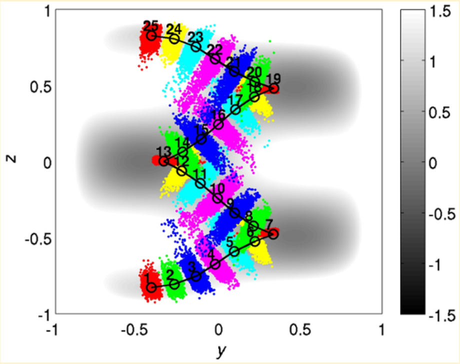
THE JOURNAL OF PHYSICAL CHEMISTRY B
Article
pubs.acs.org/JPCB

A Simple and Accurate Method To Calculate Free Energy Profiles and Reaction Rates from Restrained Molecular Simulations of Diffusive Processes

Victor Ovchinnikov,*[†] Kwangho Nam,*[‡] and Martin Karplus*[†,§]

[†]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States
[‡]Department of Chemistry, Umeå University, Umeå, Sweden, 901 87
[§]Laboratoire de Chimie Biophysique, ISIS, Université de Strasbourg, 67000 Strasbourg, France

ABSTRACT: A method is developed to obtain simultaneously free energy profiles and diffusion constants from restrained molecular simulations in diffusive systems. The method is based on low-order expansions of the free energy and diffusivity as functions of the reaction coordinate. These expansions lead to simple analytical relationships between simulation statistics and model parameters. The method is tested on 1D and 2D model systems; its accuracy is found to be comparable to or better than that of the existing alternatives, which are briefly discussed. An important aspect of the method is that the free energy is constructed by integrating its derivatives, which can be computed without need for overlapping sampling windows. The implementation of the method in any molecular simulation program that supports external umbrella potentials (e.g., CHARMM) requires modification of only a few lines of code. As a demonstration of its applicability to realistic biomolecular systems, the method is applied to model the α-helix ↔ β-sheet transition in a 16-residue peptide in implicit solvent, with the reaction coordinate provided by the string method. Possible modifications of the method are briefly discussed; they include generalization to multidimensional reaction coordinates [in the spirit of the model of Ermak and McCammon (Ermak, D. L.; McCammon, J. A. *J. Chem. Phys.* **1978**, *69*, 1352−1360)], a higher-order expansion of the free energy surface, applicability in nonequilibrium systems, and a simple test for Markovianity. In view of the small overhead of the method relative to standard umbrella sampling, we suggest its routine application in the cases where umbrella potential simulations are appropriate.



1. INTRODUCTION

Complex molecular systems, such as biological macromolecules, are frequently described by computer models with rugged potential energy surfaces, on which configurational transitions between local minima are "rare events" that occur on time scales of 1 ms or longer. Because such transitions underlie many processes of biological significance (e.g., enzyme catalysis,[2,3] allosteric regulation of protein activity,[4] DNA replication,[5,6] refolding of misfolded proteins,[7] ATP synthesis and transport by molecular motors,[8−10] and muscle contraction),[11] their understanding is of broad scientific interest.

Computer modeling of transitions in macromolecules by direct simulation is possible only in select cases using special purpose hardware.[12,13] Many algorithms to enhance the sampling of rare events by molecular simulations have therefore been developed; they include umbrella sampling,[14,15] metadynamics,[16] transition path sampling,[17,18] adaptive biasing force,[19,20] adiabatic MD,[21] targeted MD,[22] temperature-accelerated MD,[23] orthogonal space random walk,[24] milestoning,[25−27] string method,[28,29] and Markov state modeling.[30,31] The goals of such methods are (i) to provide an atomistically detailed or, alternatively, a coarse-grained model of the transition process, (ii) to compute the relative thermodynamic stability of the metastable states along the transition pathway (i.e., the potential of mean force (PMF)), and (iii) to compute the rate(s) of transitions between the metastable states. A typical computational approach starts with an initial model of the complex transition pathway obtained with, e.g., constrained pathway minimization,[32−35] targeted[22,36] or steered[37] dynamics. The initial pathway is subsequently optimized using, e.g., transition path sampling[17,18] or the string method.[28] Using the model of the transition as a guide, a reaction coordinate (RC) is determined (or simply assumed). While the equilibrium potential of mean force corresponding to the RC is usually straighforward to compute with enhanced sampling approaches,[14−16,20,21,24,29] the calculation of transition rates is generally more difficult because it requires information about the temporal dynamics of the system in addition to the configurational densities.

Special Issue: J. Andrew McCammon Festschrift

Received: February 29, 2016
Revised: April 28, 2016
Published: May 2, 2016

DOI: 10.1021/acs.jpcb.6b02139

Our main interest is the modeling of large-scale conformational transitions in proteins[36,38] and in nucleic acids,[39] for which reaction coordinates are likely to be collective (see, e.g., ref 40). Because such reaction coordinates typically exhibit diffusive behavior,[41] we focus on the rate determination via the overdamped Kramers (Smoluchowski) model[42,43] with position-dependent diffusivity. The main contribution of this manuscript is to provide a very simple and accurate method to parametrize the Smoluchowski equation for an arbitrary 1D reaction coordinate, using molecular simulations with simple restraint potentials. However, the ideas can be useful more generally for parametrizing other continuous models from simulations.

In section 2.1, the Smoluchowski model for the reaction coordinate is reviewed. Section 2.2 describes the main features and limitations of existing methods to parametrize the Smoluchowski model; readers familiar with the topic may go directly to section 2.3, in which the present method is derived. In section 3, we demonstrate the accuracy of the method using 1D and 2D model systems, apply it to a conformational transition in a 16-residue miniprotein in implicit solvent, and illustrate a simple procedure to discretize the reaction coordinate efficiently. The developments are summarized in section 4, where we also propose possible extensions of the method. Some technical aspects are presented in the appendices.

2. THEORY

2.1. Smoluchowski Model for the 1D Reaction Coordinate.

We consider a molecular system of N atoms with coordinates $\hat{r} = \{r_1, ..., r_{3N}\}$ and a corresponding potential energy function $E(\hat{r})$. System configurations are sampled from the canonical ensemble using MD simulations. We assume that a 1D differentiable function $x(\hat{r}) \in [0, 1]$ is identified *a priori*; x describes the progress of a chemical or physical reaction of the system, in the sense that (1) for some $\epsilon \ll 1$, $0 < x < \epsilon$, respectively, and $1 - \epsilon < x < 1$ correspond to *reactant* and *product* states and (2) for any two configurations $\hat{r}_1$ and $\hat{r}_2$, $x(\hat{r}_1) < x(\hat{r}_2)$ implies that the system with configuration $\hat{r}_2$ is *closer* to the product state than the system with configuration $\hat{r}_1$. Such a function is commonly called a reaction coordinate (RC), and can be thought of as a 1D coarse-grained model of the reaction in the $3N$-dimensional system (in the sense that the progress of the reaction in the system is completely specified by the evolution of the RC without need to consider the atomic motions). The term *closer* refers to proximity in a kinetic sense, e.g., the mean first passage time to reach the reactant state,[44] or the probability of reaching the product state before the reactant state, i.e., the committor function.[17,18,45−48] Because this study is not concerned with the determination of RCs but rather with parametrizing its evolution as a dynamic variable using samples from molecular simulations, we mention only some of the methods used to define approximate RCs, e.g., the string method[29,49] (used in this study), likelihood maximization,[48] variational optimization,[46] and genetic algorithms.[47]

Following Smoluchowski[42] and other authors,[50−53] we assume that x is a stochastic variable whose probability density $P(x, t)$ evolves from an initial condition $P_0(x)$ at $t_0 < t$ according to the diffusion equation

$$\frac{\partial P}{\partial t} = \frac{\partial}{\partial x}\left\{D\left[\beta\frac{\partial F}{\partial x}P + \frac{\partial P}{\partial x}\right]\right\} \quad (1)$$

where $\beta^{-1} = k_BT$. For simplicity, henceforth we will use dots and primes to denote derivatives w.r.t. t and x, respectively (e.g., eq 1 becomes $\dot{P} = \{D[\beta F'P + P']\}'$). Equation 1 is the Fokker−Planck equation (FPE) or Smoluchowski equation for 1D Brownian motion in the high-friction limit on the free energy surface $F(x)$ with a position-dependent diffusion constant $D(x)$. It can be derived from Kramers equation via the inverse friction expansion.[54] Equation 1 can be written as

$$\begin{aligned} \dot{P} &= -J' \\ J &= -De^{-\beta F}[e^{\beta F}P]' \end{aligned} \quad (2)$$

which shows that, for reflective (zero-flux, $J = 0$) boundary conditions, the Boltzmann distribution $C \exp[-\beta F(x)]$ is the stationary solution, independently of D. As noted by others,[50,52,53] the suitability of eq 1 for describing reactive dynamics in complex systems is strongly system-dependent. For example, writing the FPE in drift-diffusion form[54]

$$\begin{aligned} \dot{P} &= -[D_1P]' + [D_2P]'' \\ D_1 &= D' - \beta DF' \\ D_2 &= D \end{aligned} \quad (3)$$

one has the corresponding Langevin equation (LE) for the evolution of x

$$\begin{aligned} \dot{x} &= -[\beta DF' - D'] + \sqrt{2D}\eta \\ \overline{\eta} &= 0 \\ \overline{\eta(t)\eta(s)} &= \delta(t - s) \end{aligned} \quad (4)$$

where $\eta(t)$ is a Gaussian-distributed white noise process, $\overline{(\cdot)}$ denotes ensemble or time averaging, and δ is the Dirac distribution. Equation 4 describes the evolution of a massless and memoryless Brownian particle under the influence of the potential F. The memoryless (Markov) assumption implicit in eq 1 is particularly restrictive. To determine whether the evolution of a given system is Markovian, one can employ a variety of tests, e.g., checking whether the autocorrelation function for the coordinate x follows single-exponential decay,[55] comparing empirically computed transition probabilities to predictions of the Chapman−Kolmogorov equation,[56] or testing the sensitivity of computed D and F to the trajectory sampling interval.[41] Although models that include the effects of positional memory may be more appropriate for some cases,[57] as in the generalized Langevin equation,[58,59] they require characterization of the memory function,[25,60,61] and are not considered here. It is encouraging, however, that the memoryless FPE in eq 1 has been found to reproduce folding dynamics of coarse-grained models for proteins.[41] In the remainder of this section, we assume that eq 1 is an acceptable model for the reaction dynamics, and focus our attention on the determination of $D(x)$ and $F(x)$. Since D and F, together with the initial and boundary conditions, completely determine the evolution of P, all statistical properties of the original reaction are also determined by them. For example, the (unimolecular) rate constants are given by

$$\begin{aligned} k_f^{-1} &= T_{0\to1} = \int_0^1 \int_0^y D^{-1}(y)e^{\beta[F(y)-F(z)]}\,dz\,dy \\ k_b^{-1} &= T_{1\to0} = \int_0^1 \int_y^1 D^{-1}(y)e^{\beta[F(y)-F(z)]}\,dz\,dy \end{aligned} \quad (5)$$

where $T_{a\to b}$ is the mean first passage time from state a to state b.

2.2. Review of Existing Methods for Computing *D* and *F*.

To compute F and D from simulation, one may recall that the FPE may be derived from the Kramers−Moyal (KM) expansion of transition probabilities,[54] which relates the nth order coefficients $D_{n>0}$ to the moments of P *via*

$$D_n(x) = \frac{1}{n!}\lim_{\tau\to 0}\frac{1}{\tau}\int (y-x)^n P(y,\, t+\tau|x,\, t)\, \mathrm{d}y \tag{6}$$

$$= \frac{1}{n!}\lim_{\tau\to 0}\frac{1}{\tau}\,\overline{[y(t+\tau)-y(t)]^n}\Big|_{y(t)=x} \tag{7}$$

For a system evolving via the Langevin equation, the coefficients D_1 and D_2 in the KM expansion truncated to second order are given in eq 3. Equation 7 corresponds to an average over one or several stochastic trajectories $y(t)$, which can be computed from simulations using small time steps τ, as done in ref 56. Because the $\tau \to 0$ limit cannot be realized in computer simulations, one can also compute the transition probability $P(y,\, t+\tau|x,\, t)$ for small τ. Assuming that τ is sufficiently small that D_1 and D_2 can be considered constant, we can solve eq 1 with the initial condition $P(z, t) = \delta(z-x)$ using the Fourier transform to obtain

$$P(y,\, t+\tau|x,\, t) = \frac{\exp\left(-\frac{[y-x-D_1(x,t)\tau]^2}{4D_2(x,t)\tau}\right)}{[4\pi D_2(x,\, t)\tau]^{1/2}} \tag{8}$$

Equation 8 is a Gaussian distribution with mean $\overline{y} = x + D_1\tau$ and variance $\overline{y^2} - \overline{y}^2 = 2D_2\tau$, so that eq 3 therefore implies

$$D(x) = \frac{1}{2\tau}\overline{[\mathrm{d}y(\tau) - \overline{\mathrm{d}y}(\tau)]^2} \tag{9}$$

$$F'(x) = -\frac{1}{\beta D}\left[\frac{\overline{\mathrm{d}y}(\tau)}{\tau} - D'\right] \tag{10}$$

where $\mathrm{d}y(\tau) \equiv y(t+\tau) - y(t)$, and the averages are conditional on $y(t) = x$, as in eq 7. Thus, in principle, eqs 9 and 10 allow simultaneous extraction of F and D from one or more MD (or Langevin dynamics) trajectories. After D is computed on a sufficiently fine mesh x_i via eq 9, D' can be obtained by finite differences, and F, by integrating eq 10 using numerical quadrature. Alternatively, computing D' can be avoided by integrating eq 10 analytically, i.e.,

$$F(x) = F(0) - \int_0^x \frac{\overline{\mathrm{d}y}(\tau)}{\beta\tau D}\Bigg|_{y(t)=x^*} \mathrm{d}x^* + \frac{1}{\beta}\log\frac{D(x)}{D(0)} \tag{11}$$

(where log denotes the natural logarithm), and applying quadrature to the remaining integral.

Unfortunately, the apparent simplicity of eqs 9 and 10 is not consistently realized in practice. First, it is clear that computing F from eq 10 reliably requires accurate and precise values of D. However, computing D from eq 9 is problematic. Equation 8 was derived for small τ, but its range of validity depends on the spatial variation of D_1 in the vicinity of x (e.g., F''). Thus, different choices for τ may produce different estimates for D. In principle, one could perform MD with a very small time step $\mathrm{d}t_{\mathrm{MD}}$ and take $\tau = \mathrm{d}t_{\mathrm{MD}}$. However, decreasing the time step requires longer wall-clock times to obtain converged averages. In addition, although we assumed memoryless dynamics *a priori* to motivate the development, that assumption might only be reasonably accurate for some $\tau > \tau_c$ (where τ_c is a system-dependent threshold, e.g., trajectory decorrelation time).[25,41,55] In that case, parametrizing the FPE based on a larger τ (over which the system loses some of its memory) would lead to a more faithful description of long-time dynamics of the system.

A different issue with using eqs 9 and 10 directly concerns the computation of D and F in regions with high F, e.g., those correponding to high FE barriers. These regions will be sampled poorly by unbiased MD simulations in view of the Boltzmann distribution. To address the sampling problem, biased simulation methods are typically used to enhance sampling in the desired ranges of the reaction coordinate x *via* restraints,[14−16,53,62−64] constraints,[65,66] or trajectory reinitialization in regions of interest.[25,67−69] All of the above methods are well-suited for the computation of the free energy, but only a few permit a simultaneous computation of the diffusion coefficient,[53,62,68] although some provide kinetic information via master equation approaches.[25,64,66,67,69,70]

In restrained simulations with a harmonic umbrella potential $1/2 \times k[x(\hat{r}) - x_0]^2$, $D(x)$ can be approximated from the positional autocorrelation function (see Appendix C) as[51,52,62,71]

$$D(x_0) = (\beta k)^{-2}\left[\int_0^\infty \overline{x(t)x(0)}\, \mathrm{d}t\right]^{-1} \tag{12}$$

Practical issues that affect the accuracy of eq 12 are the choice of the upper integration limit as well as the integration quadrature itself (since the correlation function will generally be noisy),[52] and the need to use k large enough that the effective free energy landscape for x is harmonic (Appendix C). The use of eq 12 for rate computations appears to be limited.[51,52,62,71]

Hummer[52] proposed a method to determine F and D from the same set of simulations using likelihood optimization via Bayesian inference. The x-space is divided into N bins, and one or more MD trajectories are sampled at a prescribed time interval. The transitions between bins that occur within the time interval are used to construct a likelihood function of transition rates between the bins. An optimized transition rate matrix is obtained by sampling the likelihood function using Monte Carlo simulations, and F and D are calculated from the matrix. This approach has been extended by use of cubic splines to represent F and D for improved accuracy, and to include the effects of time-dependent biasing forces.[53] The likelihood method is a significant improvement over the correlation method of eq 12, but its implementation is not straightforward because it requires computing finite-time propagators via a matrix exponential, and a separate MC sampling step; in addition, it is likely to be sensitive to the choice of the prior parameter distribution. A possible approach to simplify the algorithm would be to use likelihood maximization instead.[66] Another drawback is that the likelihood method is not well-suited to umbrella sampling (US) simulations, which are routinely used to sample high-energy regions in atomistic simulations (although a variant compatible with the adaptive biasing force method[19] has been given[53]). A related procedure to estimate diffusion constants from unbiased equilibrium trajectories was described by Krivov and Karplus[72] (see also ref 73). Using the short-time transition probability in eq 8 and assuming $D_1 = 0$, the authors relate the diffusion constant at position x_0 to the rate of crossing of the boundary $x = x_0$ by simulation trajectories. Although systematic tests of the accuracy of the method have not been published, the reliance on the short-time transition probability suggests that it will be sensitive to the choice of the transition time τ, as described above for eq 9. Another related method based on milestoning[25] was described by Mugnai and Elber.[68] The method is somewhat more general because it allows for a multidimensional reaction coordinate ($\vec{x}$) but does not produce accurate estimates of the free energy, for which a separate calculation is recommended.[68]

2.3. Computing F and D from MD with Flat-Bottom Restraints.

In view of the limitations of the existing methods outlined in section 2.2, we propose a simple modification of umbrella sampling for equilibrium processes, which permits accurate determination of F and D from the same simulation. The computed free energy is, in principle, exact and does not require "unbiasing" by, e.g., weighted histogram analysis (WHAM).[74] The method incurs negligible overhead compared with standard umbrella simulations, and requires very little additional simulation output. The method should be considered in cases where umbrella sampling is appropriate.

The essential idea of the present method is to restrict the x-domain over which eq 1 is to be parametrized to a window $[a, b] \equiv \mathcal{I}$, with $|b - a| = \Delta$ sufficiently small that for $x \in \mathcal{I}$, $F(x)$ is approximately linear and $D(x)$ is approximately constant. The reason for assuming a lower order approximation for D than for F is that, within the framework of eq 1, kinetic quantities depend exponentially on F but only linearly on D. To obtain D and F for an arbitrary domain, the domain is sampled in a series of such windows (see Figure 1a) and profiles of F and D over the entire domain are constructed using some type of quadrature. In this regard, the method is an instance of umbrella sampling.[14] However, the windows need not have overlap because F is obtained by integrating its derivative, which can be computed locally for each window, rather than by matching densities from different windows, as done in WHAM.[74] In this regard, the present method is similar in spirit to umbrella integration.[75]

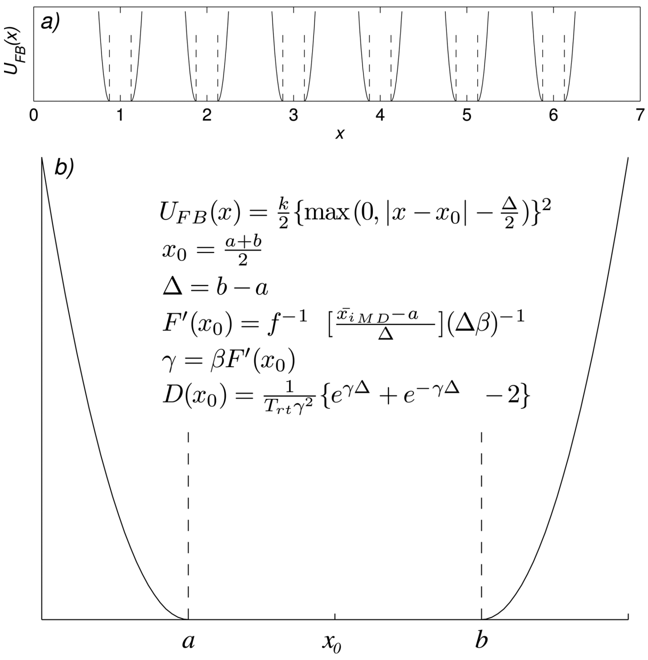

a)
$U_{FB}(x)$
0
1
2
3
4
5
6
7
x
b)
$U_{FB}(x) = \frac{k}{2}\{\max(0, |x - x_0| - \frac{\Delta}{2})\}^2$
$x_0 = \frac{a+b}{2}$
$\Delta = b - a$
$F'(x_0) = f^{-1}\ [\frac{\bar{x}_{iMD} - a}{\Delta}](\Delta\beta)^{-1}$
$\gamma = \beta F'(x_0)$
$D(x_0) = \frac{1}{T_{rt}\gamma^2}\{e^{\gamma\Delta} + e^{-\gamma\Delta}\ \ -2\}$
a
x_0
b

Figure 1. (a) The domain is discretized into multiple regions, each of which is sampled using a flat-bottom potential. (b) Close-up view of the flat-bottom restraint potential and a summary of the equations used to obtain $F(x_0)$ and $D(x_0)$ for the window centered on x_0. $T_{\rm rt}$ is the mean "roundtrip" time between the end points of the interval $[a, b]$ (eq 17), $\bar{x}_{\rm MD}$ is the average of the x_i coordinate time series sampled from simulations, conditional on $x_i \in [a, b]$, and the function f is given by eq A2 in Appendix A. The regions delimited by dashed lines are those from which simulation samples are used to compute F and D (see text).

Specifically, we choose a window of width Δ centered on x_0 (see Figure 1) such that $F''(x_0)(\Delta/2)^2 \ll 1$. Then, for each $x \in \mathcal{I}$, $F(x) \simeq F(x_0) + F'(x_0)(x - x_0)$, and, because D is assumed to be constant in the interval, eq 1 simplifies to

$$\dot{P} = D[\gamma P + P']' \tag{13}$$

where we have defined the constant $\gamma = \beta F'(x_0)$. If the samples x_i obtained from MD simulation are distributed according to the equilibrium (Boltzmann) solution of eq 13, the free energy derivative $F' = \beta^{-1}\gamma$ can be obtained from the appropriate averages of x_i ($\bar{x}_{\rm MD}$) by solving the equation

$$\bar{x}_{\rm MD} = \frac{\int_a^b x \exp(-\gamma x)\, dx}{\int_a^b \exp(-\gamma x)\, dx} \tag{14}$$

for γ, as described in Appendix A (a short MATLAB code is provided). An alternative method to compute F' is given in Appendix B; the alternative result can be compared with $\beta^{-1}\gamma$ above to ensure consistency and to assess sampling quality. Once F' is known, F is obtained by numerical integration. Using the trapezoidal rule for simplicity, we have

$$F(j) = F(1) + \sum_{i=2}^{j} \frac{F'(i-1) + F'(i)}{2}[x_{0,i} - x_{0,i-1}], \quad j = 2 \ldots M \tag{15}$$

where $F(1)$ is a normalization constant and M is the number of restraint windows, which are labeled in the order of increasing x_0's. Because the trapezoidal quadrature is second-order accurate with respect to the spacing between adjacent windows, the present method will be very efficient for systems with smooth free energy profiles, for which a small number of sparsely distributed windows will be adequate. However, for rugged profiles of F, such as those that could arise in complex conformational changes, a large number of finely spaced windows and/or a higher-order numerical quadrature may be required to determine F with sufficient accuracy. To optimize the computational cost for a general (unknown) free energy profile, we suggest that one begin with a coarse grid of uniformly spaced windows, doubling the resolution until desired accuracy is achieved. An example of such grid refinement is illustrated in section 3.4.

To obtain the diffusion coefficient using samples x_i from MD simulations, we first compute $T_{a\rightarrow b}$, the mean first passage time (MFPT) from a to b, assuming the dynamics of x obeys eq 1. Using the definition of γ, Δ, and the assumption of constant D in the interval $[a, b]$, we have[76]

$$
\begin{aligned}
T_{a\to b}(\gamma, D, \Delta) &= D^{-1}\int_a^b \int_a^y e^{\gamma(y-z)}\, dz\, dy \\
&= -\frac{1}{D\gamma}\left\{b - a - \frac{1}{\gamma}[e^{\gamma(b-a)} - 1]\right\} \\
&= -\frac{1}{D\gamma}\left[\Delta - \frac{e^{\gamma\Delta} - 1}{\gamma}\right] \\
&\simeq \frac{\Delta^2}{2D}\left[1 + \frac{\gamma\Delta}{3} + \frac{(\gamma\Delta)^2}{12} ...\right]
\end{aligned}
\tag{16}
$$

where we made explicit the dependence on D, γ, and Δ. By symmetry, $T_{b\to a}(\gamma, D, \Delta) = T_{a\to b}(-\gamma, D, \Delta)$ (e.g., reflecting the x-axis around x_0), and therefore,

$$
\begin{aligned}
T_{\rm rt} \equiv T_{a\to b} + T_{b\to a} &= \frac{1}{D\gamma^2}[e^{\gamma\Delta} + e^{-\gamma\Delta} - 2] \\
\simeq \frac{\Delta^2}{D}\left[1 + \frac{(\gamma\Delta)^2}{12} ...\right]
\end{aligned}
\tag{17}
$$

where we included the Taylor expansion up to second order in γ, and defined a mean roundtrip time $T_{\rm rt}$. Thus, the diffusion constant $D(x_0)$ for the window $[a, b]$ can be computed after γ is obtained from eq 14 as

$$
D = \frac{1}{\gamma^2 T_{\rm rt}}[e^{\gamma\Delta} + e^{-\gamma\Delta} - 2]
\tag{18}
$$

A few remarks about eqs 16 and 17 are in order. First, note that it is possible to obtain D from $T_{a\to b}$ (or $T_{b\to a}$) instead of $T_{\rm rt}$. However, unlike $T_{\rm rt}$, $T_{a\to b}$ is linear in γ for small $\gamma\Delta$. This implies greater sensitivity of the computed D to the FE derivative. Second, if $dF/dx \ll k_BT/\Delta$, the diffusion coefficient can be determined from the mean roundtrip time alone. Finally, the standard errors in D are easily obtained from those in dF/dx and $T_{\rm rt}$ (or $T_{a\to b}$) using uncertainty propagation via the chain rule.

$T_{\rm rt}$ is readily computed from simulations restrained to the window $[a, b]$. To do so, we first define a "flat-bottom" restraint potential centered on x_0 with width Δ as

$$
U_{\rm FB}(\hat{r}; x_0) = \frac{k}{2}\{\max[0, |x(\hat{r}) - x_0| - \Delta/2]\}^2
\tag{19}
$$

where $\max[a, b]$ returns the greater of a or b. Thus, $U_{\rm FB}$ is zero for $x \in [a, b] \equiv [x_0 - \Delta/2, x_0 + \Delta/2]$, and equals $k[x(\hat{r}) - (x_0 - \Delta/2)]^2/2$ and $k[x(\hat{r}) - (x_0 + \Delta/2)]^2/2$ for $x < a$ and $x > b$, respectively (see Figure 1). Using $\partial/\partial(\cdot)$ to denote derivatives w.r.t. an arbitrary quantity, it is easy to check that

$$
\begin{aligned}
\frac{\partial U_{\rm FB}}{\partial(\cdot)} &= k\{\max[0, (x - x_0) - \Delta/2] \\
&\quad + \min[0, (x - x_0) + \Delta/2]\}\frac{\partial(x - x_0)}{\partial(\cdot)}
\end{aligned}
\tag{20}
$$

$$
\begin{aligned}
&= -k\{\max[0, (x_0 - x) - \Delta/2)] \\
&\quad + \min[0, (x_0 - x) + \Delta/2)]\}\frac{\partial(x - x_0)}{\partial(\cdot)}
\end{aligned}
\tag{21}
$$

Equations 20 and 21 show that the restraint forces are computed with neglible additional effort relative to the standard harmonic potential. (The equivalence of eqs 20 and 21 is seen by interchanging $x(\hat{r})$ and x_0 in eq 19, or from the fact that $\max[0, x] = -\min[0, -x]$.) Additionally, using $\Delta = 0$ and that $x = \max[0, x] + \min[0, x]$, the energy and derivatives of the standard harmonic potential are recovered. Therefore, eqs 19 and 20 can be used as one-line modifications to existing harmonic potentials in any MD code without introducing a new functional form. If a periodic reaction coordinate with period P is used (e.g., dihedral or bond angles), then $x - x_0$ above (equivalently, $x_0 - x$) are to be evaluated modulo P to keep the difference within the domain of validity, identically to the case with standard harmonic potentials.

To compute the passage times $T_{a\to b}$ and $T_{b\to a}$ from restrained simulations, during the simulation, one records the times t at which the coordinate $x(t)$ crosses the boundaries $x = a$ or $x = b$, and the crossing direction. For example, using $\mathcal{I}^-$, $\mathcal{I}$, and $\mathcal{I}^+$ to denote the intervals $(-\infty, a)$, $[a, b]$, and (b, ∞), respectively, one records the pairs $(t, -1)$, $(t, 0)$, and $(t, 1)$ when $x(t)$ crosses into $\mathcal{I}^-$, $\mathcal{I}$, and $\mathcal{I}^+$, respectively. After all crossing events are recorded, the time record (t, i) is "pruned" to eliminate the time spent outside of $\mathcal{I}$. This step is equivalent to "reflecting" the coordinate back inside $\mathcal{I}$ after it reaches a or b. Since the average time between boundary crossings will usually be much longer than the simulation time step, it is clear that the additional output consumes a negligible amount of memory.

In practice, the pruning step is performed as follows. Starting from the beginning of the trajectory record (t_j, i_j), we find all k such that $i_k = 0$ (reentry into $\mathcal{I}$). For each k in increasing order, for all $k' \geq k$, we subtract from the time entries $t_{k'}$ the time offset $t_k - t_{k-1}$. Since the event (t_{k-1}, i_{k-1}) must correspond to an exit from $\mathcal{I}$, i.e., $i_{k-1} = \pm 1$, this offset corresponds to the time spent outside of $\mathcal{I}$. Thus, the pruning step is a simple relabeling of time values.

The passage times are computed from the pruned trajectory as follows. First, we discard all events (t_k, i_k) with $i_k = 0$ (reentry into $\mathcal{I}$), since the pruned trajectory is confined to $\mathcal{I}$. For all remaining events k, $i_k = \pm 1$, which assigns the trajectory to one of the boundaries (or milestones[25,64]), i.e., -1 for a and 1 for b. Instantaneous passage times $\mathcal{T}_{a\to b}$ are computed for each k that satisfies $i_{k-1} \neq i_k = -1$ as $\mathcal{T}_{a\to b} = \min\{t_n: n > k, i_n = 1\} - t_k$ (similarly for $\mathcal{T}_{b\to a}$). The MFPT and its variance are then computed from the sample of individual passage times. Athough the pruning approach can lead to inefficient sampling if long portions of MD trajectories need to be discarded, the inefficiency is minimized by using a high force constant in the restraint potential. Additionally, we expect that the configurations of the all-atom system $(\hat{r})$ visited when $x \notin \mathcal{I}$ contribute to faster trajectory decorrelation. This would improve the statistical independence of the individual passage events within $\mathcal{I}$. We note that similar trajectory pruning was used by Maragliano et al.[64] in Voronoi tessellation simulations.

3. RESULTS

3.1. 1D Potentials. First, we test the accuracy of eq 17 using a Brownian dynamics (BD) simulation on a 1D linear potential energy with a constant diffusion coefficient. For this test, a single simulation window is used. The BD equations are advanced with the time step dt using the forward Euler scheme,[1,76] i.e.,

$$
x^{n+1} = x^n + D_1\, dt + \sqrt{2D_2\, dt}\,\eta_n
\tag{22}
$$

where η_i is the derivative of the Wiener process (eq 4), and the drift and diffusion coefficients D_1 and D_2 are as in eq 3. Nine sets of simulations with different parameters were performed, corresponding to Table 1, with 25 identical simulations per set to compute uncertainties in the MFPTs.

Table 1. Simulation Parameters[a]

#	Δ	F'	$T_{a\to b} \pm (SE)$	$T_{b\to a} \pm (SE)$
1	0.010417	0	111.8 ± 3.5	112.4 ± 3.8
2	0.010417	1	116.8 ± 3.2	107.4 ± 2.9
3	0.010417	2	119.2 ± 3.8	103.5 ± 3.2
4	0.010417	5	132.6 ± 4.6	94.2 ± 3.1
5	0.010417	10	165.0 ± 7.5	80 ± 3
6	0.010417	25	342 ± 75	56.5 ± 8.7
7	0.010417	50	1337 ± 536	0.03 ± 8
8	0.020833	0	411 ± 20	447 ± 20
9	0.020833	1	475 ± 22	405.7 ± 18

[a] The remaining parameters were as follows: $D = 0.005$, $\beta = 10$, $\Delta t = 0.0001$ (simulation time step), $k = 3600$ (U_{FB} force constant). Each simulation was integrated for 500,000 steps and repeated 25 times to compute uncertainties. The physical units are arbitrary, and the data are plotted in nondimensional form in Figure 2.

Figure 2 shows that the MFPTs computed from the BD simulations agree with eq 17 to within the numerical uncertainty over a range of parameters. For convenience, the data plotted are nondimensional, i.e., $(T_{\mathrm{rt}} D \Delta^{-2})$ vs $(\gamma\Delta)$. In addition, the second-order approximation is accurate up to $F'\Delta \simeq k_{\mathrm{B}}T$, while the first-order approximation is accurate only for $F'\Delta \leq 0.25 k_{\mathrm{B}}T$ (see eq 17). Beyond these limits, the truncated expressions will progressively underestimate D.

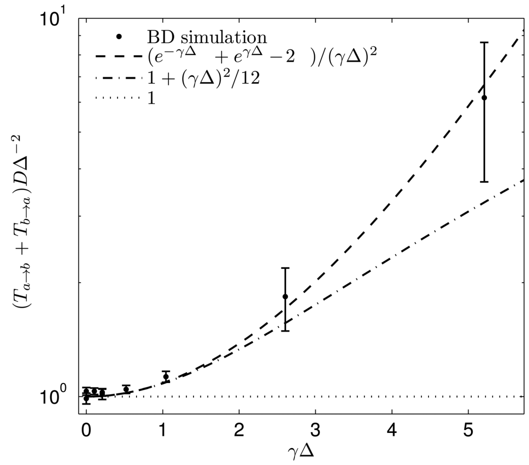


Figure 2. Comparison of MFPT computed from Brownian dynamics (BD) simulation with the analytical result in eq 17. For convenience, the data are made nondimensional. BD simulation parameters are described in the text, and listed in Table 1.

Next, for a constant diffusion coefficient, we test the accuracy of the two free energy computation methods described in Appendices A and B. We employ the 1D potential $F(x) = 1 + \cos 2x$ on the domain $[\pi/2, 3\pi/2]$ discretized using a 25-point equispaced grid, and examine the effect of the restraint force constant on the FE error. For this test, we use Langevin dynamics with $D = 0.5$ and $\beta^{-1} = 0.1$ to propagate particles with mass 0.2. The equations are integrated at $dt = 0.0005$ for 5 million steps, and snapshots are collected every 100 steps. The width of the flat-bottom window (Δ) is set to $\pi/48$, which corresponds to half of the distance between adjacent grid points, i.e., $dx = \pi/24$. Figure 3 shows that both methods are accurate to within a few percent for the test problem, provided that the force constant is sufficiently large (in this case, $dx^2k \geq 100$). The FE calculation based on averaging (eq A3) is much less sensitive to the choice of k, as expected. However, for small values of k ($dx^2k \sim 10$), the flat-bottom potential is not stiff enough to guarantee that enough samples fall inside the flat-bottom portion of the restraint window to give accurate results. It is also evident that the restraint force method (eq B5 in Appendix B) provides a better approximation to the smoothed free energy than to the true free energy, as expected from the results of the appendix, but the differences are within 1%.

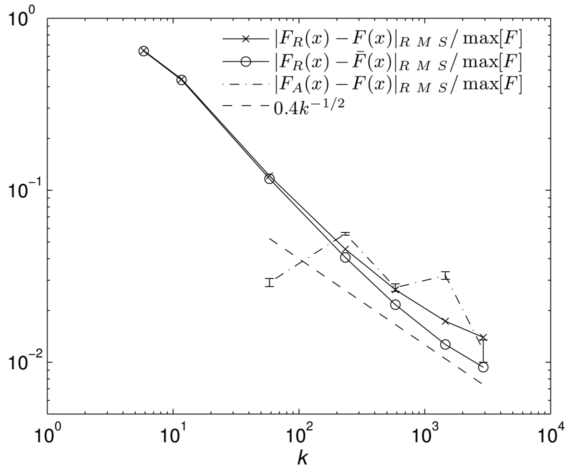


Figure 3. Normalized RMS error between the exact ($F(x)$) or smoothed ($\overline{F}(x)$) potential and the FE computed using MD restraint forces (F_R from eq B5) and the FE computed using the average x-coordinate in the window (F_A from eq A3). The force constant is normalized by the grid resolution $dx = \pi/24$, i.e., $k \to dx^2k$.

The final 1D example is the benchmark used by Hummer,[52] and later by others.[53,68] Both F and D are position-dependent with $F = 1 + \cos 2x$ and $D = 0.1(2 + \sin x)$, shown in Figure 4. For this test, we use the Euler integrator in eq 22 to perform BD simulations with various parameters listed in Table 2 together with the results. Simulations 1−6 use the present method and show the effects of the force constant, flat-bottom

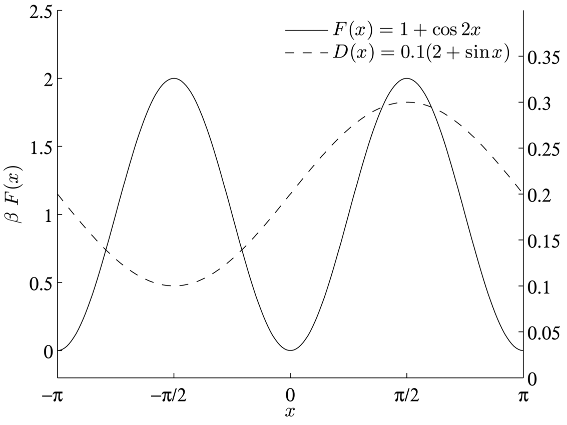


Figure 4. $F(x)$ and $D(x)$ of the 1D diffusive model used by Hummer.[52]

Table 2. Present and Previous Results of Simulations Performed on the 1D Diffusive Model[52] (See Figure 4)[a]

#	k^b	# steps	dt	Δ^b	$\lvert D(x) - D_{\mathrm{sim}}(x)\rvert_{L^2}$	$\lvert F(x) - F_{\mathrm{sim}}(x)\rvert_{L^2}$
			Present Method, eqs 18 and A3			
1	100	120 million	0.001	1	0.0051	0.025
2	100	120 million	0.001	0.5	0.0040	0.035
3	100	120 million	0.001	0.25	0.0035	0.039
4	25	120 million	0.001	1	0.0048	0.041
5	400	120 million	0.001	1	0.0029	0.026
6	100	12 million	0.0001	0.25	0.0073	0.3/0.17[c]
			Harmonic Restraint, eqs C4 and B5			
7	5	120 million	0.001	0	0.0104/0.0209[d]	0.095
8	25	120 million	0.001	0	0.0063/0.0084[d]	0.038
9	100	120 million	0.001	0	0.0165/0.0150[d]	0.055
10	400	120 million	0.001	0	0.0182/0.0182[d]	0.146
			Short-Time Solution of the FPE, eqs 9 and 10			
11	100	120 million	0.001	1	0.009/0.01/0.016[g]	0.24
12	100	120 million	0.001	0.5	0.009	0.047
			Previous Results[e]			
ref 52	n/a	100 million	0.001	n/a	0.013	0.05
ref 53	n/a	100 million	0.001	n/a	0.003	0.04
ref 68	n/a	unspecified	0.00001	n/a	0.007	n/a[f]

[a] F_{sim} and D_{sim} correspond to simulation results. The L^2 norm is computed as the RMS error over the 24-point equispaced grid. [b] The force constant k and window width Δ are normalized by the grid resolution $dx = 2\pi/24$, i.e., $k^* = k(dx^2)$, $\Delta^* = \Delta/dx$. [c] The second value was computed using restraint forces, eq B5. [d] Data are shown as "x/y" and correspond to the calculation of D from eq C4 with the upper integration limit 2τ and 5τ, respectively, where $\tau = \min\{t > 0: \overline{y(t)y(0)} = 0\}$. [e] Data was extracted from plots of $D(x)$ and $F(x)$ using the free software g3data.[77] [f] An incorrect function is shown in the reference. [g] Data shown as "$x/y/z$" correspond to τ = 5dt, 10dt, and 15dt, respectively.

window width, and simulation length on the accuracy of the F and D profiles. Simulations 7−10 employ standard harmonic restraints ($\Delta = 0$), and compute D *via* the correlation formula, eq 12, and simulations 11−12 use the short-time solution of the FPE, eqs 9 and 10. The accuracy of the previously published results is estimated from the graphs in the corresponding references.

It is clear from the results of simulations 1−12 that the present method is superior to the correlation method and the short-time solution method, the difference being most apparent in the calculation of D. In particular, the present method is not sensitive to the choice of simulation parameters (simulations 1−5). It is also noteworthy that an accurate estimate of the diffusion coefficient can be obtained from a much shorter simulation using the present method, although a small window width may be required to accumulate enough passage events (simulation 6). Only the modified likelihood method of Comer et al.[53] gives results of similar accuracy. The advantage of the present method is its simplicity and immediate applicability in umbrella sampling simulations.

3.2. 2D Potential.

In this section, we determine MFPTs between minima on a model 2D potential energy landscape. The potential energy is given by

$$E(y, z) = \sin(\pi y)\cos(2\pi z) + \frac{1}{(1.1 - y^4)(1.1 - z^4)} \qquad (23)$$

shown in Figure 5 for the domain of interest $|y|, |z| \le 1$. The first term in eq 23 creates five local minima (located near images 1, 7, 13, 19, and 25 in Figure 5; see also Table 3), and the second term is a simple confining potential that diverges along the lines $y^4 = 1.1$ and $z^4 = 1.1$, providing impenetrable walls to restrict sampling to the region of interest. We denote spatial coordinates r by (y, z) rather than (x, y) to avoid confusion with the reaction coordinate definition x used in section 2.

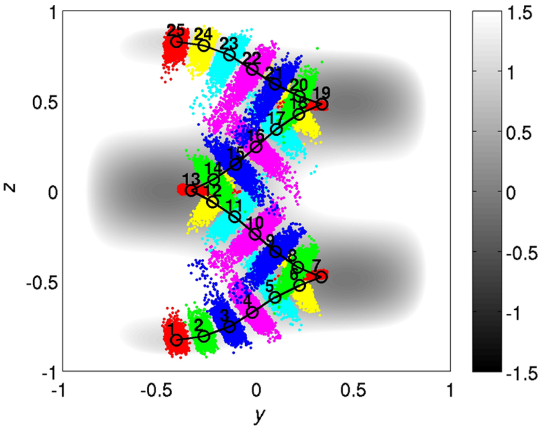


Figure 5. MD umbrella sampling in the vicinity of a minimum energy path (drawn in black) connecting images 1 and 25 along a model 2D potential. Samples corresponding to different images are plotted in (arbitrary) different colors for clarity; red: #1, 7, 13, 19, 25; green: #2, 8, 14, 20; blue: #3, 9, 15, 21; magenta: #4, 10, 16, 22; cyan: #5, 11, 17, 23; yellow: #6, 12, 18, 24. Points on the MEP are drawn as black circles. The potential energy is given by eq 23.

The reaction coordinate (x) for the 2D potential is constructed using piecewise linear segments connecting M = 25 neighboring points on the minimum energy path (MEP). To determine the MEP, we use the zero-temperature string method. Because the method was described in detail by E et al.,[35] only a summary is provided below. Starting from the straight line path (string) connecting the initial end points, $(y_1, z_1) = (-0.4, -0.9)$ and $(y_M, z_M) = (0.4, 0.9)$, steepest descent minimization was applied to each string point

$$\mathbf{r}_i^{n+1} = \mathbf{r}_i^n - h \times \nabla_{\mathbf{r}} E(\mathbf{r}) \qquad (24)$$

DOI: 10.1021/acs.jpcb.6b02139
J. Phys. Chem. B 2016, 120, 8457−8472

Table 3. Coordinates of the Discretized Minimum Energy Path in Figure 5

#	y	z	#	y	z
1	−0.41354149	−0.82797442	14	−0.22195392	0.06307774
2	−0.27246322	−0.80686235	15	−0.10728180	0.14756095
3	−0.13978017	−0.75419631	16	−0.00271479	0.24478926
4	−0.02207749	−0.67335878	17	0.10186907	0.34202615
5	0.09487859	−0.59139433	18	0.21743994	0.42530505
6	0.22104910	−0.52529109	19	0.33743802	0.48069920
7	0.33393083	−0.47836426	20	0.22023199	0.52509035
8	0.21158566	−0.42248448	21	0.09432111	0.59169826
9	0.09713014	−0.33766441	22	−0.02256344	0.67376977
10	−0.00731233	−0.24026067	23	−0.14024958	0.75463161
11	−0.11225282	−0.14344043	24	−0.27294730	0.80722673
12	−0.22780198	−0.06031313	25	−0.41405352	0.82815388
13	−0.33661052	0.00253399			

followed by linear interpolation to enforce the uniform arclength constraint $\|\mathbf{r}^n_{i+1} - \mathbf{r}^n_i\| = \text{const}$ for $i \in \{1...M - 1\}$. The minimization/interpolation procedure was repeated for 200 iterations using $h = 0.01$, and the MEP was defined as $\boldsymbol{\phi} = \mathbf{r}^{200}$; $\boldsymbol{\phi}$ is shown in Figure 5.

To compute F and D along the entire reaction coordinate using umbrella sampling, for each $\boldsymbol{\phi}_i$, we define the reaction coordinate locally. First, we define the tangent vectors

$$\begin{aligned} \mathbf{t}_i &= \frac{\boldsymbol{\phi}_{i+1} - \boldsymbol{\phi}_{i-1}}{\|\boldsymbol{\phi}_{i+1} - \boldsymbol{\phi}_{i-1}\|}, \quad M > i > 1 \\ \mathbf{t}_1 &= \frac{\boldsymbol{\phi}_2 - \boldsymbol{\phi}_1}{\|\boldsymbol{\phi}_2 - \boldsymbol{\phi}_1\|} \\ \mathbf{t}_M &= \frac{\boldsymbol{\phi}_M - \boldsymbol{\phi}_{M-1}}{\|\boldsymbol{\phi}_M - \boldsymbol{\phi}_{M-1}\|} \end{aligned} \tag{25}$$

The local reaction coordinate x_i corresponding to image $\boldsymbol{\phi}_i$ is

$$\begin{aligned} x_i(\mathbf{r}) &= \mathbf{t}_i \cdot [\mathbf{r} - \boldsymbol{\phi}_{i-1}], \quad M \geq i > 1 \\ x_1(\mathbf{r}) &= \mathbf{t}_1 \cdot [\mathbf{r} - \boldsymbol{\phi}_1] \end{aligned} \tag{26}$$

It measures the distance between $\mathbf{r}$ and $\boldsymbol{\phi}_i$ along the local tangent $\mathbf{t}_i$. The equations $x_i(\mathbf{r}) = C$ define a family of lines parametrized by the constant C that are locally perpendicular to the MEP. Because lines that correspond to different windows will intersect for nonzero string curvature (see sampling data in Figure 5), the above construction does not define a single-valued global reaction coordinate, i.e., $x(\mathbf{r})$ is not uniquely determined for some $\mathbf{r}$. A method that has been used to overcome this issue uses Voronoi tessellations.[64,66] For simplicity, in this illustration, we do not employ this method here but mitigate the problem by using an additional flat-bottom potential to restrain the distance locally perpendicular to the MEP. (These two methods for dealing with the crossing of reaction coordinate isosurfaces were compared in ref 78.) This potential is defined for $M \geq i > 1$ by

$$U^i_\perp(\mathbf{r}) = \frac{\kappa}{2} \max\{\|[\mathbf{r} - \boldsymbol{\phi}_{i-1}] \cdot [I - \mathbf{t}_i \mathbf{t}_i^T]\| - \delta_0, 0\}^2 \tag{27}$$

where $(\cdot)^T$ denotes transposition. (For $i = 1$, $\boldsymbol{\phi}_1$ is used in place of $\boldsymbol{\phi}_{i-1}$ above.) Thus, $U^i_\perp$ creates a restoring force perpendicular to segment i of the MEP when the perpendicular distance between $\mathbf{r}$ and $\boldsymbol{\phi}_i$ exceeds δ_0. Samples obtained from restrained BD simulations along the MEP using $\delta_0 = 2.5 \times \mathrm{d}\boldsymbol{\phi}$, where $\mathrm{d}\boldsymbol{\phi}$ is the distance between adjacent points on the MEP, are shown in Figure 5; simulation details are given below.

To compute the free energy derivative and diffusion coefficient in each window, we define the constants

$$x_{0,i} = \begin{cases} 0 & i = 1 \\ \|\boldsymbol{\phi}_i - \boldsymbol{\phi}_{i-1}\| = \mathrm{d}\boldsymbol{\phi} & M \geq i > 1 \end{cases} \tag{28}$$

where the second line holds because of the uniform arclength constraint. The above $x_{0,i}$ and a choice of Δ (given below) are sufficient to define the potential U_{FB} in eq 19, but computing F using eq 15 requires replacing $[x_{0,i} - x_{0,i-1}]$ by $\mathrm{d}\boldsymbol{\phi}$. This is equivalent to adding to the locally defined x_i and $x_{0,i}$ the arclength segment up to image $i - 1$, i.e., $x_i \to x_i + (i - 1)\mathrm{d}\boldsymbol{\phi}$ and $x_{0,i} \to x_{0,i} + (i - 1)\mathrm{d}\boldsymbol{\phi}$, and using eq 15 as written.

Each window around $\boldsymbol{\phi}_i$ is sampled using the BD integrator in eq 22 applied in 2D, i.e.,

$$\mathbf{r}^{n+1} = \mathbf{r}^n - \beta D \nabla_{\mathbf{r}}[E(\mathbf{r}) + U_{\text{FB}}(x_i(\mathbf{r}); x_{0,i}) + U^i_\perp(\mathbf{r})]\,\mathrm{d}t + \sqrt{2D\,\mathrm{d}t}\,\boldsymbol{\eta}_n \tag{29}$$

where the 2 × 2 diffusion tensor $\boldsymbol{D}$ is replaced by a scalar D because the former is isotropic, i.e., $\boldsymbol{D} = D\boldsymbol{I}$. The temperature β^{-1} was set to 0.2, the diffusion coefficient D was 0.01, the constants in the restraint potential U_{FB} were $k = 5 \times (\mathrm{d}\boldsymbol{\phi})^{-2}$ and $\Delta = 0.5 \times \mathrm{d}\boldsymbol{\phi}$, and those in the potential $U_\perp$ were $\kappa = 5 \times (\mathrm{d}\boldsymbol{\phi})^{-2}$ and $\delta_0 = 2.5 \times \mathrm{d}\boldsymbol{\phi}$ (normalization by the arclength increment $\mathrm{d}\boldsymbol{\phi} = 0.14$ is used for convenience). For each $\boldsymbol{\phi}_i$, eq 29 was integrated for 1 million steps with the time step $\mathrm{d}t = 0.001$. Figure 6a shows F computed from eq 18. The free energy has only slightly lower minima than the potential energy, indicating that entropy differences are minor contributors to the transition free energy. Figure 6b shows that the diffusion coefficient obtained from simulation data using eq A3 has high accuracy and precision, the average value over all windows being $\bar{D} = 0.01 \pm 0.0002$; i.e., the average value corresponds exactly to the value specified in simulation, and the standard error is 2%.

To compare the kinetics obtained from the diffusion model with those computed directly from the actual dynamical system, we calculate the MFPT from the second minimum (image 7 in Figure 5) to images 8−19 using diffusion theory (eq 5) and unbiased BD simulation. The image range 7−19 is appropriate because it spans the three long-lived metastable states in Figure 5, escapes from which are rare events. To compute the MFPTs from unbiased simulation, 25 BD trajectories are initialized

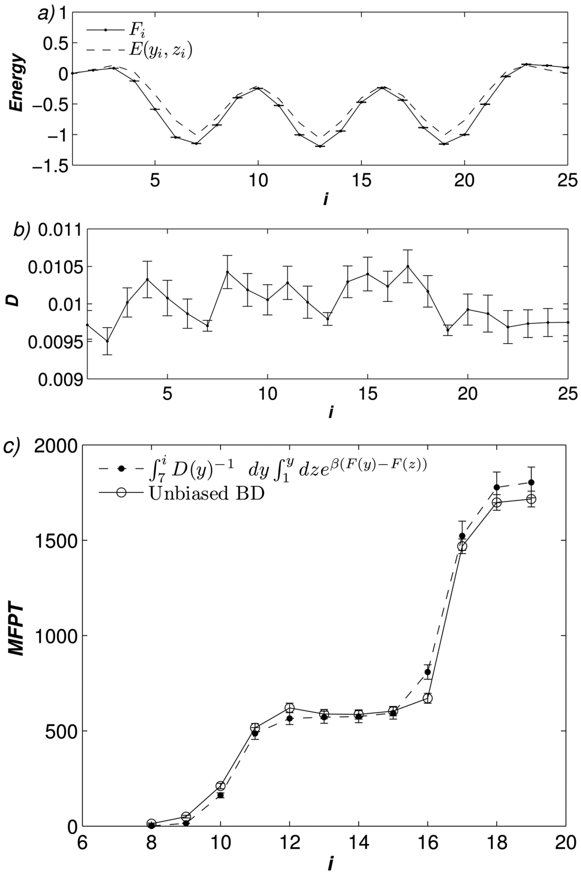


Figure 6. (a) Free energy F, (b) diffusion coefficient D, and (c) mean first passage time from image $i = 7$ to images $i \in [8...19]$ for the 2D energy landscape in Figure 5. Images 7, 13, and 19 correspond to the three long-lived metastable states along the reaction path.

from state 7, and integrated for 10 million steps each with time step 0.001, as before. Configurations are saved once in every 1000 steps, corresponding to $\delta t = 1$. Each configuration $\mathbf{r}^n$ at time step n is assigned to the image $\boldsymbol{\phi}_m$ which minimizes the distance $\|\mathbf{r}^n - \boldsymbol{\phi}_j\|$ over $1 \leq j \leq M$. This corresponds to introducing an indicator $i_n = m$ for each configuration. For consistency with the restrained simulations, configurations for which the perpendicular distance to the coresponding MEP segment, e.g., $\|[\mathbf{r}^n - \boldsymbol{\phi}_{m-1}]\cdot[I - \mathbf{t}_m\mathbf{t}_m^T]\|$ for $1 < m \leq M$, exceeds the value δ_0 prescribed in the restrained simulation are discarded, along with the corresponding indicator. MFPTs $T_{a\to b}$ are computed from the series $\{t_n = n \times \mathrm{d}t, i_n\}$ as described in section 2.3 for pruned trajectories (with the image indices $1...M$ replacing the indicators ± 1).

Figure 6c shows that the agreement between diffusion theory and direct simulation is good for all images spanning the three long-lived minima. The relatively modest differences between the two data sets are likely to be caused by the crossing of the sampling regions mentioned above, and the lack of a special treatment of path curvature.[78]

3.3. β-Hairpin from Protein G.

We apply the present method to compute the free energy profile and diffusion coefficient along a conformational transition in a 16-residue β-hairpin fragment of streptococcal protein G.[80] This miniprotein has been used previously as a realistic test system for the application of enhanced sampling methods to biomolecules.[78,79,81−83] The transition pathway considered here connects the β-sheet and α-helical conformations of the

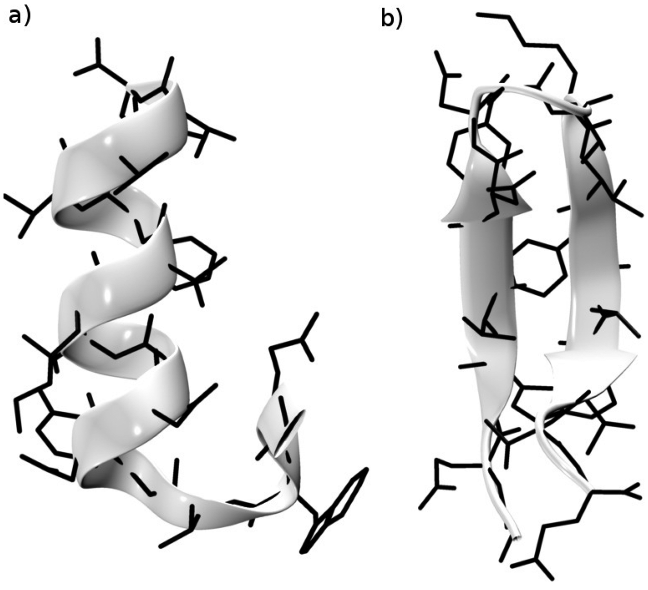


Figure 7. Equilibrated structures of the 16-residue peptide from protein G in (a) α-helical conformation and (b) β-sheet conformation. Details of the equilibration are given in the text. The N-terminal domain in panel a is at the bottom. Reprinted with permission from ref 79. Copyright 2013 American Chemical Society.

miniprotein (see Figure 7); it was investigated previously[78,79] using the FACTS solvation model.[85] In particular, the free energy difference between the end point states was computed using the path-independent confinement method[79,82] and by the string method[78,84] (see Table 4). The reaction pathway describing the transition is modeled using the finite-temperature string (FTS) method in rotation-invariant coordinates, described in detail by Ovchinnikov and Karplus.[78] The present method differs from that of ref 78 in that (1) flat-bottom restraints are used to compute the free energy and diffusion coefficients and (2) the string is defined in coarse-grained positional (rather than atomic) coordinates. The coarse-grained coordinates are selected as follows. For each residue in the miniprotein, two positional coordinates are defined, (i) the center-of mass (COM) position of the backbone atoms (which include atoms CA, O, C, N, HA, HN, OT1, and OT2)[86] and (ii) the COM position of the remaining (side chain) atoms. Details on the system preparation and on the calculation of the minimum-energy path (MEP) by the zero-temperature string method are provided in ref 78. The sequence of transition events in the MEP corresponds, essentially, to a sequential unwinding of the α-helical turns, beginning with the N-terminal end of the helix, followed by a "zipping" up of the intrastrand hydrogen bonds to form the β-sheet conformation.[78]

Table 4. Free Energy Differences between End Point States of the β-Hairpin[a]

method	$\Delta F_{\alpha\to\beta}$ (kcal/mol)
eq A3	−8.2 ± 1
FTS[78]	−8.25
SCM[79]	−6.7 ± 0.4

[a] SCM is the simplified confinement method described in ref 79. FTS is the finite-temperature string method.[84]

Using the MEP spanned by 128 replicas, the reaction coordinate for the transition is constructed locally for each

MEP image $\boldsymbol{\phi}_i$ in analogy to the 2D system (eqs 25 and 26). However, all displacement vectors are computed in mass-weighted coordinates using a rigid-body invariant metric in the local frame of $\boldsymbol{\phi}_i$ (see eqs 4, 8, and 9 in ref 78), summarized below. First, each string image in increasing order $\boldsymbol{\phi}_{i>1}$ is superposed onto the previous image $\boldsymbol{\phi}_{i-1}$. This procedure determines a rotation $A(\boldsymbol{\phi}_i, \boldsymbol{\phi}_{i-1})$ such that the distance $\|\boldsymbol{\phi}_i - B\boldsymbol{\phi}_{i-1}\|$ is the minimum over all possible rotations B (we call A a best-fit). The sequential superposition implies that, for each $\boldsymbol{\phi}_i$, the neighboring images $\boldsymbol{\phi}_{i\pm1}$ are best-fit onto $\boldsymbol{\phi}_i$. String tangents are defined by eq 25, and the local reaction coordinates are defined by

$$x_i(\mathbf{r}) = \mathbf{t}_i\cdot[A(\mathbf{r}, \boldsymbol{\phi}_i)\mathbf{r} - \boldsymbol{\phi}_{i-1}], \qquad M \geq i > 1$$

$$x_1(\mathbf{r}) = \mathbf{t}_1\cdot[A(\mathbf{r}, \boldsymbol{\phi}_1)\mathbf{r} - \boldsymbol{\phi}_1] \qquad (30)$$

where mass weighting is implied in the scalar product (see ref 78). Simulations corresponding to different images were performed simultaneously on a supercomputing cluster using 2.1 GHz AMD Opteron 6172 processors with 4 processor cores per string image, for a total of 512 cores. 40 ns of simulation per image were performed to compute the free energy and the diffusion coefficient. The constants in the restraint potential U_{FB} were k = 20 kcal/mol × $(\mathrm{d}\boldsymbol{\phi})^{-2}$ and Δ = 0.5 × $\mathrm{d}\boldsymbol{\phi}$ (normalization by the arclength increment $\mathrm{d}\boldsymbol{\phi}$ is used for convenience, as for the 2D problem).

Panels a and b of Figure 8 show F and D computed from eqs 18 and A3, respectively. The free energy curve is relatively smooth, indicating that the transition path is well resolved by the 128 replicas. The β-hairpin state is energetically more favorable than the α-helical state by ≃8.2 ± 1 kcal/mol, consistent with prior string method results,[78] and in fair agreement with a path-independent confinement method ($\Delta F_{\alpha\to\beta}$ = −6.7 kcal/mol; see Table 4). This is also the physically expected result, because the β-sheet conformation is known to be stable as part of a crystal structure.[80]

The diffusion coefficient is shown in Figure 8b in units of cm^2/s. To compute D in the units of cm, we used as the dimensional length scale the RMSD between adjacent string replicas computed from the MEP, which was 0.27 Å. (The spatial scale of D is chosen somewhat arbitrarily, since the MFPT of transition does not depend on it.) It is noteworthy that D is relatively uniform along the transition path, with the average value of $(5.27 \pm 1.55) \times 10^{-11}$ cm^2/s. The uniformity of D is not surprising, since the Jacobian of the coordinate transformation to define the coarse-grained variables (essentially, centers of mass) is expected to be nearly constant.[78] (The reason that it is not exactly constant is the use of best-fit rotations in the scalar product metric, as discussed in ref 78; the protein conformations are not identical along the path, which corresponds to differences in the rotation matrices and their derivatives along the reaction path.) For more complicated coordinates, e.g., fraction of native contacts, the diffusion coefficient is likely to be nonuniform.[41] Using the free energy profile and diffusion coefficient, we found the MFPT from the α-helical state to the β-sheet to be ∼40 s, which can be taken as the inverse of the interconversion rate via the pathway described by the MEP. To show that the variations in D have little effect on the interconversion rate, the MFPT was recomputed using the average diffusion coefficient (dashed line in Figure 8c), and shown to agree well with the original (variable-diffusion) curve. Interconversion via other pathways is likely to be important, as discussed in ref 78. Additional pathways are not considered here, however, since the present objective is to demonstrate applications of the method.

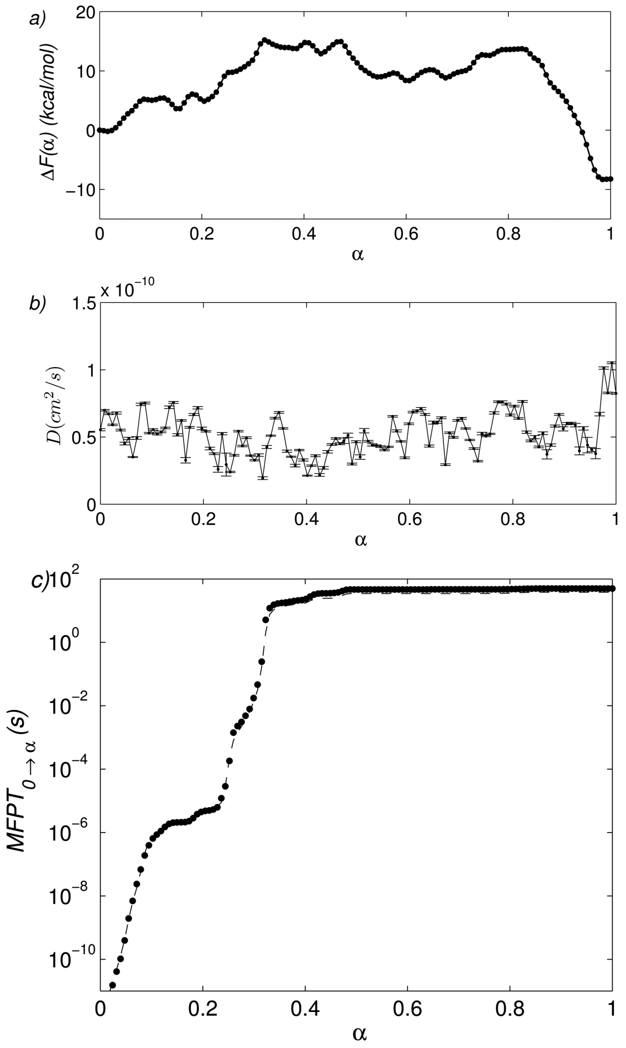


Figure 8. (a) Free energy F, (b) diffusion coefficient D, and (c) MFPT from α = 0 to 1 ≥ α > 0 for the α-helix ↔ β-sheet conformational transition in the hairpin miniprotein. In part c, the dashed line is the MFPT recomputed using the average diffusion coefficient.

3.4. Optimization of Domain Discretization. In the preceding applications, the distribution of restraint windows was chosen *a priori*. Since the free energy profile is unknown for many processes, we suggest a simple general strategy to optimize the locations of restraint centers. First, the reaction coordinate in the domain of interest is discretized using a uniform coarse grid, and the corresponding free energy profile is computed. This step is possible regardless of the choice of grid because window overlap is not needed. The resolution of the grid is then increased by inserting an additional window halfway between each pair of adjacent windows, and the free energy profile is recomputed. Note that new simulations are only required for the inserted windows. This procedure can be repeated until the free energy profile does not change from one iteration to the next to within a desired tolerance. To illustrate this approach, we compute the free energy profile for the unfolding of deca-L-alanine in a vacuum[19] using CHARMM. The CHARMM22 force field without CMAP was used, all nonbonded interations were switched off smoothly in the range 10−12 Å, and the temperature was kept at 300 K using

Langevin dynamics with friction $\gamma = 1$/ps for all atoms. All restrained simulations were performed for 2 ns with a time step of 1 fs, and $\Delta = 0.1$ Å. The reaction coordinate is taken to be the distance between the C_α atoms of the first and last residues, and the region of interest spans 12−32 Å.[19] The free energy profiles computed at different resolutions are given in Figure 9.

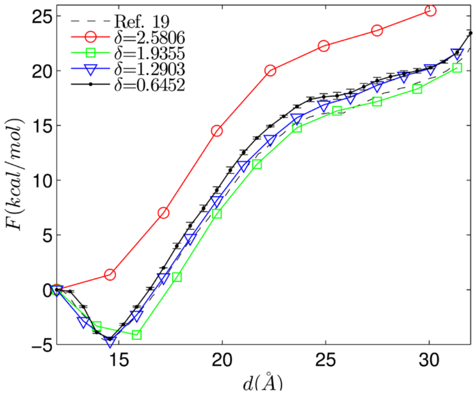


Figure 9. Free energy profile for the unfolding of deca-alanine. The error bars corresponding to the fine resolution data were computed by comparing two independent simulations started from different initial structures.

Even a coarse discretization with 8 windows separated by $\delta \sim 2.6$ Å reproduces the overall shape of the profile but does not capture the finer-scale details in the vicinity of the profile minimum. On the other hand, increasing the resolution from 16 windows ($\delta \sim 1.3$ Å) to 32 windows ($\delta \sim 0.65$ Å) does not substantially refine the value or location of the free energy minimum, or the free energy of the stretched state. An alternative initial discretization with 11 windows ($\delta \sim 1.9$ Å) provides a result that is intermediate in accuracy between those obtained with 8 and 16 windows, as expected. To optimize the computational efficiency further, one can also refine the grid only in the regions of curvature, e.g., near the minimum in Figure 9.

4. CONCLUDING DISCUSSION

We described a simple yet accurate method for computing simultaneously free energies and diffusion constants from restrained simulations within the framework of the Smoluchowski diffusion equation. The accuracy of the method stems from using a sufficiently fine discretization of the desired reaction coordinate so that the free energy profile between adjacent sampling windows can be considered linear. Essentially, the linear approximation implies a steady state analytical solution of the Smoluchowski equation, using which the free energy derivative and diffusion coefficients are easily related to samples from simulations. The sampling approach is loosely based on the idea of restrained Voronoi simulations,[64] and relies on the use of flat-bottom restraint potentials. These potentials were implemented by a generalization of harmonic restraints, requiring a one-line modification for the harmonic restraint energy and a one-line modification for the restraint forces. The standard harmonic potential is recovered when the flat-bottom region has zero width. Because the method can be viewed as an instance of umbrella sampling, it is trivially parallel, and simulations corresponding to different windows can be run on separate processor groups. Two methods of computing the free energy derivative from the restrained simulations were presented and found to be of comparable accuracy. Computation of the diffusion coefficient required only a time record of crossings in and out of the flat-bottom window, producing minimal additional simulation output.

The present method was compared to the existing methods of computing diffusion coefficients[52,53,62,71] using a model problem with position-dependent diffusivity.[52] The only method that was comparable in accuracy is that of Comer et al.,[53] which uses splines to represent the free energy and diffusion coefficient, and is more complicated to implement. The method was also applied to a transition on a model 2D landscape, using the minimum energy path as the reaction coordinate. The computed diffusion coefficients were in agreement with the exact value specified in the Brownian dynamics simulation, and the mean first passage times between metastable states computed using Smoluchowski diffusion theory were in good agreement with unbiased Brownian dynamics results. Finally, as a demonstration of the application of the method to a more physically interesting problem, we computed the free energy profile and diffusion coefficients for a conformational transition in the β-hairpin fragment of protein G,[80] for which the reaction coordinate was modeled by the finite temperature string method.[66,78]

The scope of the present study is limited to the efficient parametrization of the Smoluchowski model (SM), regardless of whether the SM is an appropriate choice to model the reaction under study. The main limitation of the model is the lack of memory in the representation of the degrees of freedom orthogonal to the reaction coordinate (bath). However, as is the case with master equation approaches,[25,67,69,70] the parametrization involves a choice of length scale (Δ). This implies that, if the parametrization length scale is chosen sufficiently large that the dynamics loses memory over distances of the order Δ, the Smoluchowski model can provide a good description of the long-time evolution of the system (typically, the regime of interest) even if the short-time predictions would not be accurate. The method can be easily extended to provide a test of Markovianity using additional (but still minimal) simulation output. One can define a rather broad flat-bottom restraint window of width Δ centered on the location of interest x_0, and record crossings in and out of smaller windows centered on x_0, of size, e.g., $\Delta/2$, in addition to crossings in and out of the main window. Computing the free energy and diffusion coefficient on the basis of statistics from the different window widths (still obtained from the same set of simulations) would reveal the sensitivity of the parameters to the choice of spatial discretization scale. The present method can also be extended to multiple dimensions by computing analytical expressions of mean exit times out of multidimensional sampling regions, using a multivariable linear approximation to the free energy in different windows. Additionally, the accuracy of the method can be improved by considering a quadratic (rather than linear) expansion of the free energy in a given simulation window. Since this would require the determination of two coefficients (curvature, in addition to the slope), another statistic would need to be sampled from MD simulation, which could be the standard deviation of the reaction cordinate, in addition to the mean. Finally, we remark that the method is developed for equilibrium processes, which is implicit in the assumption of the Boltzmann distribution as the stationary solution of the SM. However, if a nonequilibrium system admits a different (known) distribution as the stationary solution, the relation between the diffusion constant and the

DOI: 10.1021/acs.jpcb.6b02139

MFPT (eq 18) can be modified accordingly. The implementation of some of these possibilities is left to a future study.

APPENDIX A: COMPUTATION OF THE FREE ENERGY DERIVATIVE

Using a linear approximation to the free energy $F(x)$ in a reaction coordinate window $x \in [a, b]$ and setting $x_0 = (a + b)/2$, we have $F(x) = F(x_0) + F'(x_0)\,(x - x_0) + O(x - x_0)^2$. $F'(x_0)$ is easily computed from a conditional average over N_{MD} samples of x obtained from equilibrium MD simulations. Using $I(x)$ as the indicator function for the interval $[a, b]$, we define $\overline{x}_{\mathrm{MD}} = [\sum_{i=1}^{N_{\mathrm{MD}}} x_i I(x_i)]/[\sum_{i=1}^{N_{\mathrm{MD}}} I(x_i)]$; i.e., $\overline{x}_{\mathrm{MD}}$ is an average over only those samples that fall within the window $[a, b]$. Similar trajectory "pruning" has been used by Maragliano et al.[64] Since the probability density of x is assumed to evolve according to the Smoluchowski equation, which, for reflective boundary conditions at a and b, has the Boltzmann distribution as the stationary solution,[76] we can solve for $F'(x_0)$ from

$$\overline{x}_{\mathrm{MD}} = \frac{\int_a^b x \exp(-\beta F' x)\, \mathrm{d}x}{\int_a^b \exp(-\beta F' x)\, \mathrm{d}x} \tag{A1}$$

For convenience, we map the interval $[a, b]$ to $[0, 1]$ using $\Delta = b - a$ and $y = (x - a)/\Delta$ and define $\gamma = \beta F'\Delta$. Equation A1 becomes

$$\frac{\overline{x}_{\mathrm{MD}} - a}{\Delta} = \overline{y} = \frac{\int_0^1 y \exp(-\gamma y)\, \mathrm{d}y}{\int_0^1 \exp(-\gamma y)\, \mathrm{d}y} = f(\gamma)$$
$$\equiv \begin{cases} \dfrac{1}{\gamma} - \dfrac{1}{\exp(\gamma) - 1} & \text{for } \gamma \neq 0 \\ \dfrac{1}{2} & \text{for } \gamma = 0 \end{cases} \tag{A2}$$

or

$$\alpha = f^{-1}\left[\frac{\overline{x}_{\mathrm{MD}} - a}{\Delta}\right](\Delta\beta)^{-1} \tag{A3}$$

Equation A3 can be solved iteratively (an example in MATLAB[87] is given at the end of this section). First, to evaluate f numerically for $0 < |\gamma| \ll 1$, we expand the numerator and denominator in Maclaurin series

$$\frac{1}{\gamma} - \frac{1}{\exp(\gamma) - 1} = \frac{\exp(\gamma) - 1 - \gamma}{\gamma(\exp(\gamma) - 1)}$$
$$= \frac{\frac{\gamma^2}{2} + \frac{\gamma^3}{6} + ...}{\gamma^2 + \frac{\gamma^3}{2}...} \approx \frac{1 + \frac{\gamma}{3}}{2 + \gamma}$$
$$= \frac{1 + \frac{\gamma}{3}}{2}\left[1 - \frac{\gamma}{2} + \frac{\gamma^2}{4}...\right] \approx \frac{1}{2} - \frac{\gamma}{12} \tag{A4}$$

Second, we note that the inverse of f is everywhere defined because f is continuous and monotonically decreasing. To see this, note that

$$f'(\gamma) = -\frac{1}{\gamma^2} + \frac{\exp(\gamma)}{(\exp(\gamma) - 1)^2} < 0$$
$$\Leftrightarrow \frac{\exp(\gamma)}{(\exp(\gamma) - 1)^2} < \frac{1}{\gamma^2} \Leftrightarrow \exp(\gamma) < \left(\frac{\exp(\gamma) - 1}{\gamma}\right)^2$$
$$\Leftrightarrow \exp(\gamma/2) < \frac{\exp(\gamma) - 1}{\gamma}$$
$$\Leftrightarrow 1 + \frac{\gamma}{2} + \frac{\gamma^2}{4 \times 2!} + \frac{\gamma^3}{8 \times 3!}... < 1 + \frac{\gamma}{2} + \frac{\gamma^2}{3!} + \frac{\gamma^3}{4!}... \tag{A5}$$

The final inequality clearly holds for $\gamma > 0$, as can be seen by comparing the power series term by term. Because f' is symmetric, $f'(\gamma) < 0$ also holds for $\gamma < 0$; finally, $f'(0) < 0$ from eq A4. To compute statistical (sampling) errors for $\mathrm{d}F/\mathrm{d}x$, one first uses standard methods to compute the error for $\overline{x}_{\mathrm{MD}}$. (For example, one can use correlation analysis to compute the number of independent samples $\mathcal{N}$ in the MD time series x_i restricted to $[a, b]$ and define $\delta(\overline{x}_{\mathrm{MD}}) = [\mathrm{var}(x_i)/\mathcal{N}]^{1/2}$.) Using the definition of y and eqs A2 and A3, the uncertainty in $F'(x_0)$ can be taken as $\delta(F'(x_0)) = \delta(\overline{x}_{\mathrm{MD}})\beta^{-1}\Delta^{-2}|f'|^{-1}_{f(\gamma)=\overline{y}}$. The (squared) uncertainty in the free energy is obtained by integrating $\delta^2(F')$ by the trapezoid rule as used for computing F (see eq 15).

To solve eq A2 for γ, we use the `fzero` command in MATLAB[87] (`fzero` is also provided by OCTAVE,[88] which is distributed freely). First, we define the function "xave_boltzmann.m"

```
function [x,dx]=xave_boltzmann(g)
% NOTE: "%" indicates a comment
tol=1e-8 ; % threshold at which to switch to
           % Maclaurin expansion
if ( abs(g) > tol )
 x = 1/g - 1/(exp(g)-1);
 if (nargout>1) % compute derivative if more
                % than one argument given
  dx=-1/g^2 + exp(g)/(exp(g)-1)^2 ;
 end
else % use Maclaurin expansion
 x = 0.5 * (1-g/6);
 if (nargout>1)
  dx=-1/12;
 end
end
```

defining `yave` as $\overline{y}$ and `g` as γ, we obtain γ via

```
f=@(x) xave_boltzmann(x)-yave ;
g=fzero(f,0)
```

APPENDIX B: ALTERNATIVE COMPUTATION OF THE FREE ENERGY DERIVATIVE

In this section, we show that an approximation to the free energy derivative can also be computed from the average of restraint forces (eq B5). First, for the simulation window $\mathcal{I} = [a, b]$ (using the notation of Appendix A and Figure 1 in section 2), we define the smoothed free energy by

$$\begin{aligned} \overline{F}(x_0) &= -\beta^{-1} \log\left[\frac{1}{\Delta}\int_a^b \mathrm{e}^{-\beta F(x)}\right] \\ &= -\beta^{-1} \log\left[\frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} \mathrm{e}^{-\beta F(x_0+x)}\right] \end{aligned} \tag{B1}$$

In eq B1 and the following equations, the integration is performed with respect to x unless indicated otherwise, and the measure dx is omitted. Comparing $\overline{F}$ to F within the linear approximation assumed in section 2, we have

$$\begin{aligned} |\overline{F}(x_0) - F(x_0)| &= \beta^{-1}\left|\log \frac{1}{\Delta}\int_a^b \mathrm{e}^{-\beta[F(x)-F(x_0)]}\right| \\ &= \beta^{-1}\left|\log\left(\frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} \mathrm{e}^{-\gamma x}\right)\right| \\ &= \beta^{-1}\left|\log\left[\frac{1}{\gamma\Delta}(\mathrm{e}^{\gamma\Delta/2} - \mathrm{e}^{-\gamma\Delta/2})\right]\right| \\ &\simeq \beta^{-1}\left|\log\left[1 + \frac{(\gamma\Delta)^2}{24} + ...\right]\right| \simeq \beta^{-1}\frac{(\gamma\Delta)^2}{24} \end{aligned} \tag{B2}$$

which shows that $\overline{F}$ approximates F to second order accuracy in γ and Δ.

Next, we write the all-atom partition function for the restrained system

$$\begin{aligned} Q_{\mathrm{FB}}(x_0) &= \int_{\mathbb{R}^{3N}} \mathrm{e}^{-\beta[E(\hat{r})+U_{\mathrm{FB}}(x(\hat{r});x_0)]}\,\mathrm{d}\hat{r} \\ &= \int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}^{3N}} \mathrm{e}^{-\beta[E(\hat{r})+U_{\mathrm{FB}}(x(\hat{r});x_0)]}\delta(x(\hat{r}) - y)\,\mathrm{d}\hat{r} \\ &= \int_{\mathbb{R}} \mathrm{e}^{-\beta U_{\mathrm{FB}}(y;x_0)}\,\mathrm{d}y \int_{\mathbb{R}^{3N}} \mathrm{e}^{-\beta E(\hat{r})}\delta(x(\hat{r}) - y)\,\mathrm{d}\hat{r} \\ &= \int_{\mathbb{R}} \mathrm{e}^{-\beta[F(y)+U_{\mathrm{FB}}(y;x_0)]}\,\mathrm{d}y \end{aligned} \tag{B3}$$

and the corresponding free energy

$$F_{\mathrm{FB}}(x_0) = -\beta^{-1}\log[Q_{\mathrm{FB}}(x_0)\Delta^{-1}] \tag{B4}$$

where the normalization by Δ is used for consistency with eq B1. From eqs B3 and B4,

$$F'_{\mathrm{FB}}(x_0) = \frac{\overline{\mathrm{d}U_{\mathrm{FB}}}}{\mathrm{d}x_0}(x;\, x_0) \tag{B5}$$

where U'_{FB} is given by eq 20, and the overbar represents configurational averages in the restrained simulations; i.e., the derivative of the restrained free energy F_{FB} is related to the average restraint force in the simulation.

Finally, we show that F_{FB} approaches $\overline{F}$ as the restraint constant $k \to \infty$. Denoting by $H(x)$ the Heaviside step function and writing $\overline{F}(x_0)$ for $\Delta > 0$ as

$$\overline{F}(x_0) = \frac{1}{\beta}\log\left[\frac{1}{\Delta}\int_{\mathbb{R}} \mathrm{e}^{-\beta F(x)}H\left(\frac{\Delta}{2} - |x - x_0|\right)\right] \tag{B6}$$

we have

$$\begin{aligned} |F_{\mathrm{FB}}(x_0) - \overline{F}(x_0)| &= \frac{1}{\beta}\left|\log \frac{\int_{\mathbb{R}} \mathrm{e}^{-\beta[F(x)+U_{\mathrm{FB}}(x;x_0)]}}{\int_{\mathbb{R}} \mathrm{e}^{-\beta F(x)}H(\Delta/2 - |x - x_0|)}\right| \\ &= \frac{1}{\beta}\left|\log\left[1 + \frac{\int_{|x-x_0|>\Delta/2} \mathrm{e}^{-\beta[F(x)+U_{\mathrm{FB}}(x;x_0)]}}{\mathrm{e}^{-\beta\overline{F}(x_0)}}\right]\right| \\ &< \frac{1}{\beta}\left|\log\left[1 + \mathrm{e}^{\beta\overline{F}(x_0)}\int_{\mathbb{R}} \mathrm{e}^{-\beta F(x)-\beta k/2\times(x-x_0-\Delta/2)^2}\right.\right. \\ &\quad \left.\left. + e^{\beta\overline{F}(x_0)}\int_{\mathbb{R}} \mathrm{e}^{-\beta F(x)-\beta k/2\times(x-x_0+\Delta/2)^2}\right]\right| \end{aligned} \tag{B7}$$

where, in line 3, we split the partition function Q_{FB} into $\exp(-\beta\overline{F})$ and two integrals over half-harmonic wells (see Figure 1), and, in line 4, bounded the two integrals by integrals over the full harmonic wells. To within $O(1/k)$,[29] we approximate the Gaussians by δ-distributions

$$\begin{aligned} &|F_{\mathrm{FB}}(x_0) - \overline{F}(x_0)| \\ &< \frac{1}{\beta}\left|\log\left[1 + \left\{\int_{\mathbb{R}} \mathrm{e}^{-\beta F(x)}\delta(x - x_0 - \Delta/2)\right.\right.\right. \\ &\quad \left.\left.\left. + \int_{\mathbb{R}} \mathrm{e}^{-\beta F(x)}\delta(x - x_0 + \Delta/2)\right\} \times \mathrm{e}^{\beta\overline{F}(x_0)}\sqrt{\frac{2\pi}{\beta k}}\right]\right| \\ &= \frac{1}{\beta}\left|\log\left[1 + \mathrm{e}^{\beta\overline{F}(x0)}\sqrt{\frac{2\pi}{\beta k}}\{\mathrm{e}^{-\beta F(x_0+\Delta/2)}\right.\right. \\ &\quad \left.\left. + \mathrm{e}^{-\beta F(x_0-\Delta/2)}\}\right]\right| \simeq \beta^{-1}\sqrt{\frac{2\pi}{\beta k}}\mathrm{e}^{\beta\overline{F}(x0)}[\mathrm{e}^{-\beta F(x_0+\Delta/2)} \\ &\quad + \mathrm{e}^{-\beta F(x_0-\Delta/2)}] = C(x_0)k^{-1/2} \end{aligned} \tag{B8}$$

For the special case of $\Delta = 0$, which corresponds to the ordinary harmonic restraint, Maragliano et al.[29] showed that the approximation holds to $O(k^{-1})$.

APPENDIX C: COMPUTATION OF DIFFUSION COEFFICIENTS USING HARMONICALLY RESTRAINED POTENTIALS

In this section, we recall how coordinate-dependent diffusion coefficients can be estimated from equilibrium MD simulations with harmonic restraint potentials. The derivation below uses Markovianity explicitly via the FPE in eq 1. A derivation that uses a nontrivial memory kernel in a harmonic oscillator GLE can be found in ref 62.

First, we define a restraint potential $k(x - x_0)^2/2$ with k so large that $F''(x_0)[|F'(x_0)|k^{-1} + 1/\sqrt{k\beta}]^2 < \epsilon$, for some $1 \gg \epsilon > 0$. This condition guarantees that, for $x \in \mathcal{I} \equiv [x_0 + F'k^{-1} - 1/\sqrt{k\beta}),\quad x_0 + F'k^{-1} + 1/\sqrt{k\beta})]$, $|F(x) - F(x_0) - F'(x_0)\ (x - x_0)| < \epsilon$; i.e., for the majority of thermally accessible configurations, F is approximately linear.

Defining $\gamma = \beta F'(x_0)$, we assume, as in section 2, that the x-density in $\mathcal{I}$ evolves according to the Smoluchowski equation, which for the present case is

$$\dot{P}(x, t) = D[\{\beta k(x - x_0) + \gamma\}P(x) + P'(x)]' \tag{C1}$$

In the variables $k^* = kD\beta$, $y = x - x_0 + \gamma(k\beta)^{-1}$, and $P^*(y) = P(y + x_0 - \gamma(k\beta)^{-1})$, eq C1 reads

$$\dot{P}^*(y, t) = [k^* y^* P^*(y)]' + DP^{*\prime\prime}(y) \tag{C2}$$

$P^*(y, t)$ is the density for the Ornstein−Uhlenbeck process, for which the stationary position correlation is given by[54,76]

$$\overline{y(t)y(0)} = \frac{D}{k^*}\exp(-k^*t) = \frac{1}{\beta k}\exp(-D\beta kt) \tag{C3}$$

This identity can be used to compute D from position correlations obtained in restrained MD simulations. For example, integrating eq C3 over $[0, \infty)$ and rearranging gives

$$\begin{aligned} D &= (\beta k)^{-2}\left[\int_0^\infty \overline{y(t)y(0)}\,\mathrm{d}t\right]^{-1} \\ &= \left[\overline{y(t)^2}\right]^2\left[\int_0^\infty \overline{y(t)y(0)}\,\mathrm{d}t\right]^{-1} \end{aligned} \tag{C4}$$

Equation C4 and its variants have been used before to compute position-dependent diffusion coefficients.[51,52,62,71] As discussed in ref 52, and shown in section 3, a practical issue that affects the accuracy of eq C4 is the choice of the upper integration limit, which must be finite.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: ovchinnv@georgetown.edu.
*E-mail: kwangho.nam@chem.umu.se.
*E-mail: marci@tammy.harvard.edu. Phone: +1 617-495-4018. Fax: +1 617-496-3204.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

V.O. and M.K. acknowledge financial support under NIH Grant No. 5R03AI111416. K.N. was supported by grants from the Umeå University, Kempestiftelserna, and the Swedish Research Council (2015-04114). Computer resources were provided by the National Energy Resource Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, by the Faculty of Arts and Sciences Computing Group at Harvard, and by the Swedish National Infrastructure for Computing at the High Performance Computing Center North and the National Supercomputing Center.

■ REFERENCES

(1) Ermak, D. L.; McCammon, J. A. Brownian Dynamics with Hydrodynamic Interactions. *J. Chem. Phys.* **1978**, *69*, 1352−1360.

(2) Benkovic, S.; Hammes-Schiffer, S. A Perspective on Enzyme Catalysis. *Science* **2003**, *301*, 1196−1202.

(3) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. How Enzymes Work: Analysis by Modern Rate Theory and Computer Simulations. *Science (Washington, DC, U. S.)* **2004**, *303*, 186−195.

(4) Ojeda-May, P.; Li, Y.; Ovchinnikov, V.; Nam, K. Role of Protein Dynamics in Allosteric Control of the Catalytic Phosphoryl Transfer of Insulin Receptor Kinase. *J. Am. Chem. Soc.* **2015**, *137*, 12454−12457.

(5) Johnson, K. Conformational Coupling in DNA Polymerase Fidelity. *Annu. Rev. Biochem.* **1993**, *62*, 685−713.

(6) Joyce, C.; Steitz, T. Function and Structure Relationships in DNA Polymerases. *Annu. Rev. Biochem.* **1994**, *63*, 777−822.

(7) Ma, J.; Siegler, P.; Xu, Z.; Karplus, M. A Dynamic Model for the Allosteric Mechanism of GroEL. *J. Mol. Biol.* **2000**, *302*, 303−313.

(8) Karplus, M.; Gao, Y. Biomolecular Motors: The F$_1$-ATPase Paradigm. *Curr. Opin. Struct. Biol.* **2004**, *14*, 250−259.

(9) Houdusse, A.; Sweeney, H. L. Myosin Motors: Missing Structures and Hidden Springs. *Curr. Opin. Struct. Biol.* **2001**, *11*, 182−194.

(10) Nam, K.; Pu, J.; Karplus, M. Trapping the ATP Binding State Leads to a Detailed Understanding of the F1-ATPase Mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 17851−17856.

(11) Geeves, M.; Holmes, K. Structural Mechanism of Muscle Contraction. *Annu. Rev. Biochem.* **1999**, *68*, 687−727.

(12) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; et al. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341−346.

(13) Harvey, M.; Giupponi, G.; Fabritis, G. D. ACEMD: Accelerated Molecular Dynamics Simulations in the Microseconds Timescale. *J. Chem. Theory Comput.* **2009**, *5*, 1632−1639.

(14) Torrie, G.; Valleau, J. Non-Physical Sampling Distributions in Monte Carlo Free Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.

(15) Bartels, C.; Schaefer, M.; Karplus, M. Determination of Equilibrium Properties of Biomolecular Systems Using Multidimensional Adaptive Umbrella Sampling. *J. Chem. Phys.* **1999**, *111*, 8048.

(16) Laio, A.; Parrinello, M. Escaping Free Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562.

(17) Dellago, C.; Bolhuis, P.; Geissler, P. Transition Path Sampling. *Adv. Chem. Phys.* **2003**, *123*, 1−78.

(18) Bolhuis, P.; Chandler, D.; Dellago, C.; Geissler, P. Transition Path Sampling: Throwing Ropes over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291.

(19) Hénin, J.; Chipot, C. Overcoming Free Energy Barriers Using Unconstrained Molecular Dynamics Simulations. *J. Chem. Phys.* **2004**, *121*, 2904−2914.

(20) Hénin, J.; Fiorin, G.; Chipot, C.; Klein, M. L. Exploring Multidimensional Free Energy Landscapes Using Time-Dependent Biases on Collective Variables. *J. Chem. Theory Comput.* **2010**, *6*, 35−47.

(21) Rosso, L.; Tuckerman, M. An Adiabatic Molecular Dynamics Method for the Calculation of Free Energy Profiles. *Mol. Simul.* **2002**, *28*, 91.

(22) Schlitter, J.; Engels, M.; Krüger, P.; Jacoby, E.; Wollmer, A. Targeted Molecular Dynamics Simulation of Conformational Change - Application to the T→R Transition in Insulin. *Mol. Simul.* **1993**, *10*, 291−308.

(23) Maragliano, L.; Vanden-Eijnden, E. A Temperature Accelerated Method for Scampling Free Energy and Determining Reaction Pathways in Rare Events Simulations. *Chem. Phys. Lett.* **2006**, *426*, 168−175.

(24) Zheng, L.; Chen, M.; Yang, W. Random Walk in Orthogonal Space to Achieve Efficient Free-Energy Simulation of Complex Systems. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 20227−20232.

(25) Faradjian, A.; Elber, R. Computing Time Scales from Reaction Coordinates by Milestoning. *J. Chem. Phys.* **2004**, *120*, 10880.

(26) Shalloway, D.; Faradjian, K. Efficiently Computing the First Passage Time Generating Function by Steady-State Relaxation. *J. Chem. Phys.* **2006**, *124*, 054112.

(27) Kirmizialtin, S.; Nguyen, V.; Johnson, K.; Elber, R. How Conformational Dynamics of DNa Polymerase Select Correct Substrates: Experiments and Simulations. *Structure* **2012**, *20*, 618−627.

(28) E, W.; Ren, W.; Vanden-Eijnden, E. Transition Pathways in Complex Systems: Reaction Coordinates, Isocommittor Surfaces, and Transition Tubes. *Chem. Phys. Lett.* **2005**, *413*, 242−247.

(29) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. String Method in Collective Variables: Minimum Free Energy Paths and Isocommittor Surfaces. *J. Chem. Phys.* **2006**, *125*, 024106.

(30) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011−19016.

(31) Voelz, V.; Bowman, G.; Beauchamp, K.; Pande, V. Molecular Simulation of Ab Initio Protein Folding for a Millisecond Folder NTL9(1−39). *J. Am. Chem. Soc.* **2010**, *132*, 1526−1528.

(32) Elber, R.; Karplus, M. A Method for Determining Reaction Paths in Large Biomolecules: Application to Myoglobin. *Chem. Phys. Lett.* **1987**, *139*, 375−380.

(33) Fischer, S.; Karplus, M. Conjugate Peak Refinement: An Algorithm for Finding Reaction Paths and Accurate Transition States in Systems with Many Degrees of Freedom. *Chem. Phys. Lett.* **1992**, *194*, 252−261.

(34) Jónsson, G.; Jacobsen, K. In *Classical and Quantum Dynamics in Condensed Phase Simulations*; Berne, B., Ciccotti, G., Coker, D., Eds.; World Scientific: Singapore, 1998; pp 385−404.

(35) E, W.; Ren, W.; Vanden-Eijnden, E. Simplified and Improved String Method for Computing the Minimum Energy Paths in Barrier-Crossing Events. *J. Chem. Phys.* **2007**, *126*, 164103.

(36) Ovchinnikov, V.; Karplus, M. Analysis and Elimination of a Bias in Targeted Molecular Dynamics Simulations of Conformational Transitions: Application to Calmodulin. *J. Phys. Chem. B* **2012**, *116*, 8584−8603.

(37) Isralewitz, B.; Baudry, J.; Kosztin, J. G. D.; Schulten, K. Steered Molecular Dynamics Investigations of Protein Function. *J. Mol. Graphics Modell.* **2001**, *19*, 13−25.

(38) Ovchinnikov, V.; Cecchini, M.; Vanden-Eijnden, E.; Karplus, M. A Conformational Transition in the Myosin VI Converter Contributes to the Variable Step Size. *Biophys. J.* **2011**, *101*, 2436−2444.

(39) Qi, Y.; Nam, K.; Spong, M.; Banerjee, A.; Sung, R.; Zhang, M.; Karplus, M.; Verdine, G. Strandwise Translocation of a DNa Glycosylase on Undamaged DNA. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 1086−1091.

(40) Ovchinnikov, V.; Karplus, M.; Vanden-Eijnden, E. Free Energy of Conformational Transition Paths in Biomolecules: The String Method and Its Application to Myosin VI. *J. Chem. Phys.* **2011**, *134*, 085103.

(41) Best, R.; Hummer, G. Diffusion Models of Protein Folding. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16902−16911.

(42) Smoluchowski, M. Versuch Einer Mathematischen Theorie Der Koagulationskinetik Kolloider Lösungen. *Z. Phys. Chem.* **1918**, *92*, 129−168.

(43) Kramers, H. A. Brownian Motion in a Field of Force and the Diffusion Model of Chemical Reactions. *Physica* **1940**, *7*, 284.

(44) Krivov, S.; Muff, S.; Caflisch, A.; Karplus, M. One-Dimensional Barrier-Preserving Free-Energy Projections of a ß-Sheet Miniprotein: New Insights into the Folding Process. *J. Phys. Chem. B* **2008**, *112*, 8701−8714.

(45) Hummer, G. Reaction Coordinaets and Rates from Transition Paths. *J. Chem. Phys.* **2004**, *120*, 516.

(46) Best, R.; Hummer, G. Reaction Coordinates and Rates from Transition Paths. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6732−6737.

(47) Ma, A.; Dinner, A. An Automatic Method for Identifying Reaction Coordinates in Complex Systems. *J. Phys. Chem. B* **2005**, *109*, 6769−6779.

(48) Peters, B.; Trout, B. Obtaining Reaction Coordinates by Likelihood Maximization. *J. Chem. Phys.* **2006**, *125*, 054108.

(49) E, W.; Ren, W.; Vanden-Eijnden, E. String Method for the Study of Rare Events. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2002**, *66*, 052301.

(50) Berne, B.; Borkovec, M.; Straub, J. Classical and Modern Methods in Reaction Theory. *J. Phys. Chem.* **1988**, *92*, 3711−3725.

(51) Socci, N.; Onuchic, J.; Wolynes, P. Diffusive Dynamics of the Reaction Coordinate for Protein Folding Funnels. *J. Chem. Phys.* **1996**, *104*, 5860−5868.

(52) Hummer, G. Position-Dependent Diffusion Coefficients and Free Energies from Bayesian Analysis of Equilibrium and Replica Molecular Dynamics Simulations. *New J. Phys.* **2005**, *7*, 34.

(53) Comer, J.; Chipot, C.; Gonzáles-Nilo, F. D. Calculating Position-Dependent Diffusivity in Biased Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2013**, *9*, 876−882.

(54) Risken, H. *The Fokker-Planck Equation. Methods of Solution and Applications*, 2nd ed.; Springer-Verlag: Berlin, 1989.

(55) Micheletti, C.; Bussi, G.; Laio, A. Optimal Langevin Modeling of Out-Of-Equilibrium Molecular Dynamics Simulations. *J. Chem. Phys.* **2008**, *129*, 074105.

(56) van Mourik, A.; Daffertshofer, A.; Beek, P. Estimating Kramers-Moyal Coefficients in Short and Non-Stationary Data Sets. *Phys. Lett. A* **2006**, *351*, 13−17.

(57) Zuckerman, D.; Woolf, T. Transition Events in Butane Simulations: Similarities Across Models. *J. Chem. Phys.* **2002**, *116*, 2586−2591.

(58) Zwanzig, R. Memory Effects in Irreversible Thermodynamics. *Phys. Rev.* **1961**, *124*, 983.

(59) Grote, R.; Hynes, J. the Stable States Picture of Chemical Reactions. II. Rate Constants for Condensed and Gas Phase Reaction Models. *J. Chem. Phys.* **1980**, *73*, 2715−2732.

(60) Lange, O.; Grubmüller, H. Collective Langevin Dynamics of Conformational Motions in Proteins. *J. Chem. Phys.* **2006**, *124*, 214903.

(61) West, A.; Elber, R.; Shalloway, D. Extending Molecular Dynamics Time Scales with Milestoning: Example of Complex Kinetics in a Solvated Peptide. *J. Chem. Phys.* **2007**, *126*, 145104.

(62) Woolf, T.; Roux, B. Molecular Dynamics Simulation of the Gramicidin Channel in a Phospholipid Bilayer. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 11631−11635.

(63) Chipot, C.; Hénin, J. Exploring the Free Energy Landscape of a Short Peptide Using an Average Force. *J. Chem. Phys.* **2005**, *123*, 244906.

(64) Maragliano, L.; Vanden-Eijnden, E.; Roux, B. Free Energy and Kinetics of Conformational Transitions from Voronoi Tessellated Milestoning with Restraining Potentials. *J. Chem. Theory Comput.* **2009**, *5*, 2589−2594.

(65) Carter, E.; Ciccotti, G.; Hynes, J.; Kapral, R. Constrained Reaction Coordinate Dynamics for the Simulation of Rare Events. *Chem. Phys. Lett.* **1989**, *156*, 472.

(66) Vanden-Eijnden, E.; Venturoli, M. Markovian Milestoning with Voronoi Tessellations. *J. Chem. Phys.* **2009**, *130*, 194101.

(67) Glowacki, D.; Paci, E.; Shalashilin, D. Boxed Molecular Dynamics: A Simple and General Technique for Accelerating Rare Event Kinetics and Mapping Free Energy in Large Molecular Systems. *J. Phys. Chem. B* **2009**, *113*, 16603−16611.

(68) Mugnai, M.; Elber, R. Extracting the Diffusion Tensor from Molecular Dynamics Simulation with Milestoning. *J. Chem. Phys.* **2015**, *142*, 014105.

(69) Suárez, E.; Lettieri, S.; Zwier, M.; Stringer, C.; Subramanian, S.; Chong, L.; Zuckerman, D. Simultaneous Computation of Dynamical and Equilibrium Information Using a Weighted Ensemble of Trajectories. *J. Chem. Theory Comput.* **2014**, *10*, 2658−2667.

(70) Pande, V.; Beauchamp, K.; Bowman, G. Everything You Wanted to Know About Markov State Models but Were Afraid to Ask. *Methods (Amsterdam, Neth.)* **2010**, *52*, 99−105.

(71) Mamonov, A.; Kurnikova, M.; Coalson, R. Diffusion Constant of K+ Inside Gramicidin A: A Comparative Study of Four Computational Methods. *Biophys. Chem.* **2006**, *124*, 268−278.

(72) Krivov, S.; Karplus, M. Diffusive Reaction Dynamics on Invariant Free Energy Profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 13841−13846.

(73) Banushkina, P.; Krivov, S. Fep1d: A Script for the Analysis of Reaction Coordinates. *J. Comput. Chem.* **2015**, *36*, 878−882.

(74) Kumar, S.; Rosenberg, J.; Bouzida, D.; Swendsen, R.; Kollman, P. *J. Comput. Chem.* **1992**, *13*, 1011.

DOI: 10.1021/acs.jpcb.6b02139

(75) Kästner, J.; Thiel, W. Bridging the Gap Between Thermodynamic Integration and Umbrella Sampling Provides a Novel Analysis Method: "Umbrella Integration. *J. Chem. Phys.* **2005**, *123*, 144104.

(76) Gardiner, C. *Handbook of Stochastic Methods*, 3rd ed.; Springer-Verlag: Berlin, 2003.

(77) Frantz, J. g3data: Grab Graph Data, a Program for Extracting Data from Graphs. http://www.frantz.fi/software/g3data.php, 2000; Date accesed: April 25, 2016.

(78) Ovchinnikov, V.; Karplus, M. Investigations of α-Helix↔ β-Sheet Transition Pathways in a Miniprotein Using the Finite-Temperature String Method. *J. Chem. Phys.* **2014**, *140*, 175103.

(79) Ovchinnikov, V.; Cecchini, M.; Karplus, M. a Simplified Confinement Method (SCM) for Calculating Absolute Free Energies and Free Energy and Entropy Differences. *J. Phys. Chem. B* **2013**, *117*, 750−762.

(80) Gronenborn, A.; Filpula, D.; Essig, N.; Achari, A.; Whitlow, M.; Wingfield, P.; Clore, G. A Novel, Highly Stable Fold of the Immunoglobulin Binding Domain of Streptococcal Protein G. *Science* **1991**, *253*, 657−661.

(81) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. Free Energy Landscape for β Hairpin Folding from Combined Paralle Tempering and Metadynamics. *J. Am. Chem. Soc.* **2006**, *128*, 13435−13441.

(82) Cecchini, M.; Krivov, S.; Spichty, M.; Karplus, M. Calculation of Free-Energy Differences by Confinement Simulations. Application to Peptide Conformers. *J. Phys. Chem. B* **2009**, *113*, 9728−9740.

(83) Spichty, M.; Cecchini, M.; Karplus, M. Conformational Free-Energy Difference of a Miniprotein from Nonequilibrium Simulations. *J. Phys. Chem. Lett.* **2010**, *1*, 1922−1926.

(84) Vanden-Eijnden, E.; Venturoli, M. Revisiting the Finite Temperature String for the Calculation of Reaction Tubes and Free Energies. *J. Chem. Phys.* **2009**, *130*, 194103.

(85) Haberthür, U.; Caflisch, A. FACTS: Fast Analytical Continuum Treatment of Solvation. *J. Comput. Chem.* **2008**, *29*, 701−715.

(86) Best, R.; Zhu, X.; Shim, J.; Lopes, P.; Mittal, J.; Feig, M.; MacKerell, A., Jr. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone $\boldsymbol{\phi}$, ψ Ans Side-Chain χ_1 and χ_2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257−3273.

(87) *MATLAB*, version 7.10.0 (R2010a); The MathWorks Inc.: Natick, MA, 2010.

(88) Eaton, J. W.; Bateman, D.; Hauberg, S.; Wehbring, R. GNU Octave Version 4.0.0 Manual: A High-Level Interactive Language for Numerical Computations; 2015.

DOI: 10.1021/acs.jpcb.6b02139