Efficient Estimation of Rare-Event Kinetics

Benjamin Trendelkamp-Schroer[†] and Frank Noé[*]

Institut für Mathematik und Informatik, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany

(Received 24 September 2014; revised manuscript received 22 July 2015; published 28 January 2016)

The efficient calculation of rare-event kinetics in complex dynamical systems, such as the rate and pathways of ligand dissociation from a protein, is a generally unsolved problem. Markov state models can systematically integrate ensembles of short simulations and thus effectively parallelize the computational effort, but the rare events of interest still need to be spontaneously sampled in the data. Enhanced sampling approaches, such as parallel tempering or umbrella sampling, can accelerate the computation of equilibrium expectations massively, but sacrifice the ability to compute dynamical expectations. In this work we establish a principle to combine knowledge of the equilibrium distribution with kinetics from fast "downhill" relaxation trajectories using reversible Markov models. This approach is general, as it does not invoke any specific dynamical model and can provide accurate estimates of the rare-event kinetics. Large gains in sampling efficiency can be achieved whenever one direction of the process occurs more rapidly than its reverse, making the approach especially attractive for downhill processes such as folding and binding in biomolecules. Our method is implemented in the PyEMMA software.

DOI: 10.1103/PhysRevX.6.011009 Subject Areas: Chemical Physics, Statistical Physics

I. INTRODUCTION

A wide range of biological or physicochemical systems exhibit rare-event kinetics, consisting of rare transitions between a couple of long-lived (metastable) states. Examples are protein folding, protein-ligand association, and nucleation processes. Metastability can be found in any system in which states of minimum energy are separated by barriers higher than the average thermal energy.

A thorough understanding of such systems encompasses the kinetics of the rare events, e.g., rates and transition pathways. Obtaining reliable estimates for such systems is notoriously difficult: The simulation time needs to exceed the longest waiting time, resulting in a sampling problem.

In recent years, Markov state models (MSMs) [1–8] and their practical applicability through software [9,10] have become a key technology for computing kinetics of complex rare-event systems. A well-constructed MSM separates the kinetically distinct states and captures their transition rates or probabilities. With a suitable choice of state space discretization and lag time, kinetics can be approximated with high numerical accuracy [8,11]. MSMs can be straightforwardly interpreted and analyzed with Markov chain theory and transition path theory [12,13]. This was demonstrated, for example, for protein folding [14,15] or protein-ligand binding [16,17].

MSMs can somewhat alleviate the sampling problem by virtue of the fact that they can be estimated from short simulations produced in parallel [14,15,18], thus avoiding the need for single long trajectories [19]. However, the rare events of interest must be sampled in the data in order to be captured by the model. For example, in protein-ligand binding, a dissociation rate can only be computed if each step of the dissociation process has been sampled at least once.

Orders of magnitude of speed-up can be achieved with enhanced sampling methods such as umbrella sampling, replica exchange dynamics, or metadynamics [20–23]. The speed-up is achieved by coupling the unbiased ensemble of interest with ensembles at higher temperature at which the rare events occur more frequently or by using biasing potentials which allow to "drag" the system across an energy barrier. With such approaches, accurate equilibrium expectations, such as free-energy profiles, can be computed efficiently, but the dynamical properties of the unbiased ensemble, such as transition rates, relaxation time scales, and transition pathways, are generally not available.

A common approach to reconstruct the kinetics from the free-energy profiles is to employ rate theories such as transition state theory or Kramers or Smoluchowski-Langevin models [24–27]. Such dynamical models introduce additional assumptions that cannot be self-consistently validated because the predicted dynamics is not present in the data.

A much more advanced approach was recently introduced in Ref. [28], where a MSM-based estimator for the equilibrium distribution using transition counts harvested

[*]Corresponding author.
frank.noe@fu-berlin.de
[†]benjamin.trendelkamp-schroer@fu-berlin.de

from simulations at different thermodynamic states was developed. This approach allows us to mix, in principle, umbrella sampling simulations and direct molecular dynamics and compute the rates from the transition matrix of the unbiased ensemble. However, the coefficient matrices used to connect the biased and unbiased transition matrices require a specific dynamical model to be formulated (such as Brownian dynamics in the free-energy coordinate), making this approach essentially a rate model. In general, key assumptions underlying rate models are usually the existence of a time-scale separation and approximate Markovianity on a single or few reaction coordinates—assumptions that are unlikely to hold for complex multistate systems describing, e.g., biomolecular dynamics.

The recently introduce transition-based reweighting analysis methods (TRAM) [29–31] permit one to rigorously combine direct molecular dynamics and enhanced sampling methods towards full thermodynamics and kinetics without assuming any restrictive rate model. However, a current limitation is that in order to extract unbiased kinetics, the transition events need to be evaluated at a common and sufficiently large lag time τ at all thermodynamic states. This requirement is not consistent with efficient umbrella sampling or replica-exchange MD simulations that typically employ very short simulation snippets.

Finally, computation of kinetic quantities without rate models is also possible with path sampling methods, such as transition path sampling [32], milestoning [33], transition interface sampling [34], and multistate transition interface sampling [35]. A challenge is that these approaches are essentially two-state methods. The transition end states must be defined *a priori* and all relevant rare events must be distinguishable in the reaction coordinates, cores, or milestones that the method operates on.

Here, we construct a general simulation approach that enables the computation of kinetic observables related to slow processes without having to explicitly sample the rare events. It is based upon a very simple but general idea: Simulations are often constructed in such a way that they obey microscopic reversibility or at least a generalization thereof [36,37]. In this case, for any partition of state space into sets $i, j, \ldots$ and any choice of the lag time τ, we have the detailed balance relation

$$\pi_i p_{ij}(\tau) = \pi_j p_{ji}(\tau), \tag{1}$$

where $p_{ij}(\tau)$ is the probability of making a transition from set i to set j within a time τ and π_i is the equilibrium probability of set i. Suppose we have knowledge about the equilibrium probabilities π_i, π_j from an enhanced sampling simulation. Then, only the larger one of the two transition probabilities—p_{ij} or p_{ji}—needs to be sampled while the less probable event can be reconstructed by Eq. (1).

Speaking in terms of a network of states, a direct analysis or an analysis via Markov state models requires all states to be connected in both directions (strongly connected). The presented method allows us to relax this requirement if an estimate of the equilibrium probabilities is given—now all states need to be connected in only one direction (weakly connected).

The slow rate exhibits a functional dependence on the transition probabilities of the slow event. By virtue of the detailed balance condition, a reliable estimate of the transition probabilities for the frequent event entails a reliable estimate for the transition probabilities of the slow event—resulting in a reliable estimate of the slow rate.

While the inference procedure is trivial for a two-state system where three of the four components in Eq. (1) are known exactly, it is far from trivial for a system with many states and when some or all estimates are subject to statistical uncertainty. Here, we establish a systematic inference scheme for combining multistate estimates of the equilibrium probabilities (π_i) with sampling data of at least the "downhill" transition probabilities p_{ij}. Our approach is built upon the framework of reversible Markov models [38,39], where Eq. (1) is enforced between all pairs of states. As a consequence, our estimates do not invoke any additional dynamical model, are accurate within a suitable state space discretization [8,11], and are precise in the limit of sufficient sampling.

In contrast to the dynamic weighted histogram analysis method (DHAM) [28] and TRAM methods, we use equilibrium probabilities previously estimated from enhanced sampling simulations as additional input parameters for the estimation of MSM transition probabilities. Standard reweighting schemes used to obtain the equilibrium probabilities do not usually assume a dynamical model to obtain the desired unbiased probabilities.

The estimation procedure can reduce the sampling problem tremendously for processes with some long-lived states and some other states from which the system relaxes rapidly. This case is ubiquitous in metastable systems, because long-lived states are connected by short-lived transition states. But even long-lived states usually have very different lifetimes: For example, many ligands or inhibitors bind to their protein receptor with nanomolar concentrations, meaning that the transition probabilities leading to the associated state are orders of magnitude higher than the dissociation probabilities. The present reversible Markov model approach provides the basis for estimating the kinetics and mechanisms of protein-drug dissociation by combining the much more rapid association trajectories with suitable enhanced sampling methods, such as Hamiltonian replica exchange [40] or umbrella sampling [18,41].

The methods described here are implemented in PyEMMA [42]. Tutorials for the maximum-likelihood and Bayesian estimation of Markov models given equilibrium distributions using the examples in this paper can be found in the Supplemental Material [43].

II. THEORY

A. Markov state models

Classical dynamics, governed by Newton's equations in the case of an isolated system and by Langevin equations for systems at constant temperature [44], gives rise to a transfer operator $\mathcal{P}$ propagating a phase-space density from time t to time $t+\Delta t$ [45–47]. Numerical solutions for Newton or Langevin equations can be obtained for complex systems with many degrees of freedom, but a direct numerical assessment of the transfer operator is, in most cases, prohibited due to the curse of dimensionality.

Markov state models bridge this gap, estimating the transfer operator on a suitably defined state space partition,

$$\Omega = \{s_1, \ldots, s_n\}, \tag{2}$$

using trajectories obtained by direct numerical simulation [8,11].

MSMs model the jump process between states of this partition by a Markov chain. Observed transitions between pairs of states i and j are collected in a count matrix $C = (c_{ij})$, and the likelihood for the observed counts for a given transition matrix $P = (p_{ij})$ is given by

$$\mathbb{P}(C|P) \propto \prod_i \left(\prod_j p_{ij}^{c_{ij}} \right). \tag{3}$$

While the likelihood functions allow us to determine the maximum likelihood estimator $\hat{P}$ optimizing the likelihood function for a given observation C over the set of all possible models P, it does not specify the uncertainty of a chosen model.

For a finite amount of observation data there will in general be a whole ensemble of models compatible with the given data. In order to specify uncertainties and determine statistical errors of estimated quantities, we need to infer the posterior probability of a model for a given observation. An application of Bayes's formula yields

$$\underbrace{\mathbb{P}(P|C)}_{\text{posterior}} \propto \underbrace{\mathbb{P}(P)}_{\text{prior}} \underbrace{\mathbb{P}(C|P)}_{\text{likelihood}}. \tag{4}$$

For a uniform prior, i.e., no *a priori* knowledge about the model, the posterior probability is given as a product of Dirichlet distributions:

$$\mathbb{P}(P|C) \propto \prod_i \left(\prod_j p_{ij}^{c_{ij}} \right). \tag{5}$$

B. Inference using a given equilibrium distribution

There are many methods that allow us to efficiently estimate the equilibrium vector, even in situations in which a direct estimation from a finite observation of the Markov chain is unfeasible due to the metastable nature of the system [20–23,48,49]. In such situations it is often possible to alter the system dynamics in a controlled way such that the artificial dynamics equilibrates more rapidly than the original one. The desired equilibrium distribution of the original dynamics can then be related to the equilibrium distribution estimated from the altered process [50–54].

In the following, we show how such prior knowledge about the equilibrium distribution can be used to improve the estimates of kinetic observables in systems with rare events.

We are again given a finite observation of a Markov chain in terms of the count matrix C. Assume we are additionally given the equilibrium distribution π for our system of interest and we know that the transition probabilities of the chain fulfil detailed balance for the given equilibrium distribution,

$$\pi_i p_{ij} = \pi_j p_{ji}. \tag{6}$$

Then, we can express the posterior probability for our model via Eq. (4). Prior knowledge about the equilibrium distribution π in combination with the detailed balance assumption formally entails the following prior distribution on the posterior ensemble:

$$\mathbb{P}(P|\pi) = \prod_{i<j} \delta(\pi_i p_{ij} - \pi_j p_{ji}). \tag{7}$$

According to Eq. (4), the constrained posterior is

$$\mathbb{P}(P|C,\pi) \propto \mathbb{P}(P|\pi)\mathbb{P}(C|P). \tag{8}$$

The effect of the prior Eq. (7) is a restriction of the posterior to the subspace of transition matrices fulfilling detailed balance with respect to the fixed equilibrium distribution π.

C. Maximum likelihood estimate given an equilibrium distribution

We can also use prior knowledge of the equilibrium distribution to constrain the maximum likelihood estimate $\hat{P}$ to the set of matrices obeying Eq. (6) for a given equilibrium distribution π. This results in the following convex constrained optimization problem:

$$\begin{aligned} \text{minimize} \quad & -\sum_{i,j} c_{ij} \log p_{ij} \\ \text{subject to} \quad & p_{ij} \geq 0, \\ & \sum_j p_{ij} = 1, \\ & \pi_i p_{ij} = \pi_j p_{ji}, \end{aligned} \tag{9}$$

which can be solved using a fixed-point iteration method outlined in Ref. [55].

One can show that the solution to Eq. (9) can be written as

$$p_{ij}^{*} = \frac{(c_{ij} + c_{ji})\pi_j}{\lambda_i^{*}\pi_j + \lambda_j^{*}\pi_i}. \tag{10}$$

The Lagrangian parameters λ_i^{*} are obtained by iterating the following self-consistent equation to convergence:

$$\lambda_i^{(n+1)} = \sum_j \frac{(c_{ij} + c_{ji})\pi_j \lambda_i^{(n)}}{\lambda_i^{(n)}\pi_j + \lambda_j^{(n)}\pi_i}. \tag{11}$$

D. Inference using an equilibrium distribution with uncertainty

An equilibrium distribution estimate usually carries a finite sampling error which should be accounted for when inferring a reversible transition matrix from data. From a Bayesian viewpoint, we have to combine two sources of evidence: the observed count matrix C from standard equilibrium simulations and the data from enhanced or biased sampling methods E used to estimate the equilibrium distribution.

An error model for the estimation of uncertainty in the equilibrium distribution assesses the posterior of equilibrium distributions given the enhanced sampling data, $\mathbb{P}(\pi|E)$. Recent methods for the uncertainty quantification of reversible MSMs with fixed equilibrium distribution allow us to sample the posterior $\mathbb{P}(P|\pi, C)$ in Eq. (8).

The posterior for transition matrices under the combined evidence $\mathbb{P}(P|C, E)$ can be formally decomposed as

$$\mathbb{P}(P|C, E) = \int d\pi \mathbb{P}(P|C, \pi, E)\mathbb{P}(\pi|C, E). \tag{12}$$

Assuming that the direct effect of the enhanced sampling information E is negligible in the posterior of transition matrices with given equilibrium distribution,

$$\mathbb{P}(P|C, \pi, E) \approx \mathbb{P}(P|C, \pi), \tag{13}$$

and that the direct effect of observed transition counts C is unimportant compared to the enhanced sampling data used to obtain π from a standard reweighting scheme,

$$\mathbb{P}(\pi|C, E) \approx \mathbb{P}(\pi|E), \tag{14}$$

we model the uncertainty encoded in the desired posterior by inserting the two approximations Eqs. (13) and (14) into Eq. (12):

$$\mathbb{P}(P|C, E) \approx \int d\pi \mathbb{P}(P|C, \pi)\mathbb{P}(\pi|E). \tag{15}$$

Approximate sampling from $\mathbb{P}(P|C, E)$ can now be achieved by drawing a random sample $\pi^{(1)}, \ldots, \pi^{(M)}$ distributed according to a given error model, $\pi^{(k)} \sim \mathbb{P}(\pi|E)$, and generating a sample of transition matrices $P_1^{(k)}, \ldots, P_N^{(k)}$ from the constrained posterior $P_i^{(k)} \sim \mathbb{P}(P|C, \pi^{(k)})$ for each of the $\pi^{(k)}$. The sample $P_1^{(1)}, \ldots, P_N^{(1)}, \ldots, P_1^{(M)}, \ldots, P_N^{(M)}$ will then be approximately distributed according to $\mathbb{P}(P|C, E)$.

In Ref. [39] we presented a Markov chain Monte Carlo approach to sample reversible transition matrices fulfilling detailed balance with respect to a fixed equilibrium distribution. This method, however, has suffered from poor acceptance probabilities. In Ref. [55], we outline a method to efficiently generate samples from the constrained posterior using a Gibbs sampling algorithm that we also use here.

For given vector (π_i) detailed balance Eq. (6) enforces a linear dependence between the transition matrix element p_{ij} and the element p_{ji}. As an immediate consequence, the standard error of both elements for a sample generated from the posterior $\mathbb{P}(P|C)$ has to be equal:

$$\frac{\sqrt{\mathbb{V}(p_{ji})}}{\mathbb{E}(p_{ji})} = \frac{\sqrt{\mathbb{V}(p_{ij})}}{\mathbb{E}(p_{ij})}. \tag{16}$$

We show how this can be used in order to significantly improve various estimates in situations in which $p_{ij} \ll p_{ji}$.

III. RESULTS

Here, we demonstrate the usefulness of Eq. (16) via a comparison of the standard error for kinetic quantities depending on rare events that are either estimated from a Markov model of the direct unbiased simulation [unconstrained posterior Eq. (5)], or from a combination of direct simulations and enhanced sampling data [constrained posterior Eq. (8) or constrained posterior with uncertain equilibrium distribution Eq. (15)].

A. Finite state space Markov chain

Consider a three-state Markov chain with the following transition matrix:

$$P = \begin{pmatrix} 1 - 10^{-b} & 10^{-b} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 10^{-b} & 1 - 10^{-b} \end{pmatrix}. \tag{17}$$

The parameter $b > 0$ can be thought of as the height of an energy barrier between states one and three. The corresponding equilibrium distribution is given by

$$\pi = (1 + 10^{-b})^{-1}\left(\frac{1}{2}, 10^{-b}, \frac{1}{2}\right)^{T}. \tag{18}$$

The pair (π, P) satisfies the detailed balance equation (6).

Any process starting in state one has an exponentially small probability of crossing over to state three. In fact, a chain starting in state one can reach state three only via state

two, but the probability to go from state one to state two is exponentially small in the barrier height b. The reversed process, going from state two to state one, occurs much faster. The same applies to state three and state two. The eigenvalues of this matrix are

$$\lambda_1 = 1, \qquad \lambda_2 = 1 - 10^{-b}, \qquad \lambda_3 = -10^{-b},$$

and the slowest time scale in the system is given by

$$t_2 = -\frac{1}{\log \lambda_2} \approx 10^b.$$

It is apparent from $t_2 \approx p_{12}^{-1}$ that estimates of t_2 and of p_{12} have similar standard errors. The standard error ϵ for a matrix element p_{ij} for sampling from the unconstrained posterior Eq. (5) is

$$\epsilon(p_{ij}) = \frac{1}{\sqrt{c_{ij}}}.$$

For $b = 4$ and a single chain of length $N \approx 7 \times 10^4$ steps starting in state one, we can on average expect $c_{12} = 4$ resulting in a relative standard error of 50%. In order to decrease the error down to 1% we would need to run a chain of length $N \approx 100^2 \times 10^4 = 10^8$ steps. This is clearly an unsatisfactory situation and we would like to reduce the required simulation effort to reach a given error level as much as possible.

In comparison for an ensemble of M short chains of length L, with $L \ll 10^b$, starting in state two one will on average observe a transition from state two to state one for every second chain, $c_{21} = M/2$, so that a relative error of 1% for p_{21} can be achieved for $M \approx 10^4$, with $L \ll 10^b$, so that the total simulation effort can be reduced by orders of magnitude.

We do not have explicit expressions for the standard error of matrix elements p_{ij} when sampling from the restricted ensemble enforcing detailed balance with respect to a given equilibrium distribution. It is, however, conceivable that the standard errors of p_{21} can be reduced in the same way. The relation Eq. (16) guarantees that a small error for p_{21} will also result in a small error for the rare-event quantity p_{12}.

Figure 1 shows the standard error of t_2 versus the total simulation effort. The error for a single long chain is estimated from a sample of transition matrices generated from the unconstrained posterior. The error for the ensemble of short chains is estimated from a sample of transition matrices generated from the constrained posterior using the algorithm outlined in Ref. [55]. From Fig. 1 it is apparent that using *a priori* information about the equilibrium distribution in combination with an ensemble of short simulations started from the unstable state results in a 3 orders of magnitude smaller simulation effort when trying to estimate t_2 with a prescribed error. In particular,

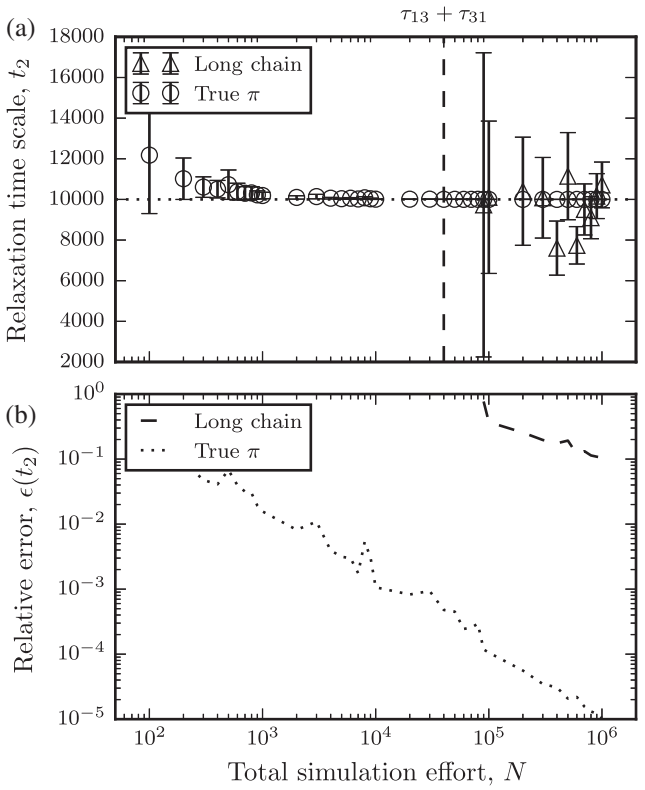

FIG. 1. Mean and standard error of the largest implied time scale t_2 total simulation effort N for metastable three-state system with barrier parameter $b = 4$. (a) Convergence of the mean value, using either a single long trajectory starting in one of the metastable states or the equilibrium distribution together with an ensemble of short chains relaxing from the transition state. The latter approach allows us to obtain a reliable estimate before the average waiting time for a single rare event $\tau_{13} + \tau_{31}$ has elapsed. The comparison of the estimated standard error (b) indicates a 3 orders of magnitude speed-up when estimating the rare-event sensitive quantity t_2 using the equilibrium distribution in combination with short relaxation trajectories.

estimation of the rare-event kinetics can be conducted orders of magnitude before a direct simulation would even encounter a single transition event.

This effect is even more pronounced when choosing $b = 9$, so that estimation via long trajectories, which need to sample the rare event, is hopeless. Using short trajectories starting in the transition state in combination with the equilibrium distribution, one can accurately estimate t_2 with a total simulation effort of $N = 10^3$ steps; cf. Fig. 2. That is 6 orders of magnitude before on average even a single rare event would have been observed.

B. Double-well potential

Let us now go to an example where the Markov state model is an approximation of the true dynamics. We employ Brownian dynamics in a double-well potential defined by

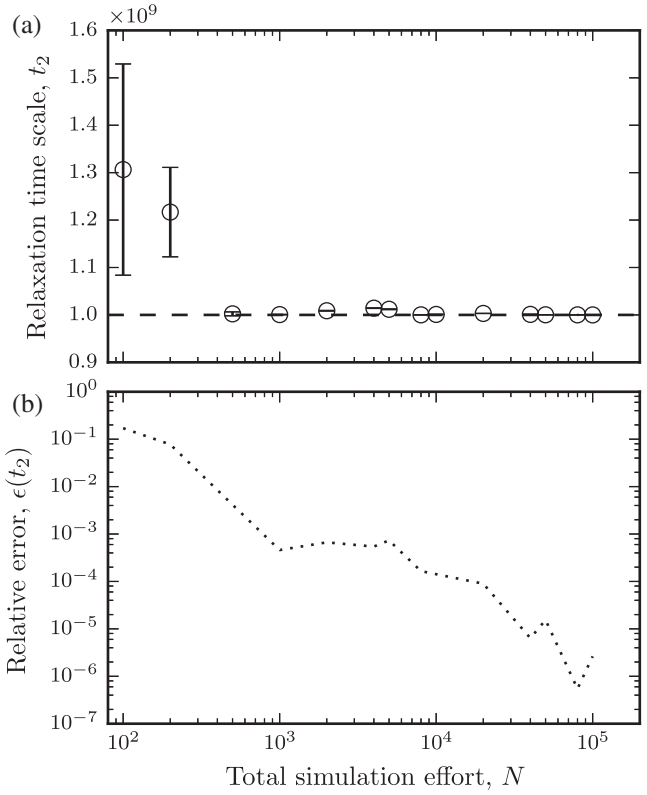

FIG. 2. Mean and standard error of the largest relaxation time scale t_2 total simulation effort N for metastable three-state system with barrier parameter $b = 9$. (a) Convergence of the mean value using short trajectories relaxing from the transition state. A correct estimate can be obtained 6 orders of magnitude before a single rare event would have occurred on average. (b) Standard error of the estimate. The estimation using long trajectories is unfeasible.

$$V(x) = (x^2 - \sigma^2)^2 + \delta\sigma\left(\frac{1}{3}x^3 - \sigma^2 x\right). \tag{19}$$

The two minima of the potential at $\pm\sigma$ are separated by a maximum at $-\delta\sigma/4$; cf. Fig. 3. The dynamics is governed by the following stochastic differential equation (SDE):

$$dX_t = -\nabla V(X_t) + \sqrt{2\beta^{-1}}dW_t, \tag{20}$$

with dW_t denoting the increments of the Wiener process. The inverse temperature $\beta = (k_B T)^{-1}$ controls the intensity of the stochastic fluctuations.

Equation (20) defines a process, X_t, that samples from the canonical distribution,

$$\pi(x) = Z(\beta)^{-1} e^{-\beta V(x)}. \tag{21}$$

The temperature-dependent constant $Z(\beta)$ is the partition function ensuring correct normalization, $\int dx\pi(x) = 1$. Spectral properties of this Markov process, such as the largest implied time scale, can be computed from a

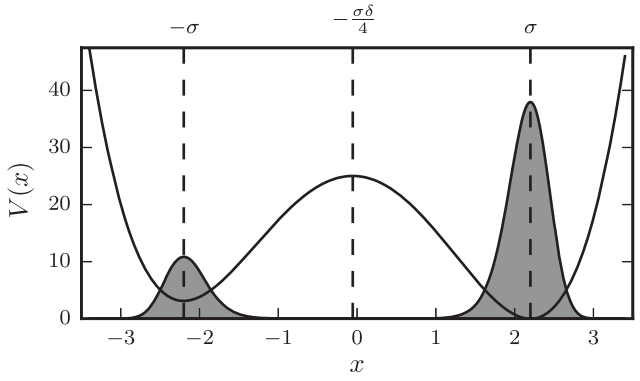

FIG. 3. Potential $V(x)$ and equilibrium distribution $\pi(x)$ for Brownian dynamics in double-well potential. The equilibrium distribution (shaded area) is scaled to fit the scale of the potential function. It can be seen that the equilibrium distribution is concentrated in the metastable regions around the two minima of the potential at $\pm\sigma$.

spatial discretization of its associated transition kernel; cf. Appendix A.

For the numerical experiment we use a double-well potential with parameters $\sigma = 2.2$ and $\delta = 0.1$. The time step for the explicit Euler scheme is $\Delta t = 10^{-3}$. The noise parameter is $\beta = 0.4$.

Spatial discretization of the transition kernel is performed with $L_x = 3.4$ and $n_x = 400$ regular subintervals. The matrix (p_{ij}) is assembled by evaluating the kernel at the midpoints of the subintervals. The largest implied time scale, $t_2 = 1.2 \times 10^6$, is computed from an eigenvalue decomposition of the assembled matrix. Mean first-passage times (MFPTs) between sets $A = [\sigma - 0.2, \sigma + 0.2]$ and $B = [-\sigma - 0.2, -\sigma + 0.2]$ are computed as $\tau_{AB} = 5.3 \times 10^6$ and $\tau_{BA} = 1.6 \times 10^6$, see Appendix B for details. Values computed from the spatial discretization are used as reference values for comparison with estimates obtained from a Markov model.

The Markov model is built using a regular grid discretization of $[-L, L]$, with $L = 3.4$ and $n = 100$ states. From an implied time-scale estimation using long trajectories with $N = 10^8$ steps, we obtain a lag time of $\tau = 10dt$.

The equilibrium distribution is estimated from umbrella sampling simulations using the weighted histogram analysis method [51,56]. Estimates are computed using $M_\pi = 20$ umbrella sampling simulations with $L_\pi = 2.5 \times 10^4$ points per umbrella, as well as from umbrella sampling simulations with $L_\pi = 5 \times 10^6$ points per umbrella. To account for the uncertainty in the estimated equilibrium distribution, we use bootstrap resampling [57] of the generated data and compute the equilibrium distribution for each resampled data set to model the ensemble of equilibrium distribution compatible with the observed umbrella sampling data.

In Fig. 4 we show the mean and standard error of the largest implied time scale t_2 versus the total simulation effort N. The total simulation effort N is composed of the

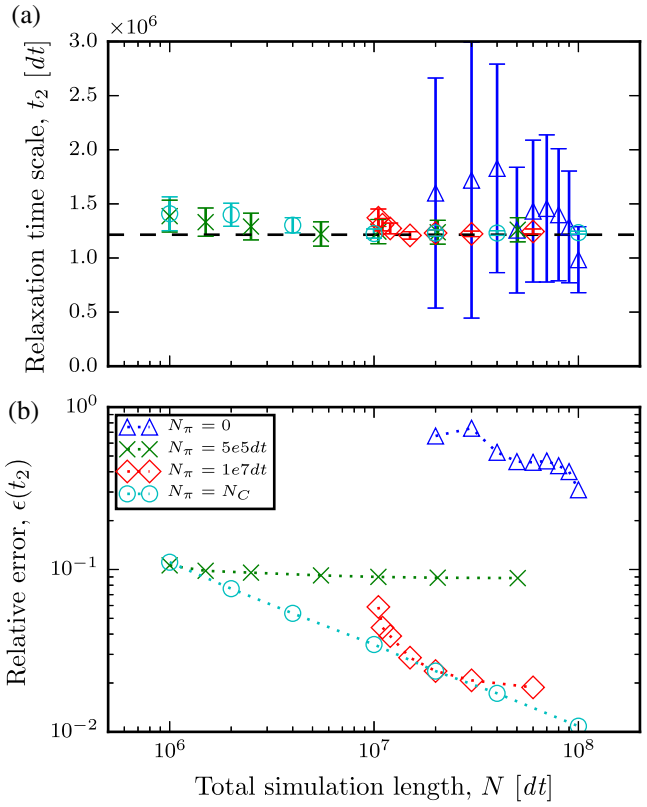

FIG. 4. Mean and standard error of largest implied time scale t_2, given total simulation effort N, for Brownian dynamics in double-well potential. (a) Convergence of the mean value, using either a single long trajectory starting in one of the metastable states or the equilibrium distribution together with an ensemble of short chains relaxing from the transition state. The latter approach allows us to obtain a reliable estimate before the average waiting time for a single rare event $\tau_{AB} + \tau_{BA}$ has elapsed. A comparison of the standard error (b) indicates a more than 2 orders of magnitude speed-up when estimating the rare-event sensitive quantity t_2. By combining short trajectories with information about the equilibrium probabilities, reliable estimates of the slowest relaxation time scale can be obtained with a total amount of simulation data that is about 1 order of magnitude smaller than the expected waiting time for a forward and backward transition across the barrier.

simulation effort spent on obtaining a count matrix from standard simulations N_C and the simulation effort spent on obtaining the equilibrium distribution from umbrella sampling simulations N_π:

$$N = N_\pi + N_C. \tag{22}$$

We compare three different approaches when estimating the mean and standard error of the largest implied time scale t_2.

(1) Generate a single trajectory starting in one of the metastable regions and compute estimates without *a priori* knowledge of the equilibrium distribution.

(2) Generate an ensemble of short trajectories starting on the barrier and compute estimates with an error model for the equilibrium distribution as prior information.

(3) Balanced sampling: Split the total simulation effort equally between umbrella simulations and short trajectories starting on the barrier, $N_\pi = N_C = N/2$. Compute estimates updating the error model for the equilibrium distribution according to the increasing amount of data available for the estimation.

Transition matrices are sampled according to Eq. (5) if no prior knowledge about the equilibrium distribution is available and from Eq. (15) if the equilibrium distribution is estimated from umbrella simulation data. For the first approach we use $M_C = 20$–100 long trajectories of length $L_C = 10^6 dt$ starting in the minimum point, $x_0 = s$, and for the second approach we use an ensemble of $M_C = 50$–5000 short trajectories of length $L_C = 10^4 dt$ starting on the barrier, $x_0 = -\delta\sigma/4$. For the second approach we estimate the equilibrium distribution from a small as well as for a large amount of umbrella sampling data in order to demonstrate the dependence of the standard error of the kinetic observable on the error in the ensemble of input equilibrium distributions.

It can be seen from Fig. 4 that for a fixed effort $N_\pi = M_\pi L_\pi$ the standard error cannot be reduced below a certain amount with increasing $N_C = M_C L_C$. This is a result of the nonzero statistical error in the estimate of the equilibrium distribution for fixed N_π. The usual $N^{-1/2}$ dependence of the standard error can be recovered for the proposed splitting $N_\pi = N_C = N/2$. Figure 4 shows the favorable scaling coefficient of such an approach leading to a more than 2 orders of magnitude faster convergence of the estimated quantity compared to using standard simulations alone. Reliable estimates of the rare-event kinetics can be obtained 1 order of magnitude simulation effort before the standard approach using long trajectories, and no information about the equilibrium probabilities can be applied at all. The finite error for the estimate of the equilibrium distribution for $N_\pi = 5 \times 10^4 dt$ and $N_\pi = 10^7 dt$ results in a saturation of the error of t_2, which can be further decreased using a more precise estimate of the equilibrium distribution from additional enhanced sampling simulations.

For metastable systems we propose the following strategy for distributing initial conditions exploiting the information from the equilibrium vector. Once all metastable sets and all kinetic barriers separating the sets have been identified using some enhanced sampling protocol, short trajectories should be started on top of all barriers or in high-energy metastable states. The length of the short trajectories needs to be sufficient to relax towards the low-energy metastable states. The method described here can be used to combine these data to an estimate of the full rare-event kinetics.

C. Alanine dipeptide

As an example for a rare-event quantity in a molecular system, we use the mean first-passage time for the C_5 to C_7^{ax} transition in the alanine-dipeptide molecule. Alanine dipeptide has been the long-serving laboratory rat of molecular dynamics [58–62]. The ϕ and ψ dihedral angles have been identified as the two relevant coordinates for the slowest kinetic processes of the system in equilibrium. The potential of mean force for the two dihedral angles is shown in Fig. 5.

One can identify five metastable regions in the free-energy landscape. The C_5 and P_{II} regions correspond to dihedral angles found in a beta-sheet conformation and the α_R and α_L regions correspond to a right- and a left-handed α-helix conformation. Reference values for the mean first-passage times between all pairs of sets are computed from the maximum likelihood estimator of Eq. (3) using a total of 10 μs of simulation data. Values can be found in Table I. For details of the computation of mean first-passage times, see Appendix B.

All computations are carried out on high-performance graphics processing unit (GPU) cards using the OpenMM simulation package [63]. The force field we use is *amber99sb-ildn* [64] and the water model we use is *tip3p* [65]. The peptide is simulated in a cubic box of 2.7-nm length including 652 solvent molecules. Langevin equations are integrated at $T = 300$ K using a time step dt of 2 fs. The potential we use for umbrella sampling simulations is $V_i(\phi) = k[1 + \cos(\phi - \phi_i - \pi)]$, with $k = 200$ kJ/mol. Umbrellas are placed at a spacing of $\phi_i - \phi_{i+1} = 9°$.

1. Analysis in ϕ and ψ dihedral angle space

We show the convergence of the largest relaxation time scale and validate the MSM constructed at a lag time of $\tau = 6$ ps via a Chapman-Kolmogorov test in Fig. 12. Convergence of the largest relaxation time indicates that the slow eigenfunctions of the associated dynamical operator are well approximated by the discrete MSM. The Chapman-Kolmogorov test explicitly checks the Markov assumption comparing self-transition probabilities computed from the MSM, parametrized at lag time τ, with direct estimates from the data at larger lag times, $n\tau$. A thorough discussion of MSM validation can be found in Ref. [8].

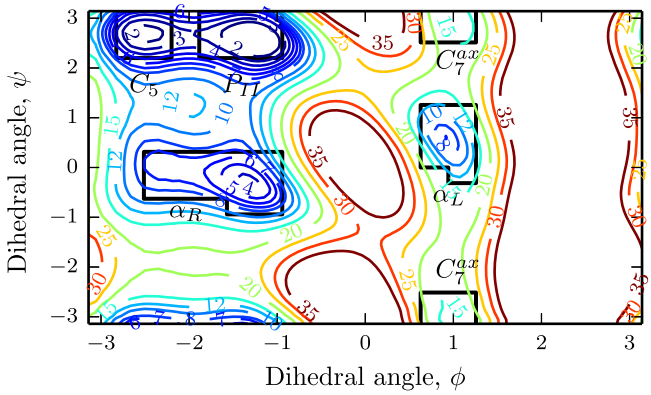

FIG. 5. Free-energy profile of alanine dipeptide as a function of the dihedral angles. Energies are given in kJ/mol. The average thermal energy k_BT at 300 K is 2.493 kJ/mol. One can identify five metastable sets on the dihedral angle torus, indicated here by black lines. There are three low-energy (high-probability) sets C_5, P_{II}, and α_R, with $\phi < 0$, and two high-energy (low-probability) sets α_R and C_7^{ax}, with $\phi > 0$.

TABLE I. Mean first-passage time (MFPT) between metastable regions of alanine dipeptide. The MFPTs have been estimated from 10 μs of simulation data using a Markov state model.

τ_{AB}/ns	C_5	P_{II}	α_R	α_L	C_7^{ax}
C_5	0	0.021	0.253	43.456	60.220
P_{II}	0.041	0	0.255	43.449	60.213
α_R	0.142	0.125	0	43.549	60.312
α_L	1.553	1.527	1.744	0	17.757
C_7^{ax}	1.559	1.533	1.745	1.221	0

In Fig. 6 we show the estimate of the mean first-passage time τ_{AB} between the C_5 and the α_L region together with the corresponding standard error $\epsilon(\tau_{AB})$ for different values of the total simulation effort N. The simulation setup is similar to the one described for the double-well potential in the previous section. Instead of starting short trajectories directly on the barrier, we start them from the metastable α_L region. Figure 6 shows that, by combining umbrella sampling data and short trajectories relaxing from a metastable region with low probability (high free-energy) towards a metastable state with high probability (low free-energy), we are able to estimate the reference value, $\tau_{AB} = 43$ ns, for the C_5 to α_L transition with a total simulation effort of 70 ns if short downhill trajectories are used in combination with umbrella sampling data. Utilizing information about the equilibrium distribution in combination with short simulations that do not have to sample the rare event, we are able to achieve a standard error with almost an order of magnitude less simulation effort compared to an ensemble of long trajectories. The observed eightfold speed-up is in good agreement with the expected speed-up given by

$$\frac{\tau_{AB}}{L},$$

with $\tau_{AB} = 43$ ns the MFPT for the slow "up-hill" transition from C_5 to α_L and $L = 5$ ns the length of individual short trajectories.

The present approach of estimating rare-event kinetics is more powerful than traditional rate theories because quantities that can be estimated can be much more complex than only rates. As a reversible Markov model is estimated, full mechanisms, such as the ensemble of transition pathways from one state to another state, can be computed.

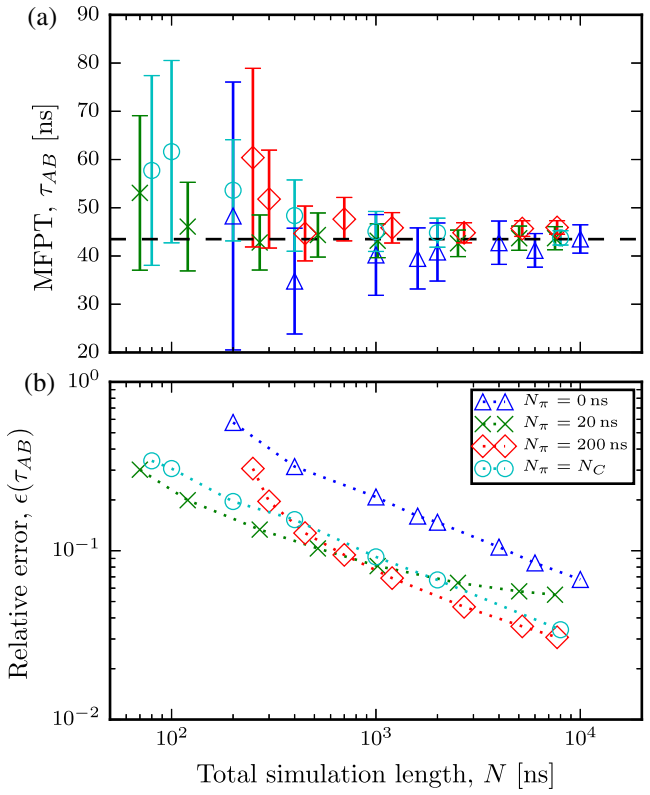

FIG. 6. Mean and standard error of mean first passage time (MFPT) τ_{AB}, total simulation effort N, for alanine-dipeptide MSM on the ϕ, ψ dihedral angles. The mean first-passage time τ_{AB} of the C_5 to α_L transition is used as an observable for a rare-event process. (a) Convergence of the mean value is shown for a small number of long chains starting in the C_5 region (blue), an ensemble of short chains starting in the α_L region combined with different amounts of umbrella sampling simulations (green, red, light blue). The correct value of the C_5 to α_L transition, $\tau_{AB} = 43$ ns, can be obtained for a total simulation effort of $N = 70$ ns when short "down-hill" simulations are used in combination with umbrella sampling data. (b) The standard error shows almost 1 order of magnitude speed-up when estimating the kinetic characteristic of a rare event τ_{AB} using short trajectories in combination with umbrella sampling simulations compared to using long trajectories and no additional information about the equilibrium distribution.

To illustrate this we compute the committor probability function, cf. Appendix C, from C_5 to α_L using both estimates Fig. 7. We see that information about the equilibrium distribution results in nearly the same committor function as the one estimated using an order of magnitude larger simulation effort.

2. Analysis in the ϕ coordinate alone

The method we present can also work if only information about the slowest degree of freedom is used. In Fig. 8, we show the free-energy profile for the ϕ dihedral angle. An energetic barrier clearly separates the low free-energy region, $\phi < 0$, from the high free-energy region, $\phi > 0$.

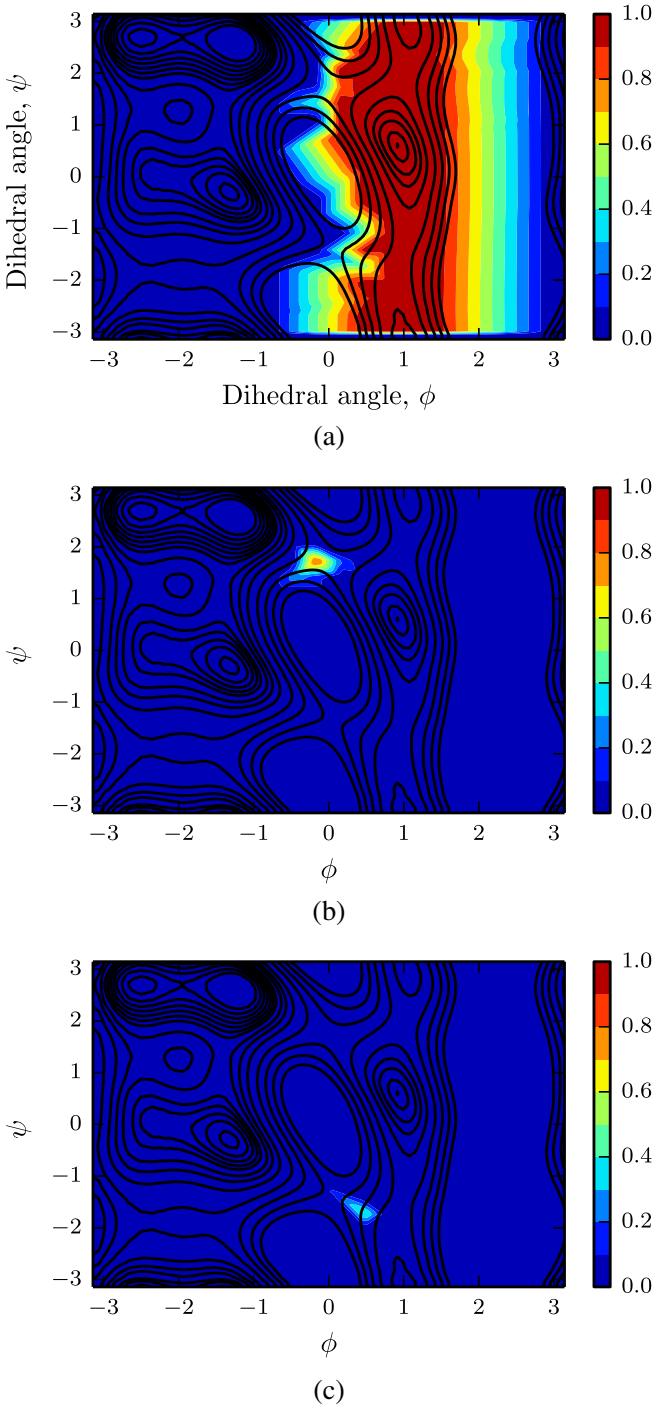

FIG. 7. Forward committor $q^+(x)$ for transition from C_5 to α_L region. (a) Nonreversible reference estimate for $N = 10$ μs of simulation data. Dark contour lines indicate the free-energy profile. (b) Difference between the reference estimate and a nonreversible estimate for $N = 1$ μs of simulation data. There is a large error in the transition region due to insufficient sampling in the short simulation. (c) Distance for an estimate using a combination of umbrella sampling and standard simulation data with $N = N_\pi + N_C = 960$ ns. There is no significant error in the transition region; the small error close to the second saddle is probably due to insufficient sampling of this region by the reference simulation.

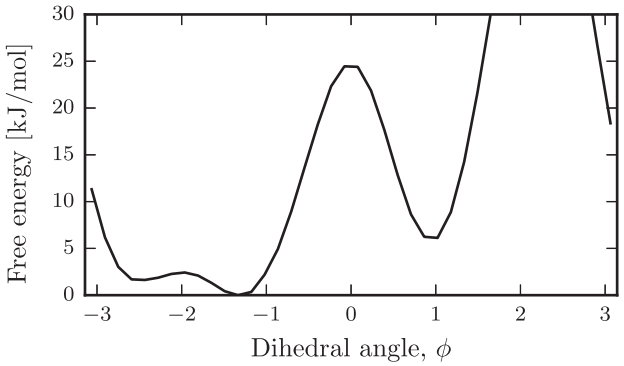

FIG. 8. Free-energy profile for alanine dipeptide as a function of the ϕ dihedral angle. One can identify three metastable sets: two low-energy (high-probability) sets with $\phi < 0$ and a single high-energy (low-probability) set with $\phi > 0$.

Crossing events from $\phi < 0$ to $\phi > 0$ are rare, leading to a sampling problem if kinetic quantities associated with barrier crossings need to be estimated. Again, we show convergence of the largest relaxation time scale t_2 and a Chapman-Kolmogorov test for a MSM estimated at lag time $\tau = 15$ ps, Fig. 13.

In Fig. 9, we show that the correct mean first-passage time for the C_5 to α_L transition can also be recovered from the MSM of the ϕ angle alone. This demonstrates that the method we present is robust with respect to the choice of microstates. Choosing a slightly larger lagtime, $\tau = 15$ ps, for the ϕ MSM allows us to recover the correct mean first-passage times despite the fact that information about the ψ dihedral angle is completely neglected. The MSM for ϕ dihedral angle is still a good approximation to the true kinetics if the discretization and the lag time are suitably matched. A thorough discussion of approximation errors for MSMs can be found in Refs. [8,11].

D. Vesicle model

As a final example, we consider the diffusive motion of a colloid that can reversibly attach to a surface via $m = 0, \ldots, M$ tethers. A biological example of such a system is a neuronal vesicle that can attach to a plasma membrane by soluble N-ethylmaleimide-sensitive-factor attachment receptor (SNARE) protein complexes. The diffusion in the solvent is free, but the attachment of tethers restricts the location of the vesicle to a vicinity of the membrane. The restriction is stronger the more tethers are attached. Attachment of the vesicle to the membrane is a fast process, but the dissociation from the membrane is an extremely rare event. We show that the mean first-passage time for dissociation can be reliably estimated despite the fact that a non-Markovian coordinate, the membrane-vesicle distance, is used.

Figure 10 shows the energy for the different vesicle attachment modes. For $m > 0$, attachment of the vesicle to the membrane is governed by a harmonic potential close to the membrane. For $x > 2$, all attachment modes are

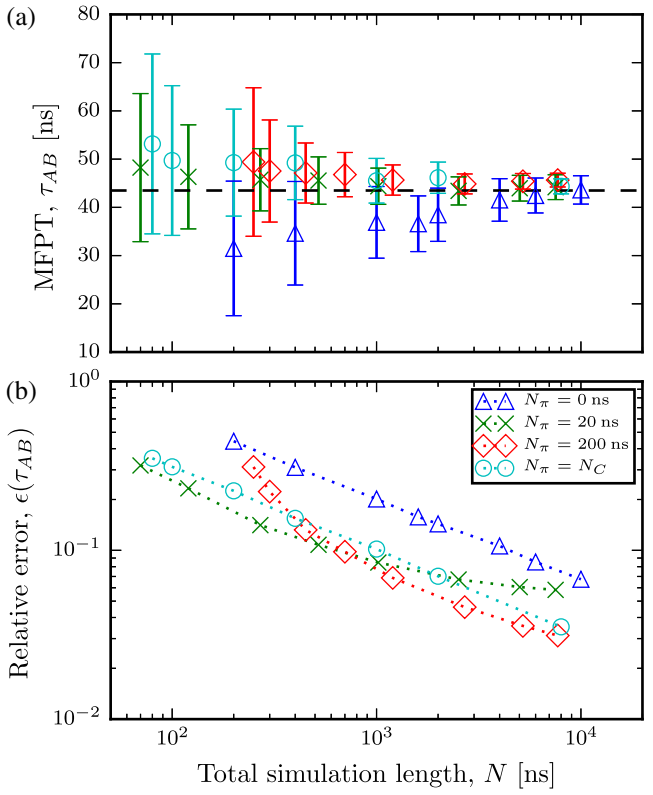

FIG. 9. Mean and standard error of MFPT τ_{AB}, total simulation effort N, for alanine-dipeptide MSM on the ϕ dihedral angle alone. The mean first-passage time τ_{AB} of the transition from the low free-energy region, $A = \{\phi | -162° < \phi < -54°\}$, to the high free-energy region, $B = \{\phi | 36° < \phi < 72°\}$, is used as an observable for a rare-event process. (a) Convergence of the mean value is shown for a small number of long chains starting in the A region (blue) and for an ensemble of short chains starting in the B region combined with different amounts of umbrella sampling simulations (green, red, light blue). The correct value, $\tau_{AB} = 43$ ns, for the C_5 to α_L transition can be obtained even if no information about the ψ dihedral angle is used in the construction of the MSM. (b) The standard error shows almost 1 order of magnitude speed-up when estimating the kinetic characteristic of a rare event τ_{AB} using short trajectories in combination with umbrella sampling simulations compared to using long trajectories and no additional information about the equilibrium distribution.

energetically equal corresponding to a breaking of the m tethers once the distance between the vesicle and the membrane exceeds a certain threshold. The association of the vesicle has to overcome a small energetic barrier, modeling a weak repulsion of the untethered vesicle.

The state of the vesicle is given by the pair (x, m), where x is the vesicle membrane distance and m denotes the number of tethers attached. A discretization of the vesicle membrane distance with $0 = x_1 < \ldots < x_d = 4$ allows us to describe the vesicle dynamics by a Markov chain on a finite state space with $(M + 1)d$ microstates. The equilibrium distribution of the chain is given as

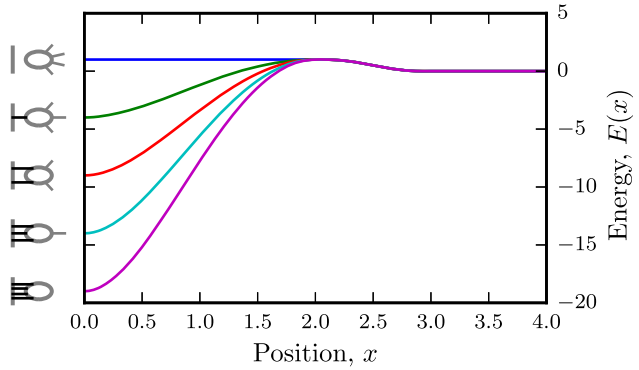

FIG. 10. Energy landscape for the different attachment modes, $m = 0, 1, \ldots, 4$

$$\pi = [\pi^{(0)}(x_1), \ldots, \pi^{(M)}(x_d)], \tag{23}$$

with entries given in terms of the usual Gibbs-Boltzmann distribution:

$$\pi^{(m)}(x_i) \propto e^{-E^{(m)}(x_i)}. \tag{24}$$

$E^{(m)}(x)$ is the energy of a vesicle at x with m tethers attached; cf. Fig. 10 and Eq. (D1).

The transition matrix $P = (p_{ij})$ for the vesicle dynamics is now constructed as follows. We encode random walk probabilities in a proposal matrix $Q = (q_{ij})$. The particle moves from x_i to x_{i-1} or x_{i+1} with probability $1/3$; if the particle remains at its current position x_i, it can attach, $m \to m+1$, or detach, $m \to m-1$, a tether with probability $1/3$ so that the overall proposal probability for attachment or detachment is $1/9$. To account for the energetic differences of the microstates, we use the Metropolis-Hastings acceptance criterion to modulate the proposal probabilities and obtain the desired transition probabilities via

$$p_{ij} = \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\}, \qquad i \neq j. \tag{25}$$

Correct normalization is ensured by setting $p_{ii} = 1 - \sum_{j \neq i} p_{ij}$. As a result of Eq. (25), the constructed transition matrix P automatically fulfills the detailed balance condition Eq. (6) with respect to the desired equilibrium distribution.

The mean first-passage time for the dissociation of the vesicle is $\tau_{AB} = 8.56 \times 10^9$, the mean first-passage time for association, $\tau_{BA} = 1.59 \times 10^3$, is orders of magnitude smaller. The mean first-passage time for dissociation of a vesicle with the maximum number of tethers attached is $\tau_{AB} = 3.83 \times 10^{10}$, so that the system dynamics cannot be described in terms of the subspace with $m = 4$ tethers. This indicates that the dissociation kinetics is effectively non-Markovian along the x coordinate.

The dissociation time τ_{AB} can reliably be estimated even if no information about the mode of attachment is available.

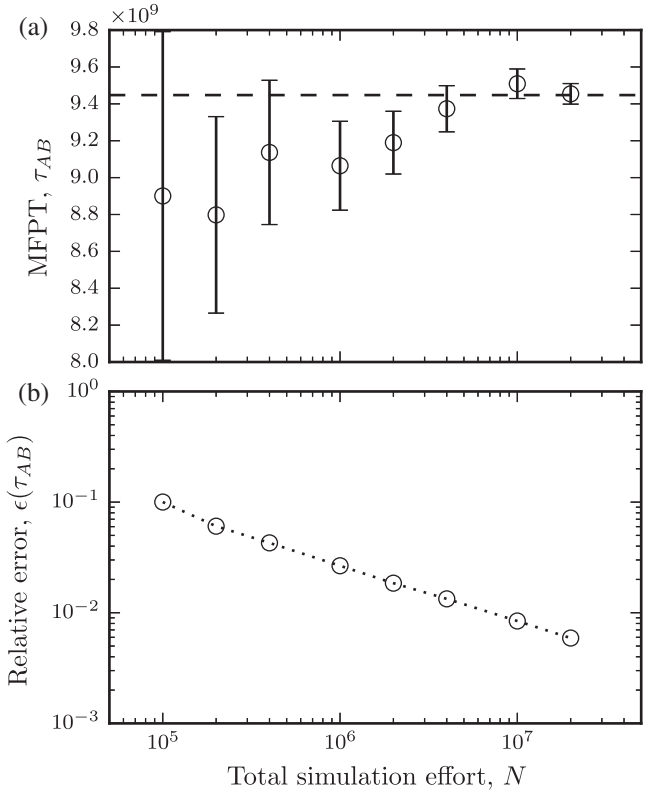

FIG. 11. Mean and standard error of mean first-passage time of vesicle dissociation on a coarse-grained non-Markovian state space. (a) Convergence of the mean value, (b) standard error. Estimates are obtained for an ensemble of association trajectories starting in the high-energy region and relaxing towards the low-energy region in combination with the coarse-grained equilibrium distribution. The dissociation time can be estimated orders of magnitude before a single dissociation event would have been observed.

If only information about the position of the vesicle is available, then the state space of the $(M+1)d$ distinct microstates is coarse grained into d distinct sets, each containing $(M+1)$ microstates corresponding to the $M+1$ possible tethering modes at position x. The coarse-grained equilibrium distribution $\tilde{\pi}$ is obtained by summing the full equilibrium distribution π over all possible tethering modes. If short association trajectories starting in the region $x > 2$ are combined with the coarse-grained equilibrium distribution, the dissociation time can again be estimated orders of magnitude before a single dissociation event would on average be observed despite the fact that the MSM is built on a coordinate that is inherently non-Markovian. In Fig. 11, we show the mean and standard error for the MFPT of vesicle dissociation for a MSM build at a lag time of $\tau = 60$ with $d = 40$ microstates.

In Fig. 14, we again show convergence of the largest relaxation time and the Chapman-Kolmogorov test for a MSM constructed at lag time $\tau = 60$. The MSM is estimated solely from short association trajectories starting

in the high-energy region using the coarse-grained equilibrium distribution $\tilde{\pi}$ to obtain a reversible maximum likelihood transition matrix from Eq. (9). The total simulation effort, $N = 2 \times 10^7$, we use to obtain the MSM and perform the validation is again orders of magnitude smaller than the expected dissociation time.

IV. CONCLUSION

We describe a principle that allows rare-event kinetics to be efficiently estimated without having to assume a rate model. Our approach is applicable when the kinetic properties of interest can be computed from a Markov state model discretization of the system. Note that this approach is qualitatively different from assuming a specific rate theory, such as transition state theory or Kramers model, because MSMs are a numerical approximation method of the full kinetics and can be made arbitrarily accurate in the limit of a good state space discretization [8], whereas a specific rate model needs to apply by design and can usually not be self-consistently validated.

The key idea of the our approach is to use enhanced sampling methods to obtain reliable estimates of the equilibrium distribution in combination with direct simulations of the fast downhill processes. These data are combined rigorously in a reversible Markov model. Our approach can deliver estimates of kinetic properties, including rates, passage times, as well as complex quantities, such as committor functions and transition path ensembles, while achieving enormous speed-ups compared to a direct simulation.

We illustrate our method using two toy models, an explicit-solvent MD simulation of alanine dipeptide with about 2000 degrees of freedom and a model for reversible attachment of a vesicle to a membrane. In these examples, the kinetics of the rare events could be computed using between 1 and 6 orders of magnitude less simulation time than needed with a direct simulation approach that has to wait for the rare events to happen spontaneously.

In general, the present approach will be efficient whenever the rare event occurs between low-probability and high-probability states. A very important example of this class is computational drug design, where the binding of the drug compound occurs relatively fast [18], while the unbinding may be many orders of magnitude slower. Yet the unbinding kinetics have been shown to be critical for drug efficacy [66].

We demonstrate in two applications that the approach can compute kinetics from non-Markovian projections of the data: by using only the ϕ coordinate in alanine dipeptide, and by using only the distance coordinate in the vesicle attachment model. A requirement is that the resolved coordinates are slow compared to the nonresolved coordinates. However, this requirement is not overly restrictive, as the same requirement applies for the enhanced sampling simulations, such as umbrella sampling, employed to obtain an estimate for the equilibrium distribution.

While the applications in the present paper use reversible Markov model estimates in such a way that the enhanced sampling simulation and the unbiased "downhill" simulations visit the same state space, the principle we explore here can be generalized beyond this case. States visited only in one but not in the other simulation can be modeled by appropriate uninformative priors on the respective variables, e.g., uniform prior in the equilibrium distributions of states not visited in an umbrella sampling simulation.

A general framework to reconcile direct MD and enhanced MD simulations is the transition-based reweighting analysis method framework [29–31]. In order to apply the TRAM framework to the current setting, a hybrid TRAM method must be developed that can mix kinetic simulations (with an estimation lag time τ) and simulations that contain trajectories shorter than τ, such as those used in umbrella sampling or replica exchange molecular dynamics (REMD).

The present inference principle can be exploited in an adaptive sampling framework [67,68] to optimally distribute the computational effort between enhanced sampling and unbiased molecular dynamics simulations.

Our method is implemented in the python-based Markov modeling software PyEMMA [42], and demonstrated via IPython notebook tutorials in the Supplemental Material [43].

ACKNOWLEDGMENTS

We are grateful to Feliks Nüske for stimulating discussions. This work was funded by the Deutsche Forschungsgemeinschaft (DFG) Grants No. NO825/3-1 and No. SFB 740 TP D7 (B. T.-S.), and a European Research Council (ERC) starting grant pcCell (F. N.).

APPENDIX A: TRANSITION KERNEL FOR THE EULER METHOD

The solution of Eq. (20) with initial position $X_0 = x_0$ on $[0, T]$ is usually carried out by choosing a regular discretization of the time interval

$$0 = t_0 < t_1 < \ldots < t_N = T,$$

with $\Delta t = t_k - t_{k-1}$, for all $k = 1, \ldots, N$. The evolution of the stochastic process is then approximated by the following time-stepping scheme:

$$X_{t+\Delta t} = X_t - \nabla V(X_t)\Delta t + \sqrt{2\beta^{-1}}\eta, \qquad \text{(A1)}$$

with $X_0 = x_0$ and η being a $\mathcal{N}(0, \Delta t)$ distributed random variable. The time-stepping scheme Eq. (A1) is known as the Euler method or the Euler-Maruyama method [69].

For this simple time-stepping scheme, the transition kernel of the resulting Markov chain is given by

$$p_{\Delta t}(x,y) = \frac{1}{\sqrt{2\pi}\Delta t 2/\beta}\exp\left(-\frac{(y-x+\nabla V(x)\Delta t)^2}{2(\sqrt{\Delta t}\sqrt{2/\beta})^2}\right), \tag{A2}$$

with $x = X_t$ and $y = X_{t+\Delta t}$. $p_{\Delta t}(x,y)$ is a Gaussian distribution with mean $\mu = x - \nabla V(x)\Delta t$ and variance $\sigma^2 = 2\Delta t/\beta$.

The transition probability $P_{\Delta t}(B|A)$ between two sets A, B can be computed from

$$P_{\Delta t}(B|A) = \frac{\int_A dx\pi(x)\int_B dy p_{\Delta t}(x,y)}{\int_A dx\pi(x)}. \tag{A3}$$

Choosing L such that $p_{\Delta t}(x,y)$ is effectively zero outside of $[-L,L]$, we pick a spatial discretization

$$-L = x_0 < x_1 < \ldots < x_N = L, \tag{A4}$$

with a regular spacing $\Delta x = x_k - x_{k-1}$, for $k = 1,\ldots,N$, such that $p_{\Delta t}(x,y)$ and $\pi(x)$ are approximately constant on subintervals $S_i = (x_k, x_{k+1}]$. In this case, we have

$$\int_{x_i}^{x_{i+1}} dx\mu(x) \approx \mu(x_k)\Delta x$$

and

$$\int_{x_i}^{x_{i+1}} dx\mu(x)\int_{x_j}^{x_{j+1}} dy p(x,y) \approx \mu(x_i)p(x_i,x_j)(\Delta x)^2.$$

We can approximate the matrix elements $p_{ij} = P(S_j|S_i)$ as

$$p_{ij} \approx p(x_i,x_j)\Delta x$$

and compute spectral properties from the matrix (p_{ij}) using standard eigenvalue solvers.

APPENDIX B: MEAN FIRST-PASSAGE TIMES BETWEEN METASTABLE REGIONS

The covered material can be found in many introductory books to stochastic processes; cf. Ref. [70].

For a stochastic process (X_t) on a state space Ω, the first hitting time T_B of a set $B \subseteq \Omega$ is defined as

$$T_B = \inf\{t \geq 0 | X_t \in B\}. \tag{B1}$$

The mean first-passage time $\tau_{x,B}$ to the set B starting in state $x \in \Omega$ is the following expectation value:

$$\tau_{x,B} = \mathbb{E}_x(T_B). \tag{B2}$$

For a Markov chain on a finite state space $\Omega = \{1,\ldots,n\}$ with transition matrix $(p_{x,y})$, the mean first-passage time can be computed from the following system of equations:

$$\tau_{x,B} = \begin{cases} 0 & x \in B \\ 1 + \sum_{y\in\Omega} p_{x,y}\tau_{y,B} & x \notin B. \end{cases} \tag{B3}$$

Assuming that the chain has equilibrium distribution (μ_x), we define the mean first-passage time $\tau_{A,B}$ from set A to set B as the μ-weighted average of all mean first-passage times to B when starting in a state $x \in A$,

$$\tau_{A,B} = \sum_{x\in A}\mu_x\tau_{x,B}. \tag{B4}$$

Computing the mean first-passage time between two sets for a Markov chain on a finite state space with given transition matrix thus amounts to finding the equilibrium distribution together with the solution of a linear system of equations—both of which can be achieved using standard numerical linear algebra libraries.

APPENDIX C: COMMITTOR FUNCTIONS

Committor functions are introduced in the context of transition path theory [12] and are a central object for the characterization of transition processes between two metastable sets.

Let (X_t) again be a stochastic process on a state space Ω and let A, $B \subseteq \Omega$ be two metastable sets. The forward committor $q^{(+)}(x)$ is the probability that the process starting in x will reach the set B first, rather than the set A,

$$q^{(+)}(x) = \mathbb{P}_x(T_A < T_B). \tag{C1}$$

Again, T_S denotes the first hitting time of a set S.

For a Markov chain on a finite state space with transition matrix P, the forward committor solves the following boundary value problem [13]:

$$\begin{aligned} \sum_j l_{ij}q_j^{(+)} &= 0, \qquad i \in X(A\cup B), \\ q_i^{(+)} &= 0, \qquad i \in A, \\ q_i^{(+)} &= 1, \qquad i \in B. \end{aligned} \tag{C2}$$

$L = P - I$ is the corresponding generator matrix of the Markov chain.

Computing the committor for a finite state space again amounts to solving a linear system of equations.

APPENDIX D: VESICLE POTENTIAL

The potential for the vesicle model is given by

$$E^{(m)}(x) = \begin{cases} 1+m(-5+5x^2-2.5x^3+0.3125x^4) & 0\leq x<2 \\ 1+8(x-2)^2-8(x-2)^3 & 2\leq x<2.5 \\ 0.5-8(x-2.5)^2+8(x-2.5)^3 & 2.5\leq x<3 \\ 0 & 3\leq x<4. \end{cases} \tag{D1}$$

APPENDIX E: MSM VALIDATION

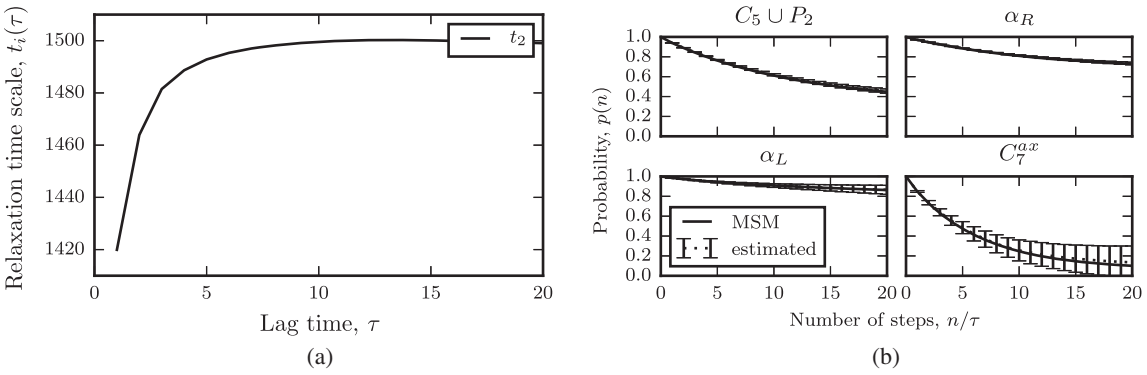

FIG. 12. (a) Implied time-scale test. Convergence of the largest relaxation time scale t_2 indicates a good Markov model fit; i.e., the slow eigenfunction of the associated dynamical operator is well approximated. (b) The Chapman-Kolmogorov test validates the Markov assumption by comparing the evolution of self-transition probabilities predicted by the MSM parametrized at lag time τ with direct estimates from the data at larger lag times $n\tau$.

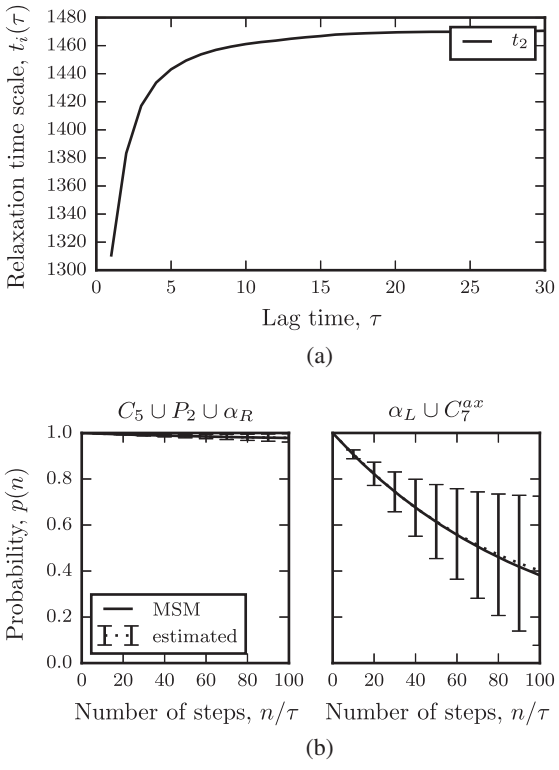

FIG. 13. (a) Implied time-scale test. Convergence of the largest relaxation time scale t_2 indicates a good Markov model fit; i.e., the slow eigenfunction of the associated dynamical operator is well approximated. (b) The Chapman-Kolmogorov test validates the Markov assumption by comparing the evolution of self-transition probabilities predicted by the MSM parametrized at lag time τ with direct estimates from the data at larger lag times $n\tau$.

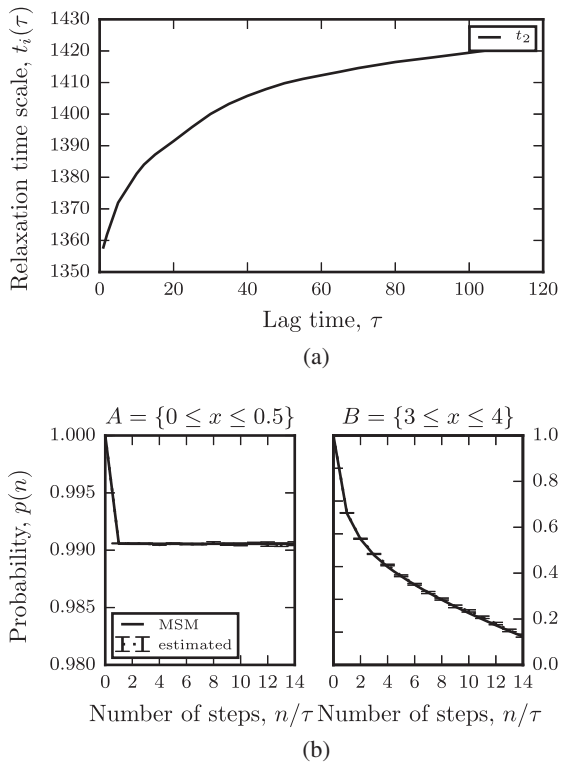

FIG. 14. (a) Implied time-scale test. Convergence of the largest relaxation time scale t_2 indicates a good Markov model fit; i.e., the slow eigenfunction of the associated dynamical operator is well approximated. (b) The Chapman-Kolmogorov test validates the Markov assumption by comparing the evolution of self-transition probabilities predicted by the MSM parametrized at lag time τ with direct estimates from the data at larger lag times $n\tau$. Values are obtained from an ensemble of short trajectories starting in the high-energy region utilizing the equilibrium distribution in the estimation of the maximum likelihood estimator (MLE) transition matrix; cf. Eq. (9).

We show the implied time-scale test and the Chapman-Kolmogorov test for the alanine dipeptide MSMs, Fig. 12 and Fig. 13, and for the vesiclemodel MSM, Fig. 14.

[1] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, *A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo*, J. Comput. Phys. **151**, 146 (1999).

[2] W. C. Swope, J. W. Pitera, and F. Suits, *Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory*, J. Phys. Chem. B **108**, 6571 (2004).

[3] N. Singhal, C. D. Snow, and V. S. Pande, *Using Path Sampling to Build Better Markovian State Models: Predicting the Folding Rate and Mechanism of a Tryptophan Zipper Beta Hairpin*, J. Chem. Phys. **121**, 415 (2004).

[4] J. D. Chodera, K. A. Dill, N. Singhal, V. S. Pande, W. C. Swope, and J. W. Pitera, *Automatic Discovery of Metastable States for the Construction of Markov Models of Macromolecular Conformational Dynamics*, J. Chem. Phys. **126**, 155101 (2007).

[5] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, *Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States*, J. Chem. Phys. **126**, 155102 (2007).

[6] A. C. Pan and B. Roux, *Building Markov State Models along Pathways to Determine Free Energies and Rates of Transitions*, J. Chem. Phys. **129**, 064107 (2008).

[7] N. V. Buchete and G. Hummer, *Coarse Master Equations for Peptide Folding Dynamics*, J. Phys. Chem. B **112**, 6057 (2008).

[8] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, *Markov Models of Molecular Kinetics: Generation and Validation*, J. Chem. Phys. **134**, 174105 (2011).

[9] M. Senne, B. Trendelkamp-Schroer, A. S. J. S. Mey, C. Schütte, and F. Noé, *EMMA: A Software Package for Markov Model Building and Analysis*, J. Chem. Theory Comput. **8**, 2223 (2012).

[10] K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, *MSMBuilder2: Modeling Conformational Dynamics at the Picosecond to Millisecond Scale*, J. Chem. Theory Comput. **7**, 3412 (2011).

[11] M. Sarich, F. Noé, and C. Schütte, *On the Approximation Quality of Markov State Models*, Multiscale Model. Simul. **8**, 1154 (2010).

[12] W. E and E. Vanden-Eijnden, *Towards a Theory of Transition Paths*, J. Stat. Phys. **123**, 503 (2006).

[13] P. Metzner, C. Schütte, and E. Vanden-Eijnden, *Transition Path Theory for Markov Jump Processes*, Multiscale Model. Simul. **7**, 1192 (2009).

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

[15] V. A. Voelz, G. R. Bowman, K. A. Beauchamp, and V. S. Pande, *Molecular Simulation of Ab Initio Protein Folding for a Millisecond Folder NTL9(1–39)*, J. Am. Chem. Soc. **132**, 1526 (2010).

[16] M. Held, P. Metzner, J.-H. Prinz, and F. Noé, *Mechanisms of Protein-Ligand Association and Its Modulation by Protein Mutations*, Biophys. J. **100**, 701 (2011).

[17] S. Gu, D.-A. Silva, L. Meng, A. Yue, and X. Huang, *Quantitatively Characterizing the Ligand Binding Mechanisms of Choline Binding Protein Using Markov State Model Analysis*, PLoS Comput. Biol. **10**, e1003767 (2014).

[18] I. Buch, T. Giorgino, and G. de Fabritiis, *Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations*, Proc. Natl. Acad. Sci. U.S.A. **108**, 10184 (2011).

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

[20] G. M. Torrie and J. P. Valleau, *Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling*, J. Comput. Phys. **23**, 187 (1977).

[21] H. Grubmüller, *Predicting Slow Structural Transitions in Macromolecular Systems: Conformational Flooding*, Phys. Rev. E **52**, 2893 (1995).

[22] Y. Sugita and Y. Okamoto, *Replica-Exchange Molecular Dynamics Method for Protein Folding*, Chem. Phys. Lett. **314**, 141 (1999).

[23] A. Laio and M. Parrinello, *Escaping Free-Energy Minima*, Proc. Natl. Acad. Sci. U.S.A. **99**, 12562 (2002).

[24] H. Eyring, *The Activated Complex in Chemical Reactions*, J. Chem. Phys. **3**, 107 (1935).

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

[26] R. B. Best and G. Hummer, *Coordinate-Dependent Diffusion in Protein Folding*, Proc. Natl. Acad. Sci. U.S.A. **107**, 1088 (2010).

[27] P. Tiwary and M. Parrinello, *From Metadynamics to Dynamics*, Phys. Rev. Lett. **111**, 230602 (2013).

[28] E. Rosta and G. Hummer, *Free Energies from Dynamic Weighted Histogram Analysis Using Unbiased Markov State Model*, J. Chem. Theory Comput. **11**, 276 (2015).

[29] H. Wu and F. Noé, *Optimal Estimation of Free Energies and Stationary Densities from Multiple Biased Simulations*, Multiscale Model. Simul. **12**, 25 (2014).

[30] H. Wu, A. Mey, E. Rosta, and F. Noé, *Statistically Optimal Analysis of State-Discretized Trajectory Data from Multiple Thermodynamic States*, J. Chem. Phys. **141**, 214106 (2014).

[31] A. Mey, H. Wu, and F. Noé, *xTRAM: Estimating Equilibrium Expectations from Time-Correlated Simulation Data at Multiple Thermodynamic States*, Phys. Rev. X **4**, 041018 (2014).

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

[33] A. K. Faradjian and R. Elber, *Computing Time Scales from Reaction Coordinates by Milestoning*, J. Chem. Phys. **120**, 10880 (2004).

[34] T. S. van Erp, D. Moroni, and P. G. Bolhuis, *A Novel Path Sampling Method for the Calculation of Rate Constants*, J. Chem. Phys. **118**, 7762 (2003).

[35] W.-N. Du and P. G. Bolhuis, *Adaptive Single Replica Multiple State Transition Interface Sampling*, J. Chem. Phys. **139**, 044105 (2013).
[36] W. F. Van Gunsteren and H. J. C. Berendsen, *A Leap-Frog Algorithm for Stochastic Dynamics*, Mol. Simul. **1**, 173 (1988).
[37] M. Tuckerman, B. J. Berne, and G. J. Martyna, *Reversible Multiple Time Scale Molecular Dynamics*, J. Chem. Phys. **97**, 1990 (1992).
[38] S. Sriraman, I. G. Kevrekidis, and G. Hummer, *Coarse Master Equation from Bayesian Analysis of Replica Molecular Dynamics Simulations*, J. Phys. Chem. B **109**, 6479 (2005).
[39] F. Noé, *Probability Distributions of Molecular Observables Computed from Markov Models*, J. Chem. Phys. **128**, 244103 (2008).
[40] K. Wang, J. D. Chodera, Y. Yang, and M. R. Shirts, *Identifying Ligand Binding Sites and Poses Using GPU-Accelerated Hamiltonian Replica Exchange Molecular Dynamics*, J. Comput.-Aided Mol. Des. **27**, 989 (2013).
[41] M. Souaille and B. Roux, *Extension to the Weighted Histogram Analysis Method: Combining Umbrella sampling with Free Energy Calculations*, Comput. Phys. Commun. **135**, 40 (2001).
[42] www.pyemma.org.
[43] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.6.011009 for IPython notebook tutorials.
[44] R. Zwanzig, *Nonlinear Generalized Langevin Equations*, J. Stat. Phys. **9**, 215 (1973).
[45] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, Appl. Opt. **25**, 3145 (1986).
[46] C. Schütte, Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules, http://publications.mi.fu-berlin.de/89/.
[47] C. Schütte and W. Huisinga, in *Proceedings of Equadiff 99: International Conference on Differential Equations: Berlin, 1999*, edited by B. Fiedler, K. Groger, and J. Sprekels (World Scientific, Singapore, 2000), Chap. 234, pp. 1247–1262.
[48] F. Wang and D. P. Landau, *Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States*, Phys. Rev. Lett. **86**, 2050 (2001).
[49] S. Trebst and M. Troyer, *Ensemble Optimization Techniques for Classical and Quantum Systems*, in Computer Simulations in Condensed Matter: From Materials to Chemical Biology, edited by M. Ferrario, G. Ciccotti, and K. Binder (Springer, New York, 2006), Vol. 1.
[50] C. H. Bennett, *Efficient Estimation of Free Energy Differences from Monte Carlo Data*, J. Comput. Phys. **22**, 245 (1976).
[51] A. M. Ferrenberg and R. H. Swendsen, *Optimized Monte Carlo Data Analysis*, Phys. Rev. Lett. **63**, 1195 (1989).
[52] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, *Multidimensional Free-Energy Calculations Using the Weighted Histogram Analysis Method*, J. Comput. Chem. **16**, 1339 (1995).
[53] Z. Tan, *On a Likelihood Approach for Monte Carlo Integration*, J. Am. Stat. Assoc. **99**, 1027 (2004).
[54] M. R. Shirts and J. D. Chodera, *Statistically Optimal Analysis of Samples from Multiple Equilibrium States*, J. Chem. Phys. **129**, 124105 (2008).
[55] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé, *"Estimation and Uncertainty of Reversible Markov Models"* (to be published).
[56] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, *The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method*, J. Comput. Chem. **13**, 1011 (1992).
[57] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap* (Chapman and Hall, New York, 1994).
[58] B. M. Pettitt and M. Karplus, *The Potential of Mean Force Surface for the Alanine Dipeptide in Aqueous Solution: A Theoretical Approach*, Chem. Phys. Lett. **121**, 194 (1985).
[59] A. G. Anderson and J. Hermans, *Microfolding: Conformational Probability Map for the Alanine Dipeptide in Water from Molecular Dynamics Simulations*, Proteins **3**, 262 (1988).
[60] D. J. Tobias and C. L. Brooks III, *Conformational Equilibrium in the Alanine Dipeptide in the Gas Phase and Aqueous Solution: A Comparison of Theoretical Results*, J. Phys. Chem. **96**, 3864 (1992).
[61] J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill, *Long-Time Protein Folding Dynamics from Short-Time Molecular Dynamics Simulations*, Multiscale Model. Simul. **5**, 1214 (2006).
[62] W.-N. Du, K. A. Marino, and P. G. Bolhuis, *Multiple State Transition Interface Sampling of Alanine Dipeptide in Explicit Solvent*, J. Chem. Phys. **135**, 145102 (2011).
[63] P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts, and V. S. Pande, *OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation*, J. Chem. Theory Comput. **9**, 461 (2013).
[64] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, *Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field*, Proteins **78**, 1950 (2010).
[65] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *Comparison of Simple Potential Functions for Simulating Liquid Water*, J. Chem. Phys. **79**, 926 (1983).
[66] P. J. Tummino and R. A. Copeland, *Residence Time of Receptor-Ligand Complexes and Its Effect on Biological Function*, Biochemistry **47**, 5481 (2008).
[67] G. R. Bowman, D. L. Ensign, and V. S. Pande, *Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models*, J. Chem. Theory Comput. **6**, 787 (2010).
[68] S. Doerr and G. De Fabritiis, *On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations*, J. Chem. Theory Comput. **10**, 2064 (2014).
[69] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations* (Springer, New York, 1992), Vol. 23.
[70] P. G. Hoel, S. C. Port, and C. J. Stone, *Introduction to Stochastic Processes* (Waveland Press, Long Grove, 1986).