JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Simultaneous Computation of Dynamical and Equilibrium Information Using a Weighted Ensemble of Trajectories

Ernesto Suárez,[†,∥] Steven Lettieri,[†,∥] Matthew C. Zwier,[‡] Carsen A. Stringer,[§] Sundar Raman Subramanian,[†] Lillian T. Chong,[‡] and Daniel M. Zuckerman*[,†]

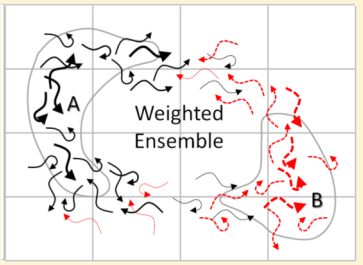
[†]Department of Computational and Systems Biology, University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, Pennsylvania 15260, United States

[‡]Department of Chemistry, University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, Pennsylvania 15260, United States

[§]Gatsby Computational Neuroscience Unit, University College London, Gower St, London WC1E 6BT, United Kingdom

Supporting Information

ABSTRACT: Equilibrium formally can be represented as an ensemble of uncoupled systems undergoing unbiased dynamics in which detailed balance is maintained. Many nonequilibrium processes can be described by suitable subsets of the equilibrium ensemble. Here, we employ the "weighted ensemble" (WE) simulation protocol [Huber and Kim, *Biophys. J.* **1996**, *70*, 97−110] to generate equilibrium trajectory ensembles and extract nonequilibrium subsets for computing kinetic quantities. States do not need to be chosen in advance. The procedure formally allows estimation of kinetic rates between arbitrary states chosen after the simulation, along with their equilibrium populations. We also describe a related history-dependent matrix procedure for estimating equilibrium and nonequilibrium observables when phase space has been divided into arbitrary non-Markovian regions, whether in WE or ordinary simulation. In this proof-of-principle study, these methods are successfully applied and validated on two molecular systems: explicitly solvated methane association and the implicitly solvated Ala4 peptide. We comment on challenges remaining in WE calculations.



1. INTRODUCTION

Although it is textbook knowledge that the functions of biomacromolecules are strongly coupled to their conformational motions and fluctuations,[1] computer simulation of such motions has been a challenge for decades.[2] Typically, distinct algorithms are employed to estimate equilibrium quantities (e.g., refs 3 and 4) and dynamical properties (e.g., refs 5−10). In principle, a single long dynamics trajectory would be sufficient to determine both equilibrium and dynamical properties,[11] but such simulations remain impractical for most systems of interest.

Aside from straightforward simulations, more technical approaches that can yield both equilibrium and dynamical simulation, sometimes under minor assumptions, have drawn increasing attention. A number of approaches employ Markov state models (MSMs) as part their overall computational strategy. On the basis of replica exchange molecular dynamics (REMD),[12,13] it is possible to extract kinetic information from continuous trajectory segments between exchanges and thereby construct an MSM.[13] The adaptive seeding method (ASM) similarly builds an MSM based on trajectories seeded from states discovered via REMD or another of the so-called generalized ensemble (GE) algorithms.[14] MSMs have also been used in combination with short, off-equilibrium simulations to construct the equilibrium ensemble of folding pathways of a protein.[15]

Another general strategy is to employ a series of non-intersecting interfaces that interpolate between states of interest selected in advance. Milestoning generates and analyzes transitions between interfaces assuming prior history does not affect the distribution of trajectories.[16,17] Transition interface sampling (TIS)[18,19] and its variants also analyze such transitions and can yield free energy barriers in addition to rates while accounting for some history information.[20] Forward flux sampling (FFS) again samples interface transitions: it accounts for history information and can yield rates and equilibrium information.[7,21]

The "weighted ensemble" (WE) simulation strategy[5] (see Figure 1), which has a rigorous basis as a path-sampling method,[22] has also been suggested as an approach for computation of both equilibrium and nonequilibrium properties.[23,24] Although WE was originally developed as a tool for characterizing nonequilibrium dynamical pathways and rates (e.g., refs 5, 25−28), the strategy was extended to steady-state conditions including equilibrium.[23] The simultaneous computation of equilibrium and kinetic properties using WE was demonstrated with configuration space separated into two states by a dividing surface[24] and later for arbitrary states

Special Issue: Free Energy Calculations: Three Decades of Adventure in Chemistry and Biophysics

Received: December 9, 2013
Published: March 3, 2014

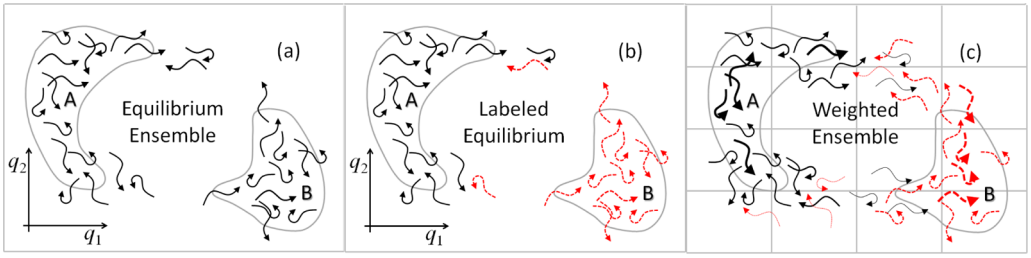


Figure 1. Equilibrium in different representations. (a) Ensemble of trajectories with arrow tips indicating the instantaneous configuration and tails showing recent history in the space of two schematic coordinates q_1 and q_2. States A and B, shown in gray, are two arbitrary regions of phase space. (b) Dissection into two subsets based on whether a trajectory was most recently in state A (black solid arrows, the "α" steady state) or state B (red dashed, the "β" steady state). (c) Statistically equivalent ensemble of weighted trajectories, with arrow thickness suggesting weight. Configuration space has been divided into cells ("bins") which each containing an equal number of trajectories.

defined in advance of a simulation.[29] In contrast to many other advanced sampling strategies, WE generates an ensemble of continuous trajectories, all at the physical condition (e.g., temperature) of interest.

Here, we further develop the capability of WE simulation to calculate equilibrium and nonequilibrium quantities simultaneously in several ways that may be important for future studies of increasingly complex systems. (i) The approach described below permits the calculation of rates between arbitrary states, which can be defined *after* a simulation has been completed. In a complex system, the most important physical states, including intermediates, generally will not be obvious prior to simulation. Further, the present approach opens up the possibility to use rate calculations to aid in the state-definition process. (ii) The non-Markovian analysis described here enables unbiased rate calculations in the typical case where "bins" used by WE simulation do not exhibit Markovian behavior. The analysis is general and can be applied outside the WE context, including the analysis of ordinary long trajectories. (iii) The non-Markovian analysis can improve the efficiency of WE simulations by yielding accurate estimates of observables from shorter simulations. The analysis is based on a previously suggested decomposition of the equilibrium ensemble into two nonequilibrium steady states.[9,20,21,30,31]

Generally speaking, WE provides an attractive basis for complex simulations. WE is easily parallelizable because it employs multiple trajectories and was recently used with 3500 cores.[32] Because there is no need to "catch" trajectories at precise transition interfaces, WE algorithms lend themselves to a scripting-like implementation which has been employed to study a wide range of stochastic systems via regular molecular dynamics,[28] Monte Carlo,[26] the string strategy,[33] and Gillespie-algorithm dynamics of chemical kinetic networks.[34]

2. THEORETICAL FORMULATION

WE simulation uses multiple simultaneous trajectories, with weights that sum to one, that are occasionally coupled by replication or combination events every τ units of time.[5] The coupling events typically are governed by a static partition of configuration space into "bins" (Figure 1c), although dynamical/adaptive bins may be used.[22] In the case of static bins, when one or more trajectories enters an unoccupied bin, those trajectories are replicated so that their count conforms to a (typically) preset value, M. Replicated "daughter" trajectories inherit equal shares of the parent's weight. If more than M trajectories are found to occupy a bin, trajectories are combined statistically in a pairwise fashion until M remain, with weight from pruned trajectories assigned to others in the same bin. These procedures are carried out in such a way that dynamics remain statistically unbiased.[22] This study does not adjust weights according to previously developed reweighting procedures[23] during the simulation. Rather, the WE simulations described here are long enough to permit relaxation to the equilibrium state.

2.1. Direct Calculation of Observables. Once the equilibrium state is reached in a WE simulation, meaning that there is a detailed balance of probability flow between any two states, equilibrium observables such as state populations or a potential of mean force can be calculated simply by summing trajectory weights in the corresponding regions of phase space. We term this "direct" estimation of observables.

To calculate rates, the equilibrium set of trajectories (Figure 1a) is decomposed into two steady states as shown in Figure 1b: the α steady state consisting of trajectories more recently in A than B, and the β steady state with those most recently in B;[9,31] these were denoted "AB" and "BA" steady states, respectively, in ref 31. Trajectories are "labeled" according to the last state visited, i.e., classified as α or β, during a WE simulation or in a postsimulation analysis ("post-analysis"). The *direct* rate k_{AB} estimate is computed from the probability arriving at the final state[4,7,9,20,23,35] via

$$k_{AB} = \frac{1}{\mathrm{MFPT}(A \rightarrow B)} = \frac{\mathrm{Flux}(A \rightarrow B|\alpha)}{p(\alpha)} \quad (1)$$

where MFPT is the mean-first-passage time, $\mathrm{Flux}(A \rightarrow B|\alpha)$ is the probability per unit time arriving at state B in the α steady state, and $p(\alpha)$ is the total probability in the α steady state. By construction $p(\alpha) + p(\beta) = 1$. Normalizing by $p(\alpha)$ effectively excludes the reverse steady state, and the rate calculation only "sees" the unidirectional α steady state as in ref 23. An expression analogous to eq 1 applies for k_{BA}. Also note that the effective first order rate constant, defined by $\mathrm{Flux}(A \rightarrow B|\alpha)/p_A^{eq}$, can be determined from equilibrium WE simulation because P_A^{eq} can be directly computed by summing weights in A.

We note that analogous direct calculation of observables can be performed from an equilibrium ensemble of unweighted (i.e., "brute force") trajectories by assigning equal weights to each.

2.2. Non-Markovian Matrix Calculation of Observables. Beyond the direct estimates of observables based on trajectory weights, we also generalize previous matrix formulations for nonequilibrium steady states[9,30,36] into an equilibrium formulation that explicitly accounts for the embedded steady states (as in Figure 1b,c). These non-Markovian matrix estimates are tested below and may prove

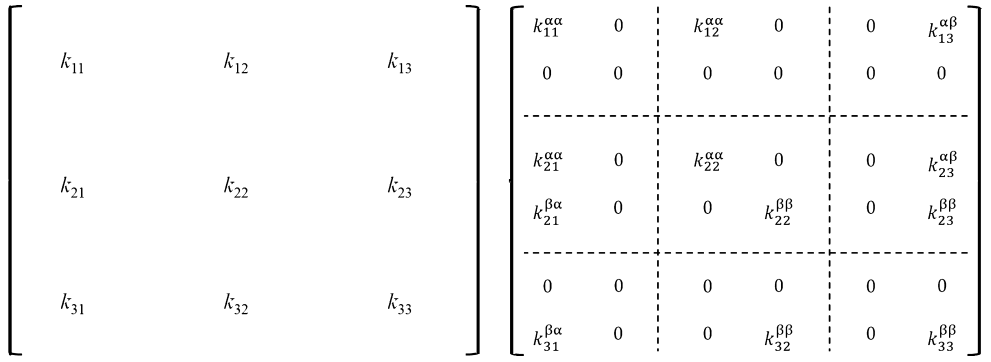


Figure 2. Constructing a labeled rate matrix for unbiased calculations. For purposes of illustration, here state A consists solely of bin 1 and state B solely of bin 3. Left: A traditional rate matrix with history-blind elements. The rate k_{ij} gives the conditional probability for transitioning from bin i to bin j in a fixed time increment, regardless of previous history. Right: The labeled rate matrix accounting for history. The element $k_{ij}^{\mu\nu}$ is the conditional probability for the i to j transition for trajectories initially in the μ subensemble which transition to the ν subensemble, where μ and ν are either α or β. The labeled rate matrix correctly assigns the α and β subpopulations of each bin, whereas the traditional matrix may not.

important for future WE studies using shorter simulations, as described in the Discussion.

Our matrix approach explicitly uses the decomposition of the equilibrium population into α and β components for each bin i:

$$p_i^{\mathrm{eq}} = p_i^{\alpha} + p_i^{\beta} \tag{2}$$

which implies $p(\alpha) = \sum_i p_i^{\alpha}$ and $p(\beta) = \sum_i p_i^{\beta}$. We called this a "labeled" analysis. Thus, with N bins, a set of $2N$ probabilities is required rather than N. Similarly, a $2N \times 2N$ rate matrix is required: $k_{ij}^{\mu\nu}$, where μ and ν can be either the α or β subsets of trajectories. See Figure 2. Each of the previously considered k_{ij} rate elements is thus decomposed into four history-dependent elements which account for whether the particular trajectory was last in state A or B and whether the trajectory transitions between the α and β subsets. The analysis assumes states consist strictly of one or more bins, but this is always possible in a post-analysis without a loss of generality. In other words, given the flexibility we have when we define the bins, it is not a real limitation that the states have to be strictly constituted by bins.

We wish to emphasize that this analysis is "non-Markovian" because we are explicitly including history information (i.e., α and β labels) in the new $2N \times 2N$ rate matrix. Once the matrix is built, the steady state observables are obtained using the same mathematical formalism that would be used in a regular Markov model. However, the matrix should be seen as a tool of linear algebra and not as embodying any physical assumptions.

Note that more than half the $k_{ij}^{\mu\nu}$ elements are zero. For example, consider a bin in the "intermediate" region (neither A nor B), such as bin 2 in Figure 2. In this region, an α trajectory cannot change into a β trajectory, nor vice versa; hence rates for these processes are zero. Similarly, an α trajectory in the intermediate region which enters a bin in B must turn into a β trajectory, so the rate will always be zero to the α components of bins in B.

The non-Markovian results below stem from the division into α and β steady states, but several steps are required.

First, rates among bins are estimated in a post-analysis as

$$k_{ij}^{\mu\nu} = \frac{\langle \omega_{ij}^{\mu\nu} \rangle_2}{\langle \omega_i^{\mu} \rangle} \tag{3}$$

where $\omega_{ij}^{\mu\nu}$ is the probability flux, for a given iteration, from bin i to j of trajectories only with initial and final "labels" μ and ν, respectively, while ω_i^{μ} is the population labeled as μ which is initially in i. The subscript "2" in the numerator indicates that the rate $k_{ij}^{\mu\nu}$ is estimated to be nonzero only when more than one transition is observed; after the second event, all events are included, from the first one, to avoid bias. The requirement for two transitions was found to greatly enhance numerical stability in estimating fluxes and rates between macroscopic states: rates estimated from single events exhibit large fluctuations.

Notice that eq 3 is a *ratio of averages* and differs from the average ratio $\langle \omega_{ij}^{\mu\nu}/\omega_i^{\mu} \rangle$, which might seem equally or more "natural." However, our data show that eq 3 yields unbiased estimates, while the average ratio may not (data not shown). The difference between the two estimators indicates that transitions are correlated with trajectory weights. Perhaps more importantly, the average ratio places less importance on high weight transitions due to the instantaneous normalization—and so, in a time-average sense, may be incorrect. That is, low-weight transitions count as heavily as high-weight events, which evidently biases the rate estimate. In the ratio of averages, high-weight events appropriately count more.

To obtain "macroscopic" rates between states consisting of arbitrary sets of bins (noting that arbitrary bins can be employed in a post-analysis), we calculate "labeled" fluxes for use in eq 1 via

$$\begin{aligned} \mathrm{Flux}(\mathrm{A} \rightarrow \mathrm{B}|\alpha) &= \sum_{i,j} p_i^{\alpha} k_{ij}^{\alpha\beta} \\ \mathrm{Flux}(\mathrm{B} \rightarrow \mathrm{A}|\beta) &= \sum_{i,j} p_i^{\beta} k_{ij}^{\beta\alpha} \end{aligned} \tag{4}$$

The labeled bin populations p_i^{α} and p_i^{β} are obtained from the steady-state solution of the labeled rate matrix $K = \{k_{ij}^{\mu\nu}\}$.

A summary of the "labeled" or non-Markovian matrix procedure for estimating rates between arbitrary states is as follows. First, we obtain the labeled rate matrix $K = \{k_{ij}^{\mu\nu}\}$ using eq 3 to average interbin transitions. Second, we solve the matrix problem $K^T p_{\mathrm{SS}} = p_{\mathrm{SS}}$, yielding the steady state solution p_{SS}. Notice that the *equilibrium* bin populations can be computed by eq 2. Then, the steady state solution p_{SS} along with the labeled rate matrix elements are used to calculate the α flux entering state B and the β flux entering A (eq 4). Finally, the MFPT values are obtained from eq 1. In the graphs below, each non-Markovian estimate shown is from the matrix solution using the $k_{ij}^{\mu\nu}$ rates calculated based on all data obtained until the given iteration of the simulation.

The non-Markovian matrix formulation exhibits a number of desirable properties: (i) Unlike with unlabeled (i.e., implicitly Markovian) analysis, kinetic properties will be unbiased as shown below. (ii) Solution of both the α and β steady states is performed simultaneously via a standard Markov-state-like analysis of the $k_{ij}^{\mu\nu}$ rate matrix. By contrast, if the α and β steady states are independently solved within a Markov formalism, there can be substantial ambiguity in how to assign feedback from the target to initial state when the initial state consists of more than one bin. (iii) The labeled formulation guarantees, by construction, the flux balance intrinsic to equilibrium, namely, $\mathrm{Flux}(\mathrm{A} \rightarrow \mathrm{B}|\alpha) = \mathrm{Flux}(\mathrm{B} \rightarrow \mathrm{A}|\beta)$. (iv) The analysis can be performed using arbitrary bins (and states defined as sets of these bins). It is not necessary to employ the bins originally used to run the WE simulation because a post-analysis can calculate rates among any regions of configuration space. (v) The analysis is equally applicable to ordinary brute-force simulations.

2.3. Markovian Matrix Calculation of Observables. For reference, we also perform a traditional Markov analysis of the trajectories, which will prove to yield biased rate estimates because most divisions of configuration space (e.g., WE bins) are not true Markovian states.

The Markov analysis proceeds *without* labeling the trajectories. Elements of the rate matrix are estimated as

$$k_{ij} = \langle w_{ij} \rangle_2 / \langle w_i \rangle \tag{5}$$

where the subscript "2" again means that we only estimate a rate as nonzero once at least two transitions from i to j have occurred. Bin populations are then computed by solving for the steady-state solution of the Markov matrix with elements k_{ij}.

The computation of an MFPT requires the use of source (A) and sink (B) states. This task is automatically performed within the labeled formalism previously described. Hence, we determine Markovian macroscopic rates by substituting the Markovian k_{ij} for all nonzero elements of the $k_{ij}^{\mu\nu}$. We emphasize that this is merely an accounting trick to establish sources and sinks and simultaneously measure both A-to-B and B-to-A fluxes/rates.

We perform a smoothing operation on the macroscopic Markovian rates because otherwise the data are fairly noisy. The MFPT results shown for the Markovian matrix analysis are running averages based on the last 50% of the estimates (where each estimate is from the matrix solution using k_{ij} estimates from all data obtained until the particular iteration). We confirmed numerically that such smoothing did not contribute bias to any of the MFPT estimates.

3. MODEL SYSTEMS AND SIMULATION DETAILS

Weighted ensemble simulations were performed on two systems: the alanine tetrapeptide (Ala4) solvated implicitly and a pair of explicitly solvated methane molecules. All simulations were performed at 300 K with a stochastic thermostat (Langevin thermostat). Friction constants of 5.0 and 1.0 ps^{-1} were used for Ala4 and methane systems, respectively. The molecular dynamics time step used for all systems was Δt = 2 fs. An iteration is defined to be the simultaneous propagation of all trajectories in the ensemble for some amount of time, τ. In these studies, a value of $\tau = 2500\Delta t$ is used for Ala4 and $\tau = 250\Delta t$ for the methane−methane system.

For Ala4, the all-atom AMBER ff99SB force field[37] with implicit GB/SA solvent and no cutoff for the evaluation of nonbonded interactions was simulated using the AMBER 11 software package.[38] The Hawkins, Cramer, and Truhlar[39,40] pairwise generalized Born model is used, with parameters described by Tsui and Case[41] (option igb=1 in AMBER 11 input file). The progress coordinates were selected and "binned" using a 10 × 10 partition of a 2D space. A dihedral distance $D = ((1/N)\sum_i d_i^2)^{1/2} \in [0,180]$ with respect to a reference set of torsions is used in the first dimension, where N is the number of torsional angles considered and d_i is the circular distance between the current value of the ith angle and our reference, i.e., the smaller of the two arclengths along the circumference. This dimension was divided every 14° from 0 to 126° and then a final partition covering the space (126,180]). In the second dimension, a regular RMSD, using only heavy atoms, is measured with respect to an α-helical structure. In this case, the space was divided every 0.4 Å from 0 to 3.6 Å and then a final partition covering the space [3.6,∞). Values and coordinates for the references used to compute the order parameters are given in the Supporting Information (SI).

The methane molecules were simulated using the GROMACS 4.5 software package[42] with the united-atom GROMOS 45a3 force field[43] and dodecahedral periodic box of TIP3P water molecules[44] (about 900 water molecules in a 34 × 34 × 24 Å box). van der Waals interactions were switched off smoothly between 8 and 9 Å; real-space electrostatic interactions were truncated at 10 Å. Long range electrostatic interactions were calculated using particle mesh Ewald (PME) summation. The single progress coordinate was the distance r between the two methane molecules, following ref 28. The coordinate $r \in [0,\infty)$ Å was partitioned with a bin spacing of 1 Å from 0 to 16 Å and a last bin covering the space $r \in [16,\infty)$ Å.

For the post analysis of methane, different bins were used to demonstrate the flexibility of the approach. The coordinate $r \in [0,\infty)$ Å was partitioned so that the first bin is the space $r \in [0,5)$ Å, then a bin spacing of 2 Å was used from 5 to 17, while the last bin covers the space $r \in [17,\infty)$ Å.

The results shown below include *all* data generated in all trajectories: no transient or relaxation period has been omitted.

4. RESULTS

4.1. Ala4. For Ala4, populations and MFPTs are estimated using WE and compared to independent measurements based on ordinary "brute force" (BF) simulation. Rates are estimated in both directions between the two sets of states A1,B1 and A2,B2 shown in Figure 3 (see SI to visualize representative structures). The second set is less populated and consequently expected to be more difficult to sample. Figure 3 also shows the bin definitions used in the post-analysis, which were the same as those used during the WE simulation. However, as we shall see in our second system, we can use any partition of the space for the post analysis.

The data shown below are based on the same total simulation times in BF and WE. The BF estimates and confidence intervals are based on a single long trajectory of 3.0 μs where thousands of transitions between states were observed. Five independent WE simulations were run, each employing a total of 3.0 μs accounting for all the trajectories. The use of independent WE runs permits straightforward error analysis for comparison with BF.

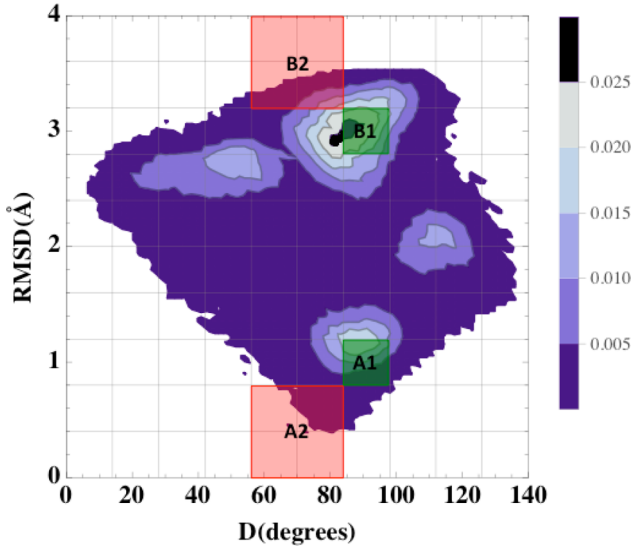


Figure 3. The Ala4 free energy surface. The surface is projected onto two coordinates: $D = ((1/N)\sum_i d_i^2)^{1/2} \in [0,180]$ from one reference structure (see SI) and the RMSD with respect to an ideal α-helix. The surface was computed using 3.0 μs of ordinary "brute force" simulation. The set of states A1,B1 is highlighted in green, while the second set A2,B2 is highlighted in red. The grid shows bins that were used both for WE simulation and for the post-analysis calculation of observables via the non-Markovian matrix formulation.

4.1.1. Direct Estimation of Observables via WE. As described above, "direct" WE measurements sum trajectory weights for population and flux calculations. Figures 4 and 5 show direct estimates for both equilibrium and kinetic quantities for both sets of states. WE estimates as a function of simulation time are compared to 95% confidence intervals for BF simulation.

As with all observables, data from five independent WE simulations are shown. The final/rightmost point from each run is the estimate using all data from the run and thus is based on a total simulation time equal to that of BF (3 μs). The spread of the rightmost WE data points therefore can be compared with the BF confidence interval to gauge statistical quality.

The mean values of the direct estimates are in agreement with BF confidence intervals in all cases. In some cases, the spread of WE estimates is significantly less than that for BF prior to the full extent of WE simulation. Each nanosecond of "molecular time" in Figures 4 and 5 (i.e., single-trajectory time) corresponds to approximately 200 ns of total simulation in a single WE run accounting for all trajectories. Hence, in some cases, considerably less WE simulation is required for an estimate of the same statistical quality as resulted from the full BF simulation of 3.0 μs.

4.1.2. Non-Markovian Matrix Analysis. We also show results of the non-Markovian matrix analysis for select observables. Figure 6 shows that the non-Markovian analysis yields unbiased estimates of the same equilibrium and nonequilibrium properties calculated with direct estimates. (Results for other observables, like the population of A1 and the A1→B1 MFPT, not shown, exhibit qualitatively similar agreement.) The agreement contrasts with a purely Markovian matrix formulation, which does not account for the "labeling" described above, which can yield statistically biased estimates for kinetic quantities (see methane results, below). Unbiased matrix-based estimates are important when reweighting is used in WE[23] as noted in the Discussion. Reweighting was not used in the present study, however.

4.2. Methane. In the methane system, WE simulation is used to measure first-passage times based on a range of state definitions. For a complex system, analyzing the sensitivity of the MFPT to state definitions could aid in the definition of states.

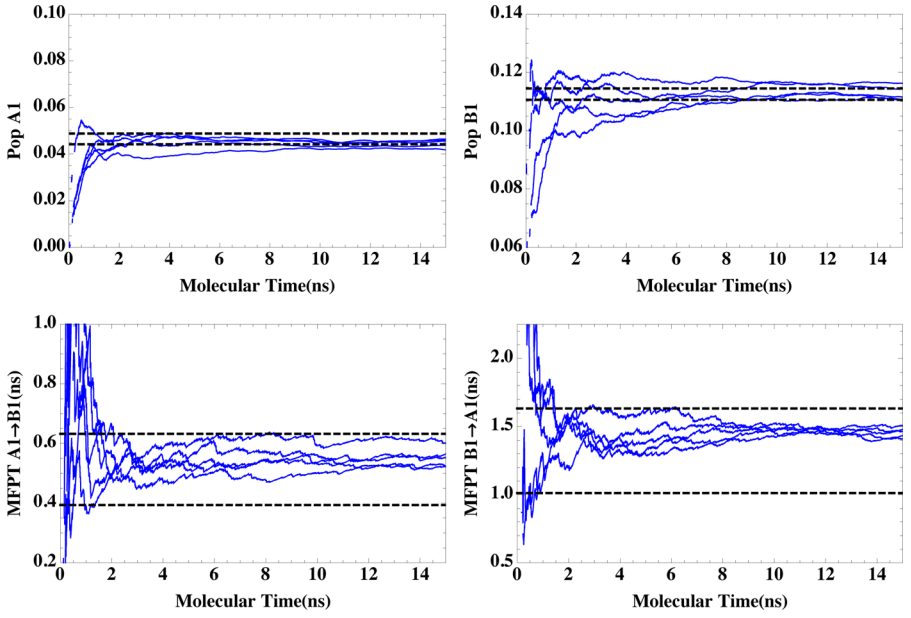


Figure 4. Direct WE estimates for populations and mean first passage times (MFPTs) for Ala4 states A1,B1 from Figure 3. Five independent WE runs are shown, each based on 3.0 μs of total simulation time. Dashed lines indicate roughly a 95% confidence interval based on 3.0 μs of brute force simulation. Each nanosecond of molecular (single-trajectory) time corresponds to approximately 200 ns of WE simulation including all trajectories in a single run.

dx.doi.org/10.1021/ct401065r | J. Chem. Theory Comput. 2014, 10, 2658−2667

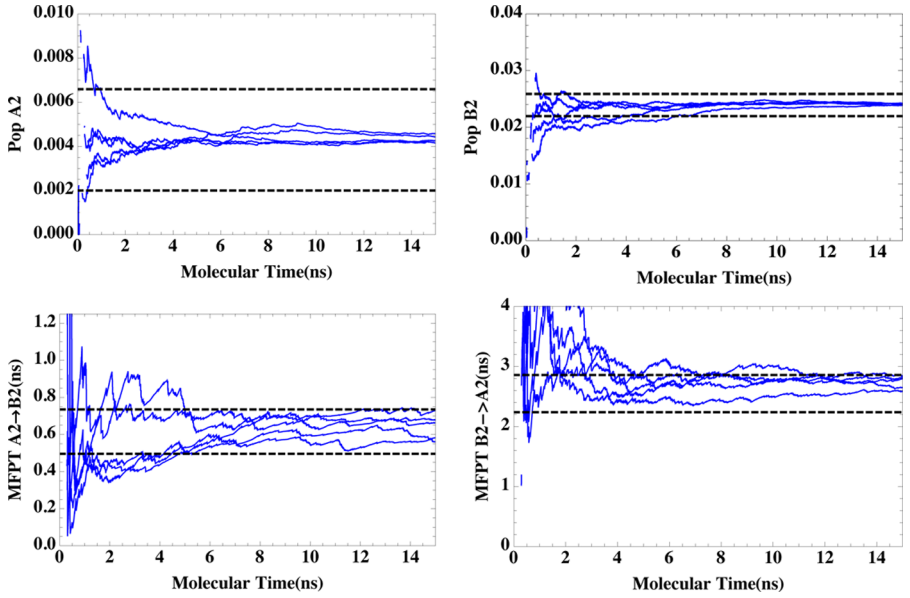


Figure 5. Direct WE estimates for populations and mean first passage times for Ala4 states A2,B2 from Figure 3. Five independent WE runs are shown, each based on 3.0 μs of total simulation time. Dashed lines indicate roughly a 95% confidence interval based on 3.0 μs of brute force simulation. Each nanosecond of molecular time corresponds to approximately 200 ns of WE simulation accounting for all trajectories in a single run.

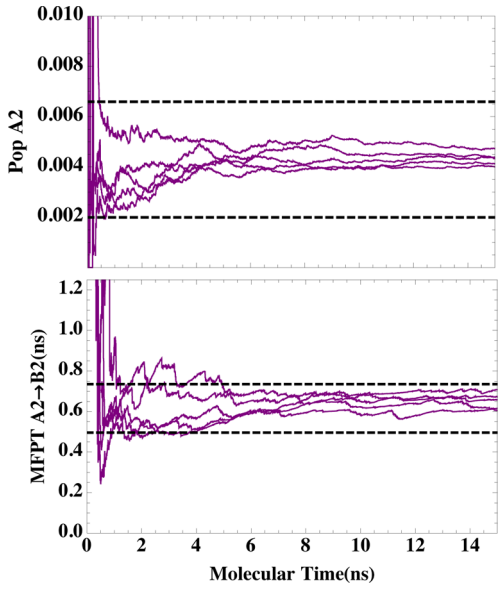


Figure 6. Population of A2 and mean first passage time for Ala4 from A2 to B2, estimated by the non-Markovian matrix analysis of WE data. Dashed lines indicate roughly a 95% confidence interval from brute force simulation, as in Figures 4 and 5. The states are defined in Figure 3.

The MFPT was estimated directly, as well as by both non-Markovian and Markovian matrix analysis. To assess statistical uncertainty, once again five independent WE simulations were run. The bins used for post-analysis differ from those used in the original WE simulation, as a matter of convenience—underscoring the flexibility of the approach.

Figure 7 shows passage times measured as a function of the boundary position for the unbound state. The boundary of the bound state A was held fixed at a separation of 5 Å while the definition of the unbound state was varied from 5 to 17 Å. The passage times were measured in increments of 2 Å and compared with BF results as shown in Figure 7. The BF confidence intervals are based on a single long trajectory of 0.4 μs, the same total simulation time used in each WE simulation.

Figure 7 shows that both direct and non-Markovian matrix estimates are in agreement with BF confidence intervals.

For fixed state definitions, Figure 8 shows the evolution of state populations MFPTs, as was done for Ala4. We fix the movable boundary position in Figure 7 (inset), defining state B as all configurations with $r > 11$ Å.

The performance of the non-Markovian matrix estimates are particularly noteworthy in Figure 8. The matrix estimates converge faster than direct estimates to the exact results for the state populations. Presumably, this is because the direct approach requires relaxation of the full probability distribution to equilibrium, whereas the matrix approach requires only relaxation of the distribution with each bin (in order to obtain accurate interbin rates $k_{ij}^{\mu\nu}$).

In contrast to the unbiased MFPT estimates obtained by both direct and non-Markovian analysis, the Markov analysis can be significantly biased for the MFPT. Figure 9 shows that applying the Markovian analysis (section 2.3) leads to MFPT estimates clearly outside the BF confidence interval. Data in the SI show that the use of a more sophisticated model such as a maximum-likelihood estimator for reversible Markov models[45] yields similar results and does not correct the bias.

Equilibrium properties, however, can be estimated without bias in a Markovian analysis because history dependence is immaterial. Figure 9 also illustrates correct (equilibrium) population estimates based on the Markovian analysis.

5. DISCUSSION

To our knowledge, this is the first weighted ensemble (WE) study using the original Huber and Kim algorithm[5] to simultaneously calculate both equilibrium and nonequilibrium quantities. The present study estimates observables (populations and MFPTs) based on arbitrary states defined in a postsimulation analysis, permitting the examination of different state definitions and their effects on observables. Two qualitatively different estimation schemes were examined,

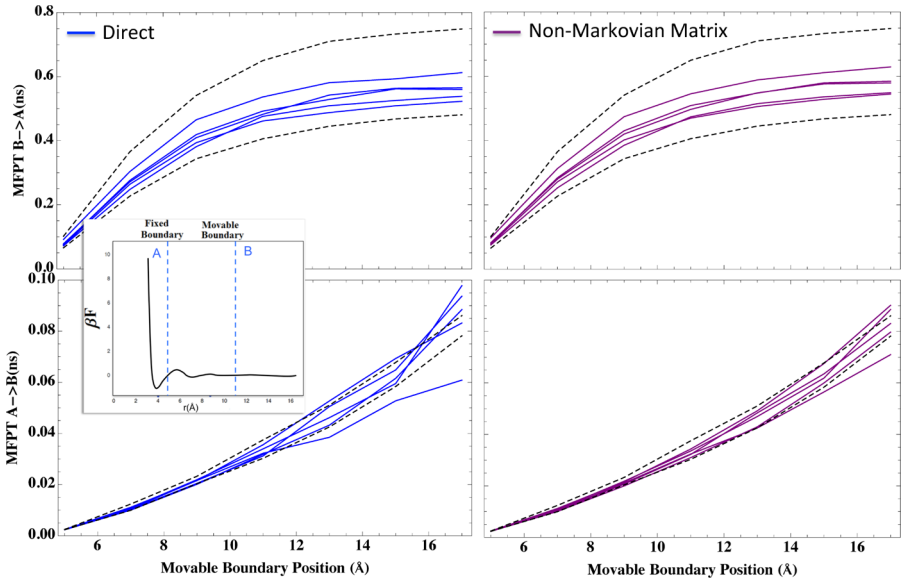


Figure 7. The mean first passage time for methane association (B to A) and dissociation (A to B) measured "directly" and from the non-Markovian matrix analysis from WE simulation as a function of the boundary of state A. The inset displays the PMF along with the definitions of the unbound and bound states, indicated by B and A, respectively. Dashed lines indicate roughly a 95% confidence interval based on 0.4 μs of brute force simulation.

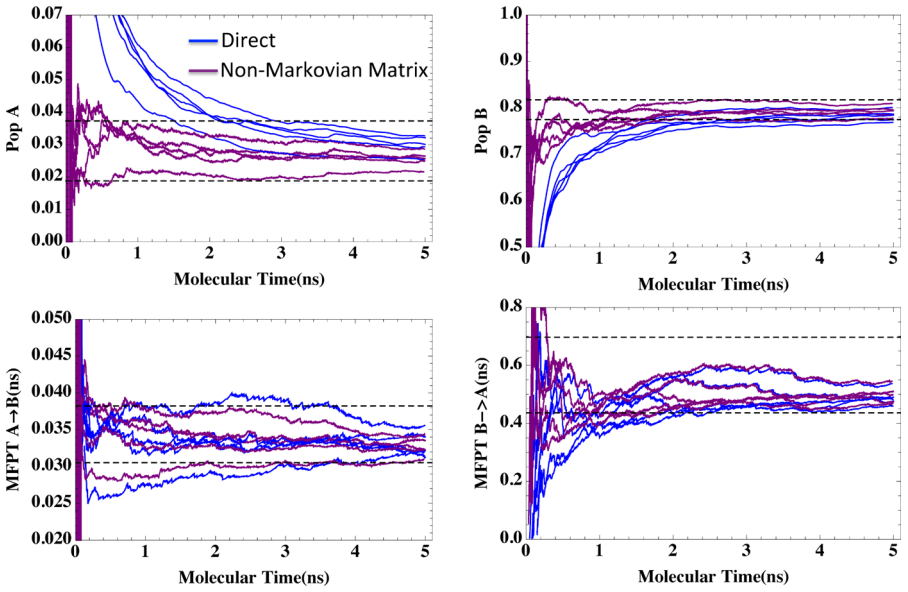


Figure 8. Methane association/dissociation observables. Direct and non-Markovian WE estimates for populations and mean first passage times (MFPTs) are plotted vs molecular time. Five independent WE runs are shown, each based on 0.4 μs of total simulation time. Dashed lines indicate roughly a 95% confidence interval based on 0.4 μs of brute force simulation. Each nanosecond of molecular time corresponds to approximately 80 ns of WE simulation accounting for all trajectories in a single run. The bound state (A) is defined by distances less than 5 Å, and B is defined by distances greater than 11 Å.

including a non-Markovian rate-matrix formulation which shows promise for reducing transient initial-state bias (a bias which is intrinsic to direct estimation of observables based on weights). Both schemes showed substantial efficiency gains for some observables even in the test systems which appear to lack significant energy barriers in their configurational landscapes. All results were validated using independent "brute force" simulations. Nevertheless, as described below, the present data do point to further challenges likely to be exhibited by larger, more complex systems.

Flexibility in State Choice. One key feature of the WE implementation studied here is the ability to investigate a range of state choices. As computer simulations tackle systems of growing complexity, it seems increasingly unlikely that states chosen prior to a study will prove physically or biochemically relevant. Indeed, it is already the case that specialized algorithms are invoked to identify physical states, separated by the slowest time scales, from existing trajectories.[46,47] With WE simulation, as suggested by our methane data, one can adjust state boundaries to minimize the sensitivity of rates to those boundaries.

A possible concern with postsimulation state construction is the need to store a potentially large set of coordinates to ensure sufficient flexibility in post analysis. However, modern hardware

dx.doi.org/10.1021/ct401065r | J. Chem. Theory Comput. 2014, 10, 2658−2667

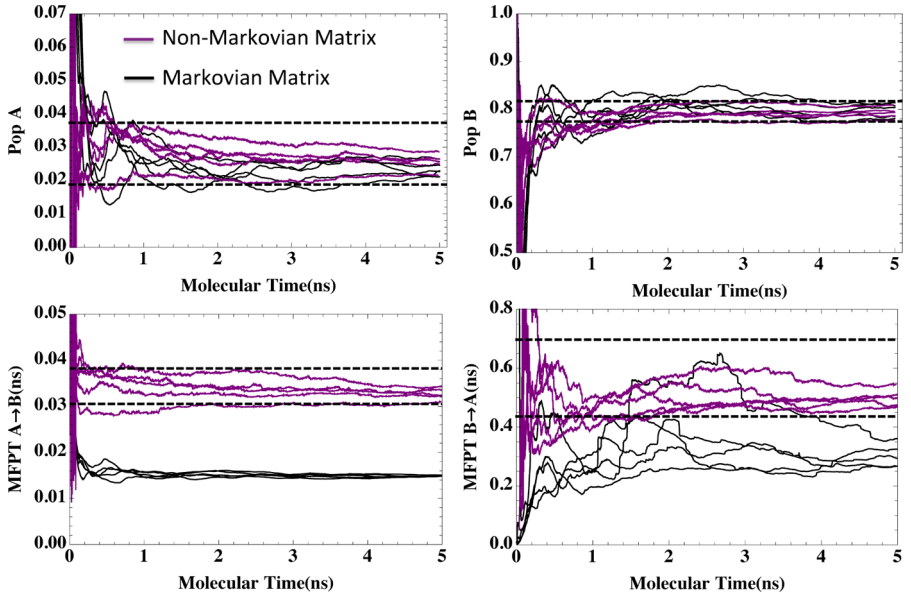


Figure 9. Populations of A (r < 5 Å) and B (r > 11 Å) and MFPTs for the methane system, estimated by the non-Markovian matrix analysis and the Markovian analysis without history information. Dashed lines indicate roughly a 95% confidence interval from brute force simulation based on 0.4 μs of total simulation time.

should be sufficient for most cases of interest. As an illustration, storage of $\{x,y,z\}$ coordinates for 1000 heavy atoms in a WE run of 1000 iterations using 1000 trajectories would require ~10 GB.

Simultaneous Calculation of Nonequilibrium and Equilibrium Observables. The estimation of both equilibrium and kinetic properties from relatively short simulations is an important goal of current methods development, including for WE.[24,29] Here, we have demonstrated as a "proof of principle" that WE simulation can do this efficiently (compared to brute force simulation), without bias, in parallel, and with flexibility in defining states. Given the relatively fast time scales (nanosecond scale) characterizing the present systems, it is somewhat surprising that WE is better than brute-force simulation for some of the observables and never worse. Previous studies suggest that WE has the potential for greater efficiency in more complex systems.[27,28,48]

Non-Markovian Behavior. Many of our results employ a non-Markovian analysis. Once a configuration space is discretized (e.g., bins in WE simulation), one expects in general that transitions among such discrete regions will not be Markovian. To take the simplest example, in a 1D system, whether a trajectory enters a finite-width bin from the left or right will affect the probability to make a transition in a given direction. So generally, discretized systems are non-Markovian, even when the underlying continuous dynamics are Markovian.

Reweighting and the Matrix Formulation. This study compared estimation of equilibrium and nonequilibrium observables using the original WE algorithm and via post-analysis. As mentioned in the Introduction, the occasional rescaling of weights to match an equilibrium or nonequilibrium steady-state condition[23] was not used to avoid any potential complications.

Our data clearly show that a standard Markovian analysis of WE simulation is inadequate (Figure 9), since WE bins typically are not Markovian. Additional information—history dependence, as embodied in the α/β labeling scheme—is needed to obtain unbiased results. Inclusion of history information in the matrix analysis means it is intrinsically "non-Markovian" regardless of the linear algebra employed.

Future work will incorporate the rate estimation and non-Markovian matrix schemes developed here, as well as possibly the simpler Markovian scheme shown in section 2.3. Our data (Figure 8) suggest that these could be very successful in bringing a WE simulation closer to a specified steady state. But it is an open question whether reweighting simulations will prove superior to the type of post-analysis suggested here. Importantly, data presented here indicate that some rate estimators could lead to biased estimates for populations, which, in turn, would bias a reweighted simulation.

One practical future approach, suggested by the work of Darve and co-workers,[49] could be to define preliminary states in advance to aid sampling transitions in both directions and then to subject the data to the same post analysis performed here to examine additional state definitions besides the initial choices.

Limitations and Future Work. The present study has not addressed some of the intrinsic limitations of the WE approach, which are the related issues of correlations among trajectories (due to the replication and merging events) and sampling "orthogonal" coordinates not divided up by WE bins. In the systems examined here, there was sufficient sampling in orthogonal dimensions to obtain excellent agreement with brute force results in all cases. However, significant future effort will be required to address correlations and orthogonal sampling, the latter being a problem common to methods which preselect coordinates such as multiple-window umbrella sampling[36,50,51] and metadynamics.[52−54]

6. CONCLUSIONS

In this proof-of-principle study, the parallel weighted ensemble (WE) approach has been applied to measure equilibrium and kinetic properties from a single simulation in small but nontrivial molecular systems. Importantly, populations and rates could be measured for arbitrary states chosen after the simulation. For all tested observables, unbiased estimates were obtained, as validated by independent brute-force simulations.

In a number of instances, WE was significantly more efficient—yielding estimates of a given statistical quality in less overall computing time compared to simple simulation, including all trajectories. In this sense, not only is WE a parallel method but it can exhibit "super-linear scaling;" e.g., 100 cores can yield desired information more than 100 times faster than single-core simulation.

We also developed a non-Markovian matrix approach for analyzing WE or brute-force trajectories, capable of yielding unbiased results, sometimes faster than direct estimates of observables from WE. The non-Markovian formulation also yields simultaneous estimates of equilibrium and nonequilibrium observables based on an arbitrary division of phase space, which is not possible in a standard Markovian analysis.

The approaches tested here will need to be further developed and tested in more complex systems.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Reference coordinates for the order parameters in Ala4, visualization of the states used in Ala4 (A1, A2, B1, and B2), and a comparison of the regular Markov model vs the Maximum Likelihood Estimator for reversible Markov models are shown. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ddmmzz@pitt.edu.

Author Contributions

‖Equal contributions

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Josh Adelman for insightful discussions, as well as the financial support from the NIH (Grant No. P41 GM103712) and the NSF (Grant Nos. MCB-0643456, MCB-1119091 and MCB- 0845216).

■ REFERENCES

(1) Berg, J. M.; Tymoczko, J. L.; Stryer, L. *Biochemistry*, 5th ed.; Freeman: New York, 2002.

(2) Zuckerman, D. M. *Annu Rev. Biophys.* **2011**, *40*, 41−62.

(3) Swendsen, R. H.; Wang, J.-S. *Phys. Rev. Lett.* **1986**, *57*, 2607−2609.

(4) Zheng, L.; Chen, M.; Yang, W. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 20227−20232.

(5) Huber, G. A.; Kim, S. *Biophys. J.* **1996**, *70*, 97−110.

(6) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291−318.

(7) Allen, R. J.; Warren, P. B.; ten Wolde, P. R. *Phys. Rev. Lett.* **2005**, *94*, No. 018104.

(8) Warmflash, A.; Bhimalapuram, P.; Dinner, A. R. *J. Chem. Phys.* **2007**, *127*, 154112−8.

(9) Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *131*, No. 044120.

(10) Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880−10889.

(11) 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.; Wriggers, W. *Science* **2010**, *330*, 341−346.

(12) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.

(13) Buchete, N.-V.; Hummer, G. *Phys. Rev. E: Stat. Nonlin. Soft Matter Phys.* **2008**, *77*, 030902.

(14) Huang, X.; Bowman, G. R.; Bacallado, S.; Pande, V. S. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19765−19769.

(15) Noe, F.; Schtte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011−19016.

(16) Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880−10889.

(17) West, A. M. A.; Elber, R.; Shalloway, D. *J. Chem. Phys.* **2007**, *126*, 145104−14.

(18) van Erp, T. S.; Moroni, D.; Bolhuis, P. G. *J. Chem. Phys.* **2003**, *118*, 7762−7774.

(19) Moroni, D.; Bolhuis, P. G.; van Erp, T. S. *J. Chem. Phys.* **2004**, *120*, 4055−4065.

(20) Moroni, D.; van Erp, T. S.; Bolhuis, P. G. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2005**, *71*, 056709.

(21) Valeriani, C.; Allen, R. J.; Morelli, M. J.; Frenkel, D.; ten Wolde, P. R. *J. Chem. Phys.* **2007**, *127*, 114109.

(22) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. *J. Chem. Phys.* **2010**, *132*, 054107.

(23) Bhatt, D.; Zhang, B. W.; Zuckerman, D. M. *J. Chem. Phys.* **2010**, *133*, 014110.

(24) Bhatt, D.; Bahar, I. *J. Chem. Phys.* **2012**, *137*, 104101.

(25) Rojnuckarin, A.; Kim, S.; Subramaniam, S. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 4288−4292.

(26) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 18043−18048.

(27) Bhatt, D.; Zuckerman, D. M. *J. Chem. Theory Comput.* **2010**, *6*, 3527−3539.

(28) Zwier, M. C.; Kaus, J. W.; Chong, L. T. *J. Chem. Theory Comput.* **2011**, *7*, 1189−1197.

(29) Darve, E.; Ryu, E. In *Innovations in Biomolecular Modeling and Simulations: Vol. 1*; Schlick, T., Ed.; Royal Society of Chemistry: London, 2012; Chapter Computing Reaction Rates in Bio-molecular Systems Using Discrete Macro-states, pp 138−206.

(30) Dickson, A.; Warmflash, A.; Dinner, A. R. *J. Chem. Phys.* **2009**, *131*, 154104.

(31) Bhatt, D.; Zuckerman, D. M. *J. Chem. Theory Comput.* **2011**, *7*, 2520−2527. PMCID: PMC3159166.

(32) Zwier, M. C.; Kaus, J. W.; Adelman, J. L.; Pratt, A. J.; Zuckerman, D. M.; Chong, L. T. 2014. Manuscript submitted for publication.

(33) Adelman, J. L.; Grabe, M. *J. Chem. Phys.* **2013**, *138*, 044105.

(34) Donovan, R. M.; Sedgewick, A. J.; Faeder, J. R.; Zuckerman, D. M. *J. Chem. Phys.* **2013**, *139*, 115105.

(35) Zuckerman, D. M. *Statistical Physics of Biomolecules: An Introduction*; CRC Press: Boca Raton, FL, 2010.

(36) Dickson, A.; Maienschein-Cline, M.; Tovo-Dwyer, A.; Hammond, J. R.; Dinner, A. R. *J. Chem. Theory Comput.* **2011**, *7*, 2710−2720.

(37) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712−725.

(38) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *Amber 11*; University of California: San Francisco.

(39) Hawkins, G.; Cramer, C.; Truhlar, D. *J. Phys. Chem.* **1996**, *100*, 19824−19839.

(40) Hawkins, G.; Cramer, C.; Truhlar, D. *Chem. Phys. Lett.* **1995**, *246*, 122−129.

(41) Tsui, V.; Case, D. *Biopolymers* **2001**, *56*, 275−291.

(42) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.

(43) Schuler, L. D.; Daura, X.; Van Gunsteren, W. F. *J. Comput. Chem.* **2001**, *22*, 1205−1218.

(44) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(45) 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.
(46) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101−17.
(47) Zhang, X.; Bhatt, D.; Zuckerman, D. M. *J. Chem. Theory Comput.* **2010**, *6*, 30483057.
(48) Adelman, J. L.; Dale, A. L.; Zwier, M. C.; Bhatt, D.; Chong, L. T.; Zuckerman, D. M.; Grabe, M. *Biophys. J.* **2011**, *101*, 2399−2407.
(49) Abdul-Wahid, B.; Yu, L.; Rajan, D.; Feng, H.; Darve, E.; Thain, D.; Izaguirre, J. A. Folding proteins at 500 ns/hour with Work Queue. 2012 IEEE 8th International Conference on E-Science, e-Science 2012.
(50) Haydock, C.; Sharp, J. C.; Prendergast, F. G. *Biophys. J.* **1990**, *57*, 1269−1279.
(51) Dickson, A.; Warmflash, A.; Dinner, A. R. *J. Chem. Phys.* **2009**, *130*, 074104−12.
(52) Vargiu, A. V.; Ruggerone, P.; Magistrato, A.; Carloni, P. *Nucleic Acids Res.* **2008**, *36*, 5910−5921.
(53) Bussi, G.; Laio, A.; Parrinello, M. *Phys. Rev. Lett.* **2006**, *96*, 090601.
(54) Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. *J. Phys. Chem. B* **2006**, *110*, 3533−3539.