# Markov models of molecular kinetics: Generation and validation

Jan-Hendrik Prinz,[1] Hao Wu,[1] Marco Sarich,[1] Bettina Keller,[1] Martin Senne,[1] Martin Held,[1] John D. Chodera,[2] Christof Schütte,[1] and Frank Noé[1,a)]
[1]*FU Berlin, Arnimallee 6, 14195 Berlin, Germany*
[2]*California Institute of Quantitative Biosciences (QB3), University of California, Berkeley, 260J Stanley Hall, Berkeley, California 94720, USA*

(Received 4 August 2010; accepted 22 February 2011; published online 4 May 2011)

Markov state models of molecular kinetics (MSMs), in which the long-time statistical dynamics of a molecule is approximated by a Markov chain on a discrete partition of configuration space, have seen widespread use in recent years. This approach has many appealing characteristics compared to straightforward molecular dynamics simulation and analysis, including the potential to mitigate the sampling problem by extracting long-time kinetic information from short trajectories and the ability to straightforwardly calculate expectation values and statistical uncertainties of various stationary and dynamical molecular observables. In this paper, we summarize the current state of the art in generation and validation of MSMs and give some important new results. We describe an upper bound for the approximation error made by modeling molecular dynamics with a MSM and we show that this error can be made arbitrarily small with surprisingly little effort. In contrast to previous practice, it becomes clear that the best MSM is not obtained by the most metastable discretization, but the MSM can be much improved if non-metastable states are introduced near the transition states. Moreover, we show that it is not necessary to resolve all slow processes by the state space partitioning, but individual dynamical processes of interest can be resolved separately. We also present an efficient estimator for reversible transition matrices and a robust test to validate that a MSM reproduces the kinetics of the molecular dynamics data.  [doi:10.1063/1.3565032]

## I. INTRODUCTION

Conformational transitions are essential to the function of proteins and nucleic acids. These transitions span large ranges of length scales, timescales, and complexity, and include folding,[1,2] complex conformational rearrangements between native protein substates,[3,4] and ligand binding.[5] Experiments have borne out the decade-old proposal that biomolecular kinetics are complex, often involving transitions between a multitude of long-lived, or "metastable" states on a range of different timescales.[6] With the ever increasing time resolution of ensemble kinetics experiments and the more recent maturation of sensitive single-molecule techniques in biophysics, experimental evidence supporting the near-universality of the existence of multiple metastable conformational substates and complex kinetics in biomolecules has continued to accumulate.[7–13] Enzyme kinetics has been shown to be modulated by interchanging conformational substates.[14] Protein folding experiments have found conformational heterogeneity, hidden intermediates, and the existence of parallel pathways.[15–20]

While laboratory experiments can resolve both fast kinetic processes and, in the case of single-molecule experiments, heterogeneity of some of these processes, the observations are always indirect; only spectroscopically resolvable probes can be monitored, and inherent signal-to-noise issues generally require sacrificing either time resolution (in single molecule experiments) or the ability to resolve heterogeneity of populations (in ensemble experiments). As a result, molecular dynamics (MD) simulations are becoming increasingly accepted as a tool to investigate structural details of molecular processes and relate them to experimentally resolved features.[21–23]

Traditionally, MD studies often involved "look and see" analyses of a few rare events via molecular movies. Although visually appealing, these analyses may be misleading as they do not supply the statistical relevance of such observations in the ensemble, and may miss rare but important events altogether. Another frequent approach, especially common in protein folding analyses, is to project the dynamics onto one or two user-defined order parameters (such as the root mean square distance [RMSD] to a single reference structure, radius of gyration, principal components, or selected distances or angles) with the notion that these order parameters allow the slow kinetics of the molecule to be resolved. While the ability to directly visualize the results of such projections on chemically intuitive order parameters is appealing, these projection techniques have been shown to disguise the true and often complex nature of the kinetics by artificially aggregating kinetically distinct structures and hiding barriers, thus creating a distorted and often overly simplistic picture of the kinetics.[24–26]

In order to resolve complex kinetic features such as low-populated intermediates, structurally similar metastable states, or structurally distinct parallel pathways, it is essential to employ analysis techniques that are sensitive to such details. While some reduction of high-dimensional biomolecular

a)Author to whom correspondence should be addressed. Electronic mail: frank.noe@fu-berlin.de.

 

dynamics, perhaps obtained from large quantities of MD trajectory data, is certainly necessary to generate a humanly understandable analysis, such reduction must be guided by the specific structural or kinetic information in these data, rather than by the subjectivity of the analyst. A natural approach toward modeling the kinetics of molecules is by first partitioning the conformation space into discrete states.[25–35] Although this step could still disguise information when lumping states that have an important distinction, it is clear that a "sufficiently fine" partitioning will be able to resolve "sufficient" detail.[36] Subsequent to partitioning, transition rates or probabilities between states can be calculated, either based on rate theories,[4,27,37] or based on transitions observed in MD trajectories.[24,26,34,35,38–40] The resulting models are often called transition networks, master equation models or Markov (state) models (MSM), where "Markovianity" means that the kinetics are modeled by a memoryless jump process between states.

This paper focuses on "Markov models" (abbreviated here by "MSM"[41]), which model the kinetics with an $n \times n$ transition probability matrix that contains the conditional probabilities that the system will, given that it is in one of its $n$ discrete substates, be found in any of these $n$ discrete substates a fixed time $\tau$ later. An essential feature of a MSM is that it abandons the view of the single trajectories and replaces it by an ensemble view of the dynamics.[42,43] Consider an experiment that traces the equilibrium dynamics of an ensemble of molecules starting from a distribution that is out of equilibrium, such as a laser-induced temperature-jump experiment.[44] Here the sequence of microscopic events occurring during the trajectory of any individual molecule may be of little relevance, as these individual trajectories all differ in microscopic detail. Instead, the relevant physical details are statistical properties of this ensemble: time-dependent averages of spectroscopically observable quantities, statistical probabilities quantifying with which conformationally similar states are populated at certain times and probabilities of how many trajectories follow similar pathways. All of these statistical properties can be easily computed from Markov models, as these models already encode the ensemble dynamics.[22,45] At the same time, because it is sometimes helpful in aiding the development of human intuition, individual realizations of almost arbitrary length can be easily obtained, simply by generating a random state sequence according to the MSM transition probabilities.

Because only *conditional* transition probabilities between discretized states are needed to construct a Markov model, the computational burden can be divided among many processors using loosely coupled parallelism, facilitating a "divide and conquer" approach. Trajectories used to estimate these transition probabilities only need to be long enough to reach *local* equilibrium within the discrete state, rather than exceed *global* equilibrium relaxation times that may be orders of magnitude longer. In other words, the dependency between simulation length and molecular timescales is largely lost; microsecond- or millisecond-timescale processes can be accurately modeled despite the model having been constructed from trajectories orders of magnitude shorter.[22,46] Moreover, assessment of the statistical uncertainty of the model can be used to adaptively guide model construction, reaching the desired statistical precision with much less total effort than would be necessary with a single long trajectory.[22,47,48]

Finally, computation of statistical quantities of interest from Markov models is straightforward, and includes:

- Time-independent properties such as the stationary, or equilibrium, probability of states or free energy differences between states.[22,25,49]
- Relaxation timescales that can be extracted from experimental kinetic measurements using various techniques such as laser-induced temperature jumps, fluorescence correlation spectroscopy, dynamic neutron scattering, or NMR.[22,25]
- Relaxation functions that can be measured with nonequilibrium perturbation experiments or correlation functions that can be obtained from fluctuations of single molecule equilibrium experiments.[22,45]
- Transition pathways and their probabilities, e.g., the ensemble of protein folding pathways.[22,50]
- Statistical uncertainties for all observables.[45,47,48,51]
- The precision and accuracy with which MSMs reproduce the true kinetics can be tested to verify the modeling error and remains small.[22,52]

In this paper we summarize the current state of the art of theory and methodology for MSM generation and validation, and fill some important methodological gaps.

Section II discusses the essential properties of the true full-dimensional continuous dynamics and how these properties may be affected by details of the simulation. Section III examines the effect of *discretizing* the state space to produce a discrete-state Markov chain approximation to the true dynamics. This is the key *numerical approximation* step, and we give a detailed analysis of the error incurred in doing so, as well as ways this error can be controlled. Finally, Section IV describes strategies for estimation of the Markov model with finite quantities of MD simulation data, the *statistical* step in building a Markov model. Sections II and III develop Markov models from a theoretical perspective, and practitioners may wish to skip directly to Sec. IV, where generation and validation of Markov models from actual trajectory data are discussed.

The main novelty of the present study is a detailed analysis of the discretization error (Sec. III), i.e., the effect of lumping state space points into discrete sets on the accuracy of reproducing quantities of the original continuous dynamics. We give quantitative upper bounds for the approximation error of the time evolution and the relaxation timescales of the slow dynamical processes. It is shown that this error can be made arbitrarily small with surprisingly little effort. In contrast to previous practice,[38–40,52] it is seen that the best MSM, in the sense of minimizing this discretization error, is not obtained by the most metastable discretization; instead the accuracy of the MSM can be improved if nonmetastable states are introduced near the transition states. Moreover, it is shown that it is not necessary to resolve all slow processes by the state space partitioning, but individual dynamical processes of interest can be described separately. These insights provide a

theoretical basis for the development of efficient adaptive discretization methods for MSMs.

Additionally, we provide a new estimator for transition matrices for reversible dynamics, i.e., Markov models that fulfill detailed balance, which is more efficient than the reversible estimators presented previously.[49,51,53] Detailed balance is expected for molecular processes taking place in thermal equilibrium[54] and using this property in the estimation of MSMs will generally enhance the model quality as unphysical models are excluded. Finally, we take up the topic of validating MSMs. Several past studies have attempted to find robust tests for the "Markovianity" of the true dynamics projected onto the discrete state space,[40,55] a concept which has been proven problematic both formally and practically. Here, we instead suggest a simple and robust direct test of the error of the model in reproducing the observed dynamics.

## II. ANALYSIS OF THE CONTINUOUS DYNAMICS

This section reviews the continuous dynamics of a molecular system in thermal equilibrium, and introduces the dynamical propagator, whose approximation is our primary concern. While this section is important for understanding the subsequent formal theory of discretization (Sec. III), practitioners wishing only to learn how to construct such models may skip directly to the discussion of Markov model estimation (Sec. IV).

### A. Continuous dynamics

A variety of simulation models that all yield the same stationary properties, but have different dynamical behaviors, are available to study a given molecular model. The choice of the dynamical model must therefore be guided by both a desire to mimic the relevant physics for the system of interest (such as whether the system is allowed to exchange energy with an external heat bath during the course of dynamical evolution), balanced with computational convenience (e.g., the use of a stochastic thermostat in place of explicitly simulating a large external reservoir).[56] Going into the details of these models is beyond the scope of the present study, and therefore we will simply state the minimal physical properties that we expect the dynamical model to obey.

Consider a state space $\Omega$ which contains all dynamical variables needed to describe the instantaneous state of the system. $\Omega$ may be discrete or continuous, and we treat the more general continuous case here. For molecular systems, $\Omega$ usually contains both positions and velocities of the species of interest and surrounding bath particles. $\mathbf{x}(t) \in \Omega$ will denote the dynamical process considered, which is continuous in space, and may be either time-continuous (for theoretical investigations) or time-discrete (when considering time-stepping schemes for computational purposes). For the rest of the paper, we will assume that $\mathbf{x}(t)$ has the following properties:

1. $\mathbf{x}(t)$ is a Markov process in the full state space $\Omega$, i.e., the instantaneous change of the system $(d\mathbf{x}(t)/dt$ in time-continuous dynamics and $\mathbf{x}(t+\Delta t)$ in time-discrete dynamics with time step $\Delta t)$, is calculated based on $\mathbf{x}(t)$ alone and does not require the previous history. As a result of Markovianity in $\Omega$, the transition probability density $p(\mathbf{x}, \mathbf{y}; \tau)$ is well defined:

$$p(\mathbf{x}, \mathbf{y}; \tau)\, d\mathbf{y} = \mathbb{P}[\mathbf{x}(t+\tau) \in \mathbf{y} + d\mathbf{y} \mid \mathbf{x}(t) = \mathbf{x}] \qquad \mathbf{x}, \mathbf{y} \in \Omega,\ \tau \in \mathbb{R}_{0+}, \tag{1}$$

i.e., the probability that a trajectory started at time $t$ from the point $\mathbf{x} \in \Omega$ will be in an infinitesimal region $d\mathbf{y}$ around a point $\mathbf{y} \in \Omega$ at time $t+\tau$. Such a transition probability density for the diffusion process in a one-dimensional potential is depicted in Fig. 1(b). When $p(\mathbf{x}, \mathbf{y}; \tau)$ is a smooth probability density the stochastic transition probability to a set $A \subseteq \Omega$ is also well defined and formally given by integrating the transition probability density over region $A$:

$$p(\mathbf{x}, A; \tau) = \mathbb{P}[\mathbf{x}(t+\tau) \in A | \mathbf{x}(t) = \mathbf{x}] = \int_{\mathbf{y} \in A} d\mathbf{y}\ p(\mathbf{x}, \mathbf{y}; \tau). \tag{2}$$

2. $\mathbf{x}(t)$ is ergodic, i.e., the space $\Omega$ does not have two or more subsets that are dynamically disconnected, and for $t \to \infty$ each state $\mathbf{x}$ will be visited infinitely often. The fraction of time that the system spends in any of its states during an infinitely long trajectory is given by its unique stationary density (invariant measure) $\mu(\mathbf{x}) : \Omega \to \mathbb{R}_{0+}$ that corresponds to the equilibrium probability density for some associated thermodynamic ensemble (e.g., NVT, NpT). For molecular dynamics at constant temperature $T$, the dynamics above yield a stationary density $\mu(\mathbf{x})$ that is a function of $T$, namely, the Boltzmann distribution

$$\mu(\mathbf{x}) = Z(\beta)^{-1} \exp\left(-\beta H(\mathbf{x})\right), \tag{3}$$

with Hamiltonian $H(\mathbf{x})$ and $\beta = 1/k_B T$ where $k_B$ is the Boltzmann constant and $k_B T$ is the thermal energy. $Z(\beta) = \int d\mathbf{x} \exp\left(-\beta H(\mathbf{x})\right)$ is the partition function. By means of illustration, Fig. 1(a) shows the stationary density $\mu(\mathbf{x})$ for a diffusion process on a potential with high barriers.

3. $\mathbf{x}(t)$ is reversible, i.e., $p(\mathbf{x}, \mathbf{y}; \tau)$ fulfills the condition of *detailed balance*:

$$\mu(\mathbf{x})\ p(\mathbf{x}, \mathbf{y}; \tau) = \mu(\mathbf{y})\ p(\mathbf{y}, \mathbf{x}; \tau), \tag{4}$$

i.e., in equilibrium, the fraction of systems transitioning from $\mathbf{x}$ to $\mathbf{y}$ per time is the same as the fraction of systems transitioning from $\mathbf{y}$ to $\mathbf{x}$. Note that this "reversibility" is a more general concept than the time-reversibility of the dynamical equations, e.g., encountered in Hamiltonian dynamics. For example, Brownian dynamics on some potential are reversible as they fulfill Eq. (4), but are not time-reversible in the same sense as Hamiltonian dynamics are, due to the stochasticity of individual realizations. Although detailed balance is not essential for the construction of Markov models, we will subsequently assume detailed

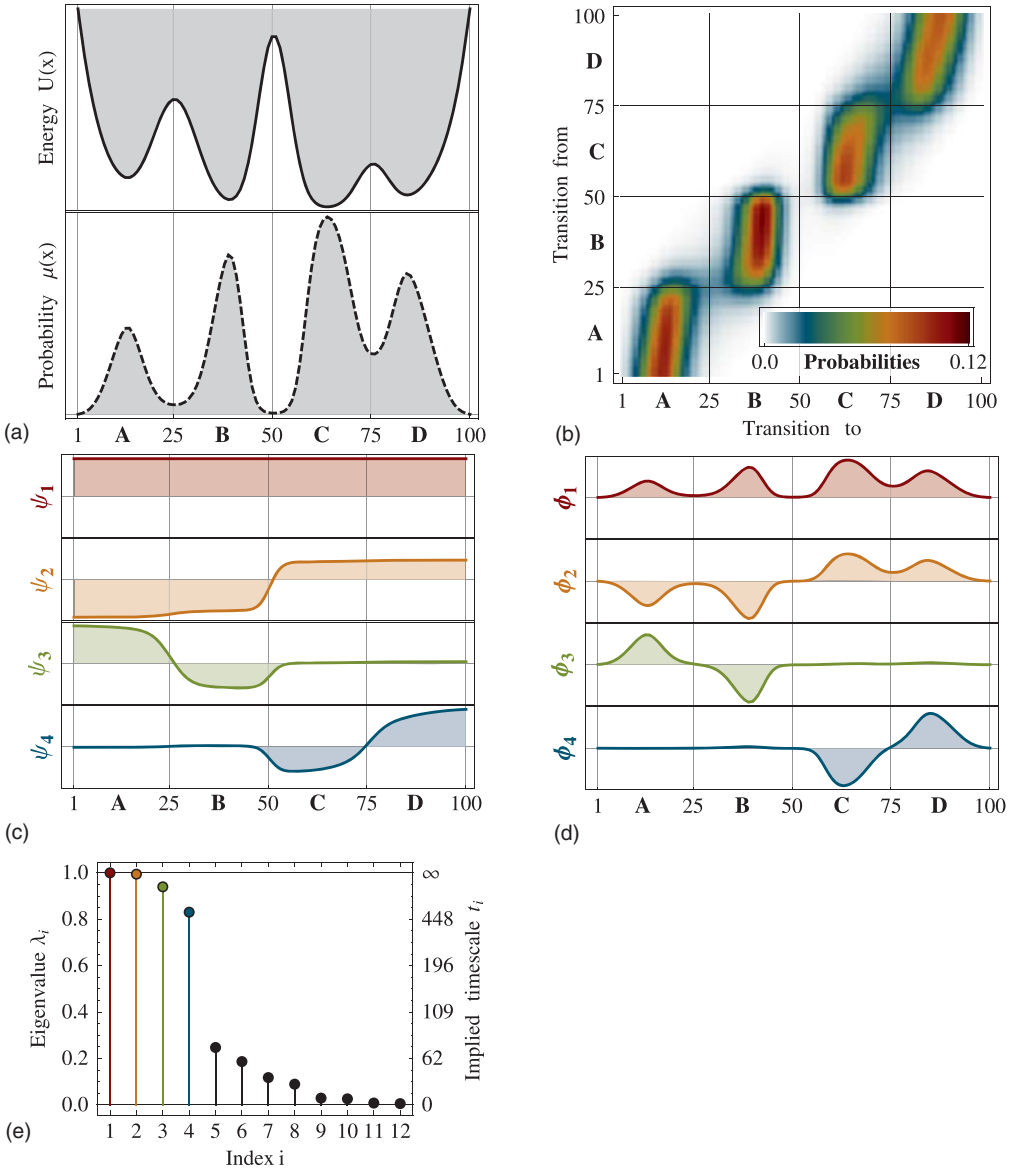


FIG. 1. (a) Potential energy function with four metastable states and corresponding stationary density $\mu(\mathbf{x})$. (b) Density plot of the transfer operator for a simple diffusion-in-potential dynamics defined on the range $\Omega = [1, 100]$ [see supplementary material (Ref. 65)], black and red indicates high transition probability, white zero transition probability. Of particular interest is the nearly block-diagonal structure, where the transition density is large within blocks allowing rapid transitions within metastable basins, and small or nearly zero for jumps between different metastable basins. (c) The four dominant eigenfunctions of the transfer operator, $\psi_1, \ldots, \psi_4$, which indicate the associated dynamical processes. The first eigenfunction is associated with the stationary process, the second to a transition between $A + B \leftrightarrow C + D$, and the third and fourth eigenfunction to transitions between $A \leftrightarrow B$ and $C \leftrightarrow D$, respectively. (d) The four dominant eigenfunctions of the transfer operator weighted with the stationary density, $\phi_1, \ldots, \phi_4$. (e) Eigenvalues of the transfer operator, the gap between the four metastable processes ($\lambda_i \approx 1$) and the fast processes is clearly visible.

balance as this allows much more profound analytical statements to be made. The rationale is that we expect detailed balance to be fulfilled in equilibrium molecular dynamics based on a simple physical argument: for dynamics that have no detailed balance, there exists a set of states which form a loop in state space which is traversed in one direction with higher probability than in the other direction. This means that one could design a machine which uses this preference of direction in order to produce work. However, a system in equilibrium is driven only by thermal energy, and conversion of pure thermal energy into work contradicts the second law of thermodynamics. Thus, equilibrium molecular dynamics must be reversible and fulfill detailed balance.

The above conditions do not place overly burdensome restrictions on the choice of dynamical model used to describe equilibrium dynamics. Most stochastic thermostats are consistent with the above assumptions, e.g., Andersen[57] (which can be employed with either massive or per-particle collisions, or coupled to only a subset of degrees of freedom), Hybrid Monte Carlo,[58] overdamped Langevin (also called Brownian or Smoluchowski) dynamics,[59,60] and stepwise-thermalized Hamiltonian dynamics.[40] When simulating solvated systems,

a weak friction or collision rate can be used; this can often be selected in a manner that is physically motivated by the heat conductivity of the material of interest and the system size.[57]

We note that the use of finite-timestep integrators for these models of dynamics can sometimes be problematic, as the phase space density sampled can differ from the density desired. Generally, integrators based on symplectic Hamiltonian integrators (such as velocity Verlet[61]) offer greater stability for our purposes.

While technically, a Markov model analysis can be constructed for any choice of dynamical model, it must be noted that several popular dynamical schemes violate the assumptions above, and using them means that one is (currently) doing so without a solid theoretical basis, e.g., regarding the boundedness of the discretization error analyzed in Sec. III below. For example, Nosé-Hoover and Berendsen are either not ergodic or do not generate the correct stationary distribution for the desired ensemble.[62] Energy-conserving Hamiltonian dynamics, even when considering a set of trajectories that are in initial contact with a heat bath, is not ergodic and therefore has no unique stationary distribution. While it is possible that future work will extend the present theoretical analysis to these and other models of dynamics, we currently advise practitioners to choose a model which unambiguously fulfills these conditions, yet provides physically reasonable kinetics.

### B. Transfer operator approach and the dominant spectrum

At this point we shift from focusing on the evolution of individual trajectories to the time evolution of an ensemble density. Consider an ensemble of molecular systems at a point in time $t$, distributed in state space $\Omega$ according to a probability density $p_t(\mathbf{x})$ that is different from the stationary density $\mu(\mathbf{x})$. If we now wait for some time $\tau$, the probability distribution of the ensemble will have changed because each system copy undergoes transitions in state space according to the transition probability density $p(\mathbf{x}, \mathbf{y}; \tau)$. The change of the probability density $p_t(\mathbf{x})$ to $p_{t+\tau}(\mathbf{x})$ can be described with the action of a continuous operator. From a physical point of view, it seems straightforward to define the *propagator* $\mathcal{Q}(\tau)$ as follows:

$$p_{t+\tau}(\mathbf{y}) = \mathcal{Q}(\tau) \circ p_t(\mathbf{y}) = \int_{\mathbf{x}\in\Omega} d\mathbf{x}\; p(\mathbf{x}, \mathbf{y}; \tau)\; p_t(\mathbf{x}). \quad (5)$$

Applying $\mathcal{Q}(\tau)$ to a probability density $p_t(\mathbf{x})$ will result in a modified probability density $p_{t+\tau}(\mathbf{x})$ that is more similar to the stationary density $\mu(\mathbf{x})$, to which the ensemble must relax after infinite time. An equivalent description is provided by the *transfer operator* $\mathcal{T}(\tau)$,[42] which has nicer properties from a mathematical point of view. $\mathcal{T}(\tau)$ is defined as[63]:

$$u_{t+\tau}(\mathbf{y}) = \mathcal{T}(\tau) \circ u_t(\mathbf{y}) = \frac{1}{\mu(\mathbf{y})} \int_{\mathbf{x}\in\Omega} d\mathbf{x}\; p(\mathbf{x}, \mathbf{y}; \tau)\; \mu(\mathbf{x})\; u_t(\mathbf{x}). \quad (6)$$

$\mathcal{T}(\tau)$ does not propagate probability densities, but instead functions $u_t(\mathbf{x})$ that differ from probability densities by a factor of the stationary density $\mu(\mathbf{x})$, i.e.,

$$p_t(\mathbf{x}) = \mu(\mathbf{x}) u_t(\mathbf{x}). \quad (7)$$

The relationship between the two densities and operators is shown in the scheme below:

$$\begin{array}{lll} p_t & \xrightarrow{\mathcal{Q}(\tau)} & p_{t+\tau} \\ \downarrow \cdot\mu^{-1} & & \uparrow \cdot\mu \\ u_t & \xrightarrow{\mathcal{T}(\tau)} & u_{t+\tau} \end{array} \quad \begin{array}{l} \text{probability densities} \\ \\ \text{densities in } \mu\text{-weighted space.} \end{array}$$

Alternatively to $\mathcal{Q}$ and $\mathcal{T}$ which describe the transport of densities exactly by a chosen time-discretization $\tau$, one could investigate the density transport with a time-continuous operator $\mathcal{L}$ called the *generator* which is the continuous basis of rate matrices that are frequently used in physical chemistry[31,64] and is related to the Fokker–Planck equation.[54] Here, we do not investigate $\mathcal{L}$ in detail, but describe some of its basic properties in the supplementary material.[65]

Equation (6) is a formal definition. When the particular kind of dynamics is known it can be written in a more specific form.[42] However, the general form (6) is sufficient for the present analysis. The continuous operators have the following general properties:

- Both $\mathcal{Q}(\tau)$ and $\mathcal{T}(\tau)$ fulfill the Chapman–Kolmogorov equation

$$p_{t+k\tau}(\mathbf{x}) = [\mathcal{Q}(\tau)]^k \circ p_t(\mathbf{x}), \quad (8)$$

$$u_{t+k\tau}(\mathbf{x}) = [\mathcal{T}(\tau)]^k \circ u_t(\mathbf{x}), \quad (9)$$

  where $[\mathcal{T}(\tau)]^k$ refers to the $k$-fold application of the operator, i.e., $\mathcal{Q}(\tau)$ and $\mathcal{T}(\tau)$ can be used to propagate the evolution of the dynamics to arbitrarily long times $t + k\tau$.
- $\mathcal{Q}(\tau)$ has eigenfunctions $\phi_i(\mathbf{x})$ and associated eigenvalues $\lambda_i$ [see Figs. 1(c) and 1(e)]:

$$\mathcal{Q}(\tau) \circ \phi_i(\mathbf{x}) = \lambda_i \phi_i(\mathbf{x}), \quad (10)$$

  while $\mathcal{T}(\tau)$ has eigenfunctions $\psi_i(\mathbf{x})$ with the same corresponding eigenvalues:

$$\mathcal{T}(\tau) \circ \psi_i(\mathbf{x}) = \lambda_i \psi_i(\mathbf{x}). \quad (11)$$

- When the dynamics are reversible, all eigenvalues $\lambda_i$ are real-valued and lie in the interval $-1 < \lambda_i \leq 1$[42]. Moreover, the two types of eigenfunctions are related by a factor of the stationary density $\mu(\mathbf{x})$:

$$\phi_i(\mathbf{x}) = \mu(\mathbf{x}) \psi_i(\mathbf{x}), \quad (12)$$

  and their lengths are defined by the normalization condition that the scalar product (see Table I) is unity for all corresponding eigenfunctions: $\langle \phi_i, \psi_i \rangle = 1$ for all $i = 1, \ldots, m$ (see Table I for definition of scalar product). Due to reversibility, noncorresponding eigenfunctions are orthogonal: $\langle \phi_i, \psi_j \rangle = 0$ for all $i \neq j$. When $\mathcal{T}(\tau)$ is approximated by a reversible transition matrix on a discrete state space, $\phi_i(\mathbf{x})$ and $\psi_i(\mathbf{x})$ are approximated by the left and right eigenvectors of that transition matrix, respectively [compare Figs. 1(c) and 1(d)].

TABLE I. Important symbols.

| Symbol | Meaning |
|---|---|
| $\Omega$ | Continuous state space (positions and momenta) |
| $\mathbf{x}(t)$ | Continuous state in $\Omega$ (positions and momenta) at time $t$ |
| $\mu(\mathbf{x})$ | Continuous (in state space) stationary density of $\mathbf{x}$. |
| $p(\mathbf{x})$ | Continuous (in state space) probability density. |
| $\tau$ | Lag time, time resolution of the model. |
| $p(\mathbf{x}, \mathbf{y}; \tau)$ | Transition probability density to $\mathbf{y} \in \Omega$ after time $\tau$ given the system in $\mathbf{x} \in \Omega$. |
| $\mathcal{T}(\tau)$ | Transfer operator, propagates the continuous dynamics for a time $\tau$. |
| $m$ | Number of dominant eigenfunctions/eigenvalues considered. |
| $\psi(\mathbf{x})$ | Eigenfunctions of $\mathcal{T}(\tau)$. |
| $\phi(\mathbf{x})$ | Density-weighted eigenfunctions of $\mathcal{T}(\tau)$. |
| $\chi_i(\mathbf{x})$ | Degree of membership of $\mathbf{x}$ to discrete state $i$. |
| $S_1, \ldots, S_n$ | Discrete sets which partition state space $\Omega$. |
| $\mu_i(\mathbf{x})$ | Local stationary density restricted to discrete state $i$. |
| $\langle f, g \rangle$ | Scalar product $\langle f, g \rangle = \int f(\mathbf{x})\, g(\mathbf{x})\, d\mathbf{x}$. |
| $\langle f, g \rangle_\mu$ | Weighted scalar product $\langle f, g \rangle_\mu = \int \mu(\mathbf{x})\, f(\mathbf{x})\, g(\mathbf{x})\, d\mathbf{x}$. |
| $n$ | Number of discrete states. |
| $\boldsymbol{\pi}$ | Discrete stationary density in $\mathbb{R}^n$. |
| $\mathbf{p}(t)$ | Discrete probability vector in $\mathbb{R}^n$ at time $t$. |
| $\mathbf{C}(\tau)$ | Transition count matrix (row-dominant) in $\mathbb{R}^{n\times n}$, elements $c_{ij}(\tau)$ count the number of $i \to j$ transitions during lag time $\tau$. |
| $\mathbf{T}(\tau)$ | Discrete transition matrix (row-stochastic) in $\mathbb{R}^{n\times n}$, elements $T_{ij}(\tau)$ give the $i \to j$ transition probability during lag time $\tau$. |
| $\hat{\mathbf{T}}(\tau)$ | Estimate of $\mathbf{T}(\tau)$ from trajectory data. |
| $\boldsymbol{\psi}_i$ | $i$th right eigenvector of $\mathbf{T}(\tau)$ in $\mathbb{R}^n$. |
| $\boldsymbol{\phi}_i$ | $i$th left eigenvector of $\mathbf{T}(\tau)$ in $\mathbb{R}^n$. |

- Since both operators are continuous, they possess a continuous spectrum of eigenvalues. By convention, we only distinguish a finite number of $m$ *dominant* eigenvalue/eigenfunction pairs and sort them by nonascending eigenvalue, i.e., $\lambda_1 = 1 > \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_m$, while the remainder of the spectrum is confined within in a ball of radius $r \leq \lambda_m$ centered on 0.
  There is one eigenvalue $\lambda_1 = 1$ that has the greatest norm (i.e., it is simple and dominant). The associated eigenfunction corresponds to the stationary distribution $\mu(\mathbf{x})$ [see Fig. 1(d), top]:

$$\mathcal{Q}(\tau) \circ \mu(\mathbf{x}) = \mu(\mathbf{x}) = \phi_1(\mathbf{x}), \tag{13}$$

  and the corresponding eigenfunction of $\mathcal{T}(\tau)$ is a constant function on all state space $\Omega$ [see Fig. 1(c), top]:

$$\mathcal{T}(\tau) \circ \mathbf{1} = \mathbf{1} = \psi_1(\mathbf{x}), \tag{14}$$

  due to the relationship $\phi_1(\mathbf{x}) = \mu(\mathbf{x})\psi_1(\mathbf{x}) = \mu(\mathbf{x})$.

To see the significance of the other eigenvalue/eigenfunction pairs, we exploit that the dynamics can be decomposed exactly into a superposition of $m$ individual slow dynamical processes and the remaining fast processes. For $\mathcal{T}(\tau)$, this yields

$$u_{t+k\tau}(\mathbf{x}) = \mathcal{T}_{\text{slow}}(k\tau) \circ u_t(\mathbf{x}) + \mathcal{T}_{\text{fast}}(k\tau) \circ u_t(\mathbf{x}), \tag{15}$$

$$= \sum_{i=1}^{m} \lambda_i^k \, \langle u_t, \phi_i \rangle \, \psi_i(\mathbf{x}) + \mathcal{T}_{\text{fast}}(k\tau) \circ u_t(\mathbf{x}), \tag{16}$$

$$= \sum_{i=1}^{m} \lambda_i^k \, \langle u_t, \psi_i \rangle_\mu \, \psi_i(\mathbf{x}) + \mathcal{T}_{\text{fast}}(k\tau) \circ u_t(\mathbf{x}). \tag{17}$$

Here, $\mathcal{T}_{\text{slow}}$ is the *dominant*, or slowly decaying part consisting of the $m$ slowest processes with $\lambda_i \geq \lambda_m$, while $\mathcal{T}_{\text{fast}}$ contains all (infinitely many) fast processes with $\lambda_i < \lambda_m$ that are usually not of interest. The weighted scalar product appearing above is defined in Table I. This decomposition requires that subspaces $\mathcal{T}_{\text{slow}}$ and $\mathcal{T}_{\text{fast}}$ are orthogonal, which is a consequence of detailed balance. This decomposition permits a compelling physical interpretation: the slow dynamics are a superposition of dynamical processes, each of which can be associated with one eigenfunction $\psi_i$ (or $\phi_i$) and a corresponding eigenvalue $\lambda_i$ [see Figs. 1(c)–1(e)]. These processes decay faster with increasing time index $k$. In the long-time limit where $k \to \infty$, only the first term with $\lambda_1 = 1$ remains, recovering the stationary distribution $\phi_1(\mathbf{x}) = \mu(\mathbf{x})$. All other terms correspond to processes with eigenvalues $\lambda_i < 1$ and decay over time, thus the associated eigenfunctions correspond processes that decay under the action of the dynamics and represent the dynamical rearrangements taking place while the ensemble relaxes toward the equilibrium distribution. The closer $\lambda_i$ is to 1, the slower the corresponding process decays; conversely, the closer it is to 0, the faster.

Thus the $\lambda_i$ for $i = 2, \ldots, m$ each corresponds to a physical timescale, indicating how quickly the process decays or transports density toward equilibrium [see Fig. 1(e)]:

$$t_i = -\frac{\tau}{\ln \lambda_i}, \tag{18}$$

which is often called the $i$th implied timescale.[40] Thus, Eq. (15) can be rewritten in terms of implied timescales as:

$$u_{t+k\tau}(\mathbf{x}) = 1 + \sum_{i=2}^{m} \exp\left(-\frac{k\tau}{t_i}\right) \langle u_t, \psi_i \rangle_\mu \, \psi_i(\mathbf{x}) + \mathcal{T}_{\text{fast}}(k\tau) \circ u_t(\mathbf{x}). \tag{19}$$

This implies that when there are gaps among the first $m$ eigenvalues, the system has dynamical processes acting simultaneously on different timescales. For example, a system with two-state kinetics would have $\lambda_1 = 1$, $\lambda_2 \approx 1$ and $\lambda_3 \ll \lambda_2$ ($t_3 \ll t_2$), while a system with a clear involvement of an additional kinetic intermediate would have $\lambda_3 \sim \lambda_2$ ($t_3 \sim t_2$).

In Fig. 1, the second process, $\psi_2$, corresponds to the slow ($\lambda_2 = 0.9944$) exchange between basins A + B and basins C + D, as reflected by the opposite signs of the elements of $\psi_2$ in these regions [Fig. 1(c)]. The next-slowest processes are the A↔B transition and then the C↔D transition, while the subsequent eigenvalues are clearly separated from the dominant

spectrum and correspond to much faster local diffusion processes. The three slowest processes effectively partition the dynamics into four metastable states corresponding to basins A, B, C, and D, which are indicated by the different sign structures of the eigenfunctions [Fig. 1(c)]. The metastable states can be calculated from the eigenfunction structure, e.g., using the Perron Cluster Cluster Analysis (PCCA) method.[30,38]

Of special interest is the slowest relaxation time, $t_2$. This timescale identifies the *worst case* global equilibration or decorrelation time of the system; no structural observable can relax more slowly than this timescale. Thus, if one desires to calculate an expectation value $\mathbb{E}[a]$ of an observable $a(\mathbf{x})$ which has a non-negligible overlap with the second eigenfunction, $\langle a, \psi_2 \rangle > 0$, a straightforward single-run MD trajectory would need to be many times $t_2$ in length in order to compute an unbiased estimate of $\mathbb{E}[a]$.

## III. DISCRETIZATION AND DISCRETIZATION ERROR

While molecular dynamics in full continuous state space $\Omega$ is Markovian by construction, the term Markov *model* is due to the fact that in practice, state space must be somehow discretized in order to obtain a computationally tractable description of the dynamics. The Markov model then consists of the partitioning of state space used together with the transition matrix modeling the jump process of the observed trajectory projected onto these discrete states. However, this jump process (Fig. 2) is no longer Markovian, as the information where the continuous process would be within the local discrete state is lost in the course of discretization. Modeling the long-time statistics of this jump process with a Markov process is an approximation, i.e., it involves a *discretization error*. In the current section, this discretization error is analyzed and it is shown what needs to be done in order to keep it small.

The discretization error is a *systematic error* of a Markov model since it causes a deterministic deviation of the Markov model dynamics from the true dynamics that persists even when the statistical error is excluded by excessive sampling. In order to focus on this effect alone, it is assumed in this section that the statistical estimation error is zero, i.e., transition probabilities between discrete states can be calculated exactly. The results suggest that the discretization error of a Markov model can be made small enough for the MSM to be useful in accurately describing the relaxation kinetics, even for very large and complex molecular systems.

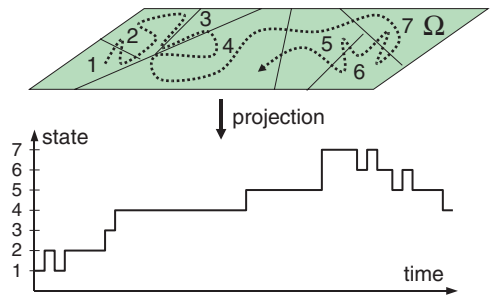


FIG. 2. Scheme: The true continuous dynamics (dashed line) is projected onto the discrete state space. MSMs approximate the resulting jump process by a Markov jump process.

In practical use, the Markov model is not obtained by actually discretizing the continuous propagator. Rather, one defines a discretization of state space and then estimates the corresponding discretized transfer operator from a finite quantity of simulation data, such as several long or many short MD trajectories that transition between the discrete states. The statistical estimation error involved in this estimation will be discussed in Sec. IV; the current section focuses only on the approximation error due to discretization of the transfer operator.

### A. Discretization of state space

Here we consider a discretization of state space $\Omega$ into $n$ sets. In practice, this discretization is often a simple partition with sharp boundaries, but in some cases it may be desirable to discretize $\Omega$ into fuzzy sets.[66] We can describe both cases by defining membership functions $\chi_i(\mathbf{x})$ that quantify the probability of point $\mathbf{x}$ to belong to set $i$ (Ref. 43) which have the property $\sum_{i=1}^{n} \chi_i(\mathbf{x}) = 1$. In the present study we will concentrate on a *crisp* partitioning with step functions:

$$\chi_i(\mathbf{x}) = \chi_i^{\text{crisp}}(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in S_i \\ 0 & \mathbf{x} \notin S_i \end{cases}. \tag{20}$$

Here we have used $n$ sets $S = \{S_1, \ldots, S_n\}$ which entirely partition state space ($\bigcup_{i=1}^{n} S_i = \Omega$) and have no overlap ($S_i \cap S_j = \emptyset$ for all $i \neq j$). A typical example of such a crisp partitioning is a Voronoi tessellation,[67] where one defines $n$ centers $\bar{\mathbf{x}}_i$, $i = 1, \ldots, n$, and set $S_i$ is the union of all points $\mathbf{x} \in \Omega$ which are closer to $\bar{\mathbf{x}}_i$ than to any other center using some distance metric [illustrated in Figs. 3(b) and 3(c)]. Note that such a discretization may be restricted to some subset of the degrees of freedom, e.g., in MD one often ignores velocities and solvent coordinates when discretizing.

The stationary probability $\pi_i$ to be in set $i$ is then given in terms of the full stationary density as:

$$\pi_i = \int_{\mathbf{x} \in S_i} d\mathbf{x}\, \mu(\mathbf{x}),$$

and the local stationary density $\mu_i(\mathbf{x})$ restricted to set $i$ [see Fig. 3(b)] is given by

$$\mu_i(\mathbf{x}) = \begin{cases} \frac{\mu(\mathbf{x})}{\pi_i} & \mathbf{x} \in S_i \\ 0 & \mathbf{x} \notin S_i \end{cases}. \tag{21}$$

These properties are local, i.e., they do not require information about the full state space.

### B. Transition matrix

Together with the discretization, the Markov model is defined by the row-stochastic transition probability matrix, $\mathbf{T}(\tau) \in \mathbb{R}^{n \times n}$, which is the discrete approximation of the transfer operator described in Sec. II B via

$$T_{ij}(\tau) = \frac{\langle \chi_j, (\mathcal{T}(\tau) \circ \chi_i) \rangle_\mu}{\langle \chi_i, \chi_i \rangle_\mu}.$$

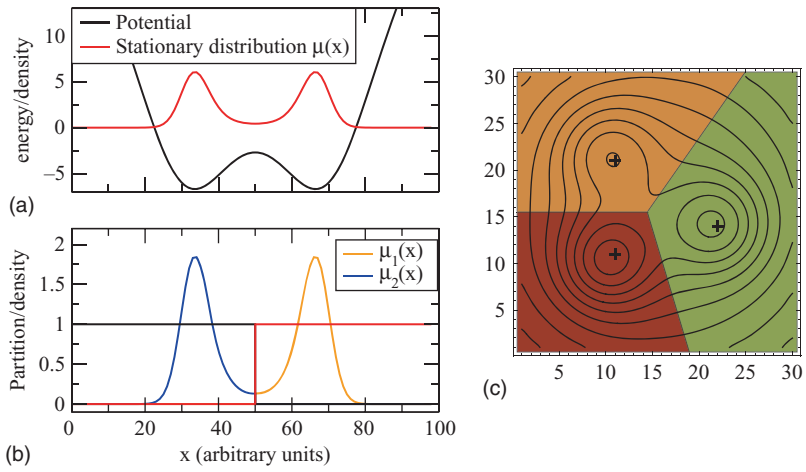


FIG. 3. Crisp state space discretization illustrated on a one-dimensional two-well and a two-dimensional three-well potential [see supplementary material for details of potential and dynamics (Ref. 65)]. (a) Two-well potential (black) and stationary distribution $\mu(\mathbf{x})$ (red). (b) Characteristic functions $v_1(\mathbf{x}) = \chi_1(\mathbf{x})$, $v_2(\mathbf{x}) = \chi_2(\mathbf{x})$ (black and red). This discretization has the corresponding local densities $\mu_1(\mathbf{x})$, $\mu_2(\mathbf{x})$ (blue and yellow), see Eq. (21). (c) Three-well potential (black contours indicate the isopotential lines) with a crisp partitioning into three states using a Voronoi partition with the centers denoted (+).

Physically, each element $T_{ij}(\tau)$ represents the time-stationary probability to find the system in state $j$ at time $t + \tau$ given that it was in state $i$ at time $t$. By definition of the conditional probability, this is equal to

$$T_{ij}(\tau) = \mathbb{P}[\mathbf{x}(t+\tau) \in S_j \mid \mathbf{x}(t) \in S_i] \tag{22}$$

$$= \frac{\mathbb{P}[\mathbf{x}(t+\tau) \in S_j \ \cap \ \mathbf{x}(t) \in S_i]}{\mathbb{P}[\mathbf{x}(t) \in S_i]} \tag{23}$$

$$= \frac{\int_{\mathbf{x}\in S_i} d\mathbf{x}\, \mu_i(\mathbf{x})\, p(\mathbf{x}, S_j; \tau)}{\int_{\mathbf{x}\in S_i} d\mathbf{x}\, \mu_i(\mathbf{x})}, \tag{24}$$

where we have used Eq. (2). Note that in this case the integrals run over individual sets and only need the local equilibrium distributions $\mu_i(\mathbf{x})$ as weights. This is a very powerful feature: in order to estimate transition probabilities, we do not need any information about the global equilibrium distribution of the system, and the dynamical information needed extends only over time $\tau$. In principle, the full dynamical information of the discretized system can be obtained by initiating trajectories of length $\tau$ out of each state $i$ as long as we draw the starting points of these simulations from a local equilibrium density $\mu_i(\mathbf{x})$.[42,43,68]

The transition matrix can also be written in terms of correlation functions:[40]

$$T_{ij}(\tau) = \frac{\mathbb{E}[\chi_i(\mathbf{x}(t))\, \chi_j(\mathbf{x}(t+\tau))]}{\mathbb{E}[\chi_i(\mathbf{x}(t))]} = \frac{c_{ij}^{\mathrm{corr}}(\tau)}{\pi_i}, \tag{25}$$

where the unconditional transition probability $c_{ij}^{\mathrm{corr}}(\tau) = \pi_i T_{ij}(\tau)$ is an equilibrium time correlation function which is normalized such that $\sum_{i,j} c_{ij}^{\mathrm{corr}}(\tau) = 1$. For dynamics fulfilling detailed balance, the correlation matrix is symmetric $[c_{ij}^{\mathrm{corr}}(\tau) = c_{ji}^{\mathrm{corr}}(\tau)]$.

Since the transition matrix $\mathbf{T}(\tau)$ is a discretization of the transfer operator $\mathcal{T}$ (Refs. 36, 42, and 63; Sec. II B), we can relate the functions that are transported by $\mathcal{T}$ [functions $u_t$ in Eq. (6)] to column vectors that are multiplied to the matrix from the right while the probability densities $p_t$ [Eq. (7)] correspond to row vectors that are multiplied to the matrix from the left. Suppose that $\mathbf{p}(t) \in \mathbb{R}^n$ is a column vector whose elements denote the probability, or population, to be within any set $j \in \{1, \ldots, n\}$ at time $t$. After time $\tau$, the probabilities will have changed according to

$$p_j(t+\tau) = \sum_{i=1}^{n} p_i(t) T_{ij}(\tau), \tag{26}$$

or in matrix form

$$\mathbf{p}^T(t+\tau) = \mathbf{p}^T(t)\, \mathbf{T}(\tau). \tag{27}$$

Note that an alternative convention often used in the literature is to write $\mathbf{T}(\tau)$ as a column-stochastic matrix, obtained by taking the transpose of the row-stochastic transition matrix defined here.

The stationary probabilities of discrete states, $\pi_i$, yield the unique discrete stationary distribution of $\mathbf{T}$ for any $\tau$:

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \mathbf{T}(\tau). \tag{28}$$

All equations encountered so far are free of approximation. We wish now to model the system kinetics of long times by *approximating* the true dynamics with a Markov chain on the space of $n$ states. Using $\mathbf{T}(\tau)$ as a Markov model *predicts* that for later times, $t + k\tau$, the probability distribution will evolve as

$$\mathbf{p}^T(t+k\tau) \approx \mathbf{p}^T(t)\mathbf{T}^k(\tau), \tag{29}$$

which can only approximate the true distribution $\mathbf{p}(t + k\tau)$ that would have been produced by the continuous transfer operator, as Eq. (29) is the result of a state space discretization. The discretization error involved in this approximation thus depends on how this discretization is chosen and is analyzed in detail below. A description alternative to that of transition matrices quite common in chemical physics is using rate matrices and master equations.[31,64,69,70] The relationship between this and the current approach is discussed in the Supplementary Information.[65]

### C. Discretization error and non-Markovianity

The Markov model $\mathbf{T}(\tau)$ is indeed a model in the sense that it only approximates the long-time dynamics based on a discretization of state space, making the dynamical equation (29) approximate. Here we analyze the approximation quality of Markov models in detail and obtain a description that reveals which properties the state space discretization and the lag time $\tau$ must fulfill in order to obtain a good model.

Previous works have mainly discussed the quality of a Markov model in terms of its "Markovianity" and introduced tests of Markovianity of the process $\mathbf{x}(t)$ projected onto the discrete state space. The space-continuous dynamics $\mathbf{x}(t)$ described in Sec. II is, by definition, Markovian in full state space $\Omega$ and it can thus be exactly described by a linear operator, such as the transfer operator $\mathcal{T}(\tau)$ defined in Eq. (6). The problem lies in the fact that by performing a state space discretization, continuous states $\mathbf{x} \in \Omega$ are grouped into discrete states $s_i$ (Sec. III A), thus "erasing" information of the exact location within these states and "projecting" a continuous trajectory $\mathbf{x}(t)$ onto a discrete trajectory $s(t) = s(\mathbf{x}(t))$. This jump process, $s(t)$, is *not Markovian*, but Markov models attempt to approximate $s(t)$ with a Markov chain.

Thus, non-Markovianity occurs when the full state space resolution is reduced by mapping the continuous dynamics onto a reduced space. In Markov models of molecular dynamics, this reduction consists usually of both, neglect of degrees of freedom and discretization of the resolved degrees of freedom. Markov models typically only use atom positions, thus the velocities are projected out.[38,39] So far, Markov models have also neglected solvent degrees of freedom and have only used the solute coordinates,[22,39] and the effect of this was studied in detail.[71] Indeed, it may be necessary to incorporate solvent coordinates in situations where the solvent molecules are involved in slow processes that are not easily detected in the solute coordinates.[72] Often, Markov models are also based on distance metrics that only involve a subset of the solute atoms, such as RMSD between heavy atom or alpha carbon coordinates,[22,39,49] or backbone dihedral angles.[31,38] Possibly the strongest approximation is caused by clustering or lumping sets of coordinates in the selected coordinate subspace into discrete states.[22,31,39,49,73] Formally, all of these operations aggregate sets of points in continuous state space $\Omega$ into discrete states, and the question to be addressed is what is the magnitude of the discretization error caused by treating the non-Markovian jump process between these sets as a Markov chain.

Consider the diffusive dynamics model depicted in Fig. 4(a) and let us follow the evolution of the dynamics given that we start from a local equilibrium in basin $A$ [Fig. 4(b)], either with the exact dynamics, or with the Markov model dynamics on the discrete state space $A$ and $B$. After a step $\tau$, both dynamics have transported a fraction of 0.1 of the

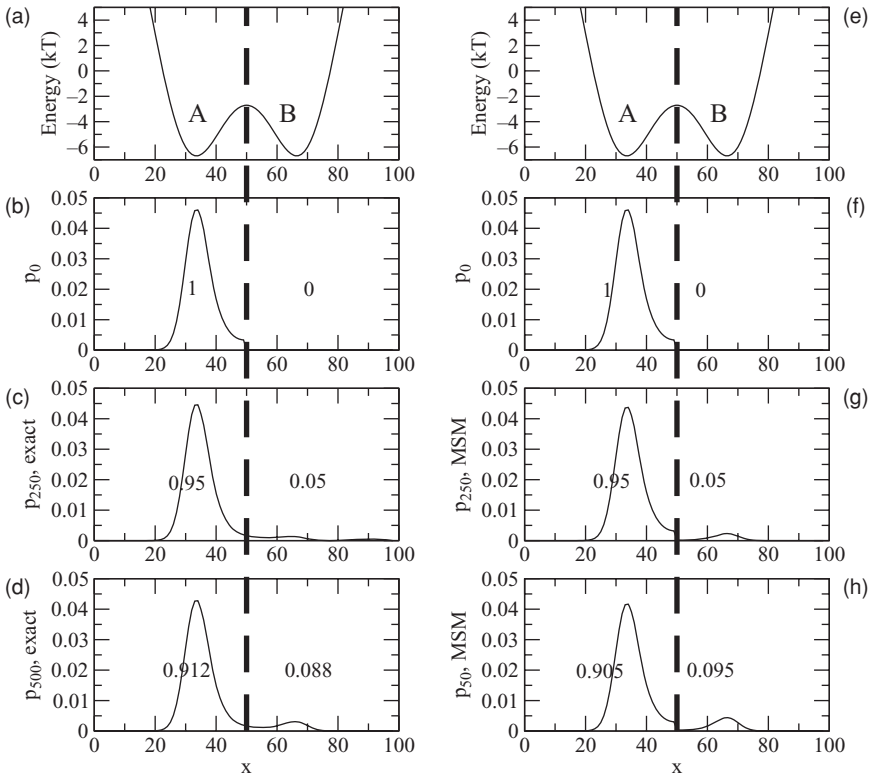


FIG. 4. Illustration of the discretization error by comparing the dynamics of the diffusion in a double-well potential (a, e) [see supplementary material for setup (Ref. 65)] at time steps 0 (b), 250 (c), 500 (d) with the predictions of a Markov model parameterized with a too short lag time $\tau = 250$ at the corresponding times 0 (f), 250 (g), 500 (h). (b, c, d) show the true density $p_t(\mathbf{x})$ (solid black line) and the probabilities associated with the two discrete states left and right of the dashed line. The numbers in (f, g, h) are the discrete state probabilities $p_i(t + k\tau)$ predicted by the Markov model while the solid black line shows the hypothetical density $p_i(t + k\tau)\mu_i(\mathbf{x})$ that inherently assumed by the Markov model by using the discrete state probabilities to correspondingly weight the local stationary densities.

ensemble to $B$. The true dynamics resolves the fact that much of this is still located near the saddle point [Fig. 4(c)]. The Markov model cannot resolve local densities within its discrete states, which is equivalent to assuming that for the next step the ensemble has already equilibrated within the discrete state [Fig. 4(g)]. This difference affects the discrete state (basin) probabilities at time $2\tau$: in the true dynamics, a significant part of the 0.1 fraction can cross back to $A$ as it is still near the saddle point, while this is not the case in the Markov model where the 0.1 fraction is assumed to be relaxed to states mostly around the minimum [compare Fig. 4(d) and (h)]. As a result, the probability to be in state $B$ is higher in the Markov model than in the true dynamics. The difference between the Markov model dynamics and the true dynamics is thus a result of discretization, because the discretized model can no longer resolve deviations from local equilibrium density $\mu_i(\mathbf{x})$ within the discrete state.

This picture suggests the discretization error to have two properties: (a) the finer the discretization, the smaller the discretization error is, and (b) when increasing the coarse-graining time, or time resolution, of our model, $\tau$, the errors for any fixed point in time $t$ should diminish, because we have less often made the approximation of imposing local equilibrium.

### D. Quantifying the discretization error

Figure 4 also suggests a practical measure to quantify the discretization error. Densities, eigenfunctions, or any other function $f(\mathbf{x})$ of the continuous state $\mathbf{x}$, are approximated through the discretization $S_1, \ldots, S_n$. Let $Q$ be the projection onto the discretization basis which produces this approximation $\hat{f}(\mathbf{x})$:

$$\hat{f}(\mathbf{x}) = Qf(\mathbf{x}) = \sum_{i=1}^{n} a_i \chi_i(\mathbf{x}), \tag{30}$$

where the coefficients are given by the projection weighted by the probability of each state:

$$a_i = \frac{\langle f, \chi_i \rangle_\mu}{\langle \mathbf{1}, \chi_i \rangle_\mu} = \frac{\int_{S_i} d\mathbf{x}\, \mu(\mathbf{x}) f(\mathbf{x})}{\int_{S_i} d\mathbf{x}\, \mu(\mathbf{x})}. \tag{31}$$

In the case of a crisp partitioning of state space, functions $f(\mathbf{x})$ are approximated by step functions that are constant within the discrete states. The approximation error that is caused by the discretization is then defined as the $\mu$-weighted Euclidean norm $\|\cdot\|_{\mu,2}$ of the difference between discretized and original function:

$$\delta_f \equiv \left\| f(\mathbf{x}) - \hat{f}(\mathbf{x}) \right\|_{\mu,2} = \left( \int_\Omega d\mathbf{x}\, \mu(\mathbf{x}) \left( f(\mathbf{x}) - \hat{f}(\mathbf{x}) \right)^2 \right)^{1/2}. \tag{32}$$

When the projection $Q$ is applied to probability densities $p(\mathbf{x})$, it effectively measures how much density is in each of the discrete states and replaces the true density within each state with a local stationary density of corresponding amplitude. This projection allows the comparison between true and Markov model dynamics to be made exactly as suggested by Fig. 4: in both cases we start with an arbitrary initial density projected onto discrete states, $Qp_0(\mathbf{x})$, then transport this density either with the true or with the Markov model dynamics for some time $k\tau$. Subsequently, the densities are again projected onto discrete states by $Q$ and then compared:

- The true dynamics transports the initial density $Qp_0(\mathbf{x})$ to $[\mathcal{T}(\tau)]^k Qp_0(\mathbf{x})$.
- The Markov model dynamics transports the initial density $Qp_0(\mathbf{x})$ to $Q\mathcal{T}(\tau)Qp_0(\mathbf{x})$ in one step and to $Q[\mathcal{T}(\tau)Q]^k p_0(\mathbf{x})$ in $k$ steps using the identity for projections $Q \circ Q = Q$.
- Projecting both densities to local densities and comparing yields the difference

$$\begin{aligned} \epsilon(k) &= \left\| Q[\mathcal{T}(\tau)]^k Qp_0(\mathbf{x}) - Q[\mathcal{T}(\tau)Q]^k p_0(\mathbf{x}) \right\|_{\mu,2} \\ &= \left\| [Q[\mathcal{T}(\tau)]^k Q - Q[\mathcal{T}(\tau)Q]^k] p_0(\mathbf{x}) \right\|_{\mu,2}. \end{aligned} \tag{33}$$

In order to remove the dependency on the initial distribution $p_0(\mathbf{x})$, we assume the worst case: the maximum possible value of $\epsilon(k)$ among all possible $p_0(\mathbf{x})$ is given by the operator-2-norm of the error matrix $[Q[\mathcal{T}(\tau)]^k Q - Q[\mathcal{T}(\tau)Q]^k]$, where $\|A\|_{\mu,2} \equiv \max_{\|x\|=1} \|Ax\|_{\mu,2}$ (Ref. 74), thus the Markov model error is defined as:

$$E(k) := \left\| Q[\mathcal{T}(\tau)]^k Q - Q[\mathcal{T}(\tau)Q]^k \right\|_{\mu,2}, \tag{34}$$

which measures the maximum possible difference between the true probability density at time $k\tau$ and the probability density predicted by the Markov model at the same time.

In order to quantify this error, we declare our explicit interest in the $m$ slowest processes with eigenvalues $1 = \lambda_1 < \lambda_2 \leq \lambda_3 \leq \cdots \leq \lambda_m$. Generally, $m \leq n$, i.e., we are interested in less processes than the number of $n$ eigenvectors that a Markov model with $n$ states has. We define the following quantities:

- $\delta_i := \|\psi_i(\mathbf{x}) - Q\psi_i(\mathbf{x})\|_{\mu,2}$ is the *eigenfunction approximation error*, quantifying the error of approximating the true continuous eigenfunctions of the transfer operator, $\psi_i$ (see Fig. 5 for illustration), for all $i \in \{1, \ldots, m\}$.
- $\delta := \max_i \delta_i$ is the largest approximation error among these first $m$ eigenfunctions.
- $\eta(\tau) := \frac{\lambda_{m+1}(\tau)}{\lambda_2(\tau)}$ is the *spectral error*, the error due to neglecting the fast subspace of the transfer operator, which decays to 0 with increasing lag time: $\lim_{\tau\to\infty} \eta(\tau) = 0$.

The Markov model error $E(k)$ can be proven[36] to be bounded from above by the following expression:

$$E(k) \leq \min\{2, [m\delta + \eta(\tau)]\,[a(\delta) + b(\tau)]\}\, \lambda_2^k \tag{35}$$

with

$$a(\delta) = \sqrt{m}(k-1)\delta \tag{36}$$

$$b(\tau) = \frac{\eta(\tau)}{1-\eta(\tau)}(1 - \eta(\tau)^{k-1}). \tag{37}$$

This implies two observations:

1. For long times, the overall error decays to zero with $\lambda_2^k$, where $0 < \lambda_2 < 1$, thus the stationary distribution (recovered as $k \to \infty$) is always correctly modeled, even if the kinetics are badly approximated.
2. The error during the kinetically interesting timescales consists of a product whose terms contain separately the discretization error and spectral error. Thus, the overall error can be diminished by choosing a fine discretization (where fine means it needs to closely approximate the slow eigenfunctions), and using a large lag time $\tau$.

Depending on the ratio $\lambda_{m+1}(\tau)/\lambda_2(\tau)$, the decay of the spectral error $\eta(\tau)$ with $\tau$ might be slow. It is thus interesting to consider a special case of the discretization where $n = m$ and $\delta = 0$. This would be achieved by a Markov model that uses a fuzzy partition with membership functions derived from the first $m$ eigenfunctions $\psi_1, \ldots, \psi_m$.[68] From a more practical point of view, this situation can be approached by using a Markov model with $n >> m$ states located such that they discretize the first $m$ eigenfunctions with a vanishing discretization error, and then declaring that we are *only* interested in these $m$ slowest relaxation processes. In this case we do not need to rely on the upper bound of the error from Eq. (35), but directly obtain the important result $E(k) = 0$.

In other words, a Markov model can approximate the kinetics of slow processes *arbitrarily well*, provided the discretization can be made sufficiently fine or improved in a way that continues to minimize the eigenfunction approximation error. This observation can be rationalized by Eq. (15) which shows that the dynamics of the transfer operator can be exactly decomposed into a superposition of slow and fast processes.

An important consequence of the $\delta$-dependence of the error is that the best partition is not necessarily metastable. Previous work[38–40,52] has focused on the construction of partitions with high metastability [defined as the trace of the transition matrix $\mathbf{T}(\tau)$], e.g., the two-state partition shown in [see second row in Fig. 5]. This approach was based on the idea that the discretized dynamics must be approximately Markovian if the system remained in each partition sufficiently long to approximately lose memory.[39] While it can be shown that if a system has $m$ metastable sets with $\lambda_m \gg \lambda_{m+1}$, then the most metastable partition into $n = m$ sets also minimizes the discretization error,[36] the expression for the discretization error given here has two further profound ramifications. First, even in the case where there exists a strong separation of timescales so the system has clearly $m$ metastable sets, the discretization error can be reduced *even further* by splitting the metastable partition into a total of $n > m$ sets which are not metastable. And second, even in the *absence* of a strong separation of timescales, the discretization error can be made arbitrarily small by making the partition finer, especially in transition regions, where the eigenfunctions change most rapidly [see second row in 5(b)].

Figure 6 illustrates the Markov model discretization error on a two-dimensional three-well example where two slow processes are of interest. The left panels show a metastable partition with three sets. As seen in Fig. 6(d), the discretization errors $|\psi_2 - Q\psi_2|(\mathbf{x})$ and $|\psi_3 - Q\psi_3|(\mathbf{x})$ are large near the transition regions, where the eigenfunctions $\psi_2(\mathbf{x})$ and $\psi_3(\mathbf{x})$ change rapidly, leading to a large discretization error. Using a random partition (Fig. 6, third column) makes the situation worse, but increasing the number of states reduces the discretization error (Fig. 6, fourth column), thereby increasing the quality of the Markov model. When states are chosen such as to well approximate the eigenfunctions, a very small error can be obtained with few sets (Fig. 6, second column)

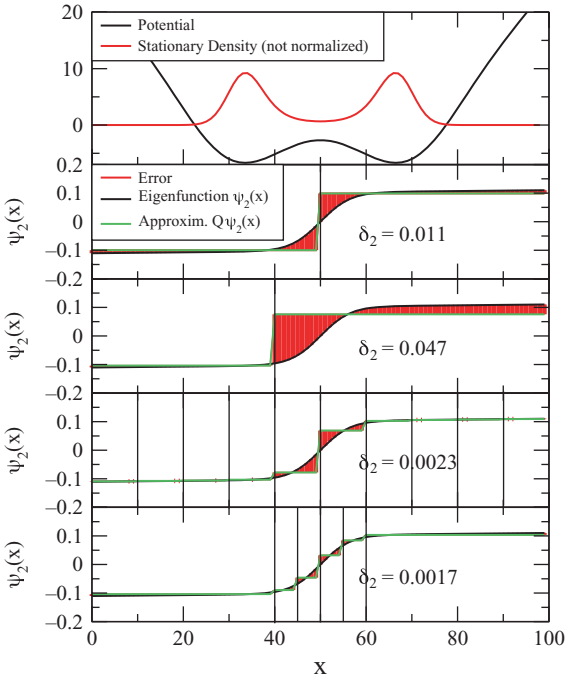


FIG. 5. Illustration of the eigenfunction approximation error $\delta_2$ on the slow transition in the diffusion in a double well (top, black line). The slowest eigenfunction is shown in the lower four panels (black), along with the step approximations (green) of the partitions (vertical black lines) at $x = 50$; $x = 40$; $x = 10, 20, \ldots, 80, 90$; and $x = 40, 45, 50, 55, 60$. The eigenfunction approximation error $\delta_2$ is shown as red area and its norm is printed.

These results suggest that an adaptive discretization algorithm may be constructed which minimizes the $E(k)$ error. Such an algorithm could iteratively modify the definitions of discretization sets as suggested previously,[39] but instead of maximizing metastability it would minimize the $E(k)$ error which can be evaluated by comparing eigenvector approximations on a coarse discretization compared to a reference evaluated on a finer discretization.[36]

One of the most intriguing insights from both Eq. (15) and the results of the discretization error is that, if for a given system only the slowest dynamical processes are of interest, it is sufficient to discretize the state space such that the first few eigenvectors are well represented (in terms of small approximation errors $\delta_i$). For example, if one is interested in processes on timescales $t^*$ or slower, then the number $m$ of eigenfunctions that need to be resolved is equal to the number of implied timescales with $t_i \geq t^*$. Due to the perfect decoupling of processes for reversible dynamics in the eigenfunctions [see Eqs. (16) and (17)], no gap after these first $m$ timescales of interest is needed. Note that the quality of the

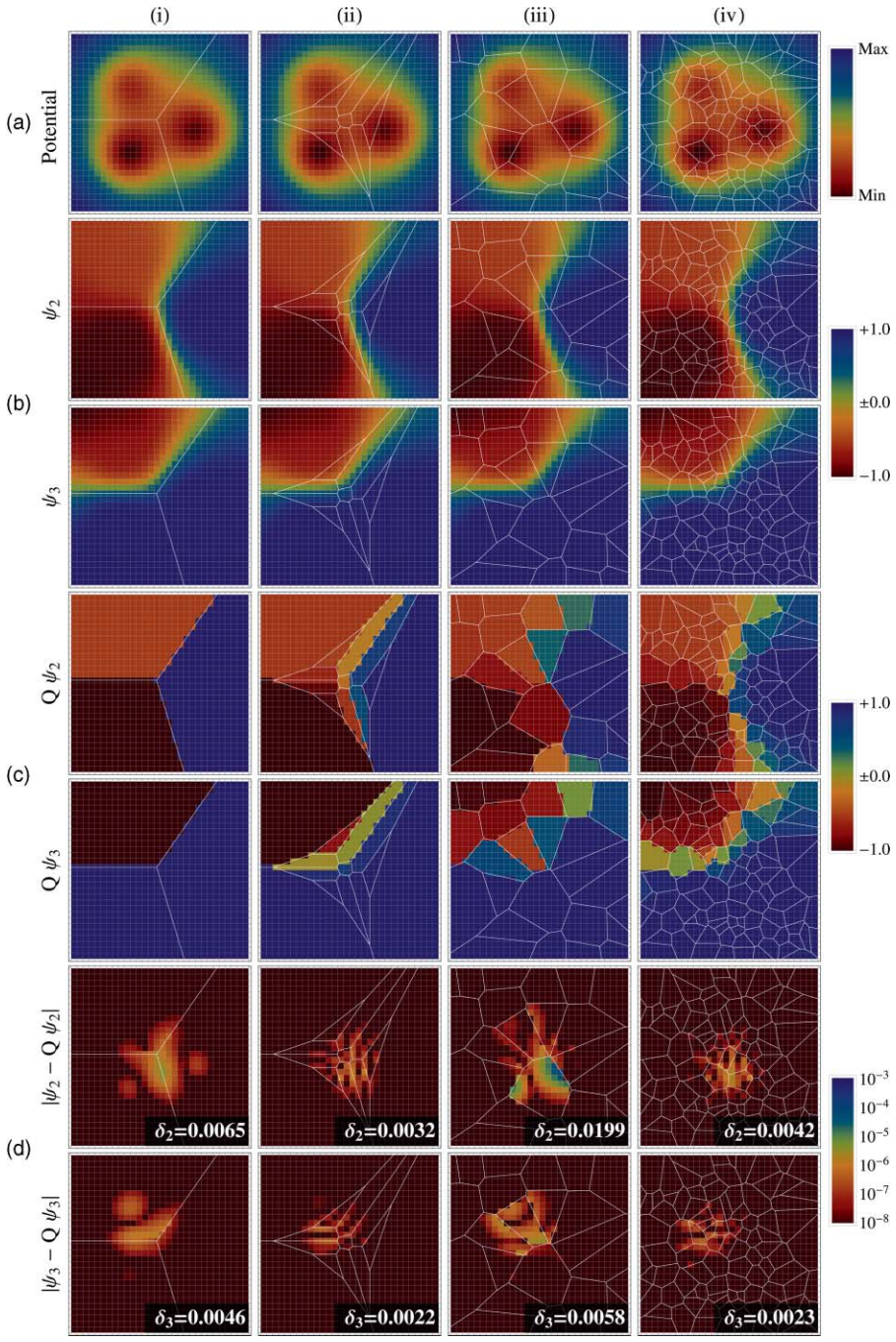


FIG. 6. Illustration of the eigenfunction approximation errors $\delta_2$ and $\delta_3$ on the two slowest processes in a two-dimensional three-well diffusion model [see supplementary material for model details (Ref. 65)]. The columns from left to right show different state space discretizations with white lines as state boundaries: (i) three states with maximum metastability, (ii) the metastable states were further subdivided manually into 13 states to better resolve the transition region, resulting in a partition where no individual state is metastable, (iii)/(iv) Voronoi partition using 25/100 randomly chosen centers, respectively. (a) Potential, (b) the exact eigenfunctions of the slow processes, $\psi_2(\mathbf{x})$ and $\psi_3(\mathbf{x})$, (c) the approximation of eigenfunctions with discrete states, $Q\psi_2(\mathbf{x})$ and $Q\psi_3(\mathbf{x})$, (d) approximation errors $|\psi_2 - Q\psi_2|(\mathbf{x})$ and $|\psi_3 - Q\psi_3|(\mathbf{x})$. The error norms $\delta_2$ and $\delta_3$ are given.

Markov model does not depend on the dimensionality of the simulated system, i.e., the number of atoms. Thus, if only the slowest process of the system is of interest (such as the folding process in a two-state folder), only a one-dimensional parameter, namely, the level of the dominant eigenfunction, needs to be approximated with the clustering, even if the system is huge. This opens a way to discretize state spaces of very large molecular systems.

### E. Approximation of eigenvalues and timescales by Markov models

One of the most interesting kinetic properties of molecular systems are the intrinsic timescales of the system. They can be both experimentally accessed via relaxation or correlation functions that are measurable with various spectroscopic techniques,[44,75–77] but can also be directly calculated

from the Markov model eigenvalues as implied timescales [Eq. (18)]. Thus, we investigate the question how well the dominant eigenvalues $\lambda_i$ are approximated by the Markov model, which immediately results in estimates for how accurately a Markov model may reproduce the implied timescales of the original dynamics. Consider the first $m$ eigenvalues of $\mathcal{T}(\tau)$, $1 = \lambda_1(\tau) > \lambda_2(\tau) \geq \ldots \geq \lambda_m(\tau)$, and let $1 = \hat{\lambda}_1(\tau) > \hat{\lambda}_2(\tau) \geq \ldots \geq \hat{\lambda}_m(\tau)$ denote the associated eigenvalues of the Markov model $\mathbf{T}(\tau)$. The rigorous mathematical estimate from Ref. 78 states that

$$\max_{j=1,\ldots,m} |\lambda_j(\tau) - \hat{\lambda}_j(\tau)| \leq (m-1)\,\lambda_2(\tau)\,\delta^2, \tag{38}$$

where $\delta$ is again the maximum discretization error of the first $m$ eigenfunctions, showing that the eigenvalues are well reproduced when the discretization well traces these eigenfunctions. In particular, if we are only interested in the eigenvalue of the slowest process, $\lambda_2(\tau)$, which is often experimentally reported via the slowest relaxation time of the system, $t_2$, the following estimate of the approximation error can be given:

$$\frac{|\lambda_2(\tau) - \hat{\lambda}_2(\tau)|}{|\lambda_2(\tau)|} \leq \delta_2^2. \tag{39}$$

As $\lambda_2(\tau)$ corresponds to a slow process, we can make the restriction $\lambda_2(\tau) > 0$. Moreover, the discretization error of Markov models based on full partitions of state space is such that the eigenvalues are always underestimated,[78] thus $\lambda_2(\tau) - \hat{\lambda}_2(\tau) > 0$. Using Eq. (18), we thus obtain the estimate for the discretization error of the largest implied timescale and the corresponding smallest implied rate, $k_2 = t_2^{-1}$:

$$\hat{t}_2^{-1} - t_2^{-1} = \hat{k}_2 - k_2 \leq -\tau^{-1} \ln(1 - \delta_2^2), \tag{40}$$

which implies that for either $\delta_2 \to 0^+$ or $\tau \to \infty$, the error in the largest implied timescale or smallest implied rate tends to zero. Moreover, since $\lambda_2(\tau) \to 0$ for $\tau \to \infty$, this is also true for the other processes:

$$\lim_{\tau \to \infty} \frac{|\lambda_j(\tau) - \hat{\lambda}_j(\tau)|}{|\lambda_j(\tau)|} = 0, \tag{41}$$

and also

$$\lim_{\delta \to 0} \frac{|\lambda_j(\tau) - \hat{\lambda}_j(\tau)|}{|\lambda_j(\tau)|} = 0, \tag{42}$$

which means that the error of the implied timescales also vanishes for either sufficiently long lag times $\tau$ or for sufficiently fine discretization. This fact has been empirically observed in many previous studies,[22,31,38–40,45,73] but can now be understood in detail in terms of the discretization error.

It is worth noting that observing convergence of the slowest implied timescales in $\tau$ is not a test of Markovianity. While Markovian dynamics implies constancy of implied timescales in $\tau$ (Refs. 38 and 40), the reverse is not true and would require the eigenvectors to be constant as well. However, observing the lag time-dependence of the implied timescales is a useful approach to choose a lag time $\tau$ at which $\mathbf{T}(\tau)$ shall be calculated, but this model needs to be validated subsequently (see Sec. IV F).

Figure 7 shows the slowest implied timescale $t_2$ for the diffusion in a two-well potential (see Fig. 5) with discretizations shown in Fig. 5. The two-state partition at $x = 50$ requires a lag time of $\approx 2000$ steps in order to reach an error of $< 3\%$ with respect to the true implied timescale, which is somewhat slower than $t_2$ itself. When the two-state partition is distorted by shifting the discretization border to $x = 40$, this quality is not reached before the process itself has relaxed. Thus, in this system two states are not sufficient to build a Markov model that is at the same time precise and has a time resolution good enough to trace the decay of the slowest process. By using more states and particularly a finer discretization of the transition region, the same approximation quality is obtained with only $\tau \approx 1500$ (blue) and $\tau \approx 500$ steps (green).

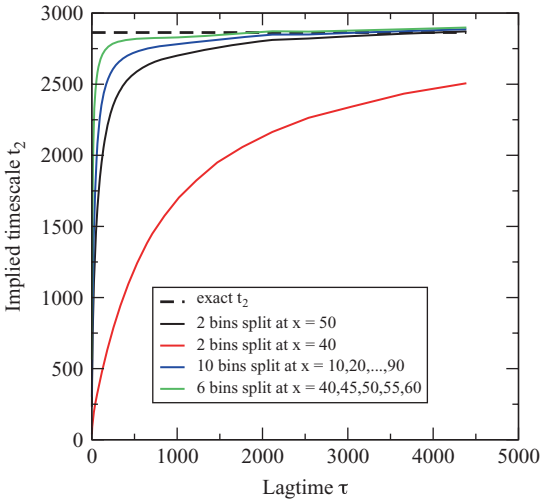


FIG. 7. Convergence of the slowest implied timescale $t_2 = -\tau / \ln \lambda_2(\tau)$ of the diffusion in a double-well potential depending on the MSM discretization. The metastable partition (black, solid) has greater error than nonmetastable partitions (blue, green) with more states that better trace the change of the slow eigenfunction near the transition state.

Figure 8 shows the two slowest implied timescales $t_2$, $t_3$ for the diffusion in a two-dimensional three-well potential with discretizations shown in Fig. 6(a). The metastable three-state partition requires a lag time of $\approx 1000$ steps in order to reach an error of $< 3\%$ with respect to the true implied timescale, which is somewhat shorter than the slow but longer than the fast timescale, while refining the discretization near the transition states achieves the same precision with $\tau \approx 200$ using only 12 states. A k-means clustering with $k = 25$ is worse than the metastable partition, as some clusters cross over the transition region and fail to resolve the slow eigenfunctions. Increasing the number of clusters to $k = 100$ improves the result significantly, but is still worse than the 12 states that have been manually chosen so as to well resolve the transition region. This suggests that excellent MSMs could be built with rather few states when an adaptive algorithm that more finely partitions the transition region is employed.

### F. Discretization methods for molecules

Macromolecular systems generally possess configuration spaces of such high dimension that grid-based methods for partitioning space become impractical. However, in many macromolecular systems such as proteins, the region, over

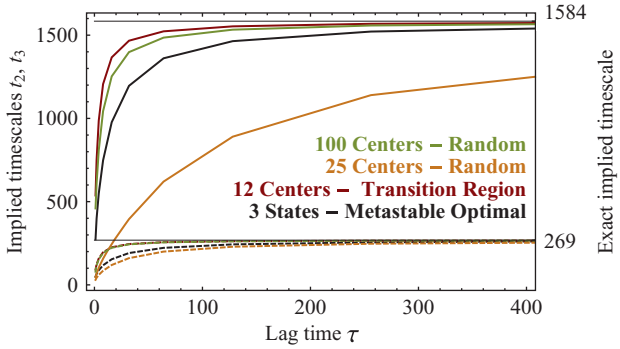


FIG. 8. Implied timescales for the two slowest processes in the two-dimensional three-well diffusion model [see Fig. 6(a) for potential and supplementary material for details (Ref. 65)]. The colors black, red, yellow, green correspond to the four choices of discrete states shown in columns 1 to 4 of Fig. 6. A fine discretization of the transition region clearly gives the best approximation to the timescales at small lag times.

which the configurational probability density is significant, defines a low-dimensional (but potentially highly nonlinear) subspace.[79] As a result, data-driven methods, where a clustering of conformations sampled by some form of molecular simulation defines the partitioning of this low-dimensional subspace, are both attractive and practical. Various combinations of distance metrics and clustering methods have been proposed. Distance metrics include Euclidean distance in backbone coordinates[22] or RMSD.[39,49] Clustering methods include manual clustering,[52] k-means clustering,[22] k-centers clustering,[49] density-based clustering,[80,81] and adaptive clustering approaches.[39] Approaches to directly discretize certain coordinates, such as the rotameric states[31,51] or the hydrogen-bond patterns,[38,73] were also made.

In the present paper we do not attempt to argue for or against a particular metric or clustering method. In theory, any metric that is able to partition *full* state space $\Omega$ more finely when the number of clusters is increased permits reduction of the eigenfunction approximation error to zero. In practice, such a metric is difficult to design and thus one often measures structural differences on a subset of coordinates (e.g., solute coordinates). In this case, the approximation of the eigenfunctions will maintain an error that must be compensated by increasing the lag time $\tau$. In practice, it is important that the metric is selected such that the molecular events under investigation can be resolved. For example, backbone

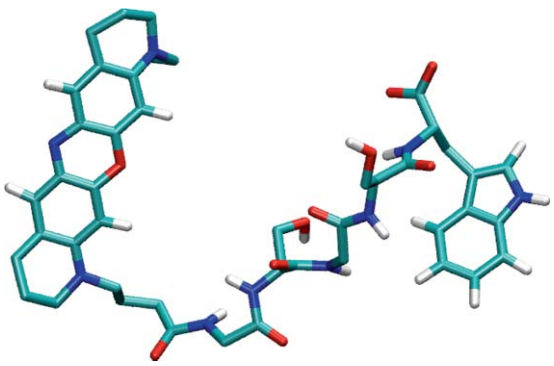

FIG. 9. Structure of the MR121-GSGS-W peptide.

rotamer angles are a poor metric when large side-chains are involved. Root mean square distance of entire protein structures might overwhelm small changes at individual degrees of freedom and therefore be unsuitable when detailed changes in the binding pocket of an enzyme are to be resolved.

However, it is interesting to see that MSMs are robust with respect to changes of the metric and the clustering method, within a significant range. This is illustrated by the following analysis: the MR121-GSGS-W peptide simulation (see Fig. 9 and supplementary material for simulation setup[65]) was clustered with a Voronoi partition in an all-atom RMSD metric, using three different methods to determine the cluster centers:

1. *k-centers clustering*.[82]
2. *Regular time clustering*: Cluster generators were picked at regular time intervals along the trajectory.
3. *Regular space clustering*: Cluster generators were chosen to be approximately equally separated in RMSD: a minimal distance $d_{\min}$ was fixed, the first trajectory frame was used as the first cluster center, then the trajectory was iterated and a frame was accepted as cluster center when its RMSD to all existing cluster centers was equal or greater than $d_{\min}$.

As the equilibrium simulation used to estimate the Markov model is a factor of 100 times longer than the slowest implied timescale we consider the estimated transition matrix from this trajectory as almost free of statistical error. The statistical issues in the estimation problem are discussed in detail in Sec. IV below.

Figure 10 shows that for all clustering methods and numbers of clusters (10, 100, or 1000) used, the slowest implied timescales converge to approximately the same values $t_2 \approx 25$ ns and $t_3 \approx 10$ ns at long lag times $\tau$. All clustering methods produce MSMs which converge for smaller values of $\tau$ when increasingly many clusters are used. This tendency can be assumed as long as sufficient statistics are available. When the number of clusters gets too large for a given amount of simulation data, statistical issues need to be considered (see Sec. IV). The differences in MSM quality between the different clustering methods for similar numbers of clusters are small. Interestingly, $k$-centers and regular space clustering do not outperform the simple method of picking cluster centers at regular time intervals. The three methods used here are relatively fast, all having a time complexity of $O(kN)$, with $k$ being the number of clusters and $N$ the number of frames in the trajectory. It is unclear whether using computationally more expensive clustering methods are able to significantly benefit the MSM construction at this stage. Our findings suggest that MSM construction from trajectory data is robust as long as sufficient data are available and a sufficient number of states are used.

## IV. ESTIMATION FROM DATA AND VALIDATION

In almost all practical cases, the transition matrix $\mathbf{T}(\tau)$ is not obtained by directly discretizing the continuous transfer operator but rather by estimation from a finite quantity of simulation data. This includes a statistical error component

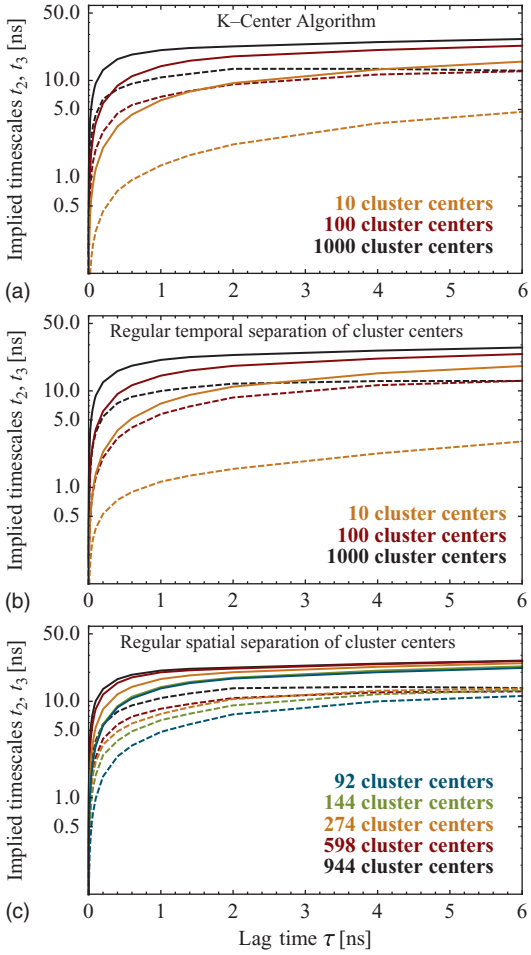


FIG. 10. Lag time dependent implied timescales $t_2$ (solid lines) and $t_3$ (dashed lines) of the slowest processes computed from Markov models of MD simulation data of the MR121-GSGSW peptide. (a) $k$-centers clustering, (b) cluster centers chosen from frames at fixed time intervals, (c) cluster centers are chosen so as to have a certain minimal distance to all others. Independent of the clustering method chosen, increasing the number of clusters enhances the convergence of implied timescales.

into the overall error in modeling the true dynamics with Markov models which will be discussed in this section. Here we assume that a state space discretization (either crisp or fuzzy) has been defined and that a trajectory is mapped onto this discrete space. We then answer the question how to estimate a Markov model based on such trajectory data.

Note that while in Sec. III we have studied only the discretization error of the Markov model without consideration of statistical issues (i.e., it was assumed the transition matrix could be computed exactly), this section only studies statistical issues without consideration of the discretization error (i.e., the discrete dynamics is now assumed to be perfectly Markovian).

### A. From trajectories to count matrix

Consider one trajectory generated at equilibrium conditions with $N$ configurations stored at a fixed time interval $\Delta t$:

$$X = [\mathbf{x}_1 = \mathbf{x}(t = 0), \mathbf{x}_2 = \mathbf{x}(t = \Delta t), \ldots, \mathbf{x}_N = \mathbf{x}(t = (N-1)\Delta t)] \tag{43}$$

and consider that a state space discretization has been defined such that each structure can be assigned to one discrete state $\mathbf{x}_k \in S_i \rightarrow s_k = i$, and the trajectory information can be simply stored as the sequence $s_1, \ldots, s_N$ of discrete states.

We also assume that $\mathbf{x}_1$ was drawn from the equilibrium density pertaining to state $s_1$, $\mu_{s_1}(\mathbf{x})$. The correct starting distribution can be generated from a global estimate of the stationary density (e.g., generated by well-converged metadynamics[83] or replica-exchange[84] simulations), or more efficiently by launching trajectories from short local equilibrium dynamics restricted to the starting density $\mu_i(\mathbf{x})$.[85] Note that in the limit of very small discrete states, this problem vanishes as $\mu_i(\mathbf{x})$ can then be well approximated by a step function.[22,65]

We can now define the discrete state count matrix $\mathbf{C}^{\text{obs}}(\tau) = [c_{ij}^{\text{obs}}(\tau)]$ at lag time $\tau$, where $\tau$ now needs to be an integer multiple of the available data resolution $\Delta t$:

$$c_{ij}^{\text{obs}}(\tau) = c_{ij}^{\text{obs}}(l\Delta t) = \sum_{k=1}^{N-l} \chi_i(\mathbf{x}_k)\chi_j(\mathbf{x}_{k+l}) \tag{44}$$

$$= |\ \{k \in \{1, \ldots, N-l\} \mid s_k = i \wedge s_{k+l} = j\}\ |, \tag{45}$$

which provides an estimator of the correlation matrix defined in Eq. (25) by

$$\hat{c}_{ij}^{\text{corr}}(\tau) = \frac{c_{ij}^{\text{obs}}(\tau)}{N-l}. \tag{46}$$

When the state space is discretized by a crisp partition, this matrix simply counts the number of observed transitions between discrete states, i.e., $c_{ij}^{\text{obs}}$ is the number of times the trajectory was observed in state $i$ at time $t$ and in state $j$ at time $t + \tau$, summed over all times $t$. If multiple trajectories are available, then the count matrices of these trajectories are simply added up.

As a shorthand notation we define the row sums of $\mathbf{C}^{\text{obs}}$:

$$c_i^{\text{obs}} = c_i^{\text{obs}}(\tau) := \sum_{k=1}^{n} c_{ik}^{\text{obs}}, \tag{47}$$

which are the total number of times the trajectory was in state $i$.

### B. Counting

We distinguish between two approaches to counting:

1. *Sampling the trajectory at lag time $\tau$*:
   Here the trajectory is sampled at lag time $\tau$ and only these sample points are used for counting:

$$c_{ij}^{\text{obs}}(\tau) = c_{ij}^{\text{obs}}(l\Delta t) = \sum_{k=1}^{\lfloor N/l \rfloor - 1} \chi_i(\mathbf{x}_{(l\cdot k)+1})\, \chi_j(\mathbf{x}_{(l\cdot k)+l+1}). \tag{48}$$

   When jump process is Markovian at $\tau$, this generates statistically independent transition counts. It is therefore straightforward to use the resulting count matrix in order

to derive expressions for the likelihood and posterior of transition matrix (see Sec. IV C below), which is important in order to obtain statistical models that do not underestimate the uncertainties.[39,45,51] A disadvantage of this approach is that much of the data are ignored, which can lead to numerical problems. In particular, states that have been actually visited or transitions that have been actually observed might be missed when subsampling the data at interval $\tau$, which may be a reason for estimators breaking down.

2. *Overlapping window count at lag time* $\tau$:
In this method we use a count window of width $\tau$ that is shifted along the time line:

$$c_{ij}^{\text{obs}}(\tau) = c_{ij}^{\text{obs}}(l\Delta t) = \sum_{k=1}^{N-l} \chi_i(\mathbf{x}_k)\chi_j(\mathbf{x}_{k+l}). \tag{49}$$

This method uses observed transitions, although nearby transitions such as $t \to t+\tau$ and $t+\Delta t \to t+\Delta t+\tau$ cannot be assumed to be statistically independent. The resulting count matrix, when assumed to consist of independent counts, will generate a posterior distribution that is too narrow in the Bayesian approaches below. However, the expectation value of $T_{ij}(\tau)$ is unbiased and thus maximum posterior estimators (Sec. IV D) are asymptotically correct, such that the window count method is preferred for this case.

At the moment it is an open question how to best make use of all observed data while at the same time using statistically independent, or at least uncorrelated counts. It appears straightforward to use the window method and then divide all counts by $l$, obtaining noninteger effective counts. However, the consequences of this approach are not fully understood because the probability distribution of transition matrices (see Sec. IV B below) becomes multimodal for counts $0 < c_{ij} < 1$. A safe approach is to use the window count method for estimating the transition matrix and sampling the trajectory at lag $\tau$ when computing count matrices for error estimators.

### C. Prior, likelihood, and posterior distribution

It is intuitively clear that in the limit of an infinitely long trajectory, the elements of the true transition matrix are given by the trivial estimator $\hat{T}_{ij}(\tau) = c_{ij}^{\text{obs}}/c_i^{\text{obs}}$, i.e., the fraction of times the transition $i \to j$ led out of state $i$. For a trajectory of limited length, the underlying transition matrix $\mathbf{T}(\tau)$ is not uniquely determined. Assuming that the matrix $\mathbf{C}^{\text{obs}}$ contains statistically independent transition counts (see discussion in Sec. IV B above), following Ref. 86, the probability that a particular $\mathbf{T}(\tau)$ would generate a sequence $s_1, \ldots, s_N$ the observed trajectory is given by the product of the individual jump probabilities, $\prod_{k=1}^{N-1} T_{s_k,s_{k+1}}$. In terms of our notation, this can be rewritten in terms of the count matrix as:

$$p(\mathbf{C}^{\text{obs}}|\mathbf{T}) = \prod_{i,j=1}^{n} T_{ij}^{c_{ij}^{\text{obs}}}. \tag{50}$$

Vice versa, the posterior probability of the transition matrix being $\mathbf{T}(\tau)$ is

$$p(\mathbf{T}|\mathbf{C}^{\text{obs}}) \propto p(\mathbf{T})p(\mathbf{C}^{\text{obs}}|\mathbf{T}) = p(\mathbf{T}) \prod_{i,j=1}^{n} T_{ij}^{c_{ij}^{\text{obs}}}, \tag{51}$$

where $p(\mathbf{T})$ is the prior probability of transition matrices before observing any data. $p(\mathbf{C}^{\text{obs}}|\mathbf{T})$ is called the likelihood. In transition matrix estimation one is interested in the most probable matrices $\mathbf{T}$, i.e., the $\mathbf{T}$'s with a large density in the posterior. The prior probability should be chosen such that it restricts the posterior to solutions that are physically meaningful in the situation where little observation data are available, but otherwise should be "weak," i.e., impose little bias (see Sec. IV E for a discussion on the choice of the prior). For computational simplicity, one typically chooses a prior that is conjugate to the likelihood, i.e., has the same functional form. This leads to the posterior:

$$p(\mathbf{T}|\mathbf{C}^{\text{obs}}) \propto \prod_{i,j=1}^{n} T_{ij}^{c_{ij}^{\text{prior}}+c_{ij}^{\text{obs}}} = \prod_{i,j=1}^{n} T_{ij}^{c_{ij}}, \tag{52}$$

with the prior count matrix $\mathbf{C}^{\text{prior}} = [c_{ij}^{\text{prior}}]$ and we have defined the total number of counts, or posterior counts $\mathbf{C} = \mathbf{C}^{\text{prior}} + \mathbf{C}^{\text{obs}}$. Since any estimator will be based on the count matrix, it is now straightforward to use a given estimator for any prior $\mathbf{C}^{\text{prior}}$. When a uniform distribution is used as a prior ($\mathbf{C}^{\text{prior}} = 0$, $p(\mathbf{T}) \propto 1$), likelihood and posterior distribution are identical.

### D. Maximum probability estimators

Based on a given probability distribution of parameters, a straightforward parameter estimator is the one that maximizes this probability distribution. Indeed, it can be shown (see supplementary material for the derivation[65]) that the maximum probability transition matrix, i.e., the maximum of Eq. (52), $\hat{\mathbf{T}} = \arg\max p(\mathbf{T} \mid \mathbf{C}^{\text{obs}})$ is given by the trivial estimator (assuming $c_i > 0$):

$$\hat{T}_{ij} = \frac{c_{ij}}{c_i}. \tag{53}$$

It turns out that $\hat{\mathbf{T}}(\tau)$, as provided by Eq. (53), is the maximum of $p(\mathbf{T}|\mathbf{C}^{\text{obs}})$ and thus also of $p(\mathbf{C}^{\text{obs}}|\mathbf{T})$ when transition matrices are assumed to be uniformly distributed *a priori*. In the limit of infinite sampling, i.e., trajectory length $N \to \infty$, $p(\mathbf{T}|\mathbf{C}^{\text{obs}})$ converges toward a Dirac delta distribution with its peak at $\hat{\mathbf{T}}(\tau)$. In this case the prior contribution vanishes:

$$\lim_{N\to\infty} \hat{T}_{ij} = \lim_{N\to\infty} \frac{c_{ij}^{\text{prior}} + c_{ij}^{\text{obs}}}{c_i^{\text{prior}} + c_i^{\text{obs}}} = \lim_{N\to\infty} \frac{c_{ij}^{\text{obs}}}{c_i^{\text{obs}}} = T_{ij}, \tag{54}$$

i.e., the estimator is "asymptotically unbiased."

Note that the estimator $\hat{\mathbf{T}}(\tau)$ does not necessarily fulfill detailed balance $\pi_i \hat{T}_{ij} = \pi_j \hat{T}_{ji}$ even if the underlying dynamics is in equilibrium and thus $\pi_i T_{ij} = \pi_j T_{ji}$ holds for the true transition matrix. In many cases it is desirable and advantageous to estimate a transition matrix that does fulfill detailed balance. There is no known closed form solution for the maximum probability estimator with the detailed balance

constraint. In Ref. 49, an iterative method was given to obtain such an estimator. Here we give a computationally more efficient algorithm to compute this estimator.

Let $x_{ij} = \pi_i T_{ij}$ be the unconditional transition probability to observe a transition $i \to j$. These variables fulfill the constraint $\sum_{i,j} x_{ij} = 1$, and the detailed balance condition is given by $x_{ij} = x_{ji}$. It is hence sufficient to store the $x_{ij}$ with $i \leq j$ in order to construct a reversible transition matrix. Let $x_i = \sum_j x_{ij}$. The maximum probability estimator is then obtained by the following iterative algorithm (see supplementary material for the proof of correctness[65]), which is iterated until some stopping criterion is met (e.g., change of $\max_{i,j}\{x_{ij}\}$ in one iteration is smaller than a given constant or the number of iterations exceeds a predefined threshold):

**Algorithm 1** Maximum probability estimator of reversible transition matrices

(1) For all $i, j = 1, \ldots, n$: initialize

$$x_{ij} = x_{ji} := c_{ij} + c_{ji}$$

$$x_i := \sum_j x_{ij}$$

(2) Repeat until stopping criterion is met

(2.1) For all $i = 1, \ldots, n$:

$$\text{update} : x_{ii} := \frac{c_{ii}(x_i - x_{ii})}{c_i - c_{ii}}$$

$$\text{update} : x_i := \sum_j x_{ij}$$

(2.2) For all $i = 1, \ldots, n-1$, $j = i+1, \ldots, n$:

$$a = c_i - c_{ij} + c_j - c_{ji}$$

$$b = c_i(x_j - x_{ij}) + c_j(x_i - x_{ij}) - (c_{ij} + c_{ji})(x_i + x_j - 2x_{ij})$$

$$c = -(c_{ij} + c_{ji})(x_i - x_{ij})(x_j - x_{ij})$$

$$\text{update} : x_{ij} = x_{ji} := \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

$$\text{update} : x_i := \sum_j x_{ij}$$

(2.3) Update $T_{ij}$, $i, j = 1, \ldots, n$:

$$T_{ij} := \frac{x_{ij}}{x_i}$$

## E. Expectation and Uncertainty

Since simulation data are finite, all validation procedures (either consistency checks or comparisons to experimental data) need to account for statistical uncertainties. For these, standard deviations or confidence intervals induced by the posterior distribution of transition matrices are of interest. It follows from well-known properties of the distribution of transition matrices[86] that the expectation value for transition matrix element $T_{ij}$ is

$$\bar{T}_{ij} = \mathbb{E}[T_{ij}] = \frac{c_{ij} + 1}{c_i + n}, \tag{55}$$

and the variance is given by

$$\mathrm{Var}[T_{ij}] = \frac{(c_{ij}+1)((c_i+n)-(c_{ij}+1))}{(c_i+n)^2((c_i+n)+1)} = \frac{\bar{T}_{ij}(1-\bar{T}_{ij})}{c_i+n+1}. \tag{56}$$

Consider a trajectory of a given molecular system which is analyzed with two different state space discretizations, one with $n = 10$ and one with $n = 1000$ and assume that a lag time $\tau$ has been chosen which is identical and long enough to provide Markov models with small discretization error for both $n$ (discussed in Sec. III). When using a uniform prior ($c_{ij} = c_{ij}^{\mathrm{obs}}$), the expectation values would be different for the two discretizations: in the $n = 1000$ case, most $c_{ij}$ are probably zero, such that the expectation value would be biased toward the uninformative $T_{ij} \approx 1/n$ matrix, and many observed transitions would be needed to overcome this bias. This behavior is undesirable. Thus, for uncertainty estimation it is suggested to use a prior which allows the observation data to have more impact. The extreme case is the so-called "null prior"[22] defined by

$$c_{ij}^{\mathrm{prior}} = -1 \qquad \forall i, j \in \{1, \ldots, n\}. \tag{57}$$

Using the null prior, the first moments of the posterior become

$$\bar{T}_{ij} = \mathbb{E}[T_{ij}] = \frac{c_{ij}^{\mathrm{obs}}}{c_i^{\mathrm{obs}}} = \hat{T}_{ij}, \tag{58}$$

$$\mathrm{Var}[T_{ij}] = \frac{c_{ij}^{\mathrm{obs}}(c_i^{\mathrm{obs}} - c_{ij}^{\mathrm{obs}})}{(c_i^{\mathrm{obs}})^2(c_i^{\mathrm{obs}}+1)} = \frac{\hat{T}_{ij}(1-\hat{T}_{ij})}{c_i^{\mathrm{obs}}+1}. \tag{59}$$

Thus, with a null prior, the expectation value is located at the likelihood maximum, or equivalently at the maximum of the posterior when a uniform prior would be used. Both expectation value and variance are independent of the number of discretization bins used. The variance of any $T_{ij}$ asymptotically decays with the number of transitions out of the state $i$, which is expected for sampling expectations from the central limit theorem.

In practice, one is often not primarily interested in the uncertainties of the transition matrix elements themselves, but rather in the uncertainties in properties computed *from* the transition matrix. Two different approaches have been suggested for this:

1. Linear error perturbation:[47,48,87] Here, the transition matrix distribution is approximated by a Gaussian and the property of interest is approximated by a first-order Taylor expansion. This results in a Gaussian distribution of the property of interest with a mean and a covariance matrix that can be computed in terms of **C**. This approach has the advantage of being deterministic, which is desirable in situations where uncertainties are used to steer an adaptive sampling procedure,[37,47,48,88] and may be implemented very efficiently. The disadvantage of this approach is that the Gaussian approximation of

the transition matrix posterior is only asymptotically valid, but easily breaks down when few counts have been observed and permits unphysical values (e.g., $T_{ij}$ outside the range [0,1]). Moreover, the property of interest is approximated linearly which can introduce a significant error when this property is nonlinear.

2. Markov chain Monte Carlo sampling of transition matrices:[45,51,89] Here, a set of transition matrices is drawn from the posterior distribution. The property of interest is then calculated for each transition matrix, and the uncertainties are directly estimated from this set. This approach requires that the true distribution is sampled often enough such that well-converged estimates of standard deviations or confidence intervals can be made. The advantage of the approach is that no assumptions are made concerning the functional form of the distribution or the property being computed. Furthermore, this approach can be straightforwardly applied to any function or property of transition matrices, including complex properties such as transition path distributions[22] without deriving the expressions necessary for the linear error perturbation analysis. Its disadvantage is that sampling may become slow for large matrices.

### F. Validation: Chapman–Kolmogorov test

We have above formulated conditions for choosing a discretization and a lag time $\tau$ that minimize the discretization error of a MSM. However, in practice it is essential to conduct a test whether lag time and discretization have been chosen such that the Markov model obtained is at least consistent with the data used to parameterize it within statistical error. In Sec. III D, the discretization error was measured as difference between Markov model propagation and true propagation in the continuous space. In practice it is easier to measure the propagation error on the discrete space directly. In particular, we are interested in checking whether the approximation,

$$[\hat{\mathbf{T}}(\tau)]^k \approx \hat{\mathbf{T}}(k\tau), \tag{60}$$

holds within statistical uncertainty. Here, $\hat{\mathbf{T}}(\tau)$ is the transition matrix estimated from the data at lag time $\tau$ (the Markov model), and $\hat{\mathbf{T}}(k\tau)$ is the transition matrix estimated from the same data at longer lag times $k\tau$. Note that when the nonreversible maximum likelihood estimator, Eq. (53), is used, this approximation is trivially exact for $k = 1$ since the Markov model was parameterized at lag time $\tau$. For all $k \gg t_2/\tau$, the approximation should always be good, as Markov models correctly model the stationary distribution, even for bad choices of $\tau$ and discretization (see Sec. III D). Thus, this test is only sensitive in ranges of $k$ greater one and smaller than the global relaxation time of the system.

There are various ways how a test of Eq. (60) could be implemented. An implementation of this test should consider the following points:

1. For large transition matrices, individual elements of $\hat{\mathbf{T}}(k\tau)$ or $[\hat{\mathbf{T}}(\tau)]^k$ can be very uncertain, and comparing $n \times n$ elements may be cumbersome. Therefore, we suggest to compare the probability of being in a given set of states, $A$, when starting from a well-defined starting distribution. This simplifies the test to few observables and allows to check the kinetics of states that are of special interest, such as folded/unfolded states or metastable states.
2. The test should be done for all times $k\tau$ for which trajectory data are available. Tests that compare Markov models that differ only one lag step ($\tau$ and $2\tau$) are likely to be unreliable as small approximation errors at short times may amplify at long times.
3. The quality of the approximation (60) should be judged within the statistical uncertainties induced by the data.

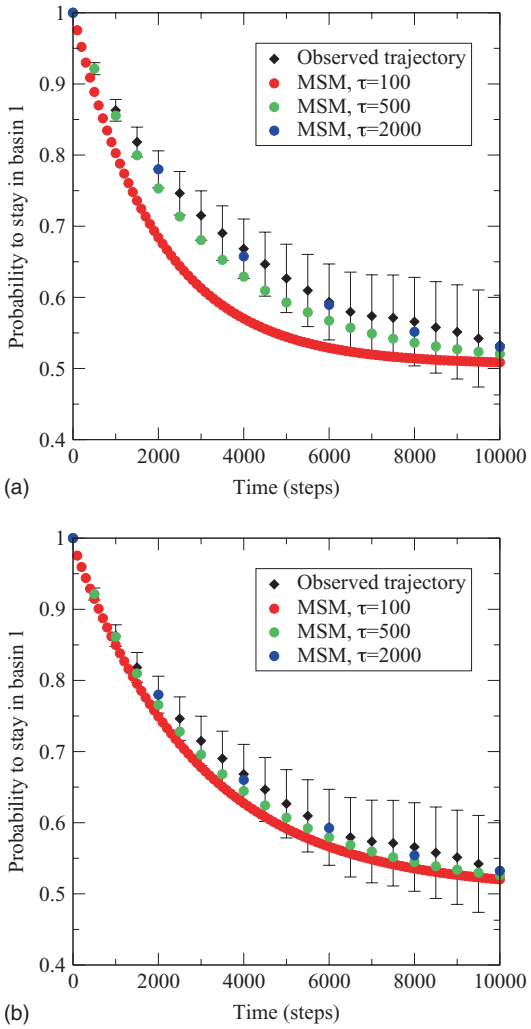


FIG. 11. Chapman–Kolmogorov test for diffusion in a two-well potential using a trajectory of length $10^6$ steps. Tested are Markov models that use lag times $\tau = 100, 500, 2000$ and (a) two-state discretization (split at $x = 50$), (b) six-state discretization (split at $x = 40, 45, 50, 55, 60$).

Here we present an implementation that takes these properties into account. Let $\boldsymbol{\pi}$ be the stationary probability of the Markov model $\hat{\mathbf{T}}(\tau)$. The corresponding stationary distribution restricted to a set $A$ is then given by

$$w_i^A = \begin{cases} \frac{\pi_i}{\sum_{j\in A}\pi_j} & i \in A \\ 0 & i \notin A \end{cases}. \tag{61}$$

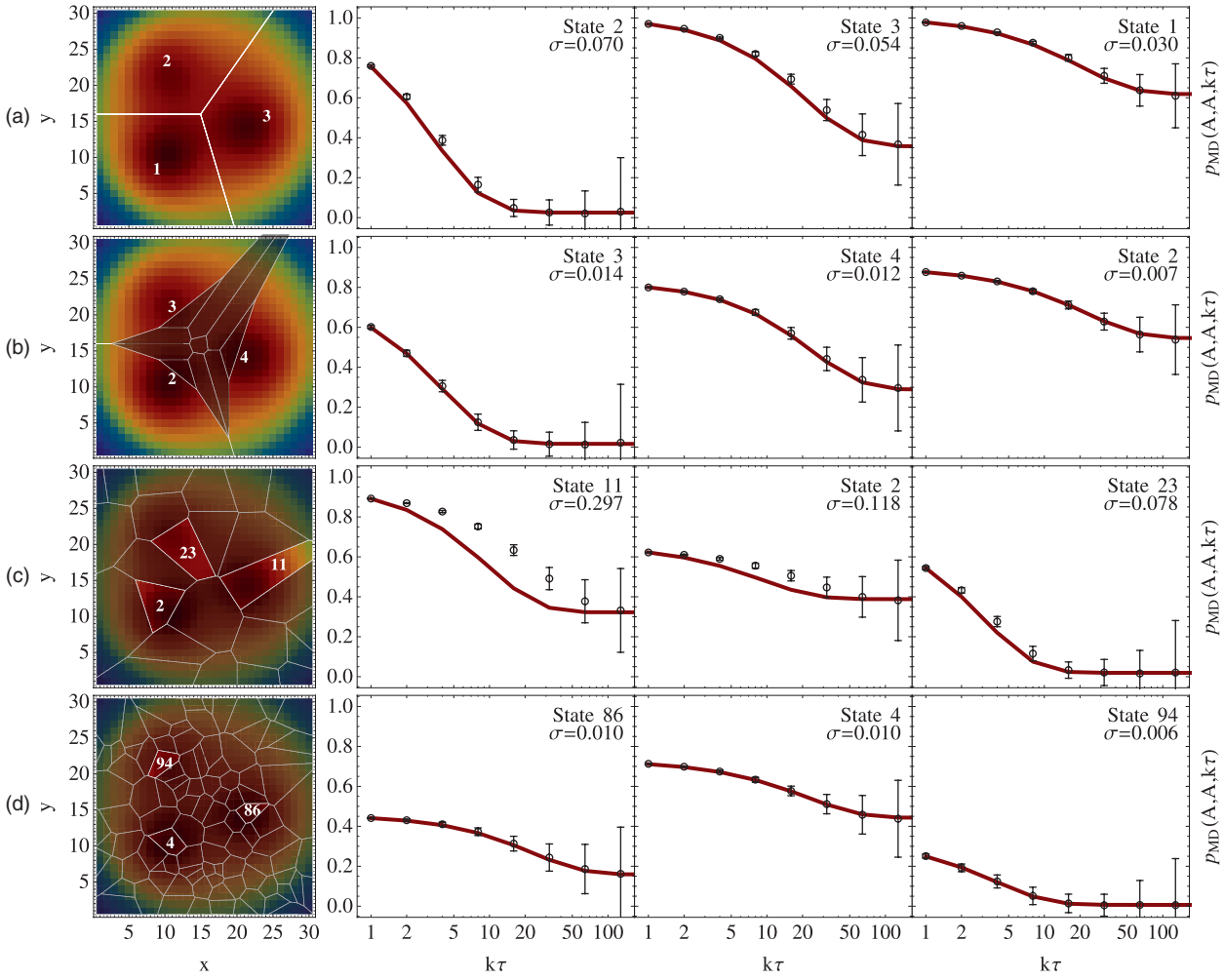


FIG. 12. Chapman–Kolmogorov test for the three-well diffusion model (see also Fig. 6). For each of four discretizations (first column, a, b, c, d), the Chapman–Kolmogorov test is shown for the three states with the greatest error (labeled with white figures in the first column). Relaxation curves from a 250 000 step trajectory, $p_{\mathrm{MD}}(A, A; k\tau)$ (black) along with the uncertainties $\epsilon_{\mathrm{MD}}(A, A, k\tau)$ are compared to the model prediction, $p_{\mathrm{MSM}}(A, A; k\tau)$ (red). The total error $\sigma$ given in the top right corners is measured as the 2-norm of the vector containing the differences $p_{\mathrm{MD}}(A, A; k\tau) - p_{\mathrm{MSM}}(A, A; k\tau)$ for time points in the range $k\tau \in [1, 128]$.

As a model test, the following "relaxation experiment" may be carried out for each set: using $\mathbf{w}^A$ as initial probability vector for each of the sets under consideration, the probability of being at that set at later times $k\tau$ is then computed according to (i) the observed trajectory data and (ii) the Markov model, and subsequently compared. The trajectory-based time-dependence of the probability to be at set $A$ after time $k\tau$ with starting distribution $\mathbf{w}^A$ is given by

$$p_{\mathrm{MD}}(A, A; k\tau) = \sum_{i \in A} w_i^A \; p_{\mathrm{MD}}(i, A; k\tau), \tag{62}$$

where $p_{\mathrm{MD}}(i, A, k\tau)$ is the trajectory-based estimate of the stochastic transition function Eq. (2):

$$p_{\mathrm{MD}}(i, A; k\tau) = \frac{\sum_{j \in A} c_{ij}^{\mathrm{obs}}(k\tau)}{\sum_{j=1}^{n} c_{ij}^{\mathrm{obs}}(k\tau)}. \tag{63}$$

Likewise, the probability to be at $A$ according to the Markov model is given by

$$p_{\mathrm{MSM}}(A, A; k\tau) = \sum_{i \in A} [(\mathbf{w}^A)^T \; \mathbf{T}^k(\tau)]_i. \tag{64}$$

Testing the validity of the Markov model then amounts to testing how well the equality

$$p_{\mathrm{MD}}(A, A; k\tau) = p_{\mathrm{MSM}}(A, A; k\tau) \tag{65}$$

holds, which is essentially a test of the Chapman–Kolmogorov property. Note that the initial distribution $w_i^A$ is simply a chosen reference distribution with respect to which the comparison is made, here chosen as in Eq. (61).

The equality (65) is not expected to hold exactly as a result of statistical uncertainties caused by the fact that only a finite number of transitions are available to estimate the true transition probabilities. To account for this, the uncertainties (one-sigma standard error) of the transition probabilities estimated from MD trajectories are computed as:

$$\epsilon_{\mathrm{MD}}(A, A; k\tau) = \sqrt{k \frac{p_{\mathrm{MD}}(A, A; k\tau) - [p_{\mathrm{MD}}(A, A; k\tau)]^2}{\sum_{i \in A} \sum_{j=1}^{n} c_{ij}^{\mathrm{obs}}(k\tau)}}. \tag{66}$$

The test then consists of assessing whether Eq. (65) holds within these uncertainties. The uncertainty of $p_{\mathrm{MSM}}(A, A, k\tau)$ can be calculated using the methods described in Sec. IV E. However, this is not necessary if the test already succeeds while using only the uncertainties $\epsilon_{\mathrm{MD}}(A, A; k\tau)$.

For illustration, we show results of this test using a $10^6$ step trajectory of a diffusion in a double-well potential (see Figure 9 and supplementary material for simulation setup[65]). Figure 11 shows the relaxation out of the left well using a two-state discretization splitting at $x = 50$ and using a six-state discretization splitting at $x = \{40, 45, 50, 55, 60\}$ [see Fig. 5 for state definitions and Fig. 11 for results]. The two-state discretization provides spurious results for $\tau = 100$, good results for $\tau = 500$, and for $\tau = 2000$ the results are excellent within the statistical uncertainty of the trajectory. For the six-state discretization even $\tau = 100$ is within the error bars while $\tau = 500$ and $\tau = 2000$ are both excellent approximations.

Figure 12 shows the corresponding results for the three-well diffusion model (see also Fig. 6 and supplementary material for model details[65]). A single 250 000 step trajectory started from the energy minimum at $\mathbf{x} = (10, 10)$ was simulated. For each of the four different discretizations shown in the first column of Fig. 12 the probability to stay in a state is shown for the three states with the largest Markov model error (highlighted in Fig. 12, left column). It is apparent that the metastable three-state discretization [Fig. 12(a)] performs well, however sacrificing metastability in order to more finely discretize the transition region generates a superior discretization [Fig. 12(b)]. The "uninformed" random 25-state clustering [Fig. 12(c)] performs worst but can be improved significantly by using more states [Fig. 12(d)]. This further supports our theoretical finding that either a clustering adapted to the eigenfunctions or using more states can improve the quality of the constructed Markov model.

Figure 13 shows the corresponding test results for the six most metastable sets of the MR121-GSGS-W peptide using a Markov model based on a Voronoi discretization using minimal RMSD to 1000 peptide configurations equally spaced in time. The lag time was set to $\tau = 2$ ns. The metastable states are determined by dominant eigenvectors and have been calculated with the PCCA+ method.[30,38] The Markov model agrees with the observed trajectory within statistical uncertainty for all metastable states.

### G. Practical approach to Markov model analysis

Markov models are becomingly increasingly popular as a tool to analyze large sets of MD trajectory data. In order to give some guidance to the practitioner, we have attached a brief walk-through for a typical Markov model analysis in the supplementary material. The analyses suggested there can be performed with the program EMMA (EMMA's Markov Model Algorithms, downloadable from `https://simtk.org/home/emma`).

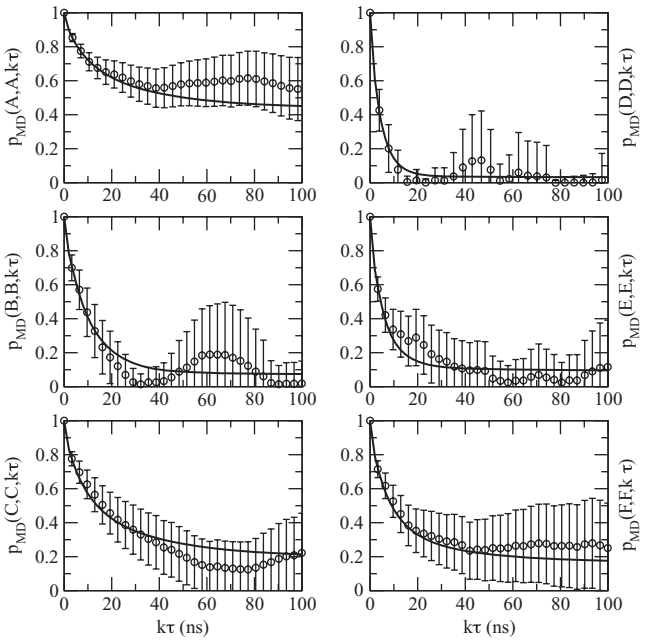


FIG. 13. Chapman–Kolmogorov test for six metastable sets $A$ to $F$ in MR121-GSGS-W. Solid curve: $p_{\mathrm{MSM}}(A, A, k\tau)$ to $p_{\mathrm{MSM}}(F, F, k\tau)$ predicted by the MSM parameterized at lag time $\tau = 2$ ns. Bullets with error bars: expectation values and uncertainties of $p_{\mathrm{MD}}(A, A, k\tau)$ to $p_{\mathrm{MD}}(F, F, k\tau)$ directly calculated from the simulation data up to 100 ns.

## V. DISCUSSION AND CONCLUSION

Markov modeling is a simulation analysis tool which is rapidly gaining popularity in the MD community. We have summarized the state of the art of generation and validation of Markov models of molecular kinetics and have filled in some important methodological gaps. Below, we summarize our discussion of this procedure, and highlight areas where further theoretical work or practical study is needed to give the approach a solid foundation.

As shown in Sec. II, any physically reasonable implementation of equilibrium molecular dynamics can be understood in terms of relaxation processes that are described by the eigenfunctions of the dynamical operator. The role of these eigenfunctions in molecular kinetics cannot be overemphasized, irrespective of whether Markov models are used or not. These eigenfunctions unambiguously yield a structural dynamical interpretation of a relaxation process. Each eigenfunction is linked to one eigenvalue with a corresponding relaxation timescale that is accessible experimentally, thus Markov models can serve as a means to interpret kinetic experimental data. From a modeling point of view, the dynamical decomposition Eq. (15) shows that these eigenfunctions define coordinates in which slow and fast dynamics can be separated exactly. Indeed, they are the only choice of coordinates for which such a separation is possible and any different attempt to model the dynamics via a projection onto slow degrees of freedom or order parameters will necessarily introduce memory terms that are challenging to deal with.[90]

One of the key insights from this work is that the discretization error made by using a Markov model on a discrete state space can be controlled by choosing the discretization and the lag time adequately (see Sec. III).

In particular, the quality of the Markov model depends on how well the discretization approximates the slowly relaxing eigenfunctions of the true dynamics. It is shown in Sec. III C how the Markov model can be used to precisely approximate only selected slow processes with relatively few discrete states slicing the state space finely in regions where the corresponding eigenfunctions change rapidly while leaving the discretization coarse in regions where only the fast eigenfunctions vary. This answers a key concern about discretization-based kinetic model approaches, namely, that for complex macromolecular systems there is no hope to enumerate all energy basins in the discrete model. The present analysis shows that this is indeed not necessary and that in principle, very few states are sufficient to obtain an excellent model. Moreover, the analysis also shows that metastable partitions suggested in previous works[38,39] are good among partitions where the number of states $n$ is allowed to be less or equal to the number of metastable states in the system, but that the approximation error can be further reduced by increasing the number of partitions, even if this means that the individual discrete states are no longer metastable.

This immediately raises the question how such a discretization can be created for a complex molecular system where the true eigenfunctions are initially unknown. This issue is not yet solved. Based on current results, it is clear that subdividing discrete states should always reduce the discretization error. Thus, when geometric clustering methods are used to subdivide state space, it is advisable to use as many clusters as possible without running into serious statistical problems. In the longer term, much better discretizations can be expected from methods that adaptively discretize in an iterative manner. For example, first an initial Markov model is created based on a geometric clustering, these clusters are then subdivided, providing a finer Markov model. The discretization error of the coarser model with respect to the finer model is computed using the error bound from Sec. III C, and it is then decided which states are kept, lumped, or split. An adaptive method based on maximizing metastability has been proposed in,[39] and a similar approach may be followed by minimizing the error bound from Sec. III C instead. In a broader sense, adaptive space discretization methods based on error bounds are commonly and successfully used in other disciplines where equations must be solved on a grid, e.g., in fluid mechanics and engineering. Moving to such approaches, MD becomes more and more a numerical analysis problem of molecular phase spaces, and may therefore benefit from the understanding of discretization methods that has been acquired in scientific computing.

We have devoted part of this work to describing how a Markov model on a given discretization can be estimated from an available data set. The main novelty in Sec. IV was the introduction of an efficient estimator for reversible transition matrices (Algorithm 1). It is recommended to use this estimator instead of the trivial nonreversible estimator in Eq. (53), because it enforces the physically reasonable detailed balance constraint, thus making more efficient use of the data and avoiding the difficulty of dealing with complex transition matrix eigenvalues and eigenvectors that typically arise from nonreversible transition matrices. As discussed in Sec. IV E, there are a number of approaches for estimating the uncertainty, i.e., the statistical estimation error of the Markov model, which is caused by the finite sample size of data used to parameterize it. The present work has treated the two types of error separately: the discretization error was examined assuming that there was no statistical error, and the statistical error was examined assuming that there was no discretization error. This represents the current state of knowledge in the field, but in reality both errors are always coupled, because a finite data set is given that is used for both defining the discretization and estimating the transition matrix. Thus, the investigation of the coupling of the two sources of error will be an important future research topic.

Although Markov model theory and methodology is now rather well developed, a number of fundamental questions remain. There is a hope that Markov models could avoid or mitigate the sampling problem by replacing single long equilibrium simulations that wait for the interesting rare events to happen by a large set short trajectories starting from different conformations that would be visited rarely in equilibrium. This immediately raises the question how relevant starting conformations can be found. This question is not specific to Markov model analyses, and it is likely that in this stage biased sampling methods such as metadynamics,[83] conformational flooding,[91] umbrella sampling,[92] targeted MD,[93] replica-exchange MD,[84] or pathway methods[35] will be useful to generate an initial exploration of conformation space from which short equilibrium simulations can then be launched. When the relevant conformations have been found and a good discretization has been obtained, it is clear that the uncertainty estimates of the Markov model can be exploited in order to pick starting points of subsequent simulations so as to adaptively reduce the uncertainty of the quantities of interest most.[47,48,94]

A more technical point that is not well understood is how to correctly weight the individual short trajectories in order to compute unbiased estimates of the transition probabilities from them [see Eq. (24)]. Since the Markov model is based on transition probabilities conditional on the starting state, there is no worry about relative weights between different discrete states. The correct weighting between states is taken care of by the Markov model, i.e., if trajectories are started from an initial distribution that is entirely out of equilibrium between states, the model will asymptotically provide the correct stationary distribution. The difficulty, however, lies in the correct weighting of trajectories within discrete states. When the size of the discrete states is not vanishingly small, the energy landscape within the states will not be approximately flat, and therefore the local equilibrium density within the states will not be flat either. Thus, when starting equilibrium trajectories from a nonequilibrium distribution, these trajectories should not contribute to the transition probability estimates with equal weights. Currently, this problem is dealt with by either discarding initial segments of all trajectories that correspond to local equilibration times or by enforcing local equilibration by picking starting conditions from simulations that are constrained to the starting states (see Sec. IV B). It would be desirable to have a simple reweighting method that allows to make use of all available data without

having to use extra simulations. This is a subject of ongoing research.

The type of Markov models investigated here, i.e., transition matrix based kinetics models between discrete state partitions of configuration space, must be viewed as one aspect within a family of conformation dynamics approaches. Rate matrix or master equation models[31,54,64] are very close in spirit, and we have mentioned connections to these models (see supplementary material[65]), making most of our present results available to these models as well. Recently, an alternative approach[31] has been proposed to obtain coarse-grained Markov or master equation models based on a noncomplete partition of state space that avoids to finely discretize the transition region. It is shown in Ref. 78 that our present analyses of the discretization error can be applied to this approach as well, only that here the eigenfunctions on the nonresolved parts of state space are effectively replaced by an interpolation based on committor functions between core sets. Generating Markov or master equation models based on rate models from an exploration of the stationary points of the energy landscape is an approach that has great tradition[27] and has been particularly successful in the analysis of Lennard-Jones or water clusters.[27,95] These models are not concerned with the same estimation problems as the present Markov models, as they are built from rate-theory based estimates (such as transition state theory) of state-to-state transition rates between the stationary points of the energy landscape, and not from trajectory statistics. However, they necessarily share the same concerns of making a discretization error by aggregating points of continuous state space into discrete model states. In a wider sense, approaches that use MD to parameterize effective stochastic equations, such as Langevin dynamics,[90,96,97] also induce models of the ensemble dynamics, such as Fokker–Planck type models. Such ensemble dynamics models generally share the advantages of Markov models over traditional MD analyses that have been discussed in the introduction. The specific advantage of Markov models is that they are on one hand asymptotically exact both in terms of discretization and estimator quality (see Sec. III and IV), and on the other hand very simple compared to models that in some way include memory. As they allow the whole arsenal of Markov chain theory to be readily accessed, the functional relationship between Markov models and most interesting molecular properties or observables has been worked out already,[22,30,42,45,48] and often has a simple and straightforwardly interpretable form. Given these advantages we expect that the popularity of Markov or similar models for the modeling of molecular kinetics will keep increasing.

## ACKNOWLEDGMENTS

We gratefully acknowledge financial support from DFG Research Center Matheon and SFB740, and DFG Grant No. 825/2. J.D.C. acknowledges support from the California Institute for Quantitative Biosciences (QB3) Distinguished Postdoctoral Fellowship.

[1]M. Jäger, Y. Zhang, J. Bieschke, H. Nguyen, M. Dendle, M. E. Bowman, J. P. Noel, M. Gruebele, and J. W. Kelly, Proc. Natl. Acad. Sci. USA **103**, 10648 (2006).
[2]A. Y. Kobitski, A. Nierth, M. Helm, A. Jäschke, and G. U. Nienhaus, Nucleic Acids Res. **35**, 2047 (2007).
[3]S. Fischer, B. Windshuegel, D. Horak, K. C. Holmes, and J. C. Smith, Proc. Natl. Acad. Sci. U.S.A. **102**, 6873 (2005).
[4]F. Noé, D. Krachtus, J. C. Smith, and S. Fischer, J. Chem. Theo. Comp. **2**, 840 (2006).
[5]A. Ostermann, R. Waschipky, F. G. Parak, and U. G. Nienhaus, Nature (London) **404**, 205 (2000).
[6]H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, Science **254**, 1598 (1991).
[7]A. Gansen, A. Valeri, F. Hauger, S. Felekyan, S. Kalinin, K. Tóth, J. Langowski, and C. A. M. Seidel, Proc. Natl. Acad. Sci. U.S.A. **106**, 15308 (2009).
[8]H. Neubauer, N. Gaiko, S. Berger, J. Schaffer, C. Eggeling, J. Tuma, L. Verdier, C. A. Seidel, C. Griesinger, and A. Volkmer, J. Am. Chem. Soc. **129**, 12746 (2007).
[9]W. Min, G. Luo, B. J. Cherayil, S. C. Kou, and X. S. Xie, Phys. Rev. Lett. **94**, 198302 (2005).
[10]E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern, Nature (London) **438**, 117 (2005).
[11]Y. Santoso, C. M. Joyce, O. Potapova, L. Le Reste, J. Hohlbein, J. P. Torella, N. D. F. Grindley, and A. N. Kapanidis, Proc. Natl. Acad. Sci. U.S.A. **107**, 715 (2010).
[12]Gebhardt, T. Bornschlögl, and M. Rief, Proc. Natl. Acad. Sci. U.S.A. **107**, 2013 (2010).
[13]B. G. Wensley, S. Batey, F. A.C. Bone, Z. M. Chan, N. R. Tumelty, A. Steward, L. G. Kwa, A. Borgia, and J. Clarke, Nature (London) **463**, 685 (2010).
[14]B. P. English, W. Min, A. M. van Oijen, K. T. Lee, G. Luo, H. Sun, B. J. Cherayil, S. C. Kou, and X. S. Xie, Nat. Chem. Bio. **2**, 87 (2006).
[15]M. O. Lindberg and M. Oliveberg, Curr. Opin. Struct. Biol. **17**, 21 (2007).
[16]K. Sridevi, J. Mol. Biol. **302**, 479 (2000).
[17]R. A. Goldbeck, Y. G. Thomas, E. Chen, R. M. Esquerra, and D. S. Kliger, Proc. Natl. Acad. Sci. U.S.A. **96**, 2782 (1999).
[18]A. Matagne, S. E. Radford, and C. M. Dobson, J. Mol. Biol. **267**, 1068 (1997).
[19]C. C. Mello and D. Barrick, Proc. Natl. Acad. Sci. U.S.A. **101**, 14102 (2004).
[20]S. A. Waldauer, O. Bakajin, T. Ball, Y. Chen, S. J. DeCamp, M. Kopka, M. Jäger, V. R. Singh, W. J. Wedemeyer, S. Weiss, S. Yao, and L. J. Lapidus, HFSP J. **2**, 388 (2006).
[21]D. D. Schaeffer, A. Fersht, and V. Daggett, Curr. Opin. Struct. Biol. **18**, 4 (2008).
[22]F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. U.S.A. **106**, 19011 (2009).
[23]W. van Gunsteren, J. Dolenc, and A. Mark, Curr. Opin. Struct. Biol. **18**, 149 (2008).
[24]S. V. Krivov and M. Karplus, Proc. Nat. Acad. Sci. U.S.A. **101**, 14766 (2004).
[25]F. Noé and S. Fischer, Curr. Opin. Struc. Biol. **18**, 154 (2008).
[26]S. Muff and A. Caflisch, Proteins **70**, 1185 (2007).
[27]D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).
[28]M. E. Karpen, D. J. Tobias, and C. L. Brooks, Biochemistry **32**, 412 (1993).
[29]I. A. Hubner, E. J. Deeds, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **103**, 17747 (2006).
[30]M. Weber, ZIB Report 03-04 (2003).
[31]N. V. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057 (2008).
[32]F. Rao and A. Caflisch, J. Mol. Bio. **342**, 299 (2004).
[33]B. de Groot, X. Daura, A. Mark, and H. Grubmüller, J. Mol. Bio. **301**, 299 (2001).
[34]V. Schultheis, T. Hirschberger, H. Carstens, and P. Tavan, J. Chem. Theory Comp. **1**, 515 (2005).
[35]A. C. Pan and B. Roux, J. Chem. Phys. **129**, 064107 (2008).
[36]M. Sarich, F. Noé, and C. Schütte, SIAM Multiscale Model. Simul. **8**, 1154 (2010).
[37]F. Noé, M. Oswald, G. Reinelt, S. Fischer, and J. C. Smith, Multiscale Model. Simul. **5**, 393 (2006).
[38]F. Noé, I. Horenko, C. Schütte, and J. C. Smith, J. Chem. Phys. **126**, 155102 (2007).
[39]J. D. Chodera, K. A. Dill, N. Singhal, V. S. Pande, W. C. Swope, and J. W. Pitera, J. Chem. Phys. **126**, 155101 (2007).

[40]W. C. Swope, J. W. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).
[41]N. Singhal, C. Snow, and V. S. Pande, J. Chem. Phys. **121**, 415 (2004).
[42]C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, J. Comput. Phys. **151**, 146 (1999).
[43]M. Weber, ZIB Report **09-27-rev** (2009).
[44]M. Jäger, H. Nguyen, J. C. Crane, J. W. Kelly, and M. Gruebele, J. Mol. Biol. **311**, 373 (2001).
[45]J. D. Chodera and F. Noé, J. Chem. Phys. **133**, 105102 (2010).
[46]V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, J. Am. Chem. Soc. **132**, 1526 (2010).
[47]N. S. Hinrichs and V. S. Pande, J. Chem. Phys. **126**, 244101 (2007).
[48]N. Singhal and V. S. Pande, J. Chem. Phys. **123**, 204909 (2005).
[49]G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, J. Chem. Phys. **131**, 124101 (2009).
[50]P. Metzner, C. Schütte, and E. V. Eijnden, Multiscale Model. Simul. **7**, 1192 (2009).
[51]F. Noé, J. Chem. Phys. **128**, 244103 (2008).
[52]J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill, Multiscale Model. Simul. **5**, 1214 (2006).
[53]S. Bacallado, J. D. Chodera, and V. Pande, J. Chem. Phys. **131**, 045106 (2009).
[54]N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, 4th ed. (Elsevier, Amsterdam, 2006).
[55]S. Park and V. S. Pande, J. Chem. Phys. **124**, 054118 (2006). J. D. Chodera, W. C. Swope, F. Noé, J.H. Prinz,
[56]J. D. Chodera, W. C. Swope, F. Noé, J.-H. Prinz, M. R. Shirts, and V. S. Pande, "*Dynamical reweighting: Improved estimates of dynamical properties from simulations at multiple temperatures*," J. Phys. Chem. (in press).
[57]H. C. Andersen, J. Chem. Phys. **72**, 2384 (1980).
[58]S. Duane, Phys. Lett. B **195**, 216 (1987).
[59]D. L. Ermak and Y. Yeh, Chem. Phys. Lett. **24**, 243 (1974).
[60]D. L. Ermak, J. Chem. Phys. **62**, 4189 (1975).
[61]W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, J. Chem. Phys. **76**, 637 (1982).
[62]B. Cooke and S. C. Schmidler, J. Chem. Phys. **129**, 164112 (2008).
[63]C. Schütte, F. Noé, E. Meerbach, P. Metzner, and C. Hartmann, in *Proceedings of the International Congress on Industrial and Applied Mathematics (ICIAM)*, edited by R. Jeltsch and G. Wanner (EMS Publishing House, Zurich, 2009), pp. 297–336.
[64]S. Sriraman, I. G. Kevrekidis, and G. Hummer, J. Phys. Chem. B **109**, 6479 (2005).
[65]See supplementary material at http://dx.doi.org/10.1063/1.3565032 for a practical approach to Markov model analysis, the model systems setup, and details about rate matrices and transition matrix estimations.
[66]M. Weber, "Meshless methods in conformation dynamics," Ph.D. thesis (Free University Berlin, 2006).
[67]M. G. Voronoi, J. Reine Angew. Math. **134**, 198 (1908).
[68]S. Kube and M. Weber, J. Chem. Phys. **126**, 024103 (2007).
[69]P. Metzner, I. Horenko, and C. Schütte, Phys. Rev. E **76**, 066702 (2007).
[70]D. Crommelin and E. V. Eijnden, Multiscale Model. Simul. **7**, 1751 (2009).
[71]B. Keller, P. Hünenberger, and W. van Gunsteren, "*An Analysis of the Validity of Markov State Models for Emulating the Dynamics of Classical Molecular Systems and Ensembles*," J. Chem. Theo. Comput. (submitted).
[72]E. Meerbach, C. Schütte, I. Horenko, and B. Schmidt, "Metastable conformational structure and dynamics: Peptides between gas phase and aqueous solution", in *Analysis and Control of Ultrafast Photoinduced Reactions*, Series in Chemical Physics, Vol. 87 (Springer, Berlin, 2007), pp. 796–806.
[73]W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, and M. Eleftheriou, J. Phys. Chem. B **108**, 6582 (2004).
[74]G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. (Johns Hopkins University Press, Baltimore, MD, 1996).
[75]C.-K. Chan, Y. Hu, S. Takahashi, D. L. Rousseau, W. A. Eaton, and J. Hofrichter, Proc. Natl. Acad. Sci. U.S.A. **94**, 1779 (1997).
[76]O. Bieri, J. Wirz, B. Hellrung, M. Schutkowski, M. Drewello, and T. Kiefhaber, Proc. Natl. Acad. Sci. U.S.A. **96**, 9597 (1999).
[77]H. Neuweiler, S. Doose, and M. Sauer, Proc. Natl. Acad. Sci. U.S.A. **102**, 16650 (2005).
[78]N. Djurdjevac, M. Sarich, and C. Schütte, "*Estimating the eigenvalue error of Markov state models*," Multiscale Model. Simul. (submitted).
[79]A. Amadei, A. B. Linssen, and H. J.C. Berendsen, Proteins **17**, 412 (1993).
[80]B. Keller, X. Daura, and W. F. van Gunsteren, J. Chem. Phys. **132** (2010).
[81]Y. Yao, J. Sun, X. Huang, G. R. Bowman, G. Singh, M. Lesnick, L. J. Guibas, V. S. Pande, and G. Carlsson, J. Chem. Phys. **130**, 144115 (2009).
[82]S. Dasgupta and P. Long, J. Comput. Syst. Sci. **70**, 555 (2005).
[83]A. Laio and M. Parrinello, Proc Natl. Acad. Sci. U.S.A. **99**, 12562 (2002).
[84]Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[85]S. Röblitz, "Statistical error estimation and grid-free hierarchical refinement in conformation dynamics, Ph.D. thesis (Free University Berlin, 2009).
[86]T. W. Anderson and L. A. Goodman, Ann. Math. Statist. **28**, 89 (1957).
[87]J.-H. Prinz, M. Held, J. C. Smith, and F. Noé, "*Efficient computation, sensitivity and error analysis of committor probabilites for complex dynamical processes*," SIAM Multiscale Model. Simul. (submitted).
[88]F. Noé, M. Oswald, and G. Reinelt, in *Operations Research Proceedings*, edited by J. Kalcsics and S. Nickel (Springer, New York, 2007), pp. 435–440.
[89]P. Metzner, F. Noé, and C. Schütte, Phys. Rev. E **80**, 021106 (2009).
[90]O. F. Lange and H. Grubmüller, J. Chem. Phys. **124**, 214903 (2006).
[91]H. Grubmüller, Phys. Rev. E **52**, 2893 (1995).
[92]G. M. Torrie and J. P. Valleau, J. Comp. Phys. **23**, 187 (1977).
[93]J. Schlitter, M. Engels, and P. Krüger, J. Mol. Graphics **12**, 84 (1994).
[94]G. R. Bowman, D. L. Ensign, and V. S. Pande, J. Chem. Theory Comput. **6**, 787 (2010).
[95]D. J. Wales, Science **271**, 925 (1996).
[96]R. Hegger and G. Stock, J. Chem. Phys. **130**, 034106 (2009).
[97]C. Micheletti, G. Bussi, and A. Laio, J. Chem. Phys. **129**, 074105 (2008).