; McGibbon, R.; Lin, Y.-S.; Pande, V. S. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17807−17813.
(19) 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**, btt055.
(20) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins* **2010**, *78*, 1950−1958.
(21) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(22) Eastman, P.; Pande, V. *Comput. Sci. Eng.* **2010**, *12*, 34−39.
(23) Paterlini, M. G.; Ferguson, D. M. *Chem. Phys.* **1998**, *236*, 243−252.
(24) Beauchamp, K.; Bowman, G.; Lane, T.; Maibaum, L.; Haque, I.; Pande, V. J. *Chem. Theory Comput.* **2011**, *7*, 3412−3419.
(25) Juraszek, J.; Bolhuis, P. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 15859−15864.
(26) Levinthal, C. *J. Chim. Phys.* **1968**, *65*, 44−45.
(27) Zwanzig, R.; Szabo, A.; Bagchi, B. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 20−22.
(28) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. *Science* **2010**, *330*, 341−346.
(29) Adams, J. *Chem. Rev.* **2001**, *101*, 2271−2290.
(30) Taylor, S.; Kornev, A. *Trends Biochem. Sci.* **2011**, *36*, 65−77.
(31) Shukla, D.; Meng, Y.; Roux, B.; Pande, V. S. *Nat. Commun.* **2014**, *5*, 3397.
(32) Weis, W. I.; Kobilka, B. K. *Curr. Opin. Struct. Biol.* **2008**, *18*, 734−740.
(33) Kohlhoff, K.; Shukla, D.; Lawrenz, M.; Bowman, G.; Konerding, D. E.; Belov, D.; Altman, R. B.; Pande, V. S. *Nat. Chem.* **2014**, *6*, 15−21.
(34) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. *J. Am. Chem. Soc.* **2011**, *133*, 9181−9183.
(35) Dror, R. O.; Pan, A. C.; Arlow, D. H.; Borhani, D. W.; Maragakis, P.; Shan, Y.; Xu, H.; Shaw, D. E. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 13118−13123.
(36) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919−11929.

DOI: 10.1021/acs.jctc.5b00031
J. Chem. Theory Comput. 2015, 11, 2412−2420