JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Stratified UWHAM and Its Stochastic Approximation for Multicanonical Simulations Which Are Far from Equilibrium

Bin W. Zhang,[†] Nanjie Deng,[‡] Zhiqiang Tan,[¶] and Ronald M. Levy*[,†]

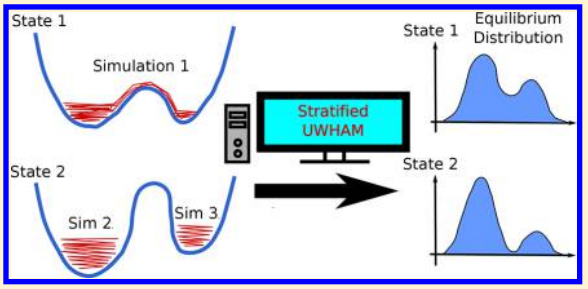
[†]Center for Biophysics and Computational Biology, Department of Chemistry and Institute for Computational Molecular Science, Temple University, Philadelphia, Pennsylvania 19122, United States

[‡]Department of Chemistry and Physical Sciences, Pace University, New York, New York 10038, United States

[¶]Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, United States

Supporting Information

ABSTRACT: We describe a new analysis tool called Stratified unbinned Weighted Histogram Analysis Method (Stratified-UWHAM), which can be used to compute free energies and expectations from a multicanonical ensemble when a subset of the parallel simulations is far from being equilibrated because of barriers between free energy basins which are only rarely (or never) crossed at some states. The Stratified-UWHAM equations can be obtained in the form of UWHAM equations but with an expanded set of states. We also provide a stochastic solver, Stratified RE-SWHAM, for Stratified-UWHAM to remove its computational bottleneck. Stratified-UWHAM and Stratified RE-SWHAM are applied to study three test topics: the free energy landscape of alanine dipeptide, the binding affinity of a host−guest binding complex, and path sampling for a two-dimensional double well potential. The examples show that when some of the parallel simulations are only locally equilibrated, the estimates of free energies and equilibrium distributions provided by the conventional UWHAM (or MBAR) solutions exhibit considerable biases, but the estimates provided by Stratified-UWHAM and Stratified RE-SWHAM agree with the benchmark very well. Lastly, we discuss features of the Stratified-UWHAM approach which is based on coarse-graining in relation to two other maximum likelihood-based methods which were proposed recently, that also coarse-grain the multicanonical data.



1. INTRODUCTION

Atomistic molecular dynamics (MD) simulations are widely used to study biological systems today to understand how structural ensembles are connected with biological functions. However, straightforward MD simulations cannot be used to study many biological problems since the time scales of transitions between functionally important states are much longer than the available simulation length determined by today's computational resources.[1−4] The desire to simulate structurally important transitions which occur on longer time scales has driven the development of simulation hardware and software.[5−7] For example, the Anton supercomputer developed by D. E. Shaw research is able to perform millisecond-scale simulations for proteins in explicit solvent.[7] The World Community Grid (WCG) projects of IBM (https://www.worldcommunitygrid.org) are able to combine the computational resources (~10^5−10^6 cpus) donated by volunteers all over the world to run molecular simulations whose goals are to develop therapies to fight cancer and AIDS. The same desire also encourages the development of enhanced sampling methods such as umbrella sampling,[8−10] replica exchange (RE) techniques,[11−16] and others.[17−25] Compared with straightforward MD simulations, those techniques show significantly better sampling efficiency on specific problems.

The Weighted Histogram Analysis Method (WHAM) is a powerful algorithm to compute free energies and expectations from multicanonical ensemble data.[26−30] Along with the popularity of enhanced sampling methods running parallel simulations at multiple thermodynamic and/or Hamiltonian states, WHAM, which is a standard analysis tool associated with those methods, has been studied by many researchers.[31−42] The most important improvement of WHAM is that a binless extension called the multistate Bennett acceptance ratio (MBAR) or unbinned WHAM (UWHAM) was introduced.[31,34,36] To avoid the requirements of very large memory and computational power to solve the UWHAM equations when the input data ensemble is large, we developed stochastic solvers for the UWHAM equations based on resampling techniques.[43,44]

When WHAM or UWHAM is applied, it is assumed that the observations generated from each thermodynamic and/or Hamiltonian state are drawn from a distribution P_α that is close to equilibrium, where P_α is determined by the Hamiltonian and/or thermostat temperature used in the simulations. However, this assumption is not fulfilled if the simulations at some thermodynamic and/or Hamiltonian states

Received: June 21, 2017
Published: September 13, 2017

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

are far away from convergence. For example, on massive but minimally communicating computational grids such as WCG, it is convenient to run multiple independent short MD simulations starting from different initial structures (which are not chosen from global equilibrium) at a single or multiple thermodynamic and/or Hamiltonian states. Here, our study focuses on how to obtain the optimal estimates of density of states, equilibrium distributions, and free energy differences for multistate simulations if the simulations at some thermodynamic and/or Hamiltonian state are far from convergence due to barrier(s) that are infrequently (or never) crossed at these states but frequently crossed at others. Simply combining all the observations of unconverged short simulations at a thermodynamic and/or Hamiltonian state as the input of that state for UWHAM introduces statistical biases even when the simulations at other thermodynamic and/or Hamiltonian states have already converged. To solve this problem, we introduce a powerful extension of UWHAM called Stratified-UWHAM. We also introduce the corresponding stochastic solver for the Stratified-UWHAM algorithm for cases where the input data ensemble is very large.

The remaining part of the paper proceeds as follows: First we review UWHAM (also called MBAR). Then we introduce Stratified-UWHAM and its stochastic solver Stratified RE-SWHAM. In the Results and Discussion section, we applied Stratified-UWHAM and Stratified RE-SWHAM to analyze the simulation data of three test systems — alanine dipeptide, a host−guest binding complex, and a Brownian particle in a two-dimensional double well potential. For the sake of simplicity, for the remainder of this paper, we refer to each of the thermodynamic and/or Hamiltonian states characterized by a specific combination of a Hamiltonian function and thermodynamic parameters as a "λ-state". We refer to each conformational structure of a biological or physical system as a "microstate" and to each free energy basin which is separated from other basins by free energy barriers as a "macrostate". A macrostate cluster means a collection of one or more free energy basins that can be transversed in the simulations.

The idea underlying Stratified-UWHAM is to coarse-grain the configurational space into macrostate clusters and divide λ-states of parallel simulations into two groups based on how well-connected the coarse-grained network is at each λ-state. The first group includes the λ-states at which the simulations are "approximately" equilibrated among macrostate clusters, namely, the fully connected λ-states. Notice that if a simulation at any λ-state is fully converged or fully globally equilibrated, running simulations at other λ-states additionally and applying UWHAM are redundant because the *true* density of states can be obtained from the fully converged simulation at that λ-state. In this study, the λ-states in the first group are those λ-states at which multiple transitions between macrostate clusters have been observed in simulations so that the coarse-grained state space is fully connected. The second group includes the λ-states at which the simulations are only locally equilibrated within each macrostate cluster, namely, the disconnected λ-states. They are also referred to as "locally equilibrated λ-states".

2. METHODS

2.1. Unbinned Weighted Histogram Analysis Method (UWHAM). To illustrate basic ideas, we first review UWHAM[36] (also called MBAR[34]). Suppose that N_α observations $\{X_i^{(\alpha)}: i = 1,...,N_\alpha\}$ are independently drawn from the αth distribution P_α

$$P_\alpha(X_i^{(\alpha)}) \sim \frac{q_\alpha(x_{\alpha i})}{Z_\alpha} \tag{1}$$

where Z_α is the partition function of the αth λ-state, $x_{\alpha i}$ is the coordinates of the microstate $X_i^{(\alpha)}$, and $q_\alpha(x_{\alpha i})$ is the unnormalized probability of observing the microstate $X_i^{(\alpha)}$ at the αth λ-state. For example, $q_\alpha(x_{\gamma i})$ equals $\exp\{-\beta_\alpha E_\alpha(x_{\gamma i})\}$ in the canonical ensemble, where $x_{\gamma i}$ is the coordinates of the ith observation observed at the γth λ-state $X_i^{(\gamma)}$. $E_\alpha(x_{\gamma i})$ is the potential energy of the microstate $X_i^{(\gamma)}$ at the αth λ-state, and β_α is the inverse temperature of the αth λ-state.

The likelihood of the simulated data is

$$\prod_{\alpha=1}^{M}\prod_{i=1}^{N_\alpha}\left[\frac{q_\alpha(u_{\alpha i})\Omega(u_{\alpha i})}{Z_\alpha}\right] \tag{2}$$

where $\Omega(u_{\alpha i})$ is the density of states of the reduced (energy) coordinate $u_{\alpha i}$ of the microstate $X_i^{(\alpha)}$. The maximum likelihood estimates (MLEs) of the density of states $\hat{\Omega}(u_{\gamma i})$ and the corresponding MLEs $\hat{Z}_\alpha$ given the data satisfy the coupled equations

$$\hat{Z}_\alpha = \sum_{\gamma=1}^{M}\sum_{i=1}^{N_\gamma} q_\alpha(u_{\gamma i})\hat{\Omega}(u_{\gamma i})$$

$$\hat{\Omega}(u_{\gamma i}) = \frac{1}{\sum_{\kappa=1}^{M} N_\kappa \hat{Z}_\kappa^{-1} q_\kappa(u_{\gamma i})} \tag{3}$$

The UWHAM estimate of the probability of observing $u_{\gamma i}$ at the αth λ-state is

$$\hat{p}_\alpha(u_{\gamma i}) = \hat{Z}_\alpha^{-1}\hat{\Omega}(u_{\gamma i})q_\alpha(u_{\gamma i}) \tag{4}$$

2.2. Stratified-UWHAM. Our new method, called Stratified-UWHAM, is based on the following conditions: the λ-states are divided into two groups, (S_1, S_2), such that

(i) simulations are approximately equilibrated among the macrostates for each of the λ-states in S_1, or more generally, the coarse-grained set of macrostates forms a connected network for each λ-state and together they form a globally connected network;

(ii) simulations are locally equilibrated within each macrostate cluster $(R_1,...,R_K)$ for each of the λ-states in S_2 but may be far from equilibrated among the macrostate clusters, or more generally, for each λ-state within S_2 the coarse-grained set of macrostates forms a disconnected network.

These conditions can be captured by a stratified model, which assumes that the set of observations $\{X_i^{(\alpha)}: i = 1,..., N_\alpha\}$ is independently drawn from P_α for each $\alpha \in S_1$, and the set of observations $\{X_i^{(\alpha)}: X_i^{(\alpha)} \in R_k, i = 1,...,N_\alpha\}$ is independently drawn from P_α *restricted* to macrostate clusters R_k for each $\alpha \in S_2$, i.e.,

$$P_\alpha(X_i^{(\alpha)}) \sim \frac{q_\alpha(x_{\alpha i})}{Z_\alpha} \text{ for } \alpha \in S_1$$

$$P_\alpha(X_i^{(\alpha)} | (X_i^{(\alpha)} \in R_k)) \sim \frac{q_{\alpha k}(x_{\alpha i})}{Z_{\alpha k}} \text{ for } \alpha \in S_2 \tag{5}$$

where $q_{\alpha k}(x) = q_\alpha(x)\delta\{x \in R_k\}$, and $\delta\{x \in A\}$ denotes the indicator function for a macrostate cluster A, and $Z_{\alpha k}$ and Z_α are the partition functions. In other words, the set of observations $\{X_i^{(\alpha)}: i = 1,...,N_\alpha\}$ is stratified into macrostate clusters $(R_1,...,R_K)$

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

for each λ-state α in S_2 such that simulations are only locally equilibrated but are not stratified for each λ-state α in S_1 where transitions between macrostates are enhanced. The likelihood of the simulated data from model 5 is

$$\prod_{\alpha\in S_2}\prod_{k}\prod_{i:X_i^{(\alpha)}\in R_k}\left[\frac{1}{Z_{\alpha k}}q_\alpha(u_{\alpha i})\Omega(u_{\alpha i})\right]\times\prod_{\alpha\in S_1}\prod_{i=1}^{N_\alpha}\left[\frac{1}{Z_\alpha}q_\alpha(u_{\alpha i})\Omega(u_{\alpha i})\right] \tag{6}$$

The method of nonparametric maximum likelihood[31] can be used for estimating the density of states and subsequently free energies and expectations.

The estimating equations from the maximization of eq 6 can be obtained in the form of UWHAM equations eq 3 but with an expanded set of λ-states. The idea is to split the K disconnected macrostate clusters of each λ-state in the S_2 group into K λ-states. Suppose there is a new λ-state which is made of the kth macrostate cluster of the γth λ-state. The Hamiltonian function of this new λ-state is set to be the same as the Hamiltonian function of the γth λ-state if the observation belongs to the kth macrostate cluster and positive infinity if the observation does not. Then all the observations in the kth macrostate cluster at the γth λ-state are treated as the observations observed at this new λ-state. This change of the Hamiltonian function and regrouping of the observations are equivalent to putting an infinite barrier covering the entire outside of the kth macrostate cluster in the conformational space. Suppose there are M_1 λ-states in the S_1 group and M_2 λ-states in the S_2 group. After the expansion of λ-states, the total number of λ-states increases from $M = M_1 + M_2$ to $M = M_1 + \sum_{\alpha=1}^{M_2} K_\alpha$ where K_α is the total number of macrostate clusters at the αth λ-state in the S_2 group. Then the MLEs of the density of states and free energy differences of eq 6 can be obtained by solving the UWHAM equations with an expanded set of λ-states

$$\hat{Z}_\alpha = \sum_{\gamma=1}^{M}\sum_{i=1}^{N_\gamma} q_\alpha(u_{\gamma i})\hat{\Omega}(u_{\gamma i}) \text{ for } \alpha\in S_1$$

$$\hat{Z}_{\alpha k} = \sum_{\gamma=1}^{M}\sum_{i=1}^{N_\gamma} q_{\alpha k}(u_{\gamma i})\hat{\Omega}(u_{\gamma i}) \text{ for } \alpha\in S_2$$

$$\hat{\Omega}(u_{\gamma i}) = \frac{1}{\sum_{\alpha\in S_1}^{M_1} N_\alpha \hat{Z}_\alpha^{-1} q_\alpha(u_{\gamma i}) + \sum_{\alpha\in S_2}^{M_2}\sum_{k=1}^{K_\alpha} N_{\alpha k}\hat{Z}_{\alpha k}^{-1} q_{\alpha k}(u_{\gamma i})} \tag{7}$$

where the unnormalized probability of an observation $u_{\gamma i}$ at a new λ-state is

$$q_{\alpha k}(u_{\gamma i}) = \begin{cases} \exp\{-\beta_\alpha E_\alpha(u_{\gamma i})\} & \text{for } u_{\gamma i}\in R_k \\ 0 & \text{for } u_{\gamma i}\notin R_k \end{cases} \tag{8}$$

for canonical ensembles. The population ratio between the mth and nth disconnected macrostate clusters of a locally equilibrated αth λ-state is estimated based on their free energy difference

$$\frac{P_{\alpha m}}{P_{\alpha n}} = \exp\left\{-\frac{\Delta F_{\alpha m,\alpha n}}{k_B T}\right\} = \frac{\hat{Z}_{\alpha m}}{\hat{Z}_{\alpha n}} = \frac{\sum_{\gamma=1}^{M}\sum_{i=1}^{N_\gamma} q_{\alpha m}(u_{\gamma i})\hat{\Omega}(u_{\gamma i})}{\sum_{\gamma=1}^{M}\sum_{i=1}^{N_\gamma} q_{\alpha n}(u_{\gamma i})\hat{\Omega}(u_{\gamma i})} \tag{9}$$

2.3. Stratified Stochastic WHAM. There is a computational bottleneck in scaling up UWHAM. At minimum, numerical solution of the UWHAM equations eq 3 requires evaluating M unnormalized density functions $q_1(X_i)$, ..., $q_M(X_i)$ at each observation X_i for $i = 1,...,N$. The total number of function evaluations is of order $\bar{n}M^2$, where $\bar{n} = N/M$ is the average sample size per distribution. These unnormalized density values need to be either computed during every iteration of the numerical solution or precomputed and stored in memory. Such a high computational cost presents a serious limitation on the use of UWHAM for large-scale simulations (for example, $M = 240$ and $N = 3.5 \times 10^7$ in our recent work[43,44]). Although Stratified-UWHAM can be applied by directly using the UWHAM software package developed before, it can require much more memory and computational time to converge because the total number of λ-states can increase substantially.

To remove the computational bottleneck, we recently developed the RE-SWHAM algorithm which solves the UWHAM equations stochastically (see ref 43 for details). A straightforward way to solve the Stratified-UWHAM equations stochastically is by performing RE-SWHAM analyses as described in ref 43 for the corresponding UWHAM equations with an expanded set of λ-states. Note that the direct outputs of RE-SWHAM are the estimates of conformational equilibrium distributions at each λ-state. The estimates of free energy differences (and the population ratios) between macrostate clusters of a locally equilibrated λ-state can then be calculated using thermodynamic cycles similarly as shown in Figure 2 and discussed in detail in Section 3.1, while applying the "free energy perturbation formula" (see eq 20 in ref 43).

We describe a different algorithm called Stratified RE-SWHAM to solve the Stratified-UWHAM equations stochastically by improving the above straightforward application of RE-SWHAM. In the original implementation of RE-SWHAM, every cycle consists of a move process and an exchange process, the same as replica exchange simulations. The move process for the next observation is chosen from the database of observations at each λ-state according to the probability $1/N_\alpha$, where N_α is the number of observations generated at that λ-state. This move process in RE-SWHAM is analogous to the move process of an explicit RE simulation when the MD simulation period per cycle is so long that the initial and final configurations of the MD simulation period are largely uncorrelated. However, when the simulations at some λ-states are only locally equilibrated within macrostate clusters and the coarse-graining results in a disconnected network of macrostate clusters, the move process in RE-SWHAM at these λ-states needs to be adjusted accordingly as follows. In the stratified RE-SWHAM analysis, the next observation is chosen from the data elements in the same connected macrostate cluster (instead of all the macrostates) with equal probability for each of the λ-states in the stratified S_2 group.

The procedure of running Stratified RE-SWHAM to analyze simulation data is as follows:

1. A database of observations is constructed for each λ-state using all the data elements observed at that λ-state. Each data element is tagged by the macrostate which it belongs to.
2. Then Stratified RE-SWHAM is run in cycles like replica exchange simulations:
   - Move: For each λ-state, one data element is selected from its database to associate with the replica at that λ-state. At the fully connected λ-states, one of the data elements is chosen with equal probability; at the

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

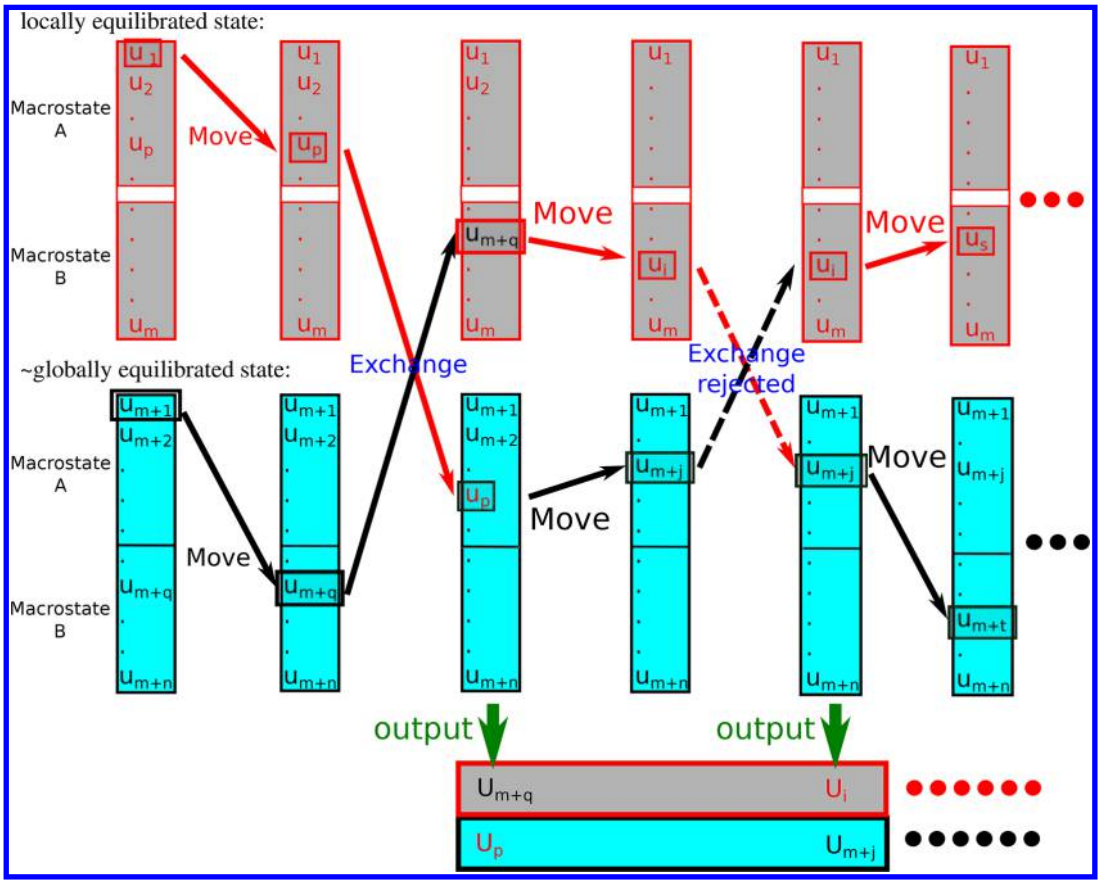


Figure 1. An illustration of the Stratified RE-SWHAM algorithm. This drawing shows two λ-states with "grey" or "cyan" color. Each λ-state has two macrostates A and B. The gray λ-state is locally equilibrated, while the simulations at the cyan λ-state are approximately equilibrated among the macrostates. The white gap between macrostates at the gray λ-state represents an uncrossable barrier for the "move" procedure during the Stratified RE-SWHAM analysis. Beforehand, we construct each λ-state a database which contains all the observations obtained from that λ-state, and each observation is tagged by the macrostate which it belongs to. As shown in the picture, the observations are separated into two subgroups A and B. Then Stratified RE-SWHAM is run in cycles, which consists of a "move" procedure and an "exchange" procedure. In the move procedure, Stratified RE-SWHAM chooses an observation to associate with the replica at each λ-state. At the cyan λ-state, the next observation is chosen from the whole database of that λ-state with equal probability. However, at the gray λ-state, the next observation is chosen from the subgroup which the previous observation belongs to with equal probability. In the exchange procedure, if the exchange attempt is accepted, in addition to the swap of the replicas, the observations associated with the replicas are also swapped to the database of the other λ-state. At the end of each cycle, the observation associated with each replica is recorded as the output of Stratified RE-SWHAM.

disconnected λ-states, one of the data elements which are in the same connected macrostate cluster as the data element previously associated with the replica at that λ-state is chosen with equal probability.

- Exchange: Replica exchange attempts are examined according to the multicanonical exchange criterion. If an exchange attempt is accepted, the replicas are swapped, and the data elements associated with the replicas are also swapped to the database of the other λ-state.
- At the end of the cycle, the data element associated with the replica at each λ-state is recorded as the output of that λ-state.

3. The output of each λ-state is the estimate of the equilibrium distribution of that λ-state. Further statistical analyses can be applied to the data ensembles generated by Stratified RE-SWHAM at those interested λ-states. Figure 1 illustrates the procedure of stratified RE-SWHAM.

Compared with RE-SWHAM with an expanded set of λ-states, Stratified RE-SWHAM does not split the λ-states in the S_2 group into multiple new λ-states. In the Appendix, we show that, without the splitting of locally equilibrated λ-states, the output of Stratified RE-SWHAM at a locally equilibrated λ-state is the estimate of the equilibrium distribution of that λ-state. In particular, the population ratios can be estimated directly as those in the estimate of the equilibrium distribution of that λ-state from the output of Stratified RE-SWHAM, without explicitly invoking the thermodynamic cycle. Therefore, in addition to all of the advantages of RE-SWHAM over UWHAM discussed in ref 43, one more benefit of using Stratified RE-SWHAM to solve the Stratified-UWHAM equations is that the number of λ-states does not increase compared with the original system.

3. RESULTS AND DISCUSSION

3.1. Example 1: Alanine Dipeptide. To illustrate the problem, first we study the free energy landscape of alanine dipeptide (AlaD) in vacuum and in implicit solvent. The Ramachandran plots of an AlaD molecule are shown in Figure 2b. In the picture, the A macrostate cluster contains the $\beta/C5$, $C7_{eq}$, and α_R free energy basins on the left side of the plot, and the B macrostate cluster contains the α_L and $C7_{ax}$ free energy

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

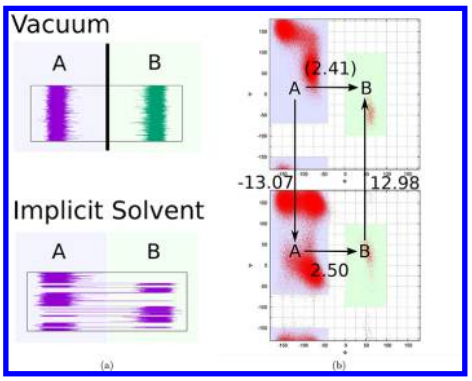


Figure 2. Thermodynamic cycle of alanine dipeptide (AlaD) in vacuum and implicit solvent. (a) It is much more computationally time-consuming to obtain the equilibrium distribution of AlaD in vacuum than AlaD in implicit solvent by brute force simulations. (b) The upper picture is the Ramachandran plot of AlaD in vacuum, and the lower picture is the Ramachandran plot of AlaD in implicit solvent. The free energy difference of the A and B macrostate clusters in vacuum equals to the sum of the other three legs.

basins on the right side of the plot. The simulation of AlaD in vacuum is ∼6 time faster per step than the simulation of AlaD in implicit solvent (OBC GB model) using GROMACS.[45,46] However, the free energy barriers between the A and B macrostate clusters are much higher for AlaD in vacuum than AlaD in implicit solvent because the electrostatic interaction screening of water is absent. Consequently, it turns out that it is much more computationally time-consuming to obtain the equilibrium distribution of AlaD in vacuum by brute force simulations. The first passage times of AlaD in implicit solvent are $\tau_{A\to B} = (78 \pm 3)$ ns and $\tau_{B\to A} = (1.33 \pm 0.04)$ ns, and the first passage times of AlaD in vacuum are $\tau_{A\to B} = (2.6 \pm 0.2)$ μs and $\tau_{B\to A} = (55 \pm 4)$ ns. In this study, the equilibrium distribution of AlaD in implicit solvent was obtained by replica exchange simulations first. Then two independent simulations of AlaD in vacuum, one starting from the A macrostate cluster and the other from B, were run. See the Supporting Information for simulation details. For both simulations, no transitions between the A and B macrostate clusters were observed during the first 100 ns. The goal is to apply UWHAM to estimate the equilibrium distributions of AlaD in vacuum based on the data generated by two 100 ns long independent simulations of AlaD in vacuum and the previously obtained equilibrium distribution of AlaD in implicit solvent. The two λ-states of this model problem will be referred to as the implicit solvent (I) state and the vacuum (V) state.

Figure 2b shows a typical thermodynamic cycle. To calculate the free energy difference between the A and B macrostate clusters at the vacuum state $\Delta F^V_{A\to B}$, the standard procedure is to calculate the two vertical legs, $\Delta F^{V\to I}_A$ and $\Delta F^{I\to V}_B$ using BAR (or UWHAM), and calculate the lower horizontal leg $\Delta F^I_{A\to B}$ using the population percentages of the two macrostate clusters at the implicit solvent state obtained by simulations.[47] Then the free energy difference presented by the upper horizontal leg can be calculated by

$$\Delta F^V_{A\to B} = \Delta F^{V\to I}_A + \Delta F^I_{A\to B} + \Delta F^{I\to V}_B \quad (10)$$

Given $\Delta F^V_{A\to B}$, the equilibrium distribution of AlaD in vacuum and the free energy difference between an AlaD molecule in vacuum and in implicit solvent can be estimated. The results obtained by using the thermodynamic cycle (eq 10) serve as the benchmark for this model problem.

On the other hand, conventional UWHAM is inappropriate to be applied straightforwardly to estimate the density of states and free energy difference between an AlaD molecule in vacuum and in implicit solvent. As mentioned previously, the two simulations of AlaD in vacuum are far from converged in 100 ns because there have been no transitions between the two macrostate clusters (β/C5, $C7_{eq}$, α_R) and (α_L, $C7_{ax}$). Simply combining the two unconverged data sets at the same λ-state does not provide an ensemble drawn from the Boltzmann distribution of that λ-state. Therefore, the corresponding UWHAM results are not correct. The difference between the conventional UWHAM estimate of $\Delta F^V_{A\to B}$ and the benchmark can be seen in Table 1. However, Stratified-UWHAM can be used to process the same data to obtain an accurate estimate of the free energy surfaces. We split the vacuum state into two λ-states and applied Stratified-UWHAM to obtain the density of states and free energy differences between λ-states for this new system with an expanded set of λ-states. The free energy difference between the A and B macrostate clusters at the vacuum state $\Delta F^V_{A\to B}$ was calculated according to eq 9. As can be seen in Table 1, $\Delta F^V_{A\to B}$ estimated by Stratified-UWHAM agrees very well with the benchmark, and the Stratified RE-SWHAM estimate also matches the benchmark within statistical error.

We continued running the two independent MD simulations at the vacuum states to obtain better converged raw data until the conventional UWHAM estimates also match the benchmark. The evolution of the conventional UWHAM and Stratified-UWHAM estimates is shown in Figure 3. As can be seen, Stratified-UWHAM converges to the benchmark within statistical error from the first data point where the simulation time is 100 ns. On the contrary, it takes several microseconds simulation time of AlaD in vacuum for the conventional UWHAM estimate to reach a similar precision level as the Stratified-UWHAM estimate. Figure 3 also shows the estimates of $\Delta F^V_{A\to B}$ based on the independent MD simulations A and B. That $\Delta F^V_{A\to B}$ converges on the same time scale when MD simulations A and B are UWHAMMed as when the simulations are considered individually reflects the fact that the macrostate clusters must be connected in simulations A and B before the two simulations can be UWHAMMed without bias. See more discussion about the convergence of Stratified-UWHAM estimates in the Supporting Information.

3.2. Example 2: β-Cyclodextrin Heptanoate Complex.

As the second example, we study the binding affinity of a host−guest system — the β-cyclodextrin heptanoate complex. The

Table 1. Free Energy Difference between the A and B Macrostate Clusters of AlaD in Vacuum $\Delta F^V_{A\to B}$ Estimated by the Thermodynamic Cycle, the Conventional UWHAM, Stratified-UWHAM, and Stratified RE-SWHAM[a]

	T cycle	Stratified-UWHAM	Stratified RE-SWHAM	UWHAM
$\Delta F^V_{A\to B}$(kcal/mol)	2.41 ± 0.04	2.42 ± 0.04	2.45 ± 0.05	0.1060 ± 0.0007

[a]Standard errors are estimated by the block bootstrap method.

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

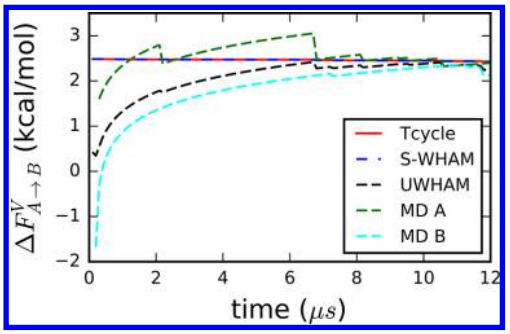


Figure 3. Dependence of the conventional UWHAM and Stratified-UWHAM estimates of the free energy difference between the A and B macrostate clusters of AlaD in vacuum $\Delta F^V_{A\to B}$ on the simulation length of the simulations of AlaD in vacuum. The red line is the benchmark — the estimates of thermodynamic cycle. The blue dashed line is the estimates of Stratified-UWHAM when the vacuum state is split into two new λ-states. The black dashed line is the estimates of the conventional UWHAM estimates when the data at the vacuum states are simply combined as the input of that λ-state. The green and cyan dashed lines are the estimate based on the independent MD simulations starting from the A and B macrostate clusters, respectively. It takes MD simulations lasting several microseconds of AlaD in vacuum for the conventional UWHAM estimates to match the benchmark compared with Stratified-UWHAM which converges within 100 ns. The Stratified-UWHAM estimates are visually identical with the benchmark.

host, β-cyclodextrin (βCD), is a frustum-shaped molecule with a hydrophobic interior core. The narrow opening end of βCD is laced with 7 primary hydroxyls, and the wide opening end is laced with 14 secondary hydroxyls. Because of its chemical nature, βCD can bind with a number of ligands, therefore serving as a classic "host" for the study of molecular recognition phenomena. The guest molecule, heptanoate, consists of a hydrophilic carboxylate group and hydrophobic alkyl groups. As the hydrophobic alkyl groups of heptanoate are nested in the cavity of βCD, the carboxylate group of heptanoate can form hydrogen bonds with either the primary or the secondary hydroxyls of βCD depending on the orientation of the heptanoate molecule. As shown in Figure 4, the β-cyclodextrin

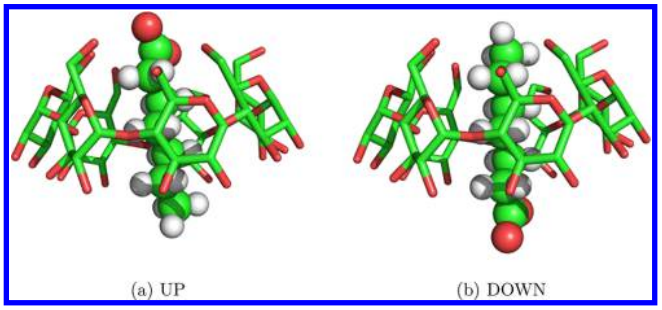


Figure 4. β-Cyclodextrin heptanoate binding complex. (a) UP macrostate. (b) DOWN macrostate.

heptanoate complex has two binding states, which will be referred to as the UP and DOWN macrostates. In our previous research,[15,43] we have studied the binding affinity of this host–guest system by using BEDAM — a free energy method based on replica exchange simulations. In BEDAM simulations, an additional parameter λ is introduced to scale the interaction between the host and the guest molecules from none to full interaction. The features of β-cyclodextrin heptanoate binding obtained using replica exchange serve as the benchmark for this test case where we employ Stratified-UWHAM to combine and analyze the results of independent (uncoupled) MD simulations at each of the λ Hamiltonian states.

We ran two sets of 72 ns independent MD simulations at 300 K of the β-cyclodextrin heptanoate complex in implicit solvent (AGBNP GB model[48]) at 16 λ-states: (0.0, 0.001, 0.002, 0.004, 0.01, 0.04, 0.07, 0.1, 0.2, 0.4, 0.6, 0.7, 0.8, 0.9, 0.95, 1.0). The λ-states are chosen to be the same as those in the previous BEDAM simulations.[15,43] At the λ = 0.0 state, there is no interaction between the ligand and the receptor, and the interaction is fully turned on at the λ = 1.0 state. However, there is no replica exchange coupling among different λ-states. We note that simulations which use computational grids typically do not employ replica exchange; this observation serves to motivate example 2. One set of independent simulations was started from the UP macrostate, and the other set was started from the DOWN macrostate. The simulation details of this example can be found in ref 15. At the seven λ-states with the largest λ values, because of the strong interaction between heptanoate and βCD molecules, it is difficult for the binding complex to switch between the UP and DOWN macrostates. During the 72 ns simulations, no transitions between the UP and DOWN macrostates were observed at the λ = 1.0, 0.95, 0.9 states, and only one or two transitions were observed at the λ = 0.8, 0.7, 0.6 states. However, when the interaction between the ligand and the receptor is further reduced (for λ values smaller than or equal to 0.2), multiple transitions occurred. See the Supporting Information for the number of transitions between macrostates during each simulation. We applied conventional UWHAM, Stratified-UWHAM, and Stratified RE-SWHAM to estimate the population percentage of each macrostate of the β-cyclodextrin heptanoate complex. To compare the equilibrium conformational ensembles estimated by different analysis methods based on the raw data from the independent simulations at each of the λ-states, we also examined the probability density of the binding energies for each conformational ensemble.

The red line in Figure 5a shows the equilibrium population percentages of the configurations in the DOWN macrostate at each λ-state as determined from the benchmark replica exchange data set. According to the benchmark, the population percentage of the DOWN macrostate starts from 50% at the λ = 0.0 state and continues increasing to the highest value 94.5% at the λ = 0.8 state. Then the population percentage of the DOWN macrostate decreases to 80.3% at the λ = 1.0 state. The DOWN macrostate is more favorable at large λ values; this comes from the larger entropy when the carboxylate group of the heptanoate molecule is located in the wide opening of the βCD molecule. Figure 5b shows the distributions of binding energy of the β-cyclodextrin heptanoate complex at the λ = 1.0 state. Although the UP macrostate is less favorable at the λ = 1 state, heptanoate and βCD can form more hydrogen bonds, resulting in more favorable (i.e., more negative) binding energy in this macrostate, because of the flexibility of the primary hydroxyls at the narrow opening end of βCD which can interact with the heptanoate carboxylate of the UP macrostate. We combined the data generated at each λ-state from the two sets (UP and DOWN) of independent simulations and applied conventional UWHAM to estimate the population percentage of the DOWN macrostate. The results shown in Figure 5a exhibit significant differences compared with the benchmark at all the λ-states whose λ value is larger than 0.2. At the λ = 1.0

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

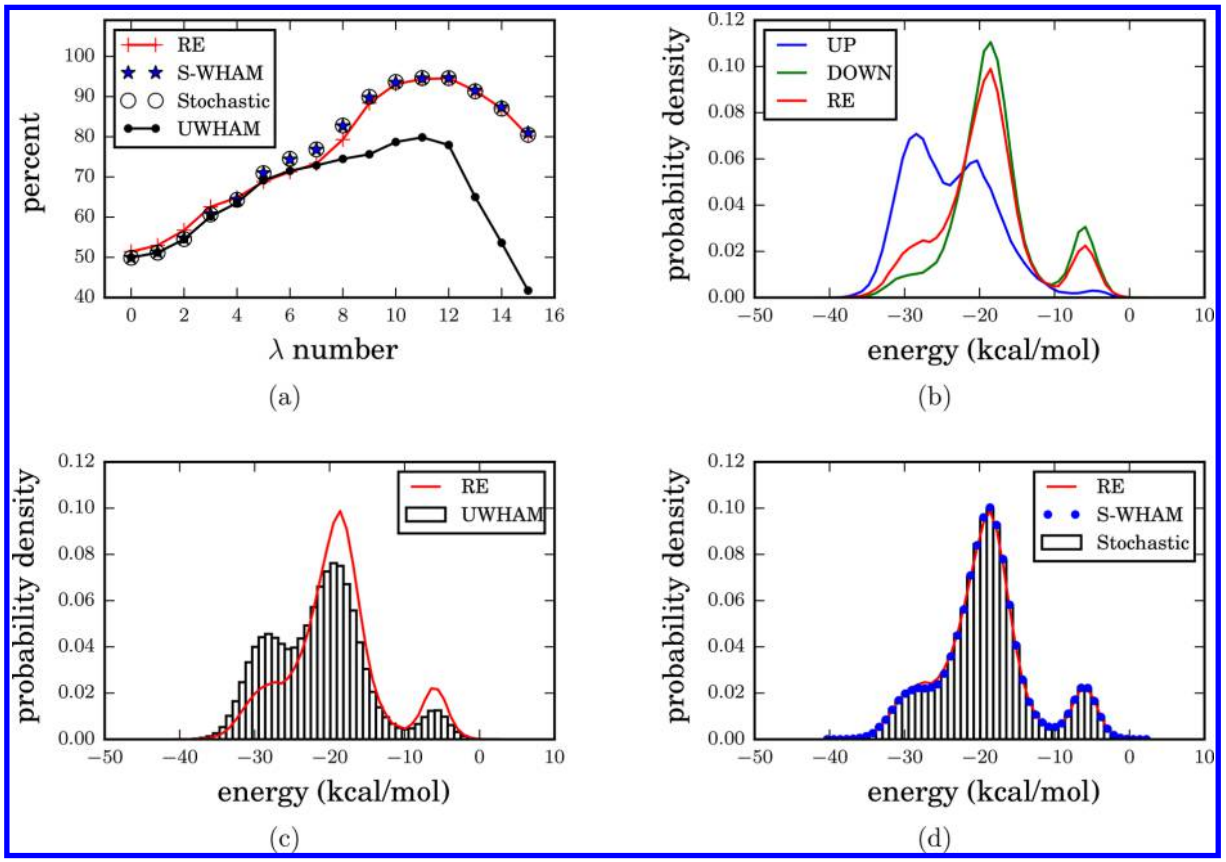


Figure 5. Population percentage of the configurations in the DOWN macrostates and probability density of binding energy at the $\lambda = 1.0$ state. The red line is the benchmark — the BEDAM (replica exchange simulation) results. (a) Comparison of the population percentages of the configurations in the DOWN macrostate estimated by Conventional UWHAM, Stratified-UWHAM, Stratified RE-SWHAM (based on the raw data generated by two sets of 72 ns independent MD simulations), and the benchmark. The blue stars are the Stratified-UWHAM estimates when the data at the largest seven λ-states are clustered into UP and DOWN macrostates. The black circles are the Stratified RE-SWHAM estimates, and the black line with dots is the conventional UWHAM estimates when the data at the unconverged λ-states are simply combined as the input of that λ-state. (b) The probability density of binding energy at the $\lambda = 1.0$ state obtained by replica exchange simulations. The blue line is the probability density of binding energy of the configurations in the UP macrostate, the green line is the probability density of binding energy of the configurations in the DOWN macrostate, and the red line is the overall probability density of binding energy at the $\lambda = 1.0$ state. (c) Comparison of the probability density of binding energy at the $\lambda = 1.0$ state estimated by conventional UWHAM to the 32 independent simulations and the benchmark. The bars show the UWHAM estimates. (d) Comparison of the probability density of binding energy at the $\lambda = 1.0$ state estimated by Stratified-UWHAM, Stratified RE-SWHAM, and benchmark. The blue dots are the Stratified-UWHAM estimates. The bars are the Stratified RE-SWHAM estimates. As can be seen by comparing the four figures, only the Stratified-UWHAM and Stratified RE-SWHAM estimates agree with the benchmark.

state, the difference between the benchmark and the conventional UWHAM estimate is as large as 38.6%. Not surprisingly, Figure 5c shows that the conventional UWHAM estimate (from the 32 independent simulations) of the distribution of the binding energies at the $\lambda = 1.0$ state does not agree with the benchmark either. Then we applied Stratified-UWHAM to analyze the data generated by the independent parallel MD simulations (Figure 5a and Figure 5d). In this case, the λ-states with the largest seven λ values are considered to be only locally equilibrated and are split into 14 new λ-states. As can be seen in Figure 5a and Figure 5d, the Stratified-UWHAM estimates of the population percentage of the DOWN macrostate at each λ-state and the distribution of the binding energies at the $\lambda = 1.0$ state agree with the benchmark very well, and the estimates obtained by the stochastic RE-SWHAM analysis are indistinguishable from the Stratified-UWHAM estimates. In the Supporting Information, we list the numerical results and uncertainties of the population percentages of the DOWN macrostate estimated by Stratified-UWHAM, Stratified RE-SWHAM, and the benchmark. We also show the comparisons of the probability density of binding energies estimated by Stratified-UWHAM, Stratified RE-SWHAM, and the benchmark at all λ-states.

3.3. Example 3: Dynamical Path Reweighting. Lastly, we study the trajectories of a Brownian particle moving in a two-dimensional double well potential. We apply UWHAM and Stratified-UWHAM to analyze the path ensembles generated by the transition path sampling method at different Hamiltonian states. Inspired by previous research,[49,50] the two-dimensional potential function is defined via

$$U(x, y)/(k_B T) = 1.25[64y^2(x^2 + y^2 - 1)^2 - \exp\{-4(x - 1)^2 + y^2\} - \exp\{-4(x + 1)^2 + y^2\} + \exp\{8(x - 1.25)\} + \exp\{-8(x + 1.25)\} + 4\exp\{-4(y + 0.25)\} + 12\exp\{-2x^2\}] \quad (11)$$

Figure 6a shows the contours of this potential. As can be seen, $U(x, y)/k_B T$ is symmetric with respect to a rotation about the y axis. The minimum of $U(x, y)$ equals $1.698 k_B T$ at $(x = \pm 1.087, y = 0.188)$. To study the transition events between these two free energy basins, we define the region where

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

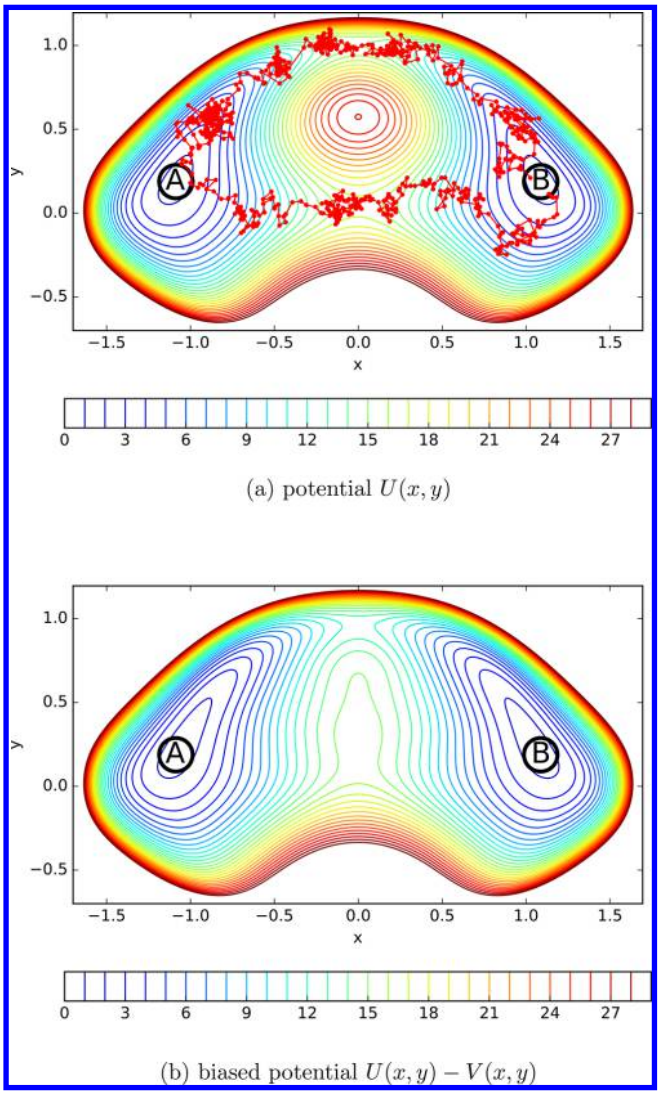


Figure 6. Two-dimensional double well potentials. (a) Contours of the double well potential $U(x, y)$ defined via eq 11. The red trajectories show two paths connecting the reactant (A) and product (B) regions. One goes through the UP channel, and the other goes through the DOWN channel. (b) Contours of the biased potential $U(x,y) - V(x, y)$, where $V(x, y)$ is defined via eq 14.

$$(x + 1.087)^2 + (y - 0.188)^2 < 1.000 \quad (12)$$

as the reactant (or A) region and

$$(x - 1.087)^2 + (y - 0.188)^2 < 1.000 \quad (13)$$

as the product (or B) region. The barrier between the reactant and product regions has two saddle points on the y axis. The upper saddle point is located at (x = 0.000, y = 1.000), where $U(x, y)$ is 15.033 k_BT, and the lower one is located at (x = 0.000, y = 0.039), where $U(x, y)$ is 16.650k_BT. Along the y axis, the maximum potential between these two saddle points is located at (x = 0.000, y = 0.574), where $U(x, y)$ is 27.024k_BT. See the Supporting Information for the cross-section of $U(x, y)$ at x = 0. The pathways connecting the reactant and product regions are separated into two distinct channels by the peak around (x = 0.000, y = 0.574). To categorize paths according to the positions where they cross the barrier between the reactant and product regions, we examine the intersection points between the path and the y axis (x_c = 0, y_c). If a path crosses the y axis multiple times, the last intersection point is used. The paths with y_c larger than 0.574 are tagged as in the UP channel, and the paths with y_c smaller than 0.574 are tagged as in the DOWN channel. In Figure 6a, we show two transition paths of a Brownian particle connecting the reactant and product regions. One transition path goes through the UP channel, and the other goes through the DOWN channel.

The TPS method is applied to sample the path ensembles connecting the reactant and product regions. In TPS simulations, the trial paths are generated by the "shooting" algorithm.[18] As previous researchers found,[49,51,52] like any conventional Monte Carlo (MC) simulations, TPS, which is a MC sampling in path space, can be trapped in local minima, namely channels. Possible solutions to this problem include combining the replica exchange algorithm with TPS[51−53] or applying different transition path sampling techniques.[49] Notice that the path channels in this example are analogous to the macrostate clusters in the previous two examples. The goal is to estimate the population percentage of the paths in each channel.

Here we show how to overcome the "trapping" problem by running independent parallel TPS at different Hamiltonian states and reweighting paths by Stratified-UWHAM. First we introduce a biasing potential to remove the peak which separates transition paths into channels

$$V(x, y)/(k_BT) = 1.25(9.5185 \exp\{-9[(y - 0.5741)^2 + x^2/3]\}) \quad (14)$$

In Figure 6b the contours of the potential $U(x,y) - \lambda V(x, y)$ with λ = 1.0 are plotted. As can be seen, at the λ = 1.0 state the peak around (x = 0, y = 0.574) is removed, and the two path channels are merged. See the Supporting Information for the cross-section of $U(x, y) - V(x, y)$ at x = 0. Then two sets of independent TPS simulations were run at λ = (0.0, 0.2, 0.4, 0.6, 0.8, 1.0) states. The initial paths of the first set of simulations are in the UP channel, and the initial paths of the second set are in the DOWN channel. Each TPS simulation generated 5 million paths connecting the reactant and product regions. At the λ = 0.0 state, no transitions of paths between the two channels were observed during the TPS simulations. In other words, at λ = 0.0 state, TPS simulations of paths started in the UP channel remain in the UP channel, while paths started in the DOWN channel remain there. The changes of y_c during each TPS simulation are shown in the Supporting Information. Then we applied conventional UWHAM, Stratified-UWHAM, and Stratified RE-SWHAM to estimate the probability percentage of the paths in the UP and DOWN channels. One of us (B.W.Z.) has applied the Weighted Ensemble (WE) algorithm to obtain the correct path ensemble for two-dimensional potentials like the one shown in Figure 6.[49,50] The WE results are used as the benchmark for this test. The simulation details for Langevin dynamics, TPS, and WE can be found in refs 54 and 23.

The red line in Figure 7a shows the population percentages of the paths in the DOWN channel at different λ-states obtained by the WE simulations. At the λ = 0.0 state, the paths in the DOWN channel make up ∼29.3% of the whole path ensemble. First we simply combined the data generated from the same λ-state and applied the conventional UWHAM to estimate the population percentage of the paths in each channel. It can be seen from Figure 7a that the conventional UWHAM estimate of the population percentage of the DOWN channel shows significant differences compared with the

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

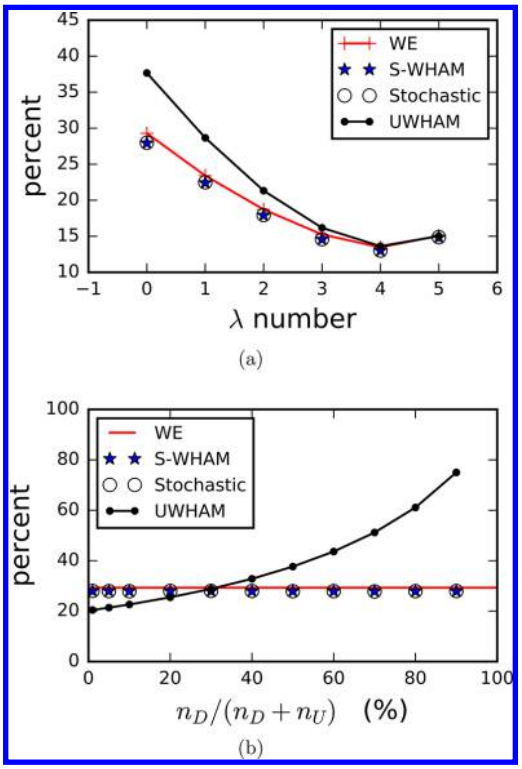


Figure 7. Population percentage of the paths in the DOWN channel. The red line is the benchmark — the results obtained by weighted ensemble (WE) simulations. The blue stars are the Stratified-UWHAM estimates when the data at the smallest three λ-states are clustered into UP and DOWN states, the black circles are the Stratified RE-SWHAM estimates, and the black line with dots is the conventional UWHAM estimates when the data at the locally equilibrated λ-states are simply combined as the input of that λ-state. (a) Comparison of the population percentages of the paths in the DOWN channel at each λ-state estimated by conventional UWHAM, Stratified-UWHAM, Stratified RE-SWHAM, and benchmark when the number of the paths in the UP and DOWN channels in the input path ensembles are equal. (b) Comparison of the population percentages of the paths in the DOWN channel at the λ = 0.0 state estimated by conventional UWHAM, Stratified-UWHAM, Stratified RE-SWHAM, and benchmark when the population ratio $n_D/(n_D + n_U)$ changes from 1% to 90%, where n_D is the number of paths in the DOWN channel at the λ = 0.0 state in the input path ensemble, and n_U is the number of paths in the UP channel at the λ = 0.0 state in the input path ensemble. As can be seen by comparing the two figures, only the Stratified-UWHAM and Stratified RE-SWHAM estimates agree with the benchmark, and the conventional UWHAM estimate for the λ = 0.0 state strongly depends on the population ratio $n_D/(n_D + n_U)$.

benchmark at the smaller λ values. The difference is negligible at the λ = 1.0 state but increases to 9% at the λ = 0.0 states. Then we applied the Stratified-UWHAM to analyze the path ensembles. For this case, the λ = 0.0, λ = 0.2, and λ = 0.4 states are considered to be the locally equilibrated λ-states, and each is split into two new λ-states. As can be seen from Figure 7a, the Stratified-UWHAM estimates of the population percentage of the paths in the DOWN channel at each λ-state agree with the WE results very well. The estimates obtained from the Stratified RE-SWHAM also match the benchmark results. In the Supporting Information, we list the numerical results and uncertainties of the population percentages of the paths in the DOWN channel estimated by Stratified-UWHAM, Stratified RE-SWHAM, and the benchmark.

When the simulations at the λ-states with a substantial barrier between the paths have not converged, the conventional UWHAM estimates of the probabilities of the UP and DOWN channel paths at these λ-states strongly depend on the number of UP and DOWN channel paths which are input to UWHAM because the conventional UWHAM always assumes the input data ensemble at each λ-state is independently drawn from the distribution described by eq 1 or eq 31. Therefore, the difference between the conventional UWHAM estimates and the true values at the λ = 0.0 state can be much larger than the case that the numbers of paths in the UP and DOWN channels generated at the λ = 0.0 state are equal (as shown in Figure 7a). To show this effect, we fixed the number of paths in the UP channel at the λ = 0.0 state in the input path ensemble n_U but changed the number of paths in the DOWN channel at the λ = 0.0 state in the input path ensemble n_D so that the population ratio $n_D/(n_D + n_U)$ ranges from 1% to 90%. Then these input path ensembles with different value of $n_D/(n_D + n_U)$ were fed to the conventional UWHAM, Stratified-UWHAM, and Stratified RE-SWHAM to estimate the population percentage of the paths in each channel at the λ = 0.0 state. The results are shown in Figure 7b. As expected, the conventional UWHAM estimates for the population percentage of the path in the DOWN channel at the λ = 0.0 state strongly depend on the ratio $n_D/(n_D + n_U)$ and change from 20% to 75% when $n_D/(n_D + n_U)$ changes from 1% to 90%, while the benchmark is ∼29.3%. On the other hand, the Stratified-UWHAM and Stratified RE-SWHAM estimates are independent of the initial condition (i.e., the ratio $n_D/(n_D + n_U)$) and agree with the benchmark.

To further compare path ensembles, we also measured the probability density of transition-event durations for each path ensemble. The definition of transition-event durations is the number of Brownian steps between the Brownian particle last leaving the reactant region and first arriving in the product region, namely the path length.[54−58] Figure 8a shows the probability density of transition-event durations of paths in each channel and overall path ensemble at the λ = 0.0 states obtained by WE simulations. As can been seen, although the paths in the DOWN channel are less favorable compared with the paths in the UP channel, their average path length is shorter. This makes sense because if a pathway goes through a steeper barrier, namely a less favorable path channel, the Brownian particle has less freedom to wander along the optimal pathway, which ends in a shorter average path length.[54] In Figure 8b, we compare the probability densities of transition-event durations at the λ = 0.0 state estimated by the conventional UWHAM, Stratified-UWHAM, and Stratified RE-SWHAM when the population ratio $n_D/(n_D + n_U)$ is 80%. As can be seen, the conventional UWHAM estimate shows a significant difference compared with the benchmark. However, the Stratified-UWHAM and Stratified RE-SWHAM estimates are indistinguishable, and both agree with the benchmark very well, which confirms that both Stratified-UWHAM and Stratified RE-SWHAM correctly estimate the weight of each individual path. In the Supporting Information, we show the comparisons of the probability density of transition-event durations estimated by Stratified-UWHAM, Stratified RE-SWHAM, and the benchmark at all λ-states when the number of paths in the UP and DOWN channels in the input path ensemble are equal.

3.4. Discussion. Stratified-UWHAM requires that the conformational space be coarse-grained. This can be done based on preliminary simulations or from biophysical knowl-

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

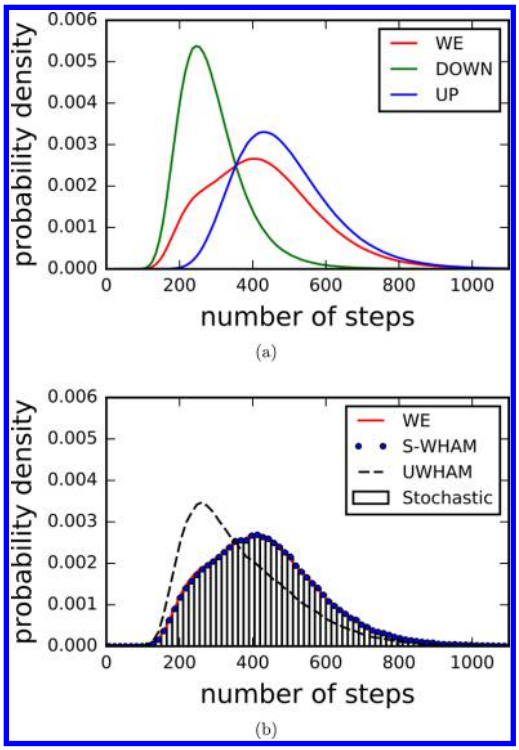


Figure 8. Probability densities of transition-event durations at the λ = 0.0 state. (a) The probability density of transition-event durations at the λ = 0.0 state obtained by WE simulations. The blue line is the probability density of transition-event durations of the paths in the UP channel, the green line is the probability density of transition-event durations of the paths in the DOWN channel, and the red line is the overall probability density of transition-event durations at the λ = 0.0 state. (b) Comparison of the probability density of transition-event durations at the λ = 0.0 state estimated by Stratified-UWHAM, Stratified RE-SWHAM, and the benchmark when the population ratio $n_D/(n_D + n_U)$ is 80%. The blue dots are the Stratified-UWHAM estimates, the bars are the Stratified RE-SWHAM estimates, and the dashed line is the conventional UWHAM estimates. The conventional UWHAM estimates show a significant difference compared with the benchmark, while Stratified-UWHAM and Stratified RE-SWHAM estimates agree with the benchmark very well.

edge, but a more general and practical method is to partition the conformational space following procedures used to construct Markov States Models. MSMs are a natural choice for the preparation of Stratified-UWHAM for the following reasons: MSMs build up a network which coarse-grains the free energy landscape. The states in the MSM network are defined based on structural (order parameters) and kinetic criteria. Each state in an MSM corresponds to a cluster of conformations that constitute a basin (or collection of basins) in the free energy landscape, and the transition rates between states in an MSM reflect the properties of the corresponding (free) energy barriers. The stratified-UWHAM S_1 and S_2 groups of λ-states can be determined by the following procedure:

- choose a set of λ-states as reference states to build the MSM using prior knowledge and/or run preliminary simulations, choosing, for reference, those biased simulations where the relaxation between the slowly equilibrating basins are enhanced.
- cluster the data from the other λ-states into MSM states using the same definition of MSMs used in the first step.
- identify disconnected macrostates or macrostate clusters based on ergodicity analyses for each λ-state in each of the biased simulations. One macrostate cluster may contain one or many basins.
- the biased λ-states whose macrostate clusters are fully connected are assigned to the S_1 group; the λ-states which include disconnected macrostate clusters are assigned to the S_2 group. For some problems, the most straightforward applications of the Stratified-UWHAM algorithm will fail when metastable basins merge or separate as the Hamiltonian function and/or thermodynamic parameters of the λ-states change. To account for this it may be necessary to build into the UWHAM stratification procedure more detailed information about the correspondence between basins at different λ-states.

Two maximum likelihood-based methods, the dynamic histogram analysis method (DHAM) and the general transition-based reweighting analysis method (TRAM),[59−62] were proposed recently to provide free energy estimates for multistate simulations when the simulations at some λ-states are only locally equilibrated. As we propose for Stratified-UWHAM, both DHAM and TRAM require building MSMs first for further analyses. In addition to providing estimates of equilibrium distributions, both DHAM and TRAM analysis methods provide estimates of the transition rates between states of the MSMs which are not accessible by the Stratified-UWHAM analysis. Here we comment on the three methods and explain some possible advantages and drawbacks of Stratified-UWHAM for estimating equilibrium populations.

DHAM calculates the estimates of transition rates between states of MSMs first. Then the equilibrium distributions are obtained by solving the eigenvalue equation for the transition matrix. Suppose there are n_b states in the MSM, then for the αth λ-state the transition matrix is $T^{(\alpha)}$, where the element $T_{ij}^{(\alpha)}$ represents the probability of the system transitioning from the ith state to the jth state during lag time Δt. The log likelihood function of observing $n_{ij}^{(\alpha)}$ transitions from the ith state to the jth state at the αth λ-state during the simulation is[59]

$$L^{(\alpha)} = \log \prod_{i=1}^{n_b} \prod_{j=1}^{n_b} (T_{ij}^{(\alpha)})^{n_{ij}^{(\alpha)}} = \sum_{i=1}^{n_b} \sum_{j=1}^{n_b} n_{ij}^{(\alpha)} \log(T_{ij}^{(\alpha)}) \tag{15}$$

DHAM supposes that the transition matrix element $T_{ij}^{(\alpha)}$ at the αth λ-states can be written as $T_{ij}^{(\alpha)} = f_j^{\alpha} c_{ij}^{(\alpha)} T_{ij}$, where $c_{ij}^{(\alpha)}$ is a bias factor, T_{ij} is the ijth elements of an unbiased transition matrix T, and f_j^{α} is a normalization factor. With this assumption, DHAM maximizes the likelihood function $L_D = \prod_{\alpha=1}^{M} L^{(\alpha)}$, where $L^{(\alpha)}$ is defined by eq 15. Notice that the transition probabilities at different λ-states are coupled by the bias factor. If the bias factors for the transition rates $\{c_{ij}^{\alpha}\}$ are known, DHAM provides better estimates of equilibrium populations than conventional UWHAM for multistate simulations when the simulations at some λ-states are far from being equilibrated.[59] However, the challenge of applying DHAM is that the bias factors $\{c_{ij}^{\alpha}\}$ are usually unknown and may be difficult to construct for arbitrary multistate simulations. In contrast, the analogous quantities in Stratified-UWHAM — the probabilities of observing a microstate at different λ-states $q_{\alpha}(u_{\gamma i})$ in eq 3 — are more readily obtained from the Hamiltonian and thermodynamic parameters of the multistate simulations.

In the TRAM method, the estimates of equilibrium distributions and transition rates of MSMs are calculated simultaneously. The maximum likelihood function of TRAM is a product of the maximum likelihood functions of binless

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

WHAM (population counts) and DHAM (transition counts).[62] Unlike Stratified-UWHAM, TRAM stratifies *every* λ-state into configuration states (macrostates) of MSMs. The local free energy of each configuration state at each λ-state is calculated during each iteration of TRAM analysis: the free energy differences between the same configuration states at different λ-states are calculated in a binless manner; the free energy differences between configuration states at each λ-state are calculated based on the transition counts and the detailed balance condition. Those calculations form multiple thermodynamic cycles like the one shown in Figure 2b. The optimal and consistent estimates of all the legs in the thermodynamic cycles, namely the free energy differences, and the transition rates are obtained simultaneously by maximizing the likelihood function. In ref 62 TRAM was applied to obtain the thermodynamic and kinetic information on a protein−ligand binding complex successfully, while the MBAR/UWHAM or WHAM analysis was found to be unfeasible or less efficient.

Stratified-UWHAM and TRAM each have strength and weakness. Stratified-UWHAM is an algorithm that focuses on equilibrium populations, not kinetics. The transition counts observed during the multistate simulations are not used to estimate the equilibrium distributions, and Stratified-UWHAM does not provide estimates for transition rates although there are methods which can infer transition rates from equilibrium distributions estimated from multicanonical simulations.[63−67] When we solve the Stratified-UWHAM equations, the λ-states in the S_1 group (fully connected λ-states) are *not* split into new λ-states so that the density of states obtained by Stratified-UWHAM is global (or globally normalized). Therefore, the existence of at least one λ-state in the S_1 group seems to be essential for applying Stratified-UWHAM. However, it is worth pointing out that this is not a requirement of Stratified-UWHAM. Suppose there is a system which has three macrostates. The simulations at one λ-state are approximately equilibrated between the first and the second macrostates, and the simulations at another λ-state are approximately equilibrated between the second and the third macrostates. If the sampled phase space of the second macrostate at these two λ-states are well overlapped,[68,69] these two λ-states together are equivalent to one approximately globally equilibrated λ-state. For such cases, either Stratified-UWHAM or Stratified RE-SWHAM can be used to obtain the global density of states. A practical criterion to validate the application of Stratified-UWHAM is that if Stratified RE-SWHAM is used to analyze the raw data, each replica shall have resampled every macrostate of every λ-state during the analysis. In other words, in Stratified RE-SWHAM, which is a multicanonical resampling analysis analogous to multicanonical simulations such as replica exchange, all the macrostates need to be fully connected when the data at all λ-states are combined in order to produce converged results.

On the other hand, as the name implies, TRAM is a transition-based reweighting analysis method. Because TRAM stratifies every λ-state, it does not depend on the population ratios of different states of the MSM at each λ-state, but approximately converged transition counts connecting states at each λ-state are essential for TRAM to obtain the global density of states. Note that unconverged transition counts can pollute the TRAM estimates, as unconverged population counts pollute the conventional UWHAM estimates as described previously in Section 3. Because each transition matrix element at each λ-state is an unknown parameter to be determined by the maximum likelihood algorithm, TRAM has thousands more variables to solve than Stratified-UWHAM. Further work on TRAM and Stratified-UWHAM may benefit from the development of a "population-plus-transition-based" reweighting algorithm which inherits the strengths of both methods.

4. CONCLUSION

We have developed a new analysis tool called Stratified-UWHAM to compute the density of states and free energies for data ensembles generated by multistate simulations when a subset of the simulations is only locally equilibrated, macrostate clusters may be disconnected at some λ-states, and their population estimates are far from equilibrium. To remove the computational bottleneck of Stratified-UWHAM, we developed a stochastic solver for the Stratified-UWHAM equations by extending the RE-SWHAM algorithm. As has been shown above, the Stratified-UWHAM equations can be solved in the form of UWHAM equations with an expanded set of λ-states, and the Stratified-UWHAM equations can be solved stochastically in the form of the original RE-SWHAM with a simple restraint introduced in the move procedure.

Stratified-UWHAM and Stratified RE-SWHAM have been applied to three model systems. First, we constructed the free energy surfaces of an alanine dipeptide molecule in vacuum by analyzing the data generated by two independent MD simulations of AlaD in vacuum starting from different macrostate clusters and the known equilibrium distributions of AlaD in implicit solvent which can be computed rapidly. Compared with Stratified-UWHAM and Stratified RE-SWHAM, the conventional UWHAM requires much longer MD simulations to produce estimates matching the benchmark within statistical error. Second, we studied the binding affinity of the β-cyclodextrin heptanoate complex by running two sets of independent MD simulations starting from different macrostates at 16 λ-states. Since the barrier between the "UP" and "DOWN" macrostates of this system is "infinitely" high at some λ-states, conventional UWHAM failed to estimate the equilibrium distribution at those λ-states correctly. However, the Stratified-UWHAM and Stratified RE-SWHAM estimates agree with the benchmark replica exchange simulation results very well. In the third example, we showed how to overcome the "trapping" problem of the transition path sampling algorithm by running TPS in a two-dimensional double well potential at multiple λ-states independently and using Stratified-UWHAM and Stratified RE-SWHAM to analyze the path ensemble. As far as we know, this is the first time the Onsager-Machlup action-based path sampling algorithm has been combined with a UWHAM type analysis tool to study kinetics.

Stratified-UWHAM requires that the conformational space be coarse-grained. For the three examples we discussed above, the coarse-graining was done based on our preliminary knowledge about the system. For an arbitrary problem, we proposed that one can partition the conformational space using Markov States Models and suggested a procedure to identify locally equilibrated λ-states and macrostate clusters. Features of Stratified-UWHAM were compared with DHAM and TRAM. Compared with TRAM and DHAM, one drawback of the current version of Stratified-UWHAM is the requirement of manually determining locally equilibrated λ-states and macrostate clusters for each λ-state. However, this is necessary in order to avoid feeding UWHAM biased information which can pollute the estimates of the density of states. Algorithms to

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

combine states of MSMs into macrostates and identify disconnected macrostate clusters based on raw simulation data can be automated.[25] We proposed a criterion to validate the application of Stratified-UWHAM: if Stratified RE-SWHAM is used to analyze the raw data, each replica shall have resampled every macrostate of every λ-state during the analysis. Unlike DHAM or TRAM, Stratified-UWHAM does not require the bias factors for transition rates at different λ-states or approximately converged transition counts between states of MSMs to obtain equilibrium distributions. Last but not least, the stochastic version of Stratified-UWHAM, Stratified RE-SWHAM, provides a practical analysis tool for multistate simulations on massive computational grids.[14]

APPENDIX

A. Stratified RE-SWHAM

Stratified RE-SWHAM is a resampling technique we developed to solve the Stratified-UWHAM equations stochastically by using the replica exchange simulation protocol (see Figure 1). Like RE simulations, at the end of each cycle, the observation associated with each replica is recorded as the output of Stratified RE-SWHAM. Here we use the alanine dipeptide problem as an example to show that the output of Stratified RE-SWHAM for a λ-state in the S_2 set which contains disconnected macrostate clusters is the estimate of the equilibrium distribution of that λ-state. Therefore, the splits of locally equilibrated λ-states are not necessary.

In the AlaD problem, the implicit solvent state (I state) is a fully-connected λ-state; the vacuum state (V state) is a locally equilibrated λ-state with two disconnected macrostate clusters. Suppose the replica at the V state is resampling the observations of the A macrostate cluster by the move procedure. During Stratified RE-SWHAM, to switch the replica at the V state to resample the observations of the B macrostate cluster requires (i) the other replica at the I state is associated with an observation in the B macrostate cluster, (ii) an exchange attempt of these two replicas is accepted. Therefore, the probability of switching the replica at the V state from resampling the A macrostate cluster to resampling the B macrostate cluster is

$$p_I^{(B)}\langle\Psi(U_I^{(A)} + U_V^{(B)} - U_V^{(A)} - U_I^{(B)})\rangle_{V^{(A)}I^{(B)}} \tag{16}$$

where $p_I^{(B)}$ is the probability that the observation associated with the replica at the I state belongs to the B macrostate cluster. $U_V^{(X)}$ and $U_I^{(X)}$ are the potential energies of an observation of the "X" macrostate cluster at the V state and the I state, respectively. Notice they are the energy values of the same microstate at different λ-states and suppose the energies are in units of k_BT. Ψ is the Metropolis function to determine the acceptance ratio[70]

$$\Psi(x) = \min(1, \exp[-x]) \tag{17}$$

The angle brackets and subscript $V^{(A)}I^{(B)}$ represent the ensemble average when the observation associated with the replica at the V state belongs to the A macrostate cluster and the observation associated with the replica at the I state belongs to the B macrostate cluster.

Similarly, if the replica at the V state is resampling the B macrostate cluster, the probability of switching the replica to resample the A macrostate cluster is

$$p_I^{(A)}\langle\Psi(U_I^{(B)} + U_V^{(A)} - U_V^{(B)} - U_I^{(A)})\rangle_{V^{(B)}I^{(A)}} \tag{18}$$

Because of the requirement of detailed balance, the forward and backward currents of a replica moving between the A and B macrostate clusters at the V state are equal, which yields

$$\begin{aligned} &p_V^{(A)}p_I^{(B)}\langle\Psi(U_I^{(A)} + U_V^{(B)} - U_V^{(A)} - U_I^{(B)})\rangle_{V^{(A)}I^{(B)}} \\ &= p_V^{(B)}p_I^{(A)}\langle\Psi(U_I^{(B)} + U_V^{(A)} - U_V^{(B)} - U_I^{(A)})\rangle_{V^{(B)}I^{(A)}} \end{aligned} \tag{19}$$

where $p_V^{(X)}$ is the probability that the observation associated with the replica at the V state belongs to the "X" macrostate cluster.

The Metropolis exchange criterion in the Stratified RE-SWHAM analyses satisfies

$$\begin{aligned} &\Psi(U_I^{(A)} + U_V^{(B)} - U_V^{(A)} - U_I^{(B)})\exp(-U_V^{(A)})\exp(-U_I^{(B)}) \\ &= \Psi(U_I^{(B)} + U_V^{(A)} - U_V^{(B)} - U_I^{(A)})\exp(-U_V^{(B)})\exp(-U_I^{(A)}) \end{aligned} \tag{20}$$

Integrating the equation over the configuration space leads to

$$\begin{aligned} &Q_V^{(A)}Q_I^{(B)}\frac{\int\Psi(U_I^{(A)} + U_V^{(B)} - U_V^{(A)} - U_I^{(B)})\exp(-U_V^{(A)})\exp(-U_I^{(B)})dq_V^{(A)}dq_I^{(B)}}{Q_V^{(A)}Q_I^{(B)}} \\ &= Q_V^{(B)}Q_I^{(A)}\frac{\int\Psi(U_I^{(B)} + U_V^{(A)} - U_V^{(B)} - U_I^{(A)})\exp(-U_V^{(B)})\exp(-U_I^{(A)})dq_V^{(B)}dq_I^{(A)}}{Q_V^{(B)}Q_I^{(A)}} \end{aligned} \tag{21}$$

where $Q_Y^{(X)}$ is the canonical configurational integral of the "X" macrostate cluster at the "Y" λ-state

$$Q_Y^{(X)} = \int\exp(-U_Y^{(X)})dq_Y^{(X)} \tag{22}$$

Eq 21 can be rewritten as

$$\frac{\langle\Psi(U_I^{(A)} + U_V^{(B)} - U_V^{(A)} - U_I^{(B)})\rangle_{V^{(A)}I^{(B)}}}{\langle\Psi(U_I^{(B)} + U_V^{(A)} - U_V^{(B)} - U_I^{(A)})\rangle_{V^{(B)}I^{(A)}}} = \frac{Q_V^{(B)}Q_I^{(A)}}{Q_V^{(A)}Q_I^{(B)}} \tag{23}$$

Namely, the log ratio of the acceptance probability $\langle\Psi(U_I^{(A)} + U_V^{(B)} - U_V^{(A)} - U_I^{(B)})\rangle_{V^{(A)}I^{(B)}}$ over $\langle\Psi(U_I^{(B)}+U_V^{(A)} - U_V^{(B)} - U_I^{(A)})\rangle_{V^{(B)}I^{(A)}}$ provides the estimate of the free energy difference between the two vertical legs in Figure 2b.

Combining eqs 19 and 23 yields

$$\begin{aligned} -k_BT\log\left(\frac{p_V^{(B)}}{p_V^{(A)}}\right) &= -k_BT\log\left(\frac{p_I^{(B)}}{p_I^{(A)}}\right) - k_BT\log\left(\frac{Q_I^{(A)}}{Q_V^{(A)}}\right) - k_BT\log\left(\frac{Q_V^{(B)}}{Q_I^{(B)}}\right) \\ &= \Delta F_{A\to B}^{I} + \Delta F_A^{V\to I} + \Delta F_B^{I\to V} \\ &= \Delta F_{A\to B}^{V} \end{aligned} \tag{24}$$

Eq 24 shows the ratio of $p_V^{(B)}$ over $p_V^{(A)}$ provides the estimate of the upper leg in Figure 2b. Because at the end of each cycle, the observation associated with the replica at the V state is recorded as the output of Stratified RE-SWHAM at the V state (see Figure 1), $(p_V^{(A)}/p_V^{(B)})$ equals the population ratio of the A macrostate cluster over the B macrostate cluster in the output of the V state. In other words, the output of Stratified RE-SWHAM at the V state is the estimate of the equilibrium distribution of the V state.

There is another subtle difference between RE-SWHAM with an expanded set of λ-states and Stratified RE-SWHAM. During the analysis of RE-SWHAM with an expanded set of λ-states, an observation in a macrostate cluster of a locally equilibrated λ-state can possibly be exchanged only with an observation in the same macrostate cluster at another λ-state

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

because of the infinite barrier covering the outside of that macrostate cluster at the corresponding expanded λ-state. Therefore, the number of observations in a macrostate cluster of a locally equilibrated λ-state stays as a constant. During the analysis of Stratified RE-SWHAM, because an observation in a macrostate cluster of a locally equilibrated λ-state is allowed to be exchanged with any observation at another λ-state if the exchange attempt is accepted (see Figure 1), the number of observations in a macrostate cluster of a locally equilibrated λ-state fluctuates by ± 1. However, if the total number of observations in each macrostate cluster at each λ-state is large, such fluctuations become negligible.

B. Onsager-Machlup Action-Based Path Ensemble

In Section 3.3, we apply Stratified-UWHAM to analyze the path ensembles of a Brownian particle moving in a two-dimensional double well potential. The stochastic dynamics of the Brownian particle in this two-dimensional space is governed by the overdamped Langevin equation

$$\frac{dx}{dt} = \frac{F_x}{\gamma} + R_x(t)$$
$$\frac{dy}{dt} = \frac{F_y}{\gamma} + R_y(t) \tag{25}$$

where F_x and F_y are the forces acting on the particle, γ is the friction constant, and $R_x(t)$ and $R_y(t)$ are the thermal noise taken from Gaussian functions with zero mean and correlation

$$\langle R(t)R(t')\rangle = \left(\frac{2k_B t}{\gamma}\right)\delta(t-t') = 2D\delta(t-t') \tag{26}$$

$D = k_B T/\gamma$ in eq 26 is the diffusion constant.

Given the two-dimensional potential $U(x, y)$, the probability of a $(N-1)$-steps path connecting the reactant region and the product region is

$$P_{\text{path}}((x_0, y_0), (x_1, y_1), \cdots, (x_N, y_N)) = \exp\left\{-\frac{U(x_0, y_0)}{k_B T}\right\} \times \left[\prod_{i=0}^{N-1} p(x_i, x_{i+1}; U)p(y_i, y_{i+1}; U)\right] \tag{27}$$

$p(x_i, x_{i+1}; U)$ and $p(y_i, y_{i+1}; U)$ in eq 27 are the single-step transition probabilities

$$p(x_i, x_{i+1}; U) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{[x_{i+1} - x_i - (1/2)(\partial U/\partial x_i)(2D\Delta t)]^2}{2\sigma^2}\right\}$$
$$p(y_i, y_{i+1}; U) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{[y_{i+1} - y_i - (1/2)(\partial U/\partial y_i)(2D\Delta t)]^2}{2\sigma^2}\right\} \tag{28}$$

where

$$\sigma^2 = 2\left(\frac{k_B T}{\gamma}\right)\Delta t = 2D\Delta t \tag{29}$$

and Δt is the time interval of a single step. By combining eq 27 and eq 28, the probability of a path can be written as a single exponential function

$$P_{\text{path}} = \exp\left\{\left[-\frac{2U(x_0, y_0)}{\gamma} - 2D(N-1)\log(2\pi\sigma^2) - \sum_{i=0}^{N-1}\frac{[x_{i+1} - x_i - (1/2)(\partial U/\partial x_i)(2D\Delta t)]^2}{2(\Delta t)^2} - \sum_{i=0}^{N-1}\frac{[y_{i+1} - y_i - (1/2)(\partial U/\partial y_i)(2D\Delta t)]^2}{2(\Delta t)^2}\right]/(2D)\right\}$$
$$= \exp\left\{\frac{-A[x(t), y(t), U(x, y)]}{2D}\right\} \tag{30}$$

where $A[x(t), y(t), U(x,y)\ ]$ is called the Onsager-Machlup action "functional".[71] Compared with the probability of a microstate of a mechanical system governed by the canonical ensemble, the action functional of a path is analogous to the potential energy of a microstate. With this understanding, many enhanced sampling methods and analysis tools which have been developed to explore the conformational space such as replica exchange and UWHAM can be applied straightforwardly to the transition path space.[53,72,73] The transition path sampling (TPS) method is a MC simulation in the path space to draw pathway X_i^α:$i = 1, \cdots, N_\alpha$ according to the distribution

$$P_\alpha(X_i^\alpha) \sim \frac{\exp\{-A[x_{\alpha i}(t), y_{\alpha i}(t), U(x, y)]/(2D)\}}{Z_\alpha} \tag{31}$$

where Z_α is the normalizing constant (analogous to the partition function of a canonical ensemble) of the αth λ-state.[17,18]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b00651.

Simulation details of AlaD in vacuum and in implicit solvent; convergence of Stratified-UWHAM estimates of the free energy difference between the A and B macrostate clusters of AlaD in vacuum; orientation of heptanoate in each independent MD simulation; population percentages of the DOWN macrostate of the β-cyclodextrin heptanoate complex at each λ-state estimated by Stratified-UWHAM, Stratified RE-SWHAM, and RE simulations; probability density of binding energies of the β-cyclodextrin heptanoate complex at each λ-state estimated by Stratified-UWHAM, Stratified RE-SWHAM, and RE simulations; cross-section of the potential function $(U(x,y) - \lambda V(x,y))$ at $x = 0$; change of the intersection point $(x_c = 0, y_c)$ in each TPS simulation; population percentages of the paths in the DOWN channel estimated by Stratified-UWHAM, Stratified RE-SWHAM, and WE simulations; probability density of transition-event durations at each λ-state estimated by Stratified-UWHAM, Stratified RE-SWHAM, and WE simulations (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ronlevy@temple.edu.

ORCID

Bin W. Zhang: 0000-0003-3007-4900

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Ronald M. Levy: 0000-0001-8696-5177

Funding

This work was supported by NIH grant (GM30580), NSF grant (1665032), and by an NIH computer equipment grant OD020095. This work also used Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation (ACI-1053575).

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Zwier, M. C.; Chong, L. T. Reaching Biological Timescales with All-Atom Molecular Dynamics Simulations. *Curr. Opin. Pharmacol.* **2010**, *10*, 745−752.
(2) Zuckerman, D. M. Equilibrium Sampling in Biomolecular Simulations. *Annu. Rev. Biophys.* **2011**, *40*, 41−62.
(3) Makarov, D. E. *Single Molecule Science: Physical Principles and Models*; CRC Press: Boca Raton, FL, 2015; Chapter 5, p 59.
(4) Dai, W.; Sengupta, A. M.; Levy, R. M. First Passage Times, Lifetimes, and Relaxation Times of Unfolded Proteins. *Phys. Rev. Lett.* **2015**, *115*, 048101.
(5) Shirts, M.; Pande, V. S. COMPUTING: Screen Savers of the World Unite! *Science* **2000**, *290*, 1903−1904.
(6) Anderson, J. A.; Lorenz, C. D.; Travesset, A. General Purpose Molecular Dynamics Simulations Fully Implemented on Graphics Processing Units. *J. Comput. Phys.* **2008**, *227*, 5342−5359.
(7) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334*, 517−520.
(8) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.
(9) Northrup, S. H.; Pear, M. R.; Lee, C. Y.; McCammon, J. A.; Karplus, M. Dynamical Theory of Activated Processes in Globular Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **1982**, *79*, 4035−4039.
(10) Dickson, A.; Dinner, A. R. Enhanced Sampling of Non-equilibrium Steady States. *Annu. Rev. Phys. Chem.* **2010**, *61*, 441−459.
(11) Swendsen, R. H.; Wang, J.-S. Replica Monte Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.* **1986**, *57*, 2607−2609.
(12) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.
(13) Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. Simulating Replica Exchange Simulations of Protein Folding with a Kinetic Network Model. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 15340−15345.
(14) Xia, J.; Flynn, W. F.; Gallicchio, E.; Zhang, B. W.; He, P.; Tan, Z.; Levy, R. M. Large-Scale Asynchronous and Distributed Multi-dimensional Replica Exchange Molecular Simulations and Efficiency Analysis. *J. Comput. Chem.* **2015**, *36*, 1772−1785.
(15) Zhang, B. W.; Dai, W.; Gallicchio, E.; He, P.; Xia, J.; Tan, Z.; Levy, R. M. Simulating Replica Exchange: Markov State Models, Proposal Schemes, and the Infinite Swapping Limit. *J. Phys. Chem. B* **2016**, *120*, 8289−8301.
(16) Yu, T.-Q.; Lu, J.; Abrams, C. F.; Vanden-Eijnden, E. Multiscale Implementation of Infinite-Swap Replica Exchange Molecular Dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 11744−11749.
(17) Pratt, L. R. A Statistical Method for Identifying Transition States in High Dimensional Problems. *J. Chem. Phys.* **1986**, *85*, 5045−5048.
(18) Dellago, C.; Bolhuis, P. G.; Chandler, D. Efficient Transition Path Sampling: Application to Lennard-Jones Cluster Rearrangements. *J. Chem. Phys.* **1998**, *108*, 9236−9245.
(19) Wang, F.; Landau, D. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Phys. Rev. Lett.* **2001**, *86*, 2050−2053.
(20) E, W.; Ren, W.; Vanden-Eijnden, E. String Method for the Study of Rare Events. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2002**, *66*, 052301.
(21) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.
(22) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. Replica Exchange with Solute Tempering: a Method for Sampling Biological Systems in Explicit Water. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 13749−13754.
(23) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. Efficient and Verified Simulation of a Path Ensemble for Conformational Change in a United-Residue Model of Calmodulin. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 18043−18048.
(24) Zheng, L.; Chen, M.; Yang, W. Random Walk in Orthogonal Space to Achieve Efficient Free-Energy Simulation of Complex Systems. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 20227−20232.
(25) *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation (Advances in Experimental Medicine and Biology)*; Bowman, G. R., Pande, V. S., Noe, F., Eds.; Springer: Dordrecht, Heidelberg, New York, London, 2013.
(26) Ferrenberg, A.; Swendsen, R. Optimized Monte Carlo Data Analysis. *Phys. Rev. Lett.* **1989**, *63*, 1195−1198.
(27) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13*, 1011−1021.
(28) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. Multidimensional Free-Energy Calculations Using the Weighted Histogram Analysis Method. *J. Comput. Chem.* **1995**, *16*, 1339−1350.
(29) Roux, B. The Calculation of the Potential of Mean Force Using Computer-simulations. *Comput. Phys. Commun.* **1995**, *91*, 275−282.
(30) Bartels, C.; Karplus, M. Multidimensional Adaptive Umbrella Sampling: Applications to Main Chain and Side Chain Peptide Conformations. *J. Comput. Chem.* **1997**, *18*, 1450−1462.
(31) Tan, Z. On a Likelihood Approach for Monte Carlo Integration. *J. Am. Stat. Assoc.* **2004**, *99*, 1027−1036.
(32) Gallicchio, E.; Andrec, M.; Felts, A. K.; Levy, R. M. Temperature Weighted Histogram Analysis Method, Replica Exchange, and Transition Paths. *J. Phys. Chem. B* **2005**, *109*, 6722−6731.
(33) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations. *J. Chem. Theory Comput.* **2007**, *3*, 26−41.
(34) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.
(35) Kim, J.; Keyes, T.; Straub, J. E. Communication: Iteration-Free, Weighted Histogram Analysis Method in Terms of Intensive Variables. *J. Chem. Phys.* **2011**, *135*, 061103.
(36) Tan, Z.; Gallicchio, E.; Lapelosa, M.; Levy, R. M. Theory of Binless Multi-State Free Energy Estimation with Applications to Protein-Ligand Binding. *J. Chem. Phys.* **2012**, *136*, 144102.
(37) Zhu, F.; Hummer, G. Convergence and Error Estimation in Free Energy Calculations Using the Weighted Histogram Analysis Method. *J. Comput. Chem.* **2012**, *33*, 453−465.
(38) Law, S. M.; Ahlstrom, L. S.; Panahi, A.; Brooks, C. L., III Hamiltonian Mapping Revisited: Calibrating Minimalist Models to Capture Molecular Recognition by Intrinsically Disordered Proteins. *J. Phys. Chem. Lett.* **2014**, *5*, 3441−3444.
(39) Meng, Y.; Roux, B. Efficient Determination of Free Energy Landscapes in Multiple Dimensions from Biased Umbrella Sampling Simulations Using Linear Regression. *J. Chem. Theory Comput.* **2015**, *11*, 3523−3529.
(40) Zhang, C.; Lai, C.-L.; Pettitt, B. M. Accelerating the Weighted Histogram Analysis Method by Direct Inversion in the Iterative Subspace. *Mol. Simul.* **2016**, *42*, 1079−1089.
(41) Thiede, E. H.; Koten, B. V.; Weare, J.; Dinner, A. R. Eigenvector Method for Umbrella Sampling Enables Error Analysis. *J. Chem. Phys.* **2016**, *145*, 084115.
(42) Ding, X.; Vilseck, J. Z.; Hayes, R. L.; Brooks, C. L., III Gibbs Sampler-Based λ-Dynamics and Rao-Blackwell Estimator for Alchem-

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

ical Free Energy Calculation. *J. Chem. Theory Comput.* **2017**, *13*, 2501−2510.
(43) Zhang, B. W.; Xia, J.; Tan, Z.; Levy, R. M. A Stochastic Solution to the Unbinned WHAM Equations. *J. Phys. Chem. Lett.* **2015**, *6*, 3834−3840.
(44) Tan, Z.; Xia, J.; Zhang, B. W.; Levy, R. M. Locally Weighted Histogram Analysis and Stochastic Solution for Large-Scale Multi-State Free Energy Estimation. *J. Chem. Phys.* **2016**, *144*, 034107.
(45) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins: Struct., Funct., Genet.* **2004**, *55*, 383−394.
(46) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1−2*, 19−25.
(47) Deng, N.; Zhang, B. W.; Levy, R. M. Connecting Free Energy Surfaces in Implicit and Explicit Solvent: An Efficient Method To Compute Conformational and Solvation Free Energies. *J. Chem. Theory Comput.* **2015**, *11*, 2868−2878.
(48) Gallicchio, E.; Levy, R. M. AGBNP: An Analytic Implicit Solvent Model Suitable for Molecular Dynamics Simulations and High-Resolution Modeling. *J. Comput. Chem.* **2004**, *25*, 479−499.
(49) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. Weighted Ensemble Path Sampling for Multiple Reaction Channels. *arXiv:0902.2772v1*, 2009.
(50) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. The "Weighted Ensemble" Path Sampling Method is Statistically Exact for a Broad Class of Stochastic Processes and Binning Procedures. *J. Chem. Phys.* **2010**, *132*, 054107.
(51) Bolhuis, P. G. Rare Events via Multiple Reaction Channels Sampled by Path Replica Exchange. *J. Chem. Phys.* **2008**, *129*, 114108.
(52) Fujisaki, H.; Shiga, M.; Kidera, A. Onsager-Machlup Action-Based Path Sampling and Its Combination with Replica Exchange for Diffusive and Multiple Pathways. *J. Chem. Phys.* **2010**, *132*, 134101.
(53) Vlugt, T. J. H.; Smit, B. On the Efficient Sampling of Pathways in the Transition Path Ensemble. *PhysChemComm* **2001**, *4*, 11.
(54) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. Transition-Event Durations in One-Dimensional Activated Processes. *J. Chem. Phys.* **2007**, *126*, 074504.
(55) Chaudhury, S.; Makarov, D. E. A Harmonic Transition State Approximation for the Duration of Reactive Events in Complex Molecular Rearrangements. *J. Chem. Phys.* **2010**, *133*, 034118.
(56) Hawk, A. T.; Konda, S. S. M.; Makarov, D. E. Computation of Transit Times Using the Milestoning Method with Applications to Polymer Translocation. *J. Chem. Phys.* **2013**, *139*, 064101.
(57) Makarov, D. E. Computational and Theoretical Insights into Protein and Peptide Translocation. *Protein Pept. Lett.* **2014**, *21*, 217−226.
(58) Pollak, E. Transition Path Time Distribution and the Transition Path Free Energy Barrier. *Phys. Chem. Chem. Phys.* **2016**, *18*, 28872−28882.
(59) Rosta, E.; Hummer, G. Free Energies from Dynamic Weighted Histogram Analysis Using Unbiased Markov State Model. *J. Chem. Theory Comput.* **2015**, *11*, 276−285.
(60) Mey, A. S. J. S.; Wu, H.; Noé, F. xTRAM: Estimating Equilibrium Expectations from Time-Correlated Simulation Data at Multiple Thermodynamic States. *Phys. Rev. X* **2014**, *4*, 041018.
(61) Wu, H.; Mey, A. S. J. S.; Rosta, E.; Noé, F. Statistically Optimal Analysis of State-Discretized Trajectory Data from Multiple Thermodynamic States. *J. Chem. Phys.* **2014**, *141*, 214106.
(62) Wu, H.; Paul, F.; Wehmeyer, C.; Noé, F. Multiensemble Markov Models of Molecular Thermodynamics and Kinetics. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, E3221−E3230.
(63) Zheng, W.; Gallicchio, E.; Deng, N.; Andrec, M.; Levy, R. M. Kinetic Network Study of the Diversity and Temperature Dependence of Trp-Cage Folding Pathways: Combining Transition Path Theory with Stochastic Simulations. *J. Phys. Chem. B* **2011**, *115*, 1512−1523.
(64) Deng, N.; Zheng, W.; Gallicchio, E.; Levy, R. M. Insights into the Dynamics of HIV-1 Protease: A Kinetic Network Model Constructed from Atomistic Simulations. *J. Am. Chem. Soc.* **2011**, *133*, 9387−9394.
(65) Han, W.; Schulten, K. Fibril Elongation by $A\beta_{17-42}$: Kinetic Network Analysis of Hybrid-Resolution Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **2014**, *136*, 12450−12460.
(66) Dixit, P. D.; Dill, K. A. Inferring Microscopic Kinetic Rates from Stationary State Distributions. *J. Chem. Theory Comput.* **2014**, *10*, 3002−3005.
(67) Dixit, P. D.; Jain, A.; Stock, G.; Dill, K. A. Inferring Transition Rates of Networks from Populations in Continuous-Time Markov Processes. *J. Chem. Theory Comput.* **2015**, *11*, 5464−5472.
(68) Wu, D.; Kofke, D. A. Phase-Space Overlap Measures. I. Fail-Safe Bias Detection in Free Energies Calculated by Molecular Simulation. *J. Chem. Phys.* **2005**, *123*, 054103.
(69) Wu, D.; Kofke, D. A. Phase-Space Overlap Measures. II. Design and Implementation of Staging Methods for Free-Energy Calculations. *J. Chem. Phys.* **2005**, *123*, 084109.
(70) Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J. Comput. Phys.* **1976**, *22*, 245−268.
(71) Wiegel, F. *Introduction to Path-integral Methods in Physics and Polymer Science*; World Scientific Publishing Co Pre Ltd.: Singapore, Philadelphia, 1986.
(72) Minh, D. D. L.; Chodera, J. D. Optimal Estimators and Asymptotic Variances for Nonequilibrium Path-Ensemble Averages. *J. Chem. Phys.* **2009**, *131*, 134110.
(73) Chodera, J. D.; Swope, W. C.; Noé, F.; Prinz, J.-H.; Shirts, M. R.; Pande, V. S. Dynamical Reweighting: Improved Estimates of Dynamical Properties from Simulations at Multiple Temperatures. *J. Chem. Phys.* **2011**, *134*, 244107.

DOI: 10.1021/acs.jctc.7b00651
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX