Simulating rare events using a weighted ensemble-based string method

Joshua L. Adelman and Michael Grabe

Citation: J. Chem. Phys. **138**, 044105 (2013); doi: 10.1063/1.4773892
View online: http://dx.doi.org/10.1063/1.4773892
View Table of Contents: http://jcp.aip.org/resource/1/JCPSA6/v138/i4
Published by the American Institute of Physics.

Simulating rare events using a weighted ensemble-based string method

Joshua L. Adelman[1,a)] and Michael Grabe[1,2,b)]
[1]*Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA*
[2]*Department of Computational & Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA*

(Received 12 October 2012; accepted 17 December 2012; published online 24 January 2013)

We introduce an extension to the weighted ensemble (WE) path sampling method to restrict sampling to a one-dimensional path through a high dimensional phase space. Our method, which is based on the finite-temperature string method, permits efficient sampling of both equilibrium and non-equilibrium systems. Sampling obtained from the WE method guides the adaptive refinement of a Voronoi tessellation of order parameter space, whose generating points, upon convergence, coincide with the principle reaction pathway. We demonstrate the application of this method to several simple, two-dimensional models of driven Brownian motion and to the conformational change of the nitrogen regulatory protein C receiver domain using an elastic network model. The simplicity of the two-dimensional models allows us to directly compare the efficiency of the WE method to conventional brute force simulations and other path sampling algorithms, while the example of protein conformational change demonstrates how the method can be used to efficiently study transitions in the space of many collective variables. © *2013 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4773892]

I. INTRODUCTION

Molecular simulations can provide deep insight into the mechanisms of physical processes, in part, because they inherently possess a spatial and temporal resolution that is unmatched by most experimental techniques. Unfortunately, many physical and biological processes such as chemical reactions, nucleation, protein conformational changes, and ligand binding occur on timescales that are inaccessible to conventional brute-force simulation methods. Due to this shortcoming, there has been broad interest in developing methods that augment conventional simulations to allow them to capture rare events in a reasonable time frame,[1–6] several of which are reviewed in Refs. 7–9. Almost all of these methods rely on enhancing sampling in a reduced set of collective coordinates that span the important regions of the high dimensional phase space. The computational effort is thereby focused on sampling transition regions, which would otherwise be visited infrequently, if at all, in conventional simulations. The success of rare event sampling often hinges on the particular progress coordinate or order parameter used to discriminate movement along the transition, and choosing the proper progress coordinate is a non-trivial task. In fact, it is likely that most processes can only be described by multiple progress coordinates, which further complicates identification of the appropriate pathway through phase space. The need to incorporate additional progress coordinates also drastically increases the computational demand since the cost scales like the power of the number of progress coordinates.

Even if a transition occurs via a convoluted set of steps requiring multiple progress coordinates to describe the process, it is generally thought that the reaction can be meaningfully represented by a chain of connected nodes or beads (referred to as images) that define a path.[10–13] Methods built around this idea provide an elegant solution to the increased phase space needed to explore multiple directions because they are able to track an arbitrary number of progress coordinates while restraining the sampling to effectively one dimension. One such realization of this approach is the string method.[13] Since its advent, a number of variants of the string method have been developed to evaluate the transition pathways of complex systems.[14,15] Additionally, the basic framework of the string method has been integrated with a variety of other sampling procedures.[16–22] While string methods have been applied to a diverse set of problems,[23–25] reactions that occur via many, conformationally distinct pathways[26,27] may be poorly suited for string methods. Studies have shown, however, as we do here, that when the transition is confined to a few reaction channels, which can be explicitly accounted for, string methods still perform well.

Here, we merge a rare event sampling method known as the weighted ensemble (WE) method[28] with a string method, and we show that this combined approach performs both accurately and efficiently. WE sampling is a rigorous method for sampling both equilibrium and non-equilibrium systems even in the presence of long-lived intermediate states.[29–31] From a practical perspective, WE sampling is easy to implement and straightforward to parallelize across large computational resources. The WE method has been previously used with Brownian dynamics simulations to study protein binding[28,32] and protein folding,[33] and it has been used with explicit solvent molecular dynamics to investigate molecular association events.[34] Our group, and others, have also extended the method to examine large-scale conformational transitions in biomolecular systems using coarse-grained models.[26,27,29]

a)Electronic mail: jla65@pitt.edu.
b)Electronic mail: mdgrabe@pitt.edu.

Augmenting WE sampling with a string method yields not only the transition pathway represented by the converged string, but also a dynamically exact representation of the path ensemble along the string, from which steady-state distributions and kinetic information can be extracted from a single simulation.

We illustrate the WE-based string method by applying it to several example systems of varying degrees of complexity. First, we examine two different nonequilibrium steady-state processes. The first of these is a Brownian particle in a unidirectional flow on a periodic two-dimensional surface.[19] Second, we examine a driven Brownian particle on a two-dimensional potential surface with intermediate metastable states along distinct forward and backward pathways.[35] Finally, we apply the method to the equilibrium conformational transition of the nitrogen regulatory protein C receiver domain using a two-state elastic network model.[16,18] All three systems have been previously studied with other string methods, facilitating a direct comparison with the WE-based algorithm. Moreover, the low dimensionality of the first two examples makes it possible to rigorously sample each system with conventional simulation to test the accuracy of our method and serve as an additional benchmark for comparing specifically against the nonequilibrium umbrella sampling (NEUS) method.[6,19,35] Both WE and NEUS methods are able to calculate steady-state rates and distributions, even in the presence of long-lived intermediates, and the clarity of the framework developed in Refs. 19 and 35 for comparing NEUS to conventional sampling has enabled us to include WE in this comparison. In all cases, the WE-based string method obtains high-fidelity estimates of steady-state distributions and rates using comparable, or less, computational resources than other string based methods. Thus, the procedure that we present here lays the groundwork for applying WE sampling to a broad range of problems in which many progress coordinates are required to efficiently partition phase space along a transition pathway.

II. METHODS

A. Weighted ensemble path sampling

Weighted ensemble sampling is a general and rigorous method for simulating rare events.[28,29] This is accomplished via a resampling protocol that partitions the progress coordinate space of the system into non-overlapping bins that can be defined in an arbitrary number of dimensions.[30] Here, resampling refers to a scheme that generates an alternative, but equivalent, statistical sample of a system's phase space, which in the context of WE sampling is accomplished as follows.

Multiple replicas of the system are initiated from some initial distribution of conformations. Each replica is assigned a weight, such that the sum of the weights of all of the replicas in the system is unity. In the simplest case this initial distribution of replicas may consist of N_{rep} replicas of equal weight with identical coordinates, but randomly assigned velocities. Each replica is then simulated independently for a short time interval τ during which it is allowed to explore conformational space following its natural dynamics. At the end of this interval, every replica is assigned to a bin based upon its instantaneous coordinates at the end of the time interval.

While assigning the configurations of the system into bins could be done using the full configurational space of the system, in practice the progress coordinates represent a set of collective coordinates in a reduced sub-space. Let $\boldsymbol{x}$ specify the configuration of the full system, and let $\boldsymbol{\theta}(\boldsymbol{x})$ be the mapping of $\boldsymbol{x}$ into the progress coordinate space. The progress coordinates must be chosen to discriminate the product and reactant states, but need not specify a true reaction coordinate.

Within each bin, the number of replicas is held constant, or nearly constant, by occasionally replicating or terminating the replicas in that bin. If an occupied bin contains fewer than the target number of replicas, one or more of the replicas are selected via a statistical procedure and are split into M new copies that each carry a fraction $1/M$ of the probability of the parent. Conversely, if a bin contains more than the target number of replicas, a culling procedure removes excess replicas. The weight of the culled replicas is redistributed to a subset of the remaining copies in that bin. These changes to the number of replicas within a bin are carried out in a way that does not bias the underlying dynamics.[28,30] Metastable regions that would typically accumulate large numbers of replicas and consume a large fraction of the computational effort required to simulate the system are not over-populated because of the culling process. Conversely, the conformational space atop a barrier between states that would normally be poorly sampled using conventional simulations is enriched with many low-weight replicas. Details of the replication and termination protocol can be found elsewhere.[28]

B. String method

As in previous WE simulations, phase space is divided into non-overlapping regions or bins defined by a set of progress coordinates. However, instead of a partition with the number of dimensions of the progress coordinate space, the bins are constructed as a string of Voronoi cells in a single dimension (the arclength along the string), embedded in the higher dimensionality progress coordinate space. The string then evolves in time based on the dynamics of the replicas to follow the principle pathway through the conformational space. Our implementation of a string method, adapted for WE sampling, closely follows the algorithm of the finite-temperature string (FTS) method suggested by Vanden-Eijnden and Venturoli.[18]

First, an initial path is constructed between two regions of phase space using a set of progress coordinates that are assumed to be sufficient to describe the transition between regions. A discretized set of evenly spaced images along the path, $\boldsymbol{\varphi}_\alpha$, partitions the progress coordinate space into bins, B_α. The subscript $\alpha = 1, \ldots, N_{\text{im}}$, where N_{im} is the total number of images along the string. The string can be thought of as a spline with the images being the nodes connecting the individual segments. In practice, we use a linear interpolation to initialize the string with evenly spaced images. Each replica specified by $\boldsymbol{x}$ is mapped into progress coordinate space via the transformation $\boldsymbol{\theta}(\boldsymbol{x})$ and then assigned to the closest bin

B_α by calculating the distance to all images $\boldsymbol{\varphi}_\alpha$. The image $\boldsymbol{\varphi}_\alpha$ is therefore the generator of the Voronoi cell corresponding to the bin B_α. Distance measurements in progress coordinate space require the use of the appropriate metric;[18] however, once the metric has been defined, assigning a replica to a bin is straightforward, regardless of the dimensionality of that space. A detailed discussion of how to estimate the metric for a set of collective coordinates is given in Ref. 18. The boundaries separating bins, which can be quite complicated in high-dimensional spaces, need not be computed.

The string images are then adaptively updated according to the following algorithm:

1. Simulate the system to explore phase space for a time T_{move}, which is generally an integral number of τ. During this period, replicas are free to move between bins, unlike in the finite-temperature string method[18] or nonequilibrium umbrella sampling.[6,19]
2. Compute the average position of all replicas in each bin over the time T_{avg} according to

$$\langle \boldsymbol{\theta}_\alpha(\boldsymbol{x}) \rangle = \frac{\sum_{i=1}^{N_r} w_i \boldsymbol{\theta}(\boldsymbol{x}_i) h_\alpha(\boldsymbol{\theta}(\boldsymbol{x}_i))}{\sum_{i=1}^{N_r} w_i h_\alpha(\boldsymbol{\theta}(\boldsymbol{x}_i))}, \tag{1}$$

   where i indicates the ith of N_r replicas simulated in this interval, w_i is the weight of replica i, and $h_\alpha(\boldsymbol{\theta}(\boldsymbol{x}_i))$ is an indicator function for Voronoi cell α, which assumes a value of 1 if replica i is in bin α, and 0 otherwise.
3. Update the images along the strings, moving them toward the average positions within each bin according to

$$\boldsymbol{\varphi}_\alpha^* = \boldsymbol{\varphi}_\alpha^n - \zeta\left(\boldsymbol{\varphi}_\alpha^n - \langle \boldsymbol{\theta}_\alpha \rangle\right) + \boldsymbol{r}_\alpha^*, \tag{2}$$

   where $\boldsymbol{\varphi}_\alpha^*$ are the updated images, $\boldsymbol{\varphi}_\alpha^n$ are the previous position of the images, $\zeta > 0$ controls how rapidly the image moves toward the current estimate of the average position in the bin, and $\boldsymbol{r}_\alpha^*$ is a smoothing term discussed below.
4. Redistribute images uniformly along the arc length of the string by fitting a piece-wise linear spline through all $\boldsymbol{\varphi}_\alpha^*$ and then spacing images equidistantly along the curve. This reparameterization of the string ensures for a given update iteration, n,

$$\left|\boldsymbol{\varphi}_{\alpha+1}^n - \boldsymbol{\varphi}_\alpha^n\right| = \left|\boldsymbol{\varphi}_\alpha^n - \boldsymbol{\varphi}_{\alpha-1}^n\right|. \tag{3}$$

   This final step prevents the images from drifting toward nearby metastable states, which would compromise the efficient sampling of the transition regions. Finally, relabel the shifted $\boldsymbol{\varphi}_\alpha^*$ as $\boldsymbol{\varphi}_\alpha^{n+1}$.

Steps 1–4 are iterated until the images defining the string are stationary.

While this rough outline defines the method, there are additional details employed in the current study. For instance, system configurations, $\boldsymbol{x}_i$, used to identify bin centers in Eq. (1) are taken as the coordinates at the end of each WE simulation of length τ. While this choice of averaging works well for the systems presented here, more or less frequent configuration snapshots should be averaged depending on the length of τ compared to the natural decorrelation time of the system within a bin. The key is to ensure statistical independence of configurations, while not averaging too infrequently. Additionally, using only the last T_{avg}/τ iterations in calculating the average position of replicas within a bin limits the effect of early sampling near the initial position of the string. This is desirable since the weight of replicas within each bin can change rapidly early in the simulation before reaching steady state, thus biasing the calculation of the average.

Smoothing the string as it evolves is essential to prevent kinks and other pathological shapes from developing.[23] We accomplish smoothing through manipulating the term $\boldsymbol{r}_\alpha^*$ in Eq. (2) using one of two different methods. First, for the protein conformational change example (Sec. III C), we define

$$\boldsymbol{r}_\alpha^* = \kappa^n(\boldsymbol{\varphi}_{\alpha+1}^* + \boldsymbol{\varphi}_{\alpha-1}^* - 2\boldsymbol{\varphi}_\alpha^*), \tag{4}$$

for $\alpha = 1 \ldots N-1$, $\kappa^n = \kappa N_{im}\zeta$, and $\boldsymbol{r}_0^* = \boldsymbol{r}_N^* = 0$. This term adds an effective elasticity to the string that prevents the image from moving to $\langle \boldsymbol{\theta}_\alpha \rangle$ if it causes large kink angles between the current image and its neighbors. The parameter κ, which is positive definite, controls how aggressively the string is smoothed.[18] Alternatively the updated positions can be smoothed using a multidimensional curve fitting procedure.[36] This procedure was used for the examples in both Secs. III A and III B. First, $\boldsymbol{\varphi}_\alpha^*$ are calculated with $\boldsymbol{r}_\alpha^* = 0$ for all α. Then a smooth continuous path connecting $\boldsymbol{\varphi}_0^*$ and $\boldsymbol{\varphi}_N^*$ is generated by fitting

$$\boldsymbol{\varphi}_\alpha^{*,\text{cur}}(\lambda) = \boldsymbol{\varphi}_0^* + (\boldsymbol{\varphi}_N^* - \boldsymbol{\varphi}_0^*)\lambda + \sum_{i=1}^{N_{\text{dim}}} \sum_{j=1}^{P} \sigma_{i,j} \sin(j\pi\lambda) \cdot \hat{e}_i, \tag{5}$$

to $\boldsymbol{\varphi}_\alpha^*$ by varying λ over the range [0, 1]. Here, N_{dim} is the dimensionality of the progress coordinate space, $\hat{e}_i$ is the unit vector of the ith progress coordinate, and $\sigma_{i,j}$ are the coefficients of P sinusoidal basis functions in each dimension. The superscript *cur* indicates that $\boldsymbol{\varphi}_\alpha^{*,\text{cur}}$ on the left hand side is calculated by the curve fitting procedure. The parameters $\sigma_{i,j}$ and λ_α are selected to minimize

$$\chi^2 = \sum_{\alpha=0}^{N_{im}} \left|\boldsymbol{\varphi}_\alpha^{*,\text{cur}}(\lambda_\alpha) - \boldsymbol{\varphi}_\alpha^*\right|^2. \tag{6}$$

Details of the optimization procedure to minimize χ^2 are given in the supplementary materials of Ref. 36.

C. Re-weighting scheme

In the original formulation of the WE method, the initial distribution of weights gradually redistributes throughout the sampled conformational space and reaches either an equilibrium or non-equilibrium steady state after some finite number of iterations.[28] In the presence of metastable intermediates, the time for the system to relax to the correct steady-state distribution of weights can be very long, which hampers the efficiency of the method.[31] It was shown previously that this relaxation time can be dramatically reduced through a reweighting procedure that used estimates of the rate of transitions between bins to solve for the steady-state distribution.[31]

Briefly, since the dynamics of the individual replicas in a WE simulation are unbiased, the transition rates between the bins can be used to estimate the elements of a transition matrix $\boldsymbol{k}$. A complete specification of the transition rates allows one to solve for the steady-state probabilities using

$$\frac{dP_i}{dt} = \sum_j k_{j,i} P_j - \sum_j k_{i,j} P_i = 0, \tag{7}$$

where P_i is the probability associated with bin i, and $k_{i,j}$ is the rate of transitions from bin i to j and $\sum_i P_i = 1$. The new estimate of P_i is then used to re-weight the replicas in bin i, where the weight is distributed among the replicas in the bin in proportion to their weights before re-weighting. For the example in Sec. III C, which involves equilibrium sampling, we use a different re-weighting procedure that uses detailed balance to constrain the estimate of the steady-state distribution.[37]

Unlike previous WE studies,[27,31] which employed a single re-weighting step based on a short period of initial sampling, we carry out multiple re-weighting steps at periodic intervals T_{rw} as suggested in Refs. 35 and 37–39. The rates are calculated over a window spanning T_{ravg} simulation steps. In practice we use a fractional window that extends from the current simulation time back T_{ravg}/τ iterations of the WE resampling procedure, which discards inaccurate contributions to the rates early in the simulation. That fraction is fixed during the re-weighting phase to $T_{\mathrm{ravg}}/N_\tau\tau$, where N_τ is the current iteration number.

When the steady-state re-weighting method is used in a simulation, we separate our simulation protocol into distinct phases. We apply the re-weighting protocol only during the initial phase, early in the simulation, since it efficiently relaxes the system away from the initial inaccurate distribution of weights toward the correct steady-state distribution. We find, however, that the standard WE method obtains a higher overall level of accuracy as steady state is approached. The reason for this increased accuracy is related to statistical net zero flux between bins emerging naturally at steady state with standard WE, rather than being enforced based on potentially inaccurate rates determined by the re-weighting scheme. Similar observations were made by Dickson *et al.*[35] for the NEUS method, where they also only used a global re-weighting scheme early in their simulations.

D. Calculating rates

It is of interest to use simulation to calculate the rates for a system to interconvert between distinct states A and B. This can be achieved from very long conventional simulations by observing many transitions between A and B and separating the $A \to B$ transitions from the $B \to A$ transitions. Explicitly, at some time t, a trajectory that had last been in state A is assigned to state A ($\mathcal{S}_A$) at time t; if it had last been in state B, it is labeled as belonging to state B ($\mathcal{S}_B$). Given this partition, the reaction rate from state $A \to B$ or from $B \to A$ is given by[38]

$$k_{A,B} = \lim_{T\to\infty} \frac{N^T_{A,B}}{T_A}, \quad k_{B,A} = \lim_{T\to\infty} \frac{N^T_{B,A}}{T_B}, \tag{8}$$

where T_A and T_B are the total time that a trajectory has been assigned to state A or B, respectively, during the interval $[0, T]$. During the same interval, $N^T_{A,B}$ is the number of times a trajectory assigned to A switched to being assigned to B, and $N^T_{B,A}$ enumerates the reverse transition.

An alternative but equivalent definition of the reaction rate based on fluxes into A and B is given by[5]

$$k_{A,B} = \frac{\overline{\Phi}_{B|\mathcal{S}_A}}{\overline{h}_A}, \; k_{B,A} = \frac{\overline{\Phi}_{A|\mathcal{S}_B}}{\overline{h}_B}, \tag{9}$$

where $\Phi_{B|\mathcal{S}_A}$ ($\Phi_{A|\mathcal{S}_B}$) is the flux into state B (A) from trajectories that originated in state A (B), and h_A (h_B) is a history dependent indicator function that is equal to 1 if the trajectory was more recently in state A (B) than in B (A), and zero otherwise. The indicator function serves to assign trajectories to either state A or B as above. Overbars indicate a time averaged quantity.

The original formulation of the weighted ensemble method, in which replicas originating from a reactant state (state A) were reintroduced into that state upon crossing the product surface (state B), effectively isolated members of the trajectory assigned to a single state.[28,31] In principle, WE does not require one to run simulations with this feedback scheme, since it properly accounts for a dynamical trajectory's history.[37] Thus, a state can be unambiguously assigned (assuming that all trajectories are initiated from A or B—otherwise there will be some transient period when a trajectory is unassigned) to every replica enabling one to calculate the requisite fluxes in Eq. (9).

Dickson *et al.*[35] introduced a dual-direction scheme that uses an extended bin space to split the forward and backward path ensembles. In the standard nomenclature of a WE simulation with N_{dim} progress coordinates, one adds an additional progress coordinate which labels each replica as either being assigned to the product or reactant state. For the Voronoi-based division of the progress coordinate space used in this study, this additional progress coordinate modifies the distance metric used to assign a set of coordinates to a bin:

$$\left\|\boldsymbol{x} - \boldsymbol{\varphi}^n_\alpha\right\| = \begin{cases} \infty & \text{if } x_{m+1} \neq \varphi^n_{\alpha,m+1} \\ \sqrt{\sum_{i=1}^{m} \left(x_i - \varphi^n_{\alpha,i}\right)^2} & \text{otherwise.} \end{cases} \tag{10}$$

A replica who had last visited A is then infinitely far from the centers of the string associated with state B. During a WE simulation then, $\overline{\Phi}_{B|\mathcal{S}_A}$ is calculated as the flux of probability from all Voronoi cells in $\mathcal{S}_A$ into the region defining state B.

E. Implementation

We have implemented all of the systems in Sec. III using open-source software written in Python. The string method was implemented as a plugin for the Weighted Ensemble Simulation Toolkit with Parallelization and Analysis (WESTPA),[40] which provides a general framework for performing and analyzing WE simulations. The complete set of tools required to simulate the example systems, analyze

the results, and generate the figures found in this study are available at https://simtk.org/home/westring.

III. EXAMPLES

A. Periodic two-dimension system

As an initial test of the string-based weighted ensemble sampling method, we apply it to a periodic two-dimensional system[19] with a potential surface defined by

$$V(x, y) = \gamma \left[x - \frac{1}{2} \sin(2\pi y) \right]^2 + \alpha \cos(2\pi y), \quad (11)$$

where x and y are the spatial coordinates and α and γ are parameters that determine the shape of the potential. The form of the potential creates a reaction pathway that depends non-trivially on both coordinates, x and y.

Periodic boundary conditions are applied in the y direction such that $y \in [0, 1)$. A constant external force ($\boldsymbol{F}_{\text{ext}} = F\hat{y}$) is applied along the y direction, which drives the system out-of-equilibrium and generates a constant flux across the periodic boundary. A particle on the potential evolves according to the over damped equation of motion in the presence of a random stochastic fluctuation. The discretized form of the equation is

$$\boldsymbol{X}(t + \delta t) = \boldsymbol{X}(t) - \frac{\delta t}{m\xi}(\nabla_{\boldsymbol{X}} V - \boldsymbol{F}_{\text{ext}}) + \delta \boldsymbol{X}^G, \quad (12)$$

where $\boldsymbol{X} = (x, y)$, δt is the time step, $\delta \boldsymbol{X}^G$ is a random displacement with zero mean and variance $2D\delta t$. The diffusion coefficient, $D = (m\beta\xi)^{-1}$ is defined in terms of the mass of the particle, m, the inverse temperature, β, and the friction coefficient, ξ.

In this work, $\delta t = 0.002$, $\xi = 1.5$, $F = 1.8$, $\beta = 4.0$, $m = 1.0$, and $\gamma = 2.25$. Parameters related to the string parameterization and WE sampling are given in Table I. The value of α modulates the height of the barrier spanning the periodic boundary. Using two different values of α, we consider a case where transitions are common ($\alpha = 1.125$) and another where the barrier is higher and transitions are rare ($\alpha = 2.25$). These parameters were selected to match the values used in Ref. 19, although they differ from the values originally reported.[41]

TABLE I. Parameters used in the WE simulations for the periodic two-dimensional potential.

	Common	Rare
α	1.125	2.25
N_{im}	20	50
N_{rep}	40	50
$\tau/\delta t$	10	10
Iterations (τ) per phase		
Phase I	1000	5000
Phase II	34 000	30 000
	Phase I	Phase II
T_{move}/τ	25	–
T_{avg}/τ	100	–
P	2	–

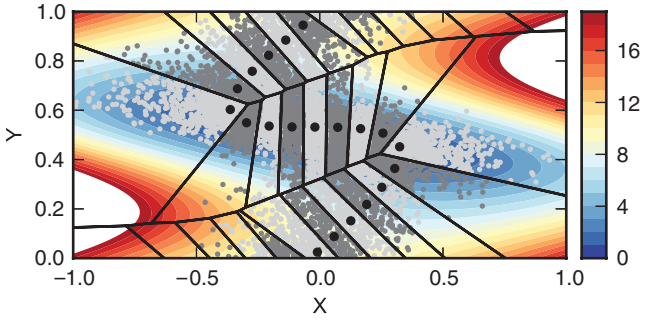

FIG. 1. The two-dimensional periodic potential surface for $\alpha = 1.125$, with the converged path of 20 centers (black dots) and corresponding Voronoi cells. Contour lines are separated by kT, and the corresponding color scale is shown in the same units. In each Voronoi cell, a random sample of the instantaneous positions visited by the replicas is shown in alternating light and dark gray dots for clarity.

Figure 1 shows the converged string for the common transition case after 1000τ. The string was initialized as a vertical line between (0, 0.05) and (0, 0.95). Updates to the string were performed using the multidimensional curve fitting procedure (Eq. (5)) with $P = 2$. For both conventional and WE simulations, we generate a projection of the steady-state probability distribution onto the y axis as the simulations progress in time. These projections are then used to analyze the convergence properties of each simulation method compared to a well-converged target distribution. The discrepancy between the target distribution and the simulated distribution accumulated to a particular time point on a logarithmic scale is

$$\text{error} = \left(\frac{1}{n} \sum_{i=1}^{n} E_i^2 \right)^{1/2}, \quad (13)$$

where

$$E_i = \begin{cases} \log P(i) - \log P_t(i) & P(i) \neq 0 \\ \log 1/T - \log P_t(i) & P(i) = 0 \end{cases}. \quad (14)$$

Here, $P(i)$ is the normalized steady-state probability distribution projected onto the y axis, $P_t(i)$ is the target distribution, and T is the total time of the complete simulation. The histograms used to construct the distributions are obtained by dividing the y axis between 0 and 1 into $n = 100$ windows of equal width. The alternative definition of the error for $P(i) = 0$ is necessary to ensure that the error is still finite when bins have yet to accumulate any samples in them. The impact of this choice is only significant for the conventional sampling case when $\alpha = 2.25$ since these simulations require a non-negligible amount of time to reach the windows at the top of the higher barrier. In both sets of WE simulations, all of the histogram windows are populated almost immediately, even if the initial estimate of the probability is inaccurate.

The final distributions show excellent agreement between the WE simulations and the long conventional target simulations for both values of α (Fig. 2). At intermediate times, Fig. 3 shows that for both choices of α, the WE simulations converge to the correct distribution more rapidly than the respective conventional simulations. For $\alpha = 1.125$,

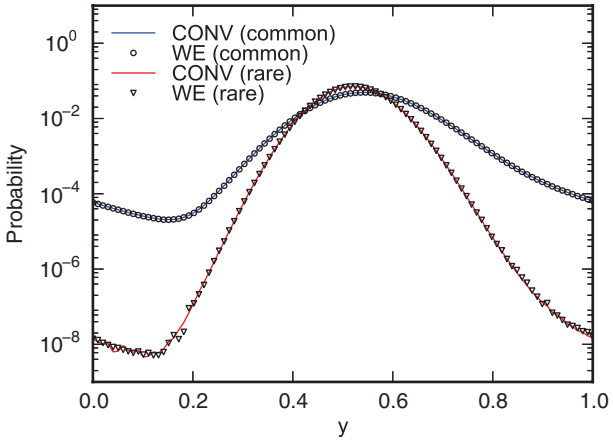

FIG. 2. Projections of the steady-state distributions for the two-dimensional periodic potential onto the y axis. Solid lines show the distributions for the common and rare parameter sets obtained using conventional sampling.

the conventional simulations take approximately five times longer to converge to the same level of accuracy as the WE simulations. For the case where $\alpha = 2.25$, the barrier is high enough that the conventional simulations require a significant number of steps to accumulate sampling in every histogram window. During this phase in which a particle in the conventional simulations has not visited every window, the error is dominated by the term in Eq. (14) corresponding to $P(i) = 0$, decreasing slowly until all of the bins have been visited. The WE simulations, conversely, populate every bin almost immediately and the total error decreases much more rapidly. As such, the methods display quite different convergence characteristics for the high barrier case, with the WE simulations converge to a specific error in at least an order-of-magnitude fewer steps than conventional simulations, and significantly faster for certain error values.

B. Two-dimensional system with two pathways

As a second example, we consider a two-dimensional ring-shaped potential with two distinct pathways[35] connecting a pair of metastable states. A transition from one state to the other requires passing through a metastable intermediate, which makes this system more difficult to sample than the model examined in Sec. III A. Many complex systems contain metastable intermediates along the transition path, so such a test becomes important in ensuring the procedure's broader applicability. The potential surface for this model is defined as

$$V(r,\theta) = \alpha(r-\gamma)^2 + \chi_1 \cos(2\theta) - \chi_2 \cos(4\theta). \quad (15)$$

Here, $r = (x^2 + y^2)^{1/2}$ and θ is the angle in radians measured counterclockwise from the x axis. Particles on this surface evolve according to Eq. (12) with the same definition of the noise term and diffusion constant. A constant external force, $\boldsymbol{F}_{\text{ext}} = -F\hat{\theta}/r$, drives the system out of equilibrium in a clockwise direction.[42] The parameters governing the dynamics and shape of the potential surface were selected to match those used in Ref. 35, where $\alpha = 3.0$, $\gamma = 3.0$, $\chi_1 = 2.25$, $\chi_2 = 4.5$, $\xi = 1.5$, $F = 7.2$, and $\delta t = 0.005$. The inverse temperature is varied between $\beta = 1.0$ and 3.0. On this surface, we define two states, A and B, which encompass the area within circles of radius $R = 1.0$ centered at $(-\gamma, 0)$ and $(\gamma, 0)$, respectively. The external force creates two distinct pathways for the forward ($A \to B$) and backward ($B \to A$) transitions (Fig. 4).

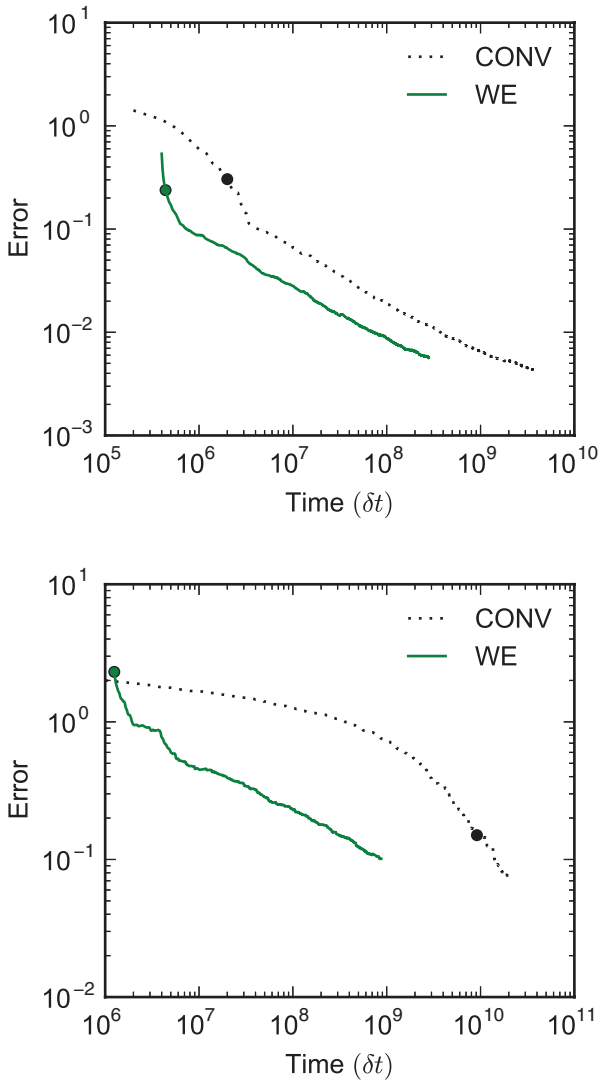

FIG. 3. Convergence of the steady-state distributions projected onto the y axis for both conventional and WE simulations for the two-dimensional periodic potential where $\alpha = 1.125$ (top) and $\alpha = 2.25$ (bottom). For both the conventional and WE simulations, the errors are averages over individual error curves calculated for ten independent simulations using Eqs. (13) and (14). The distributions for the WE simulations do not include statistics from the first 50τ. The target distribution for the shallow potential (top) is calculated from a single long conventional simulation 4.0×10^9 steps in length. For the deep potential (bottom), the target distribution was obtained by averaging ten conventional simulations of 2.0×10^{10} steps. In each case, separate simulations were used to calculate the error and the target distributions. The solid circle marks the average time at which all histogram windows were populated for the ten simulations.

Parameters associated with the WE protocol for each inverse temperature are given in Table II, and each string is initialized as a horizontal line connecting the centers of A and B. A total of N_{rep} replicas are initiated at each of the positions $(-\gamma, 0)$, $(\gamma, 0)$, $(0, -\gamma)$, and $(0, \gamma)$, and are assigned a weight

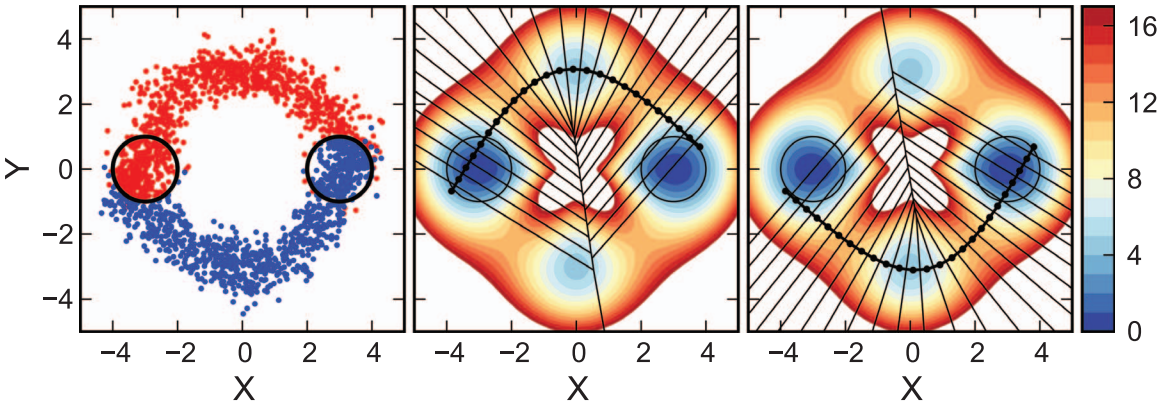

FIG. 4. The two-dimensional ring potential with two pathways. (Left panel) An instantaneous snapshot of all of the active replicas during a representative iteration, where replicas that have last visited A are shown in red, and those that last visited B shown in blue. The circles, centered at (−3,0) and (3,0), show the boundaries of the states A and B. (Center and right panels). The center and right panels show the converged strings with the corresponding Voronoi cells for the $A \rightarrow B$, and $B \rightarrow A$ transitions, respectively. The circles delineate A and B as in the left panel. Contour lines are separated by kT, and the corresponding color scale is show in the same units.

of $(1 - 3\delta)/N_{\text{rep}}$ at $(-\gamma, 0)$, and δ/N_{rep} otherwise, where $\delta = 1 \times 10^{-12}$.

The presence of metastable intermediate states centered at $(\gamma, 0)$ and $(0, -\gamma)$ along each path requires the use of the re-weighting scheme described in Sec. II C to efficiently converge to the correct non-equilibrium steady-state distribution, as our choice of initial weights starts the system far from steady state. Projections of the steady-state distribution for values of $\beta \leq 2.5$ onto the angular coordinate θ are shown for both conventional (CONV) and weighted ensemble (WE) simulation in Fig. 5. There is excellent correspondence between the distributions obtained from the conventional simulations and the WE method across the entire temperature range in terms of both the density in the metastable states as well as the barrier regions.

For this model system, we analyzed the convergence behavior of the reaction rates instead of the steady-state distribution, following Ref. 35. Separating the ensemble of replicas transiting in each direction onto two strings using the dual-direction scheme described in Sec. II D, allowed us to calculate the reaction rate using Eq. (9). We then analyzed the performance of the WE method in calculating the reaction rate between states A and B. The WE simulations are compared against conventional simulations in which the rate is calculated using Eq. (8). The error in the calculated rate is measured as

$$\text{error}(t) = |\log k(t) - \log k_t|, \tag{16}$$

where k_t is the target rate constant and $k(t)$ is the rate as measured at a time t after the start of the simulation. Target rate constants for $\beta \leq 2.5$ were calculated by averaging the rates obtained from ten conventional simulations, each longer than 200 times the mean first passage time (MFPT). For $\beta = 3.0$, the target rate constant was obtained by fitting the rates from the higher temperature target simulations to an Arrhenius form and extrapolating. Errors as a function of

TABLE II. Parameters used in the WE simulations for the two-dimensional ring potential.

β	1.0	1.5	2.0	2.5	3.0
N_{im}	50	60	80	100	100
N_{rep}	40	50	50	50	50
$\tau/\delta t$	10	10	10	10	10
Iterations (τ) per phase					
Phase I	800	800	800	800	800
Phase II	700	1700	1700	3200	4200
Phase III	3500	7500	7500	6000	20 000

	Phase I	Phase II	Phase III
T_{move}/τ	10	–	–
T_{avg}/τ	50	–	–
P	2	–	–
T_{rw}/τ	20	20	–
$T_{\text{ravg}}/N_\tau\tau$	0.5	0.5	–

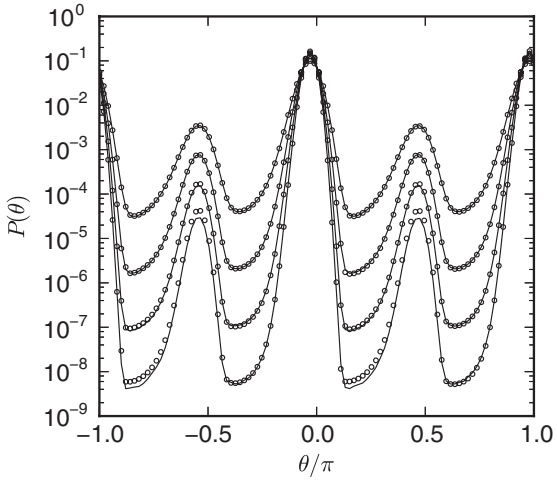

FIG. 5. Projections of the steady-state distribution for the two-dimensional ring potential onto θ for $\beta = 1.0$, 1.5, 2.0, and 2.5, obtained using conventional (lines) and weighted ensemble (circles) simulations. The probability of finding a particle in either of the two metastable intermediates at $\theta = -\pi/2$ or $\pi/2$ decreases with increasing β.

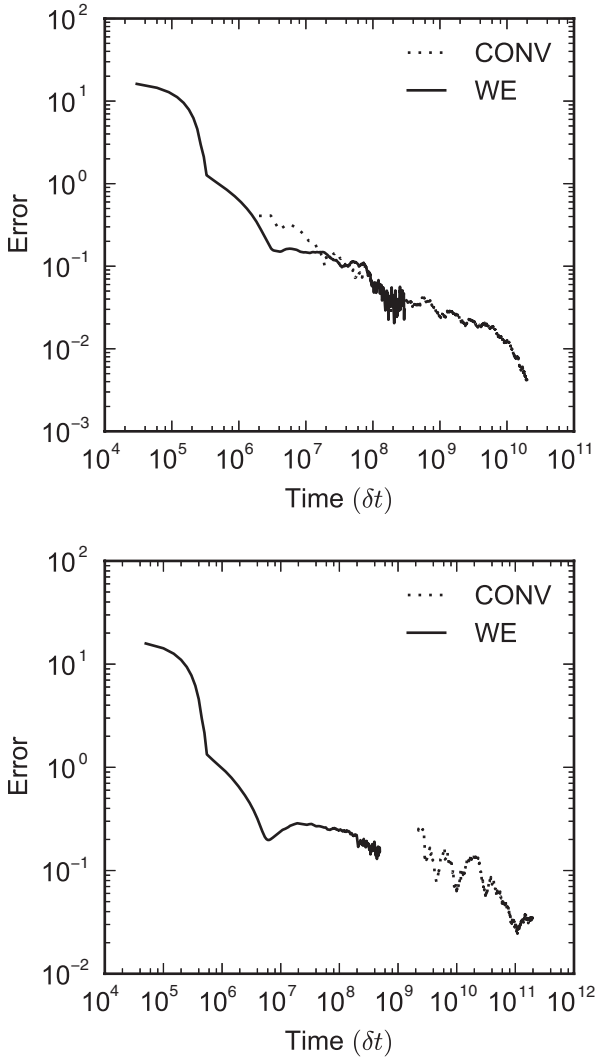

FIG. 6. Convergence of the $A \rightarrow B$ rate constant for the two-dimensional ring potential where $\beta = 1.5$ (top) and 2.5 (bottom). For both conventional and WE simulations, the errors are calculated using Eq. (16), and are the averages of the error curves of ten independent simulations. For the conventional simulations, an estimate of the error cannot be obtained until the first transition from A to B is observed.

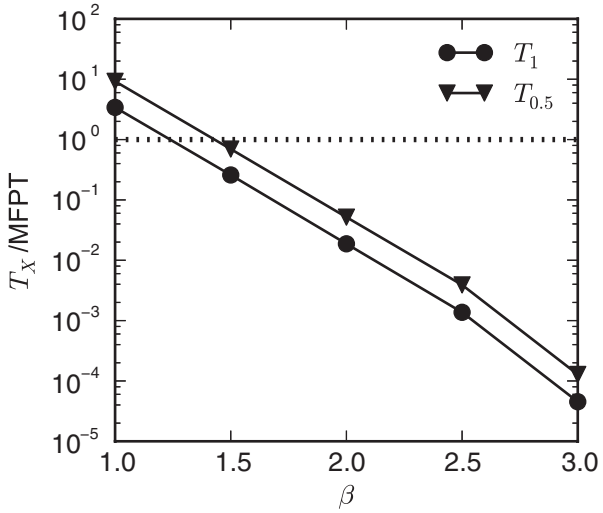

FIG. 7. Efficiency of the WE method for the two-dimensional ring potential as a function of inverse temperature, β. Efficiency is defined as the ratio of the total aggregate simulation time, T_X, to the system's natural MFPT: T_X/MFPT, where T_X is the time required to achieve a desired error, X. Thus, when the efficiency is 1 the total simulation time is the time required for the MFPT, which may be extremely long. We carried out this analysis for an order of magnitude error estimate (10^1) of the rate T_1 and a factor of three error estimate ($10^{0.5}\sim3$) of the rate $T_{0.5}$.

aggregate simulation time are plotted for $\beta = 1.5$ and 2.5 in Fig. 6 and are the root mean squared averages of errors obtained from ten simulations. In the case of the conventional simulation, errors and target rates are calculated from independent simulations. The curves corresponding to the WE simulations are a moving average over the five previous iterations, and rates are based on a moving average with a window size of 200τ.

The performance of the WE simulations over the entire temperature range is shown in Fig. 7 by plotting T_X/MFPT as a function of β. The amount of simulated time, T_X, required to obtain an error equal to X is calculated using Eq. (16). Measuring the error, X, on a logarithmic scale, T_1 corresponds to the amount of time required to obtain an order-of-magnitude estimate of the rates (error $\sim 10^1$), while $T_{0.5}$ is the time required to get an estimate that is within a factor of three of the target (error $\sim 10^{0.5} \sim 3$). When T_1/MFPT $= 1$, the amount of time required to obtain an order-of-magnitude estimate of the rate is approximately equal to the amount of time required by a conventional simulation.

Figure 7 indicates that the WE method is significantly more efficient than the corresponding conventional simulation at all temperatures except for the highest ($\beta = 1.0$), in which particles on the potential can easily cross the barriers between states. For $\beta = 1.0$, the time required to relax away from the initial distribution of weights (approximately a δ function of probability in state A) and to the correct steady state probability distribution is of the same order of time as the MFPT. The comparative efficiency of WE to conventional sampling increases with decreasing temperature (increasing barrier height); WE obtains an order-of-magnitude estimate of the rate in nearly four orders-of-magnitude less time than the MFPT for $\beta = 3.0$.

C. Elastic network model of the NTRC^r protein domain

Finally, we consider the allosteric transition between the inactive and active conformations of the nitrogen regulatory protein C receiver domain (NtrC^r). This system has been studied previously using both all-atom[43,44] and coarse-grained simulation.[16,18] For the purpose of testing the WE string method, we choose to follow the latter approach, and guided by those studies, construct a two-state elastic network model of the protein.

Elastic network models provide a reduced representation of the protein, where only the C_α atom of the backbone is explicitly included. The interaction between these coarse-grained sites is governed by harmonic potentials that stabilize one or more reference conformations, often the experimentally determined native conformation. While elastic network models lack the chemical fidelity of all-atom models, they

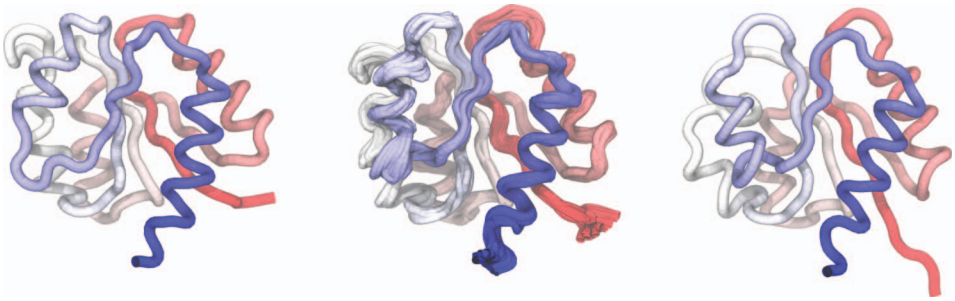

FIG. 8. The inactive (left) and active (right) conformations of the NtrC^r receiver domain, generated from PDB IDs 1DC7 and 1DC8, respectively. (Center) Ten conformations taken from the Voronoi bin corresponding most closely to the $q^+ = 0.5$ (bin index 18). Each conformation is the average of 5 randomly selected snapshots. Coloring is based on residue indices starting from blue at the N-terminus and ending with red at the C-terminus.

have been shown in some instances to recapitulate important conformational fluctuations of less simplistic models.[45–47]

Following the model construction detailed in Refs. 16 and 18, we build a two-state elastic network model of NtrC^r to study the transition from the inactive to active states of the protein upon phosphorylation. These states are generated from the 124 C_α positions in the NMR structures[48] (PDB IDs 1DC7 and 1DC8, for state A and B, respectively) and are shown in Fig. 8. The potential energy of the protein is specified by its instantaneous conformation, $\boldsymbol{x} \in R^{3M}$, where M is the number of residues in the model and $\boldsymbol{x}_i$ denotes the position of the ith residue. Specifically,

$$U(\boldsymbol{x}) = -\frac{1}{\beta_m}\ln(e^{-\beta_m U^A(\boldsymbol{x})} + e^{-\beta_m U^B(\boldsymbol{x})}) + U^R(\boldsymbol{x}), \quad (17)$$

where $U^A(\boldsymbol{x})$ and $U^B(\boldsymbol{x})$, defined below, are the individual elastic network model energies for reference states A and B, respectively. These single-well potentials are combined by exponential averaging, where the parameter β_m controls the barrier height that separates the two states and thus the rate of transitions.

$$U^A(\boldsymbol{x}) = \frac{1}{2}\sum_{ij}^{M} k_{ij} D_{ij}^A \left(\Delta x_{ij} - \Delta x_{ij}^A\right)^2, \quad (18)$$

with an analogous definition for $U^B(\boldsymbol{x})$. The distance between residues i and j, $\Delta x_{ij} = |\boldsymbol{x}_i - \boldsymbol{x}_j|$. Distances between residues calculated from the reference state structures A and B are defined as $\Delta x_{ij}^A = |\boldsymbol{x}_i^A - \boldsymbol{x}_j^A|$ and $\Delta x_{ij}^B = |\boldsymbol{x}_i^B - \boldsymbol{x}_j^B|$, respectively. The contact matrices D_{ij}^A and D_{ij}^B determine which pairs of residues are connected via harmonic linkages, and are given by

$$D_{ij}^A = \begin{cases} 1, & \Delta x_{ij}^A < d^A \\ 0, & \text{otherwise} \end{cases}, \quad (19)$$

and similarly for D_{ij}^B, where d^A is a cutoff distance. The force constants, k_{ij} are modulated by the difference in pairwise residue-residue distances in A and B as

$$k_{ij} = \min\left(\frac{\varepsilon_k}{\left(\Delta x_{ij}^A - \Delta x_{ij}^B\right)^2}, k_{\max}\right). \quad (20)$$

The final term in Eq. (17) provides a hard core repulsion that prevents steric overlap between residues, and is given by

$$U^R(\boldsymbol{x}) = \varepsilon \sum_{\substack{i,j=1,\\ i\neq j}}^{M} \left(\frac{\sigma}{\Delta x_{ij}}\right)^{12}. \quad (21)$$

The parameterization of the potential is the same as in Ref. 18: $d^{A,B} = 11.5$ Å, $\varepsilon_k = 0.5$ kcal mol^{-1}, $k_{\max} = 0.2$ kcal mol^{-1}Å^{-2}, $\varepsilon = 1.0$ kcal mol^{-1}, $\sigma = 1.0$ Å, $\beta_m = 0.02$ kcal mol^{-1}, and the masses of all particles, $m = 100$ amu. As in Ref. 18, we used a modified value of β_m that differs from the original model given in Ref. 16, where β_m was set to 0.005 kcal mol^{-1}.

We then introduce a set of collective coordinates $\boldsymbol{\theta}(\boldsymbol{x}) = R\boldsymbol{x} + \boldsymbol{X}$, which we will use to assign each conformation to an image along the string. The coordinate $\boldsymbol{\theta}$ is of the same dimensionality as our original system, but removes the degeneracies due to translation and rotation of the system. In the above definition, R is the rotation matrix and $\boldsymbol{X}$ is the translation vector that when applied to $\boldsymbol{x}$ minimizes $|\boldsymbol{\theta}(\boldsymbol{x}) - \boldsymbol{x}_{\text{ref}}|$. The distance metric in the collective coordinate space is the root mean square deviation (RMSD) of the optimally aligned conformation $\boldsymbol{\theta}(\boldsymbol{x})$ with some reference coordinates of the protein, $\boldsymbol{x}_{\text{ref}}$. Here, the RMSD is calculated using the fast quaternion-based characteristic polynomial method,[49] which does not require the explicit calculation of R. The two residues at both the C- and N-termini of the protein are highly mobile and are excluded from the RMSD calculation.

The string is initiated as the linear interpolation between A and B using 40 images ($N_{im} = 40$). Forty replicas ($N_{\text{rep}} = 40$) were initiated in the bins corresponding to A and B with each replica given equal weight. The string was free to move during the first 1000τ of the simulation and then was fixed at its converged position to collect statistics for the path. An equilibrium re-weighting procedure was applied during the first 1500τ of the simulation every 10τ. An additional 3000τ steps of the WE method were simulated with the converged string to collect statistics. The total simulation time was 4500τ.

The free energy associated with the path defined by the converged string is shown in Fig. 9. The free energy in each Voronoi bin, G_α, is calculated from the WE simulation directly as $G_\alpha = -k_B T \ln(\overline{w}_\alpha)$, where $\overline{w}_\alpha$ is the average weight assigned to bin α. Statistical errors for averages of the

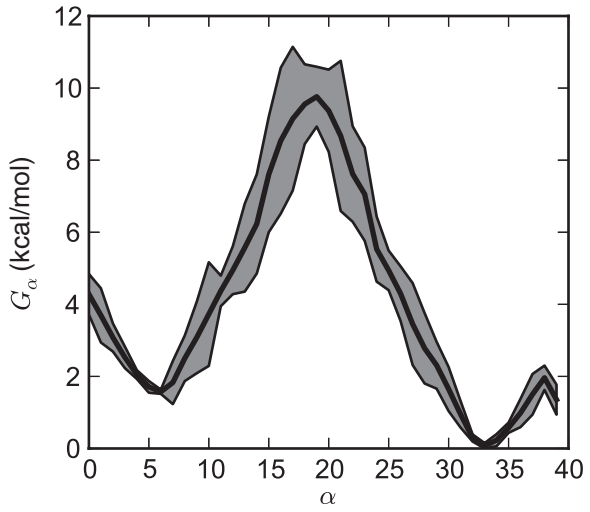

FIG. 9. Free energy G_α associated with each Voronoi tessellation along the string for the elastic network model of NtrC^r. The dark shaded region denotes the 95% confidence interval (two standard errors) for the free energy averaged over the last 3000τ of the WE simulation.

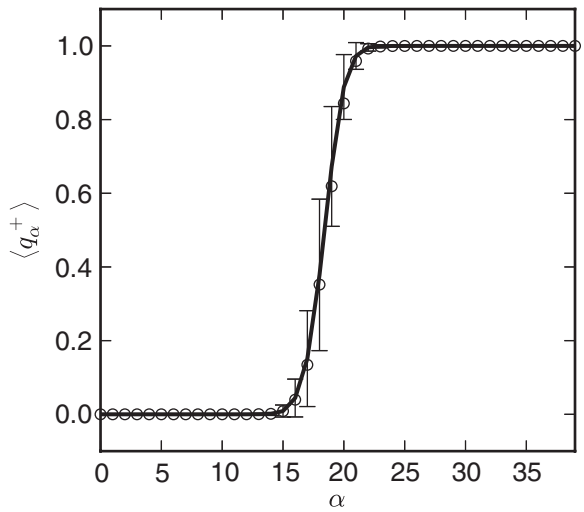

FIG. 10. Committor probability along the string for the elastic network model of NtrC^r. The average committor probability of reaching state B before state A, q^+, calculated from an ensemble of 500 conformations in each bin is shown as a solid line, with error bars equal to the standard deviation of the sample. Committors calculated from the WE simulation using Eq. (22) are shown as open circles.

timeseries data used to calculate $\overline{w}_\alpha$ were estimated by autocorrelation analysis[50] using the `timeseries` module of the pyMBAR code.[51] The free energy along the string is shown in Fig. 9 and agrees with the barrier height and relative stability between A and B to within 1 kcal/mol of the results calculated from the FTS method (Figure 11 in Ref. 18).

While we extended the simulation 3000τ beyond the phase when re-weighting was applied, an accurate estimate of the free energy difference between A and B (within the 95% confidence interval) was obtained during the first 500τ of the simulation. During this same period the barrier height differed from the converged value by less than 1.3 kcal/mol.

While the free energy shows a distinct barrier along the path separating the inactive and active states, it is important to determine whether the converged path is dynamically relevant, and if the barrier corresponds to a mechanistic transition state. To this end, we calculate the committor probability[7,52,53] for each bin, α, along the string, q_α^+, defined as the probability that fleeting trajectories initiated from that bin reach state B before state A. If the string accurately describes the dynamical reaction pathway, $q_\alpha^+ = 1/2$ should coincide with the barrier with neighboring bins rapidly asymptoting to zero and unity on either side of the barrier.

For each bin, we selected 500 random conformations drawn from the final 2500τ iterations of the WE simulation. One hundred conventional simulations were initiated from each conformation with initial velocities drawn randomly from a Boltzmann distribution at the same temperature as the WE simulations. These simulations were propagated until the trajectory reached state A or B, and the terminating state was recorded. A trajectory terminates in state A (B) when its RMSD is <2 Å from the reference conformation of A (B) and >3 Å from B (A).

Alternatively, the committor probability along the string can also be calculated directly from transition statistics gathered during the WE simulations by solving the following system of equations[54]

$$-q_i^+ + \sum_{k\in I} T_{ik} q_k^+ = -\sum_{k\in B} T_{ik} \text{ for } i \in I, \tag{22}$$

where q_i^+ is the probability that from bin i the system will reach state B before returning to state A, and $T_{i,j}$ is the probability of going from bin i to j, which is calculated directly from the WE simulation. Intermediate states, I, are those states which are not in A or B.

Committor probabilities calculated from both methods are shown in Fig. 10 and are nearly identical over all of the bins. The $q_\alpha^+ = 1/2$ bins coincide with the peak of the barrier in Fig. 9 between $\alpha = 18$ and 19. Representative conformations from these bins are shown in the center panel of Fig. 8. Taken together with the observation that the underlying distributions for each bin, $P(q_\alpha^+)$, are unimodal, the converged string appears to describe the reactive pathway between the inactive and actives states for this model of NtrC^r.

IV. DISCUSSION

We have introduced a variant of the WE method to sample transitions along a one-dimensional path embedded in the space of an arbitrary number of order parameters. As in the finite-temperature string method[18] on which our method is based, an equidistantly spaced set of images along the path define a set of Voronoi cells that are analogous to the subdivision of phase space into bins in earlier versions of the WE method. The images adaptively evolve toward the average weighted position of replicas that migrate through each bin during the WE simulation, subject to a smoothing reparameterization. The use of a dynamic set of bins that change in time during a WE simulation had been previously suggested,[28–30] and here we demonstrate its practical use to confine sampling to the region of order parameter space along the transition pathway.

Unlike in variants of the string method where the string images are used to either initiate replicas of the system,[16,21] or to perform constrained sampling along a hyperplane coincident with string,[15] here the path only serves to partition phase space. Its utility when combined with the adaptive update scheme is that it allows the WE method to enhance sampling along the transition tube between states of interest. In doing so, the string and associated Voronoi bins do not perturb the natural dynamics of the system.

The path of the string does, however, play an important role in determining the efficiency of the WE method. Partitioning progress coordinate space along a pseudo one-dimensional curve converts the computation from one where the cost scales exponentially with the number of order parameters to one which is linear in the number of images along the string. When a large number of progress coordinates are required to effectively sample a complex system, this change in the scaling behavior is likely to drastically reduce the computational cost of performing a WE simulation. Additionally, so long as the string accurately approximates the dominant path of reactive flux, the majority of transitions along the string will occur between neighboring Voronoi bins. The reduced number of relevant bin interfaces across which transitions occur should decrease statistical errors in the rate estimates used in Eq. (7). This likely increases the efficiency of the re-weighting procedure in Sec. II C compared to its use with other binning schemes previously used with WE sampling.

The effectiveness of the string method as a way of partitioning phase space is subject to the assumption that the width of the transition tube about the string is small compared to its radius of curvature.[23] This assumption is a well-documented limitation of the string method that precludes its use for systems with multiple reactive channels or many important metastable states connected via a meshwork of isolated transition pathways. The use of multiple strings to sample each pathway (using a modified version of the dual-direction scheme in Sec. III B), or pairwise between well-defined states could circumvent this limitation, although accurately positing the presence and location of multiple pathways *a priori* for complex systems in almost all cases is non-trivial.

In selecting the two driven 2D systems in Secs. III A and III B, we sought to provide a direct comparison between our method and the NEUS-based string method of Dickson *et al.*[19,35] We carefully attempted to replicate the implementation of the underlying Brownian dynamics, as well as their convergence analysis to test the efficiency of the WE method. Since the conventional dynamics in this study and the NEUS papers agree within statistical variation, the results presented in Figs. 3, 6, and 7 should be directly comparable. For both example systems, WE performs admirably and appears to show similar or better convergence characteristics than NEUS. It is important to note, however, that in this comparison, the WE simulations use a higher density of replicas than in the NEUS studies. In NEUS, a single replica is simulated per bin; in the WE simulations many replicas are run per bin, although each replica is run for a significantly shorter time than the replicas in the NEUS simulations per iteration. We generally observe a super-linear gain in efficiency over some range of increasing the number of replicas and/or bins in the system (e.g., increasing the number of bins by a factor of 2 speeds convergence by >2 times). Increasing the number of bins likely reduces the discretization errors associated with estimating the rates used in Eq. (7) during the re-weighting step early in the simulation,[55] and increasing the number of replicas reduces the variance in the inter-bin transition rates. Using multiple replicas per sampling region, which was suggested in a recent application of the NEUS to the nonequilibrium folding and unfolding of coarse-grained model of RNA,[39] in concert with a finer discretization of phase space, may ameliorate the performance differences between the two methods.

Finally, we have presented a new string method based on sampling obtained via the weighted ensemble method that appears to have many advantages over previous methods. WE sampling is easy to implement both in terms of the data structures needed to track the replicas of the system and also by not requiring any special modification of the underlying dynamics to restrain replicas to a particular region of phase space, either through momentum reversal at a boundary[18,19,35] or soft-wall restraints.[56] Additionally, it is not necessary to generate physical replicas along the initial string from the start state to the final state as it is in most other string methods. Often the initial string is generated using a simple linear interpolation, targeted molecular dynamics[16] or with a coarse-grained model,[36] all of which can lead to string images corresponding to unphysical intermediates for many systems of interest. With the WE-based method, all replicas can be initiated in the start state and the string will evolve based on the natural dynamics of the system, without ever starting replicas based on these unphysical conformations. In this regard, our method shares some similarities with the growing string method.[57] Despite this, some care should be taken when generating the initial path of the string. The calculation could still become trapped if the initial path is separated from the dominant pathway by a large barrier orthogonal to the string. Finally, as we have shown here, the WE-based string method can be drastically more efficient than conventional sampling for calculating the rates and steady-state distributions for a range of equilibrium and nonequilibrium problems. Our method also outperforms the NEUS rare-event sampling method for two of the example systems studied here, but this result may depend on the system being simulated or it may be possible to tune the NEUS parameters to increase performance. This work lays the foundation for applying the WE-based string method to simulating rare transitions in more complex and realistic systems, especially when one or a small number of progress coordinates are insufficient to fully characterize the reaction coordinate.

ACKNOWLEDGMENTS

We wish to thank Lillian Chong and Dan Zuckerman for critical reading of the manuscript, and Lillian Chong and Matt Zwier for their development of the WESTPA code's core functionality. We also thank Aaron Dinner and Alex Dickson for helpful discussions concerning the driven Brownian models used in the examples, and Eric Vanden-Eijnden and Maddalena Venturoli for sharing code related to the NtrC model. This work was supported by National Institutes of Health

(NIH) Grant No. R01-GM089740-01A1 (M.G.), NIH Grant No. T32-DK061296 (J.L.A.), and FundScience (J.L.A.).

[1]C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, "Transition path sampling and the calculation of rate constants," J. Chem. Phys. **108**, 1964–1977 (1998).
[2]T. Woolf, "Path corrected functionals of stochastic trajectories: Towards relative free energy and reaction coordinate calculations," Chem. Phys. Lett. **289**, 433–441 (1998).
[3]T. S. van Erp, D. Moroni, and P. G. Bolhuis, "A novel path sampling method for the calculation of rate constants," J. Chem. Phys. **118**, 7762–7774 (2003).
[4]A. K. Faradjian and R. Elber, "Computing time scales from reaction coordinates by milestoning," J. Chem. Phys. **120**, 10880–10889 (2004).
[5]R. J. Allen, P. B. Warren, and P. R. Ten Wolde, "Sampling rare switching events in biochemical networks," Phys. Rev. Lett. **94**, 018104 (2005).
[6]A. Warmflash, P. Bhimalapuram, and A. R. Dinner, "Umbrella sampling for nonequilibrium processes," J. Chem. Phys. **127**, 154112 (2007).
[7]P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, "Transition path sampling: throwing ropes over rough mountain passes, in the dark," Annu. Rev. Phys. Chem. **53**, 291–318 (2002).
[8]C. Dellago and P. G. Bolhuis, "Transition path sampling and other advanced simulation techniques for rare events," in *Advanced Computer Simulation Approaches for Soft Matter Sciences III*, Advances in Polymer Science Vol. 221, edited by C. Holm and K. Kremer (Springer, Berlin, 2009), pp. 167–233.
[9]M. C. Zwier and L. T. Chong, "Reaching biological timescales with all-atom molecular dynamics simulations," Curr. Opin. Pharmacol. **10**, 745–752 (2010).
[10]R. Elber and M. Karplus, "A method for determining reaction paths in large molecules: Application to myoglobin," Chem. Phys. Lett. **139**, 375–380 (1987).
[11]S. Fischer and M. Karplus, "Conjugate peak refinement: An algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom," Chem. Phys. Lett. **194**, 252–261 (1992).
[12]G. Henkelman and H. Jónsson, "Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points," J. Chem. Phys. **113**, 9978–9985 (2000).
[13]W. E, W. Ren, and E. Vanden-Eijnden, "String method for the study of rare events," Phys. Rev. B **66**, 052301 (2002).
[14]L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, "String method in collective variables: Minimum free energy paths and isocommittor surfaces," J. Chem. Phys. **125**, 024106 (2006).
[15]W. E, W. Ren, and E. Vanden-Eijnden, "Finite temperature string method for the study of rare events," J. Phys. Chem. B **109**, 6688–6693 (2005).
[16]A. C. Pan, D. Sezer, and B. Roux, "Finding transition pathways using the string method with swarms of trajectories," J. Phys. Chem. B **112**, 3432–3440 (2008).
[17]A. C. Pan and B. Roux, "Building Markov state models along pathways to determine free energies and rates of transitions," J. Chem. Phys. **129**, 064107 (2008).
[18]E. Vanden-Eijnden and M. Venturoli, "Revisiting the finite temperature string method for the calculation of reaction tubes and free energies," J. Chem. Phys. **130**, 194103 (2009).
[19]A. Dickson, A. Warmflash, and A. R. Dinner, "Nonequilibrium umbrella sampling in spaces of many order parameters," J. Chem. Phys. **130**, 074104 (2009).
[20]J. Rogal, W. Lechner, J. Juraszek, B. Ensing, and P. G. Bolhuis, "The reweighted path ensemble," J. Chem. Phys. **133**, 174109 (2010).
[21]M. E. Johnson and G. Hummer, "Characterization of a dynamic string method for the construction of transition pathways in molecular reactions," J. Phys. Chem. B **116**, 8573–8583 (2012).
[22]B. M. Dickson, H. Huang, and C. B. Post, "Unrestrained computation of free energy along a path," J. Phys. Chem. B **116**, 11046–11055 (2012).
[23]W. Ren, E. Vanden-Eijnden, P. Maragakis, and W. E, "Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide," J. Chem. Phys. **123**, 134109 (2005).
[24]T. F. Miller III, E. Vanden-Eijnden, and D. Chandler, "Solvent coarse-graining and the string method applied to the hydrophobic collapse of a hydrated chain," Proc. Natl. Acad. Sci. U.S.A. **104**, 14559–14564 (2007).
[25]V. Ovchinnikov, M. Cecchini, E. Vanden-Eijnden, and M. Karplus, "A conformational transition in the myosin VI converter contributes to the variable step size," Biophys. J. **101**, 2436–2444 (2011).
[26]D. Bhatt and D. M. Zuckerman, "Heterogeneous path ensembles for conformational transitions in semi-atomistic models of adenylate kinase," J. Chem. Theory Comput. **6**, 3527–3539 (2010).
[27]J. L. Adelman, A. L. Dale, M. C. Zwier, D. Bhatt, L. T. Chong, D. M. Zuckerman, and M. Grabe, "Simulations of the alternating access mechanism of the sodium symporter Mhp1," Biophys. J. **101**, 2399–2407 (2011).
[28]G. A. Huber and S. Kim, "Weighted-ensemble Brownian dynamics simulations for protein association reactions," Biophys. J. **70**, 97–110 (1996).
[29]B. W. Zhang, D. Jasnow, and D. M. Zuckerman, "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. **104**, 18043–18048 (2007).
[30]B. W. Zhang, D. Jasnow, and D. M. Zuckerman, "The 'weighted ensemble' path sampling method is statistically exact for a broad class of stochastic processes and binning procedures," J. Chem. Phys. **132**, 054107 (2010).
[31]D. Bhatt, B. W. Zhang, and D. M. Zuckerman, "Steady-state simulations using weighted ensemble path sampling," J. Chem. Phys. **133**, 014110 (2010).
[32]A. Rojnuckarin, D. R. Livesay, and S. Subramaniam, "Bimolecular reaction simulation using weighted ensemble Brownian dynamics and the University of Houston Brownian Dynamics program," Biophys. J. **79**, 686–693 (2000).
[33]A. Rojnuckarin, S. Kim, and S. Subramaniam, "Brownian dynamics simulations of protein folding: access to milliseconds time scale and beyond," Proc. Natl. Acad. Sci. U.S.A. **95**, 4288–4292 (1998).
[34]M. Zwier, J. Kaus, and L. Chong, "Efficient explicit-solvent molecular dynamics simulations of molecular association kinetics: Methane/methane, Na^+/Cl^-, methane/benzene, and K^+/18-crown-6 ether," J. Chem. Theory Comput. **7**, 1189–1197 (2011).
[35]A. Dickson, A. Warmflash, and A. R. Dinner, "Separating forward and backward pathways in nonequilibrium umbrella sampling," J. Chem. Phys. **131**, 154104 (2009).
[36]F. Zhu and G. Hummer, "Pore opening and closing of a pentameric ligand-gated ion channel," Proc. Natl. Acad. Sci. U.S.A. **107**, 19814–19819 (2010).
[37]S. Lettieri, M. C. Zwier, C. A. Stringer, E. Suarez, L. T. Chong, and D. M. Zuckerman, "Simultaneous computation of dynamical and equilibrium information using a weighted ensemble of trajectories," preprint arXiv:1210.3094 [physics.bio-ph] (2012).
[38]E. Vanden-Eijnden and M. Venturoli, "Exact rate calculations by trajectory parallelization and tilting," J. Chem. Phys. **131**, 044120 (2009).
[39]A. Dickson, M. Maienschein-Cline, A. Tovo-Dwyer, J. Hammond, and A. Dinner, "Flow-dependent unfolding and refolding of an RNA by nonequilibrium umbrella sampling," J. Chem. Theory Comput. **7**, 2710–2720 (2011).
[40]See http://chong.chem.pitt.edu/WESTPA for the WESTPA software package.
[41]A. Dickson, A. Warmflash, and A. R. Dinner, "Erratum: 'Nonequilibrium umbrella sampling in spaces of many order parameters' [J. Chem. Phys. **130**, 074104 (2009)]," J. Chem. Phys. **136**, 229901 (2012).
[42]A. Dickson, A. Warmflash, and A. R. Dinner, "Erratum: 'Separating forward and backward pathways in nonequilibrium umbrella sampling' [J. Chem. Phys. **131**, 154104 (2009)]," J. Chem. Phys. **136**, 239901 (2012).
[43]A. Damjanović, B. García-Moreno E, and B. R. Brooks, "Self-guided Langevin dynamics study of regulatory interactions in NtrC," Proteins **76**, 1007–1019 (2009).
[44]M. Lei, J. Velos, A. Gardino, A. Kivenson, M. Karplus, and D. Kern, "Segmented transition pathway of the signaling protein nitrogen regulatory protein C," J. Mol. Biol. **392**, 823–836 (2009).
[45]J. L. Adelman, J. D. Chodera, I.-F. W. Kuo, T. F. Miller III, and D. Barsky, "The mechanical properties of PCNA: Implications for the loading and function of a DNA sliding clamp," Biophys. J. **98**, 3062–3069 (2010).
[46]E. Lyman, J. Pfaendtner, and G. A. Voth, "Systematic multiscale parameterization of heterogeneous elastic network models of proteins," Biophys. J. **95**, 4183–4192 (2008).
[47]L. Yang, G. Song, A. Carriquiry, and R. L. Jernigan, "Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes," Structure **16**, 321–330 (2008).
[48]D. Kern, B. F. Volkman, P. Luginbühl, M. J. Nohaile, S. Kustu, and D. E. Wemmer, "Structure of a transiently phosphorylated switch in bacterial signal transduction," Nature (London) **402**, 894–898 (1999).
[49]D. L. Theobald, "Rapid calculation of RMSDs using a quaternion-based characteristic polynomial," Acta Crystallogr. A **61**, 478–480 (2005).

[50]J. Chodera, W. Swope, J. Pitera, C. Seok, and K. Dill, "Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations," J. Chem. Theory Comput. **3**, 26–41 (2007).
[51]M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," J. Chem. Phys. **129**, 124105 (2008).
[52]R. Du, V. Pande, A. Grosberg, T. Tanaka, and E. Shakhnovich, "On the transition coordinate for protein folding," J. Chem. Phys. **108**, 334 (1998).
[53]P. Geissler, C. Dellago, and D. Chandler, "Kinetic pathways of ion pair dissociation in water," J. Phys. Chem. B **103**, 3706–3710 (1999).
[54]F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations," Proc. Natl. Acad. Sci. U.S.A. **106**, 19011–19016 (2009).
[55]J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: Generation and validation," J. Chem. Phys. **134**, 174105 (2011).
[56]L. Maragliano, E. Vanden-Eijnden, and B. Roux, "Free energy and kinetics of conformational transitions from Voronoi tessellated milestoning with restraining potentials," J. Chem. Theory Comput. **5**, 2589–2594 (2009).
[57]B. Peters, A. Heyden, A. T. Bell, and A. Chakraborty, "A growing string method for determining transition states: Comparison to the nudged elastic band and string methods," J. Chem. Phys. **120**, 7877–7886 (2004).