JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

An Analysis of the Validity of Markov State Models for Emulating the Dynamics of Classical Molecular Systems and Ensembles

Bettina Keller, Philippe Hünenberger, and Wilfred F. van Gunsteren*

Laboratory of Physical Chemistry, Swiss Federal Institute of Technology Zürich, ETH Zürich, CH-8093 Zürich, Switzerland

ABSTRACT: Markov state models parametrized using molecular simulation data are powerful tools for the investigation of conformational changes in biomolecules and in recent years have gained increasing popularity. However, a Markov state model is an approximation to the true dynamics of the complete system. We show how Markov state models are derived from the generalized Liouville equation identifying the assumptions and approximations involved and review the mathematical properties of transition matrices. Using two model systems, a two-bit flipping model consisting of only four states, and molecular dynamics simulations of liquid butane, we subsequently assess the influence of the assumptions, for example, of the marginal degrees of freedom, used in the derivation on the validity of the Markov state model.

1. INTRODUCTION

The dynamics of large biomolecules encompasses processes of vastly different time scales. Fast processes, such as bond-angle vibrations, happen on the femtosecond time scale and are coupled to slow processes, such as large conformational rearrangements, which happen on the micro- to millisecond time scale. For example, the folding of an entire protein can easily take seconds. In a molecular dynamics (MD) simulation of a biomolecule at the atomic level, the integration step of the simulation is bound to the order of 1 fs.[1,2] Constraining the fast processes to a fixed value, the only exception being the bond-length vibration, will distort the dynamics of the slow processes.[3−5] Thus, a MD simulation of a folding process aims at emulating a process with a time scale of micro- to miliseconds by tracing out trajectories at a femtosecond resolution, thereby bridging time scales of 9 to 12 orders of magnitude. This together with the fact that MD simulation programs scale poorly for the parallelization to many processors makes MD simulations of biomolecular processes time-consuming. Moreover, by integrating the time evolution of each degree of freedom explicitly, the amount of detail produced by an atomic-level MD simulation is barely manageable and often far beyond what is needed for the elucidation of a particular molecular phenomenon.

The combination of MD simulation with stochastic models such as Markov state models (MSM) has the power to redress both of these shortcomings. For the construction of a MSM, the complete coordinate space of the simulated system (i.e., solute plus solvent) is split into a set of *relevant* coordinates and a set of *marginal* coordinates. In the first instance, the term "relevant coordinates" denotes any set of coordinates which is of relevance to the question which is to be investigated by the simulation. MSMs provide a means to concisely represent the dynamics of the relevant coordinates, formulated as a transition matrix with a dimension on the order of typically several hundreds to several thousands. Properties of interest, such as mean first passage times and mean life times, can be directly extracted from this transition matrix.[6,7] The conformational equilibrium distribution emerges as the first eigenvector of the transition matrix, and metastable states can be identified by grouping the states of the Markov process in such a manner that the metastability of the groups is maximized. This corresponds to maximizing the trace of the coarse-grained matrix.[8]

MSMs are associated with a time step, called *lag time*, which is typically on the order of pico- to nanoseconds. If the dynamics of the relevant coordinates are indeed Markovian at this lag time, then the MSM can be parametrized by a large number of short MD simulations.[10] This approach does not necessarily decrease the required computer time but rather, when the short simulations are run in parallel on several computers, the time one has to wait for the results.

In a MSM, the configurations of the system are mapped onto a (typically small) set of states, and the dynamics are modeled by the transition probabilities between these states. While from a mathematical point of view, the mapping corresponds to a projection and can done by a single operator multiplication,[9] in any practical application, this projection is split into two consecutive steps: (*i*) separation of the complete coordinate space into relevant and marginal coordinates and (*ii*) discretization of the relevant coordinates. These models are clearly an approximation of the true dynamics. The idea is that the influence of the marginal degrees of freedom averages out over time and that one can often find a time lag τ_{Markov} for which the deviation from Markovian behavior in the relevant coordinates is small enough to be neglected. A Markov model with a time resolution of τ_{Markov} or larger may then represent a realistic model of the true dynamics. Ultimately, the quality of the Markov model, i.e., how faithfully the model reproduces the dynamics in the relevant degrees of freedom, depends on (*i*) the interaction of the set of marginal coordinates with the set of relevant coordinates, (*ii*) the precise discretization of the relevant coordinates, and (*iii*) the statistical errors due to finite sampling of the dynamics of the system.

Since Swope et al.[11,12] presented the first extensive and detailed application of MSM to the analysis of molecular simulation data, MSMs of biomolecular systems have developed into a very active field of research.[9,13−17]

Received: January 28, 2011
Published: March 10, 2011

The extraction of metastable states from a given MSM, which is equivalent to the coarse-graining of the transition matrix, has been a major issue in the discussion of the application of MSMs. Two basic approaches have been published. One maximizes the metastability of the resulting coarse-grained states using temperature annealing schemes;[8,17] the other exploits properties of the eigenvectors of the fine-grained transition matrix to define the coarse-grained states.[7,18,19]

Recently, methods which optimize the amount of simulation data needed for the construction of a MSM have been published. These methods either rely on enhanced sampling techniques in the simulation process[16,20−24] or apply an adaptive sampling scheme which couples the start of new simulations to a quality estimate of the current MSM.[25]

A large number of publications deal with the discretization of the relevant degrees of freedom. Noé et al.[7] discretized each backbone dihedral angle along the minima of a probability distribution of this angle, thereby discretizing the conformational space according to the rotamers of the molecule. More often, however, the conformations of the molecule are mapped onto more global descriptors such as secondary structure motifs of amino acids in a peptide[13,20] or the number of intramolecular hydrogen bonds.[7] In 2007, Chodera et al.[8] published an adaptive discretization scheme in RMSD space. Jensen et al.[26] discretized the two central dihedral angles of a tetrapeptide according to the most populated regions in their Ramachandran plots and varied the positions of the boundaries. They found that the quality of the MSM is sensitive to the exact position of the boundaries. This finding is in line with the results of Sarich et al.,[9] who demonstrate analytically that the error caused by the discretization is determined by the precision with which the transition region is discretized. Moving the boundary away from the transition point impairs the quality of the MSM.

Several methods have been developed for the estimation of the statistical uncertainties in the eigenvalues and eigenvectors of the transition matrix and properties derived from the transition matrix.[27−29]

To the best of our knowledge, no systematic study of the influence of the marginal coordinates on the dynamics of the relevant coordinates has been published. Conceptually, this question is close to the discussion about the influence of the bath degrees of freedom on the solute coordinates in Brownian dynamics.[30] However, some of the assumptions (very large number of bath degrees of freedom, all coupled with the same coupling constant to the solute degrees of freedom) clearly do not apply in the context of MSM of molecular dynamics.

This publication has two objectives: (*i*) a review of the mathematical concepts and assumptions which form the basis of a stochastic model of molecular dynamics and (*ii*) an illustration of the effect of the marginal coordinates on the dynamics of the relevant coordinates. These two parts are closely linked since the properties of the marginal coordinates determine to a large extent the quality of the Markov model.

In the first part, we demonstrate how stochastic equations of motion emerge from deterministic ones when the coordinate set is split into relevant and marginal coordinates. We list the conditions a stochastic equation of motion must fulfill in order to be Markovian. We then show how the matrix formalism of transition matrices arises from a given Markovian equation of motion. Finally, we review the mathematical properties of transition matrices and link them to physical concepts such as ergodicity and equilibrium dynamics.

In the second part, we use two model systems to study how the properties of the marginal coordinates affect the assumption that the dynamics of the relevant coordinates are Markovian. The first system consists of two bits which can flip between "0" and "1". One bit represents the relevant coordinates, the other the marginal ones. We illustrate how the coupling strength and the relative speed of the two bits influence the quality of the Markov model. The second system consists of molecular dynamics simulations of a butane molecule (relevant coordinates) immersed in a solvent of butane molecules (marginal coordinates). The solvent is modeled on the one hand explicitly (at various temperatures and pressures) and on the other hand implicitly using stochastic dynamics (with various temperatures and friction coefficients). The influence of these parameters on the quality of the Markov model is demonstrated.

2. THEORY

We consider a *system* of N_x time-dependent variables and an *ensemble* of an infinite number of replicas of this system. The *configuration* of system n in the ensemble at time t is defined by a configuration vector $\mathbf{x}_n(t)$ containing the instantaneous values of the N_x variables of this system at time t. The dynamics of the ensemble is said to be *Markovian* if the individual systems obey equations of motion of the form

$$\dot{\mathbf{x}}_n(t) = f(\mathbf{x}_n(t), \mathbf{y}_s(t, s_n)) \tag{1}$$

where f and $\mathbf{y}_s$ on the right-hand side are functions that are identical for all systems in the ensemble, while s_n represents a scalar value attributed to a specific system n. Equation 1 states that the change of the configuration of the nth system at time t, $\dot{\mathbf{x}}_n(t)$, only depends on its current configuration $\mathbf{x}_n(t)$ and the current values of a set of variables $\mathbf{y}_s(t,s_n)$.

The function $\mathbf{y}_s(t,s_n)$ is used to implement the difference between deterministic and stochastic ensemble dynamics. If the ensemble dynamics are deterministic, the parameter s_n can be dropped and $\mathbf{y}_s(t,s_n) = \mathbf{y}_s(t)$ is the same for all systems; i.e., all systems in the ensemble follow the same equation of motion. If the ensemble dynamics involves a set of stochastic variables, $\mathbf{y}_s(t,s_n)$ represents a particular trajectory with index s_n in this variable space, which was drawn from a stochastic process $\mathbf{Y}(t)$. In this case, the equation of motion, eq 1, differs for each system.

In computational terms, s_n can be viewed as the seed for a pseudorandom number sequence assigned to the system. In more mathematical terms, s_n can also be viewed as defining this sequence itself, e.g., in the form of the representation of this real number by an infinite string of bits. The way in which s_n, or the derived pseudorandom number sequence, is exploited by the function $\mathbf{y}_s$, e.g., to generate the time series of stochastic Gaussian-distributed variables, need not be specified at this point. However, it is assumed that the resulting probability distribution of the stochastic variables over all systems in the ensemble at a given configuration $\mathbf{x}$ is time-invariant.

The key assumptions for Markovian ensemble dynamics are that the single-system dynamics fulfill the following conditions: (*i*) *deterministic* ($\mathbf{x}_n(t)$ is determined by the sole knowledge of $\mathbf{x}_n(0)$ and s_n), *memoryless* ($\dot{\mathbf{x}}_n(t)$, as expressed by the function f, involves no explicit dependence on $\mathbf{x}_n(t')$ with $t' < t$), and *stationary* ($\dot{\mathbf{x}}_n(t)$, as expressed by the function f, involves no explicit dependence on t); (*ii*) a common equation of motion (no stochastic component) or a set of stochastically distributed equations of motion for the different systems in the ensemble.

Note that the stochasticity property only becomes apparent at the level of the ensemble. From the point of view of a single system n, the dynamics are entirely deterministic, given the value of s_n assigned to this system. Note also that the system dynamics need not necessarily be continuous in time; i.e., $\dot{\mathbf{x}}_n(t)$ may involve Dirac delta functions in time.

The instantaneous *macrostate* of the ensemble is defined by the (normalized) *configurational probability distribution* $\rho(\mathbf{x},t)$ of the individual systems at time t in the N_x-dimensional space of the system configurations, which obeys the generalized Liouville equation:

$$\dot{\rho}(\mathbf{x},t) = \hat{\mathscr{L}}\rho(\mathbf{x},t) \tag{2}$$

$\hat{\mathscr{L}}$ is called the generalized Liouville operator or the generator. The assumptions of Markovian dynamics imply that $\hat{\mathscr{L}}$ in eq 2 is time-independent, corresponding to equilibrium dynamics. Introducing the requirement that eq 2 be valid for any arbitrary initial configurational distribution $\rho(\mathbf{x},0)$ and all times (including $t = 0$), the operator $\hat{\mathscr{L}}$ is unique, and its exact form can, at least in principle, be derived from knowledge of the function $f(\mathbf{x},\mathbf{y})$ in eq 1 and of the stochastic variable probability distribution. Different forms of the generalized Liouville equation include the Liouville equation[31,32] (Hamiltonian dynamics), the Fokker−Planck equation[6,32] (Langevin dynamics), or the Smoluchowksi equation[32] (Brownian dynamics).

By introducing an infinite set of basis functions $\phi_i(\mathbf{x})$, e.g., Dirac delta functions, covering the N_x-dimensional space of the configuration variables, the configurational probability distribution $\rho(\mathbf{x},t)$ may be rewritten as a *configurational probability vector* $\mathbf{p}(t)$ with components $p_i(t)$, which are real, non-negative, and sum up to 1. The generalized Liouville equation, eq 2, may then be translated into an equivalent matrix equation:

$$\dot{\mathbf{p}}(t) = \mathbf{K}\mathbf{p}(t) \tag{3}$$

in which the generalized Liouville matrix $\mathbf{K}$ is a rate matrix with off-diagonal elements $K_{ij} \geq 0$ representing the rate of transition from configuration point j to configuration point i and the diagonal element K_{jj} is equal to $-\sum_{i\neq j} K_{ij}$. Consequently, the elements of each of its columns add up to 0.

Equation 3 can be formally integrated in time over an interval τ ($\tau > 0$), referred to as a *lag-time* yielding

$$\mathbf{p}(t+\tau) = \mathbf{T}(\tau)\,\mathbf{p}(t) \tag{4}$$

resulting in the introduction of a corresponding *transition matrix* $\mathbf{T}(\tau)$, defined as

$$\mathbf{T}(\tau) \doteq \exp(\tau\mathbf{K}) \tag{5}$$

Equation 5 effectively introduces a time discretization of the continuous Markov process. The elements of $\mathbf{T}(\tau)$ are real and non-negative

$$T_{ij} \in \mathbb{R}, T_{ij}(\tau) \geq 0 \quad \forall\ i,j,\tau \tag{6}$$

and satisfy the normalization condition

$$\sum_i T_{ij}(\tau) = 1 \quad \forall\ j,\tau \tag{7}$$

They represent the probability of a transition from a point j to a point i in configurational space during a lag time τ. Equations 6 and 7 define a *column stochastic* matrix. From the definition of $\mathbf{T}(\tau)$, one can directly derive the Chapman−Kolmogorov equation representing the *recursivity property*

$$\mathbf{T}(\tau_1+\tau_2) = \mathbf{T}(\tau_1)\mathbf{T}(\tau_2) = \mathbf{T}(\tau_2)\mathbf{T}(\tau_1) \tag{8}$$

When formulated as

$$\mathbf{T}(n\tau) = \mathbf{T}^n(\tau) \tag{9}$$

this relation can be used as a check of whether a process with a time discretization of τ is Markovian.[6,14]

The matrix $\mathbf{T}(\tau)$ possesses N_d *eigenvalues*, $\lambda_\alpha(\tau)$, with associated left eigenvectors, $\boldsymbol{\psi}_\alpha$. The eigenvectors are formally defined within an arbitrary multiplicative factor. To make their definition unambiguous, it will be assumed that these vectors are selected such that (*i*) the sum of the two-norm of the elements of an eigenvector is always unity; (*ii*) the first nonvanishing component of an eigenvector is always real and positive. With this convention, the eigenvectors with real eigenvalues always have real components that add up to unity. Because $\mathbf{T}(\tau)$ is column stochastic, it also has the following properties:[18]

1. It possesses a special (real) left eigenvector (which will be given the index $\alpha = 1$), $\boldsymbol{\psi}_1 = N_d^{-1}\{1, 1, ..., 1\}$ associated with the eigenvalue $\lambda_1 = 1$. Therefore, it also possesses at least one corresponding (real) right eigenvector associated with this eigenvalue. Note that in the case of uncoupled Markov chains, i.e., if $\mathbf{T}(\tau)$ can be permuted into a block-diagonal form, the eigenvalue is degenerate, i.e., associated with more than one left and right eigenvector.[18]
2. Its eigenvalue spectrum has a radius of 1, i.e., $|\lambda_\alpha(\tau)| \leq 1$ for all α.

At this point, one may add a third assumption to the assumptions underlying Markovian dynamics, namely that of irreducibility. The Markovian ensemble dynamics is also *irreducible* when

$$\lim_{\tau\to\infty} \tau^{-1}\int_0^\tau d\tau' \mathrm{T}(\tau') = \mathrm{T}_{sum} > 0 \tag{10}$$

where $\mathbf{T}_{\mathrm{sum}} > 0$ is a short-hand notation for

$$\mathbf{T}_{\mathrm{sum},ij} > 0 \quad \forall\ i,j \tag{11}$$

Irreducibility implies that any configuration has a nonvanishing probability of undergoing a transition to any other configuration considering all possible lag times. It does, however, not imply that there exists a single lag time τ' for which all possible transitions have a nonvanishing transition probability.

Irreducibility is not identical with the concept of ergodicity, but it is closely linked to it. A system is *ergodic* if the time its (sufficiently long) trajectory spends in any given configuration is proportional to the probability with which this configuration is realized in the ensemble at a given time t. Then, the time average of any property A calculated along its trajectory is the same as the ensemble average of this property:

$$\langle A\rangle = \frac{1}{t}\int_0^t A(\mathbf{x}(t'))\,\mathrm{d}t' = \int A(\mathbf{x})\,\rho(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{12}$$

A system can only be ergodic if, starting from any given configuration, all other configurations can and will be reached in the course of the (sufficiently long) trajectory. Irreducibility ensures that any state can be reached from any other, but not necessarily that it will be reached; i.e., irreducibility is a necessary but not sufficient condition for ergodicity. Intuitively, the fact that a reducible transition matrix cannot lead to ergodic dynamics arises from the observation that a set of n elements in $\mathbf{p}(t)$ will never undergo transitions to the complementary set of $N_d - n$ elements in $\mathbf{p}(t+\tau)$. As a result, ensemble dynamics initiated from a specific distribution $\mathbf{p}(0)$ solely encompassing nonvanishing

elements of the former set will never generate probabilities in the latter set.

According to the Perron−Frobenius theorem,[33,34] if a column-stochastic transition matrix $\mathbf{T}(\tau)$ is irreducible, it additionally has the following properties:

3. Its eigenvalue $\lambda_1 = 1$ is nondegenerate; i.e., it is associated with a unique real right eigenvector $\boldsymbol{\psi}_1$.
4. The components of the right eigenvector $\boldsymbol{\psi}_1$ are all non-negative.

The unique right eigenvector $\boldsymbol{\psi}_1$ associated with the eigenvalue $\lambda_1 = 1$ is referred to as the *stationary* probability distribution of the ensemble dynamics and will be further noted as $\boldsymbol{\pi}$. Its properties are

$$\mathbf{T}(\tau)\boldsymbol{\pi} = \boldsymbol{\pi} \quad \forall\ \tau \tag{13}$$

and

$$\begin{aligned} \pi_i &\in \mathbb{R}\ \forall i \\ \pi_i &\geq 0\ \forall i \\ \sum_i \pi_i &= 1 \end{aligned} \tag{14}$$

Intuitively, $\boldsymbol{\pi}$ corresponds to a special probability distribution within the ensemble that is invariant upon propagation by $\mathbf{T}(\tau)$ for any lag time τ.

At this point, one may add a fourth assumption to the assumptions underlying irreducible Markovian dynamics, namely that of primitivity. A non-negative square matrix $\mathbf{A}$ is called *primitive* if there exists an integer $k > 0$ for which all elements of the matrix $\mathbf{A}^k$ are positive. A sufficient condition for a non-negative and irreducible square matrix to be primitive is that it possesses at least one nonzero element on the diagonal. If $\mathbf{T}(\tau)$ is irreducible and has at least one postive entry on its diagonal, then there is only one eigenvalue with $|\lambda_\alpha| = 1$ and this is $\lambda_1 = 1$; i.e., $\lambda_1 = 1$ is the only eigenvalue on the unit circle. The condition of primitivity ensures that in the limit of long lag times any arbitrary initial probability distribution $\mathbf{p}(0)$ converges to the stationary distribution $\boldsymbol{\pi}$:[34]

$$\lim_{\tau \to \infty} \mathbf{T}(\tau)\, \mathbf{p}(0) = \boldsymbol{\pi} \quad \forall\ \mathbf{p}(0) \tag{15}$$

As a last stipulation, we require that Markovian dynamics (as defined by eq 1) are *detailed balanced* with respect to their stationary distribution. This is the case when

$$T_{ij}(\tau)\, \pi_j = T_{ij}(\tau)\, \pi_i \quad \forall\ i,j,\tau \tag{16}$$

or introducing the diagonal matrix $\boldsymbol{\Pi}$ with elements equal to $\boldsymbol{\pi}$, i.e., $\Pi_{ij} = \pi_i \delta_{ij}$:

$$\mathbf{T}(\tau)\, \boldsymbol{\Pi} = \boldsymbol{\Pi}\, \mathbf{T}^{\mathrm{T}}(\tau) \quad \forall\ \tau \tag{17}$$

where $\mathbf{T}^{\mathrm{T}}$ denotes the transpose of the matrix $\mathbf{T}$. Detailed balance implies that the number of transitions between pairs of configurational points in a stationary ensemble (i.e., characterized by the probability distribution $\boldsymbol{\pi}$) is equal in the forward and backward directions. When the ensemble dynamics satisfy this condition, the stationary distribution $\boldsymbol{\pi}$ will be further referred to as the *equilibrium* probability distribution or Boltzmann distribution of the ensemble. Intuitively, a violation of detailed balance implies that for at least one pair of configurational points, there exists a net direct flow in the forward direction from the first point to the second that must be compensated by an equivalent net indirect flow via other points in the opposite direction to maintain the probability stationary. In the language of thermodynamics, this behavior is characteristic of a steady state rather than an equilibrium stationary situation as would be encountered, e.g., in a system where a temperature, pressure, or composition gradient is maintained. At thermodynamic equilibrium, direct flows in the forward and backward directions between all pairs of states must compensate for each other, as will be the case, e.g., in a system where temperature, pressure, and composition are homogeneous in space.

Note that if $\mathbf{T}(\tau)$ is detailed-balanced, then so is any transition matrix $\mathbf{T}(n\tau) = \mathbf{T}^n(\tau)$ with $n \in \mathbb{Z}$. Note also that a column-stochastic matrix can only be detailed-balanced with respect to a vector that is also a right eigenvector associated with the eigenvalue one. In other words, irreducible Markovian dynamics can only be detailed balanced with respect to $\boldsymbol{\pi}$ (and no other vector). If an irreducible and primitive column-stochastic transition matrix $\mathbf{T}(\tau)$ is detailed balanced with respect to its stationary distribution $\boldsymbol{\pi}$, it also has the following properties:[34]

1 All eigenvalues are real and lie in the interval $]-\mathbf{1};+\mathbf{1}]$, so that all eigenvectors are real.
2 The eigenvectors of $\mathbf{T}(\tau)$ define a complete eigenbasis being orthonormal with respect to a weighted inner product.

The detailed balance condition has a number of very pleasant implications. First, the transition matrix becomes easier to grasp in terms of physical intuition because one is relieved from the necessity to find a physical interpretation for complex eigenvectors and eigenvalues. Second, since the eigenvectors of a detailed balanced and irreducible transition matrix $\mathbf{T}(\tau)$ form a complete basis of $\mathbb{R}^{N_d}$, where N_d is the dimension of the transition matrix, any vector $\mathbf{p}(t)$ can be expressed as a linear combination of these eigenvectors:

$$\mathbf{p}(t) = \sum_\alpha k_\alpha(t)\, \boldsymbol{\psi}_\alpha = \sum_\alpha c_\alpha\, \lambda_\alpha(t)\, \boldsymbol{\psi}_\alpha \tag{18}$$

After time $n\tau$, $\mathbf{p}(t + n\tau)$ is given as

$$\mathbf{T}(n\tau)\, \mathbf{p}(t) = \mathbf{T}^n(\tau)\, \mathbf{p}(t) = \mathbf{p}(t + n\tau) \tag{19}$$

Using

$$\mathbf{p}(t + n\tau) = \sum_\alpha c_\alpha\, \lambda_\alpha^n(t)\, \boldsymbol{\psi}_\alpha \tag{20}$$

the probability distribution $\mathbf{p}(t)$ can be interpreted as consisting of modes $\{\boldsymbol{\psi}_\alpha\}$ which show a temporal behavior according to the corresponding eigenvalues $\{\lambda_\alpha\}$. More precisely, the temporal behavior is an exponential decay, as can be seen from the following transformation:

$$\begin{aligned} \lambda_\alpha(t = n\tau) = \lambda_\alpha^n(\tau) &= \lambda_\alpha^{t/\tau}(\tau) = \exp(\ln(\lambda_\alpha^{t/\tau}(\tau))) \\ &= \exp\left(\frac{t}{\tau}\ln(\lambda_\alpha(\tau))\right) = \exp\left(-\frac{t}{\mu_\alpha}\right) \end{aligned} \tag{21}$$

where $t = n\tau$ and the mean lifetime μ_α is given as

$$\mu_\alpha = -\frac{\tau}{\ln(\lambda_\alpha(\tau))} \tag{22}$$

In the context of Markov models, μ_α is typically referred to as the *implied time scale* of the decay process. The mode that corresponds to the eigenvalue $\lambda_1 = 1$ does not decay, which can be seen either by realizing that $\lambda_1^n = \lambda_1 = 1$ for any n or by noting that the argument in the exponential in eq 21 becomes 0 if $\lambda_\alpha = 1$. This result corresponds to the earlier result that the eigenvector associated with λ_1 is the equilibrium distribution: $\boldsymbol{\psi}_1 = \boldsymbol{\pi}$. The

other modes, which all correspond to an eigenvalue with $|\lambda_\alpha| < 1$, will vanish for $n \to \infty$. The smaller the value of λ_α, the faster the corresponding mode will decay.

The derivation of the implied time scales has been based on the assumption that $\mathbf{T}(\tau)$ represents a Markov process. In this case, the eigenvector expansion can be based on $\mathbf{T}(\tau)$ or any other $\mathbf{T}(n\tau)$ and will lead to the same implied time scale, i.e.,

$$\mu_\alpha = \frac{-n\tau}{|\lambda_\alpha(n\tau)|} = \text{const} \quad n = 1, 2, ... \tag{23}$$

Conversely, eq 23 can be used as a check for Markovian behavior. Plotting the implied time scales of transition matrices with various lag times $n\tau$ yields a set of constant functions if the underlying dynamics are Markovian.[7,8,11]

In the limit $\tau \to \infty$, all eigenmodes except for the stationary one have decayed, and there must be a matrix

$$\mathbf{T}_1 \doteq \lim_{\tau \to \infty} \mathbf{T}(\tau) \tag{24}$$

which immediately returns the equilibrium distribution $\boldsymbol{\psi}_1 = \boldsymbol{\pi}$ when multiplied by any arbitrary distribution $\mathbf{p}(t)$, i.e.

$$\mathbf{T}_1 \mathbf{p}(t) = \boldsymbol{\pi} \tag{25}$$

We will call the matrix $\mathbf{T}_1$ the *equilibrium matrix*.

3. MODELS

3.1. Bit-Flip Model. The coupling between the environment and a system can be studied on a simple bit-flip model consisting of two bits, S and E. Bit S represents the system and bit E the environment. Either of these bits can assume two states: ↑ ("up") or ↓ ("down"). The time scale of the dynamics of these two bits is not the same and is determined by their flipping probabilities p_S and p_E, respectively. The two bits can be coupled or uncoupled.

The complete system (consisting of S and E) has four states: ↑↑ (state 1), ↑↓ (state 2), ↓↑ (state 3), and ↓↓ (state 4), where the first arrow stands for bit S and the second for bit E. The dynamics of the complete system are modeled using a 4×4 transition matrix $\mathbf{T}_{\text{bit}}(\tau)$ which contains the transition probabilities between those four states. To obtain a transition matrix which only represents the dynamics of S, we project the matrix $\mathbf{T}_{\text{bit}}$ onto the states of bit S. This procedure intrinsically assumes that the dynamics of S are Markovian, where we identified S with $\mathbf{x}$ and E with $\mathbf{y}$. Using the implied time scales as a meausure for Markovian behavior, we can study how a coupling between E and S violates this assumption.

The transition matrix of the complete system dynamics $\mathbf{T}_{\text{bit}}(\tau)$ is constructed in the following fashion. The probability that the bit S will flip, i.e., that it will make a transition ↑→↓ or a transition ↓→↑, within time τ is given by p_S, and the probability that it will stay in its current state is given as $1 - p_S$. Consequently, the transition matrix for a single bit S (bit E not present) is given as

$$\mathbf{T}_S(\tau) = \begin{pmatrix} 1 - p_S & p_S \\ p_S & 1 - p_S \end{pmatrix} \tag{26}$$

where $\mathbf{T}_{S,11}$ represents the transition probability for ↑→↑, $\mathbf{T}_{S,21}$ the transition probability for ↑→↓, $\mathbf{T}_{S,12}$ the transition probability for ↓→↑, and $\mathbf{T}_{S,22}$ the transition probability for ↓→↓. Likewise the transition matrix for a single bit E (bit S not present) is given as

$$\mathbf{T}_E(\tau) = \begin{pmatrix} 1 - p_E & p_E \\ p_E & 1 - p_E \end{pmatrix} \tag{27}$$

where p_E denotes the probability that bit E will flip within τ. The transition for the complete system consisting of the two noninteracting bits is given by the Kronecker product (sometimes called tensor product) of $\mathbf{T}_S$ and $\mathbf{T}_E$:

$$\mathbf{T}_{\text{bit}}(\tau) = \mathbf{T}_S(\tau) \otimes \mathbf{T}_E(\tau)$$
$$= \begin{pmatrix} (1-p_S)(1-p_E) & (1-p_S)p_E & p_S(1-p_E) & p_S p_E \\ (1-p_S)p_E & (1-p_S)(1-p_E) & p_S p_E & p_S(1-p_E) \\ p_S(1-p_E) & p_S p_E & (1-p_S)(1-p_E) & (1-p_S)p_E \\ p_S p_E & p_S(1-p_E) & (1-p_S)p_E & (1-p_S)(1-p_E) \end{pmatrix} \tag{28}$$

This matrix represents transitions between states ↑↑ (state 1), ↑↓ (state 2), ↓↑ (state 3), and ↓↓ (state 4) where the first arrow stands for bit S and the second for bit E. A coupling is introduced into the dynamics of the complete system by selectively modifying elements of $\mathbf{T}_{\text{bit}}(\tau)$ and renormalizing its columns.

In order to obtain the transition matrix of bit S, one needs to project $\mathbf{T}_{\text{bit}}(\tau)$ onto the state of bit S using a projection

$$\mathbf{T}_{S,\text{proj}}(\tau) = \mathbf{P}^T \mathbf{T}_{\text{bit}}(\tau) \mathbf{P} \tag{29}$$

with

$$\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \tag{30}$$

Note that if $\mathbf{T}_{\text{bit}}$ represents the dynamics of the two uncoupled bits, i.e., if $\mathbf{T}_{\text{bit}}(\tau) = \mathbf{T}_S(\tau) \otimes \mathbf{T}_E(\tau)$, then eq 29 recovers $\mathbf{T}_S(\tau)$.

In the present application of the bit-flip model, a coupling between the environment E and the system S was introduced by multiplying the elements T_{11}, T_{14}, T_{41}, and T_{44} of $\mathbf{T}_{\text{bit}}$ by a (positive) coupling factor k and renormalizing the columns. This increases the probability of states in which the two spins are aligned (↑↑ and ↓↓). The coupling constant k was varied between 1 (no coupling) and 100 (strong coupling). The flipping probability of system S was set to $p_S = 0.100$. In order to examine the influence of the relaxation time of the environment on the system, the flipping probability of E, p_E, was varied between 0.001 (very slow dynamics, long relaxation time) and 0.150 (dynamics of the environment faster than the dynamics of the system, short relaxation time).

3.2. Butane Model. The other test system is liquid butane. We performed molecular dynamics (MD) simulations of boxes of 512 butane molecules at various temperatures ($T = 298.15$ K and $T = 400.00$ K) and densities (d: 50−500 u/nm^3, where u denotes the atomic mass unit). For each density, a box with regularly placed butane molecules was constructed using the program build_box of GROMOS++,[35] which was then heated to the target temperature over a period of 10^5 time steps (200 ps). For each system, a trajectory of 2 ns was generated using the GROMOS05 software,[35] which implements the leapfrog intergrator,[36] and the GROMOS 45A3 force field.[37] All bond lengths were constrained using the SHAKE algorithm[38] with a relative tolerance of 10^{-4}, allowing for a time step of 2 fs. Configurations of all 512 molecules were saved every 0.08 ps. The system was simulated in a rectangular box using periodic boundary conditions. The volume was kept constant, and the molecules were weakly coupled to one temperature bath of 298.15 K or 400.00 K[39] with a coupling time of 0.1 ps. We used

Table 1. Overview of the Simulations Performed

	MD setup				SD setup		
T (K)	density, ρ (u/nm^3)	number of molecules	simulation length (ns)	D (σ) (10^{-2}nm^2/ps)	γ_{friction} (1/ps)	number of molecules	simulation length (μs)
298.15	300	512	2	1.403(1.224)	5.0	1	1
298.15	345	512	2	0.7113(0.5612)	10.0	1	1
298.15	400	512	2	0.3952(0.2815)	17.9	1	1
298.15	450	512	2	0.1981(0.2009)	35.8	1	1
298.15	500	512	2	0.0561(0.0469)	126.3	1	1
400.00	50	512	2	18.18(16.31)	0.5	1	1
400.00	100	512	2	8.904(7.293)	1.1	1	1
400.00	150	512	2	5.273(5.009)	1.8	1	1
400.00	200	512	2	4.075(3.610)	2.3	1	1
400.00	250	512	2	2.630(2.502)	3.6	1	1
400.00	300	512	2	1.501(1.446)	6.3	1	1
400.00	345	512	2	1.321(1.134)	7.2	1	1

0.8 nm/1.4 nm as a twin-range cutoff and 1.4 nm as a reaction field cutoff with ε_{rf} = 1.0. The atom pair list for short-range interactions and the intermediate-range forces were updated every five steps.

We performed stochastic dynamics (SD) simulations of a single butane molecule at two different temperatures, 298.15 K and 400.00 K, using various friction coefficients, γ_{fric}: 0.5−126.3 ps^{-1}. The friction coefficients were calculated from the diffusion constants obtained from the above-described MD simulations using the relation:

$$\gamma_{\text{fric}} = \frac{k_B T}{D m_{\text{solv}}} \tag{31}$$

where k_B is the Boltzmann constant, T is the temperature of the MD simulation, D is the diffusion constant, and m_{solv} is the mass of butane (58.124 g/mol). The diffusion constant was calculated as an average of the diffusion constants of 50 arbitrarily picked butane molecules where each diffusion constant was estimated from a least-squares fit to the Einstein equation:

$$D = \lim_{t \to \infty} \frac{\langle [\mathbf{r}_0 - \mathbf{r}(t)]^2 \rangle}{2Nt} \tag{32}$$

Here, $\mathbf{r}_0$ is the center of geometry of the first configuration in the trajectory, $\mathbf{r}(t)$ is the center of geometry at time t, and N is the number of dimensions taken into account, which was set to 3.

As in the MD simulations, all bonds were constrained using the SHAKE algorithm,[38] and a time step of 2 fs was used. Each system was simulated for 1 μs, and the configuration of the molecule was saved every 0.1 ps. Vacuum boundary conditions were applied, and the temperature was maintained by the stochastic dynamic integrator.

A summary of all performed simulations, the obtained diffusion constants, and the corresponding friction coefficients is reported in Table 1 .

3.3. Generation of the Transition Matrices T(τ) for the Test System Butane. We consider a single butane immersed in a solvent of butane molecules where the solvent is modeled either explicitly using MD or implicitly using SD. The dominant degree of freedom of butane is the $C_1-C_2-C_3-C_4$-dihedral angle, which we discretize into equally sized microstates (bins). For most analyses, we use a bin size of 5° (72 bins per dihedral angle); only in Figure 6, we varied the bin size from 5° to 120° (72 to 3 bins per dihedral angle). For various values of the lag time τ (ranging from 80 fs to 100 ps), the configurations at time t = 0, τ, 2τ, 3τ, etc. are mapped onto the microstates, and the transitions from microstate i to microstate j for each combination of i and j are counted. We enforce detailed balance by counting each transition $i \rightarrow j$ also as a transition $j \rightarrow i$. This "backward counting" inherently assumes that the trajectory is in global equilibrium and the deviation from detailed balance is only due to statistical errors. The numbers of transitions are sorted into a matrix, and the columns of the matrix are normalized by the total number of transitions in each column to obtain the column-stochastic transition matrix. When constructing the transition matrix from a MD trajectory, we regard one butane molecule as solute and the remaining 511 as solvent and count the transitions of the solute. Since the choice of the solvent molecule is arbitrary, we repeat this procedure 511 times, where in each round another molecule represents the solute. The transition matrix for a single butane is then constructed from the added transition counts of all 512 evaluations of the MD trajectory.

The implied time scales μ_i of each transition matrix $\mathbf{T}(\tau)$ are calculated as

$$\mu_i(\tau) = -\frac{\tau}{\ln |\lambda_i(\tau)|} \tag{33}$$

where τ is the time step of the transition matrix and $|\lambda_i|$ is the absolute value of its ith eigenvalue. We plot the implied time scales of the dominant eigenvalues and evaluate the reference implied time scales and τ_{Markov} by visual inspection.

4. RESULTS

We use two model systems—(i) a single butane immersed in a solvent of butane and (ii) a bit-flip model as described in section 3—to illustrate some important properties of transition matrices and to study the effect of marginal degrees of freedom on the dynamics of the relevant degrees of freedom.

4.1. Colormaps of Transition Matrices. Figures 1 and 2 show colormaps of various transition matrices of the dihedral angle degree of freedom of butane. The dihedral angle has been discretized into 72 microstates of 5° per microstate, and each point in the colormaps represents a transition probability T_{ij} from microstate j to microstate i. A high transition probability is marked in red, and a transition probability close to zero is marked

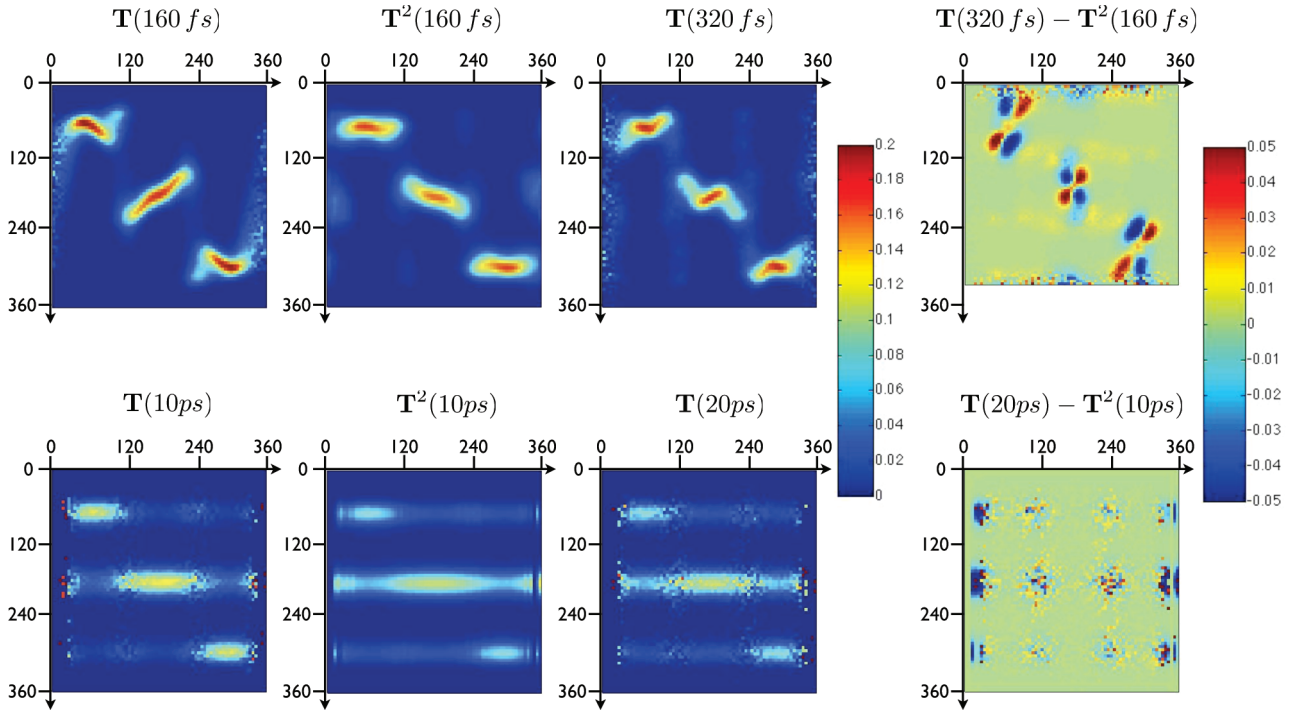

Figure 1. Transition matrices for the dihedral angle of a single butane immersed in butane (T = 298.15 K, ρ = 345 u/nm^3). Upper row: non-Markovian regime, τ = 160 fs and τ = 320 fs. Lower row: Markovian regime, τ = 10 ps and τ = 20 ps. Right-most column: diffence plots of $\mathbf{T}(2\tau) - \mathbf{T}^2(\tau)$.

in blue. The right-most column in Figure 1 and the lower row in Figure 2 show colormaps of difference matrices in which negative elements are marked blue, elements close to zero are green, and positive elements are red. Taking the first matrix in the upper row of Figure 1 as an example, one can clearly distinguish the three metastable states of butane as three red areas along the diagonal of the matrix. For any initial microstate in the *gauche* state of butane between 0° and 120°, the molecule has a high transition probability to any microstate within this state and a low transition probability to any microstate outside this state within lag time τ. The microstates between 0° and 120° are said to be "kinetically close". Analogously, microstates which correspond to the *trans* state (120−240°) are kinetically close, as are microstates which correspond to the second *gauche* state (120−360°).

4.2. Illustration of the Chapman−Kolmogorov Equation. Equation 9, according to which taking a transition matrix with lag time τ to the power n yields a matrix which is equal to a transition matrix with n times longer lag time if the dynamics are Markovian, is illustrated in Figure 1 using transition matrices of butane as an example. The first three columns show colormaps of transition matrices, and the fourth column shows colormaps of difference matrices.

Transition matrices with short lag times on the order of a few hundred femtoseconds are depicted in the upper row. The second matrix is the square of the first one with lag time τ = 160 fs and should be equal to the third matrix, which has been constructed with a lag time of $\tau = 2 \times 160$ fs = 320 fs if the dynamics are Markovian at this time resolution of the model. This is clearly not the case, as the second and the third matrix already visually differ from each other. The difference matrix correspondingly shows systematic deviations from zero. If one was to evolve a density with $\mathbf{T}^2$(160 fs), its dynamics would systematically deviate from the dynamics of the same density evolved with $\mathbf{T}$(320 fs).

The lower row shows transition matrices with longer lag times on the order of 10 ps. In this time regime, the dynamics of the dihedral angle can be approximated by a Markov process. $\mathbf{T}^2$(10 ps) and $\mathbf{T}$(20 ps) are visually similar, except for the fact that $\mathbf{T}$(20 ps) shows more noise than $\mathbf{T}^2$(10 ps). This is due to the poorer sampling for longer lag times. Accordingly, the difference matrix depicted in the right-most column shows no systematic deviations from zero but only deviations which are due to the noise in the two matrices. Note that the amplitude of noise varies with the magnitude of the transition probabilities in $\mathbf{T}^2$(10 ps) and $\mathbf{T}$(20 ps).

4.3. Illustration of the Equilibrium Matrix. Figure 2 illustrates the concept of the equilibrium matrix $\mathbf{T}_1$ defined in eq 25 using transition matrices of a butane molecule with a lag time of 5 ps as an example. For this lag time, the dynamics of the system are Markovian and the equilibrium matrix $\mathbf{T}_1$ can be constructed from the first eigenvector of the transition matrix, which is equal to the equilibrium distribution. It is depicted in the right-most panel, and as its columns are all equal to the equilibrium distribution, its colormap shows a striped pattern. When multiplied by an arbitrary initial distribution, it returns the equilibrium distribution. $\mathbf{T}_1$ does not contain any information on the kinetic proximity of groups of microstates, and metastable states cannot be extracted from this matrix.

The other two panels in the upper row of Figure 2 show the transition matrix with lag time 5 ps, $\mathbf{T}$(5 ps), and the square of it, $\mathbf{T}^2$(5 ps), which is approximately equal to $\mathbf{T}$(10 ps). At lag time τ = 5 ps (left-most panel), the three metastable states of butane are clearly discernible, implying that the equilibration time of the system is longer than 5 ps. In the middle panel, the metastable states are less discernible, and the stripe pattern emerges. At time $t = \tau$ = 10 ps, the probability of finding the system in any of the three metastable states is still slightly biased toward finding it in its initial state. But the information about the kinetic proximity of the microstates contained in the matrix is less clear. The lower

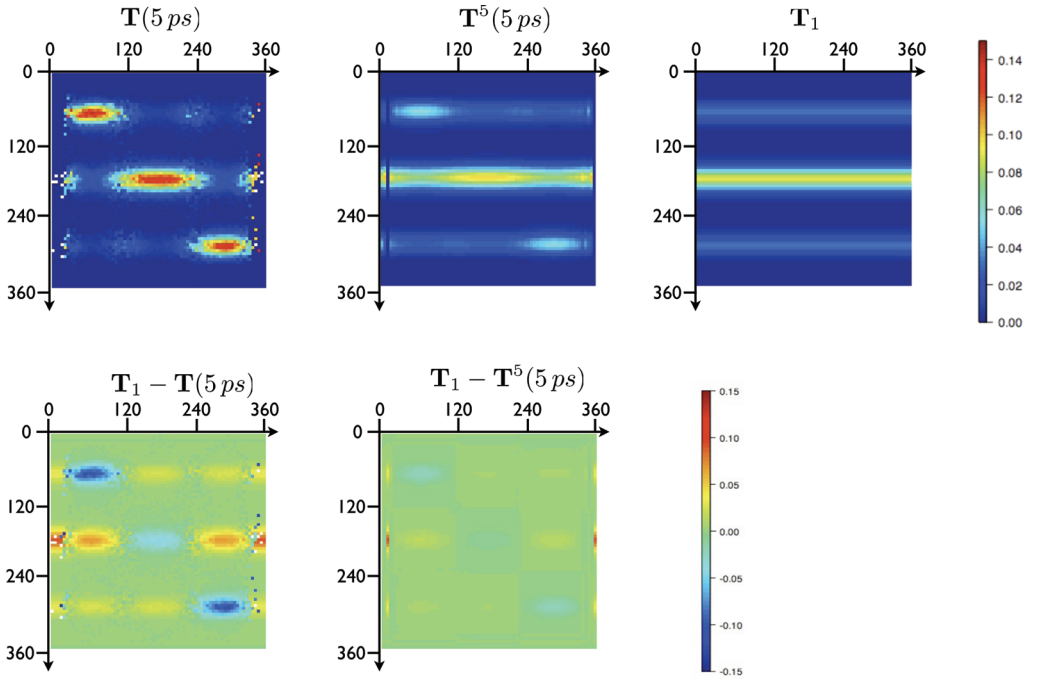

Figure 2. Transition matrices for the dihedral angle of a single butane immersed in butane (T = 298.15 K, ρ = 345 u/nm^3). Upper row: $\mathbf{T}$(5 ps), $\mathbf{T}^5$(5 ps), and equilibrium matrix $\mathbf{T}_1$ constructed from the first eigenvector ψ_1 of $\mathbf{T}$(5 ps). Lower row: difference matrices between the equilibrium matrix and $\mathbf{T}$(5 ps) and $\mathbf{T}^5$(5 ps), respectively.

row shows the difference matrices $\mathbf{T}_1-\mathbf{T}$(5 ps) and $\mathbf{T}_1-\mathbf{T}^2$ (5 ps). For τ = 5 ps (left most panel), the transition matrix shows large systematic deviations from the equilibrium matrix. For τ = 10 ps (middle panel), we see the same deviations which are, however, much smaller in absolute value.

4.4. Coupling of Marginal and Relevant Degrees of Freedom. When constructing Markov models from MD simulation data, the complete phase space is split into relevant degrees of freedom for which the model is constructed and marginal degrees of freedom which are assumed to act as stochastic forces on the relevant degrees of freedom. Depending on the time scale of the dynamics of the marginal degrees of freedom and the strength of the coupling between the marginal and the relevant degrees of freedom, this assumption can be fulfilled to a greater or lesser extent. In the bit-flip model, the relevant degrees of freedom are modeled by the bit S and the marginal degrees of freedom by the bit E. The time scale of the dynamics of E is determined by the flipping probability p_E: the higher the p_E, the faster the dynamics. The strength of the coupling is determined by the coupling constant k. In all applications of the bit-flip model presented here, the flipping probability of the system, p_S, was set to 0.100.

In Figure 3, the influence of the flipping probability on the implied time scale of the second eigenvalue of $\mathbf{T}_{S,\mathrm{proj}}$ is illustrated. The brown curve (p_E = 0.150) represents the case in which the dynamics of the environment is faster than the dynamics of the system. The implied time scale of the projected matrix rises until it reaches a plateau at about n = 150. This is the threshold in time resolution τ_{Markov} after which the dynamics of the system are Markovian until, after n = 750, the curve diverges again from a constant implied time scale. The latter deviation is caused by the fact that for a very high number of iterations the transition matrix approaches the equilibrium matrix, and the second eigenvector becomes so small that it is susceptible to numerical errors. Note that the system S has a flipping probability of 1−10 within a time

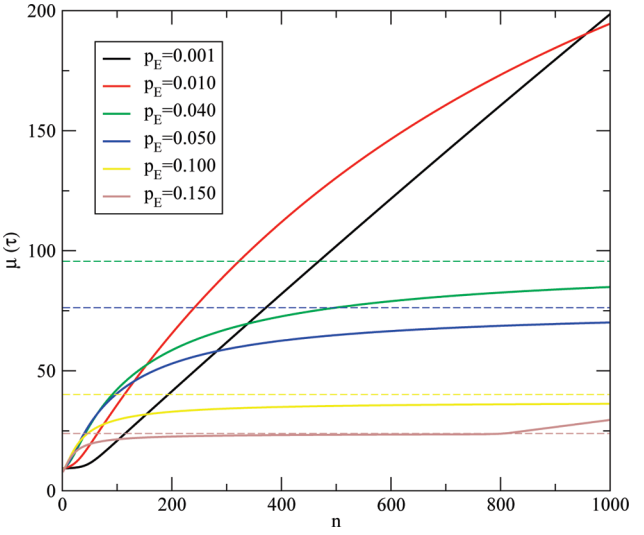

Figure 3. Implied time scale μ of the bit-flip model as a function of lag time $n\tau$ calculated from the second eigenvalue of $\mathbf{T}_{S,\mathrm{proj}}(n\tau)$ as a function of n where $\mathbf{T}_{\mathrm{bit}}$ has been constructed with varying the flipping probability, p_E, of E. Coupling constant k = 100. Flipping probability of the system p_S = 0.100. Thin dashed lines: true implied time scales, μ_2, as calculated from the respective $\mathbf{T}_{\mathrm{bit}}$. For p_E = 0.001, μ_2 = 3828.7, and for p_E = 0.010, μ_2 = 383.5, which is well beyond the region shown.

step τ. If the dynamics of the system are modeled with a time resolution of 150τ, a single iteration does not yield the probability of a *single* transition between the two states but the probability of finding the system in one of the states after a *sequence* of transitions. Each transition in this sequence has occurred under the influence of the environment; i.e., it was not Markovian, but the long lag time of the projected model provides for an ensemble of transitions in which the influence of the environment averages

dx.doi.org/10.1021/ct200069c |J. Chem. Theory Comput. 2011, 7, 1032–1044

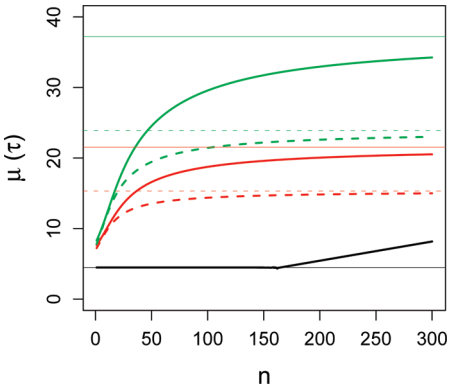

Figure 4. Implied time scale μ of the bit-flip model as a function of the lag time $n\tau$ calculated from the second eigenvalue of $\mathbf{T}_{S,\mathrm{proj}}(n\tau)$ as a function of n where $\mathbf{T}_{\mathrm{bit}}$ has been constructed with varying values of the coupling constant k and the flipping probability of the environment p_E. Flipping probability of the system $p_S = 0.100$. Solid black: $k = 1$, $p_E = 0.100$. Dashed black: $k = 1$, $p_E = 0.150$. Solid red: $k = 10$, $p_E = 0.100$. Dashed red: $k = 10$, $p_E = 0.150$. Solid green: $k = 100$, $p_E = 0.100$. Dashed green: $k = 100$, $p_E = 0.150$. The solid black and dashed black lines are on top of each other. Thin lines: corresponding true implied time scales μ_2 as calculated from the respective $\mathbf{T}_{\mathrm{bit}}$.

out. The faster the dynamics of the environment, the smaller the sequence of transitions has to be until the influence of the environment is averaged out, i.e., the smaller τ_{Markov} will be. Unless the time scale of the environment and the time scale of the system differ by orders of magnitude, a Markov model of the system dynamics does not represent the probability of a single transition within lag time τ_{Markov} but the probability of finding the system in state j at time $t = \tau_{\mathrm{Markov}}$, given that it started its sequence of transitions in state i at time $t = 0$.

The yellow curve in Figure 3 represents the case in which the dynamics of the environment have the same time scale as the dynamics of the system. The implied time scale becomes approximately constant after 250 iterations, at which time the transition matrix is, however, so close to the equilibrium matrix that the model does not contain any significant information on the dynamics of the system. Similarly, the green ($p_E = 0.040$) and the blue ($p_E = 0.050$) curves represent cases in which the dynamics of the environment have a time scale which is on the same order of magnitude as the time scale of the system but slightly larger. In this case, the implied time scale curves slowly level off but never reach a plateau region. If one encounters this type of behavior when constructing a Markov model from MD simulation data, one should consider to include more degrees of freedom into the Markov model.

Finally, the black curve ($p_E = 0.001$) and the red curve ($p_E = 0.010$) in Figure 3 represent the case in which the dynamics of the environment are much slower than the dynamics of the system. For few iterations of the complete transition matrix, $n < 30$ and $n < 10$, respectively, the environment has hardly changed, and therefore, the system does not feel its influence. The implied time scales are constant at about a value of 10. After about 10−30 iterations, the environment has changed noticeably from its state at $t = 0$ and starts to influence the dynamics of the system. However, because the dynamics of the environment are so slow, even 1000 iterations are not sufficient to provide enough statistics to average out the influence of the environment on the system, and the curve never reaches a plateau region (data not shown).

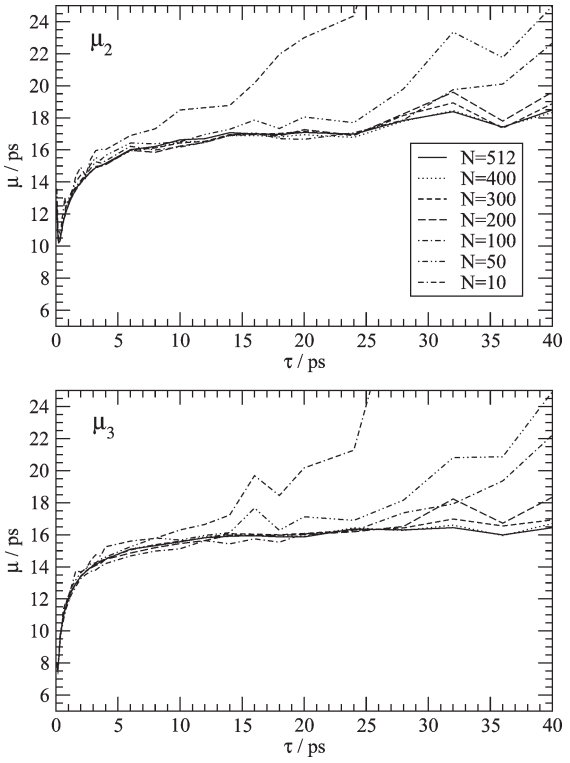

Figure 5. Implied time scale as a function of lag time τ for various numbers of data points used in the analysis. Implied time scales μ_2 and μ_3 calculated from the MD simulation of one butane molecule immersed in liquid butane at $T = 298.15$ K and $\rho = 345$ u/nm^3. The number of molecules N used for the analysis of a total of 512 molecules varied from 10 to 512.

Figure 4 shows the influence of the coupling constant on the dynamics of the system for two time scales of the environment $p_E = 0.100$ and $p_E = 0.150$. If there is no coupling, i.e., if $k = 1$, the dynamics of the system are independent of the environment, and consequently, the implied time scales are constant (black curves). As before, the deviation from constant μ after $n = 150$ is due to numerical errors. Since the transition probabilities are independent of the state of the environment, the implied time scales are also independent from the flipping probability of E. Both have a value of 4.48. Raising the coupling constants to $k = 10$ and $k = 100$ changes the implied time scale of the system, in this case, raising it. We note, however, that this might be caused by the choice of matrix elements which are modified by k. τ_{Markov} is larger the stronger the coupling between the environment and the system is. For a flipping probability of $p_E = 0.100$, $\tau_{\mathrm{Markov}} \approx 100\tau$ for $k = 10$ and $\tau_{\mathrm{Markov}} \approx 150\tau$ for $k = 100$. For a flipping probability of $p_E = 0.150$, $\tau_{\mathrm{Markov}} \approx 170\tau$ for $k = 10$ and $\tau_{\mathrm{Markov}} > 300\tau$ for $k = 100$.

4.5. Behavior of Liquid Butane. As a test system, we consider a single butane immersed in a solvent of butane at two temperatures, $T = 298.15$ K and $T = 400.00$ K, and various densities. In one set of simulations, we model the solvent explicitly with 511 butane molecules; in the other set of simulations, we model the solvent implicitly using stochastic dynamics. The dominant degree of freedom in butane is the dihedral angle between its carbon atoms $C_1-C_2-C_3-C_4$, which we use for the construction of the Markov model. All other degrees of freedom (bond-angle vibrations and solvent degrees of freedom in MD) are marginal in the model and are assumed to interact stochastically

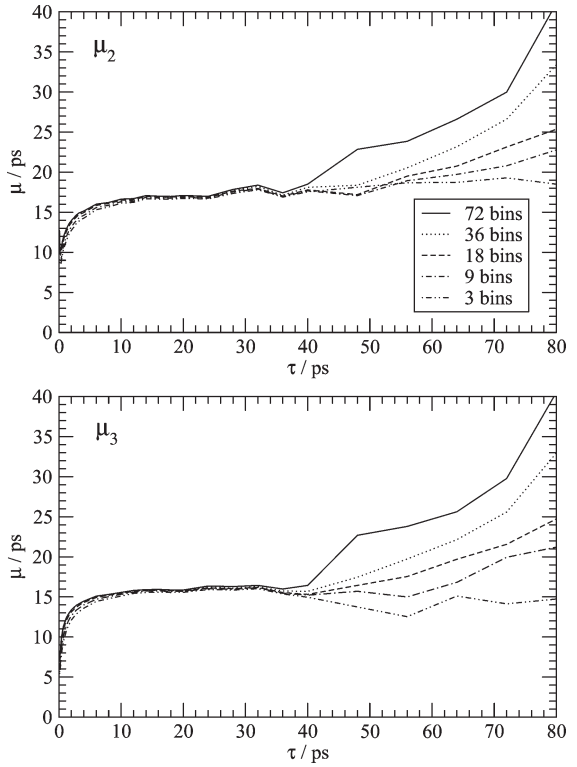

Figure 6. Implied time scale as a function of lag time τ for various resolutions of the configuration space. Implied time scales μ_2 and μ_3 calculated from the MD simulation of 1 butane molecule immersed in liquid butane at T = 298.15 K and ρ = 345 u/nm^3. The number of microstates (bins per dihedral angle) varied from 3 to 72.

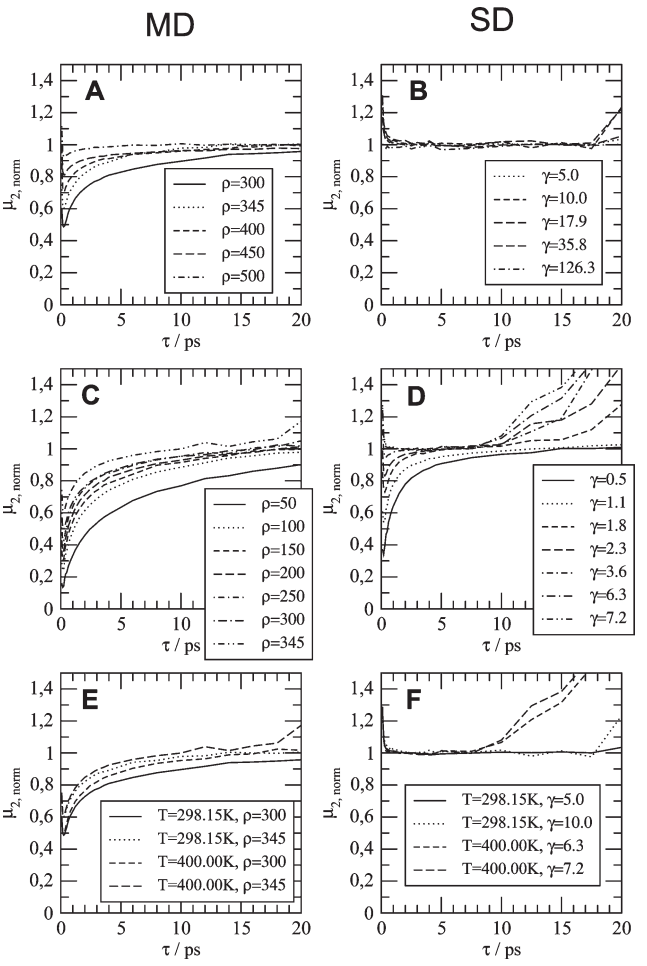

Figure 7. Normalized implied time scale $\mu_{2,\mathrm{norm}}$ (eq 34) as function of the lag time τ for various systems. Left column: MD simulations of 1 butane molecule immersed in liquid butane. Right column: SD simulations of 1 butane immersed in liquid butane modeled by the friction coefficient γ. Panel A: T = 298.15 K, density ρ varied from 300 u/nm^3 to 500 u/nm^3. Panel B: T = 298.15 K, γ varied from 5.0 ps^{-1} to 126.3 ps^{-1}. Panel C: T = 400.00 K, ρ varied from 50 u/nm^3 to 345 u/nm^3. Panel D: T = 400.00 K, γ varied from 0.5 ps^{-1} to 7.2 ps^{-1}. Panel E: Temperature and density varied. Panel F: Temperature and density varied.

with the dihedral-angle degree of freedom. The molecule has three metastable states, represented by the *gauche*+, *trans*, and *gauche*− conformation of the dihedral angle. Correspondingly, it has two dominant eigenvalues, λ_2 and λ_3 ($\lambda_1 = 1$), which we used to calculate implied time scales and to determine τ_{Markov}.

When constructing a Markov model from simulation data, the upper bound of a possible lag time is not set by the numerical accuracy with which the eigenvalue can be calculated for a transition matrix approaching the equilibrium matrix but by the extent of the sampling. Because data points are evaluated at t = 0, τ, 2τ, etc., the longer the τ, the fewer data points are available in a trajectory of a given length. Figure 5 illustrates this fact. The first panel shows the implied time scale of the second eigenvalue, and the second panel shows it for the third eigenvalue calculated from MD simulations at T = 298.15 K and a density of ρ = 345 u/nm^3. We have varied the number of data points, N, used for the construction of the Markov model by varying the number of times the MD trajectory is evaluated, where at each evaluation a different butane molecule was considered to be the solute. τ_{Markov} is not influenced by the amount of data the Markov model is built upon. It lies between τ = 5 ps and τ = 10 ps. However, the length of the plateau region is sensitive to the amount of data. The less data used, the smaller the lag time for which the implied time scales diverge from the plateau. In particular, if the trajectory is evaluated only 10 times, the implied time scales diverge before τ_{Markov} is reached.

Figure 6 illustrates how the resolution of the relevant degrees of freedom influences the implied time scale. For small lag times up to 40 ps, there is only a very small but systematic influence of the resolution of the discretization on the implied time scale curves that is discernible. More precisely, τ_{Markov} is only slightly smaller for finer discretizations. The figure also illustrates the effect of insufficient sampling of the transition probabilities. For lag times of τ = 40 ps and greater, the number of available data points becomes so small that the statistical error on the transition probabilities is too large to yield a reliable Markov model. Consequently, the curves diverge from the plateau. The effect is the greater, the finer the resolution of the model. In practice, one can improve the sampling by using a sliding window, i.e., by counting the transition from *every* time step in the simulation to the time step τ further, instead of using only the time steps at 0, τ, 2τ, 3τ, etc., as done here.

We note that our discretization is special in two aspects. First, each metastable state corresponds to a single minimum in the free-energy surface and does not consist of several substates. If by lowering the resolution of the discretization, several states are grouped into one microstate, the lag time at which the model becomes Markovian does change.[11] Second, we ensured that

Table 2. Overview of the Implied Time Scales of the Second and Third Eigenvalues Observed in the Markovian Regime of the Dynamics for Various Temperatures and Densities[a]

system			MD		SD	
T (K)	density (u/nm^3)	γ_{friction} (ps^{-1})	$\mu_{2,\text{reference}}$ (ps)	$\mu_{3,\text{reference}}$ (ps)	$\mu_{2,\text{reference}}$ (ps)	$\mu_{3,\text{reference}}$ (ps)
298.15	300	5.0	21.0	20.2	10.2	8.0
298.15	345	10.0	17.0	16.2	11.2	7.2
298.15	400	17.9	15.0	12.4	11.8	7.2
298.15	450	35.8	13.4	9.8	17.4	10.0
298.15	500	126.3	12.8	8.8	43.4	26.0
400.00	50	0.5	30.0	28.0	12.6	10.6
400.00	100	1.1	16.5	16.0	7.6	6.3
400.00	150	1.8	13.0	12.6	5.4	5.2
400.00	200	2.3	11.0	10.8	4.6	4.6
400.00	250	3.6	9.2	9.0	4.0	3.6
400.00	300	6.3	7.8	7.2	3.8	2.8
400.00	345	7.2	6.4	5.8	3.8	2.8

[a] Column MD: explicit solvent model. Column SD: implicit solvent model.

there is a microstate boundary exactly at the peak of the free-energy barrier between the metastable states. Moving this boundary away from the barrier peak will decrease the quality of the Markov model.[9,26] The error introduced by discretizing the relevant degrees of freedom has, however, a finite upper bound.[9]

In the butane test system, the marginal degrees of freedom are predominantly those of the solvent molecules, exceptions being the bond-angle degrees of freedom. Their coupling to the relevant degree of freedom, the dihedral angle, is determined by the model of the solvent, implicit or explicit; the density; and the temperature. Their influence on τ_{Markov} is examined in Figure 7.

Since model, density, and temperature do not only influence τ_{Markov} but also the implied time scales used to determine τ_{Markov}, we introduce a normalized implied time scale:

$$\mu_{i,\,\text{norm}}(\tau) = \frac{\mu_i(\tau)}{\mu_{i,\,\text{reference}}} \tag{34}$$

where $\mu_i(\tau)$ indicates the implied time scale of the ith eigenvalue and $\mu_{i,\text{reference}}$ indicates the reference implied time scale, i.e., the time scale in the Markovian regime, which we determine by visual inspection. Table 2 lists the observed reference implied time scales. The column "MD" corresponds to an explicit solvent model; the column "SD" to an implicit one. The fact that stochastic dynamics (SD) underestimate the relaxation times of a system, i.e., underestimates the implied time scales, if the fundamental assumption underlying this type of dynamics, a large heavy particle in a solvent of small light particles, is not fulfilled, is a known effect. The expectation that the system equilibrates quicker if the temperature or the density is increased is reflected in the corresponding decrease of the implied time scale. The only exceptions to this trend are the simulations with very high friction coefficients ($\gamma_{\text{friction}} = 35.8$ ps^{-1} and $\gamma_{\text{friction}} = 126.3$ ps^{-1}). In these cases, the velocity of the dihedral-angle degree of freedom is decreased so drastically at each simulation step that transitions between the metastable states are very rare. Consequently, the equilibration between these states is slow, and the implied time scales are large.

In all three rows of Figure 7, the Markovian regime is reached much earlier for the implicit solvent model than for the explicit solvent model; i.e., the influence of the marginal degrees of freedom vanishes more quickly. This can be explained by the fact that the set of marginal degrees of freedom is much smaller in transition matrices constructed from stochastic dynamics simulations. It only consists of the bond-angle degrees of freedom which equilibrate faster than the solvation shell in an explicit solvent model. Note that the emulation of the solvent by friction coefficients and stochastic kicks is by definition Markovian. In contrast to the simulations with an explicit solvent model, some of the curves in the right column of Figure 7 deviate already at small lag times from the constant regime, in particular those which correspond to small implied time scales in Table 2 . Two reasons for this are conceivable: (*i*) similar to in the bit-flip model, transition matrices with small implied time scales approach the equilibrium matrix so closely that the numerical error of the eigenvalue calculation is not negligible or (*ii*) poorly sampled transitions become more and more dominating as the modes corresponding to the second and third eigenvectors decay and cause a divergence from Markovian behavior. Panel D shows that τ_{Markov} decreases if the friction coefficient increases. The left column of Figure 7 shows how τ_{Markov} changes if the density and the temperature are varied in simulations with an explicit solvent model. Analogously to the results for stochastic dynamics, τ_{Markov} decreases if the density increases (panels A and C). At high density, the kicks among solvent molecules and among solvent molecules and the solute are more frequent than at low density, leading to a quicker equilibration of the marginal degrees of freedom. Intuitively, a high density corresponds to a high value of the flipping probability p_E in the bit-flip model. Likewise, τ_{Markov} decreases if the temperature increases (panel E). For the higher temperatures, the kicks among solvent molecules do not necessarily become more frequent but have a higher impact, which also speeds up the equilibration of the solvent degrees of freedom.

5. CONCLUSION

We have presented an overview of the assumptions which are made when mapping the equations of motion onto the central quantity in Markov models, the transition matrix. We have also reviewed the mathematical properties of transition matrices. Markov models are a powerful tool to describe the dynamics of the relevant degrees of freedom of a system provided that one

finds a partition of the degrees of freedom of the system into relevant and marginal degrees of freedom such that the marginal degrees of freedom are not strongly coupled to the relevant degrees of freedom and that the former equilibrate on much shorter time scales than the latter. For liquid butane, we find that the discretization of the relevant degrees of freedom, if the grid boundaries do not mask the free energy barriers, has only little influence on the time resolution τ_{Markov} for which the dynamics becomes Markovian. The number of data points which are used to construct the Markov model, on the other hand, has an influence on the range of lag times for which the model is Markovian: the smaller the number of data points, the earlier the system diverges from Markovian behavior.

AUTHOR INFORMATION

Corresponding Author

*Phone: +41 44 632 5501. Fax: +41 44 632 1039. E-mail: wfvgn@igc.phys.chem.ethz.ch.

ACKNOWLEDGMENT

Financial support by the National Centre of Competence in Research (NCCR; Structural Biology) and by a grant (number 200021-121913) from the Swiss National Science Foundation (SNSF) and by a grant (number 228076) from the European Research Council (ERC) is gratefully acknowledged.

REFERENCES

(1) van Gunsteren, W. F.; Berendsen, H. J. C. Algorithms for macromolecular dynamics and constraint dynamics. *Mol. Phys.* **1977**, *34*, 1311–1327.

(2) Berendsen, H. J. C.; van Gunsteren, W. F. In *Molecular-Dynamics Simulation of Statistical-Mechanical Systems*, Proceedings of the International School of Physics "Enrico Fermi", Varenna, Italy, 1985; Ciccotti, G., Hoover, W. H., Eds.; North-Holland: Amsterdam, 1986.

(3) Chandler, D.; Berne, B. J. Role of constraints on the conformational structure of *n*-butane in liquid solvents - Comment. *J. Chem. Phys.* **1979**, *71*, 5386–5387.

(4) van Gunsteren, W. F. Constrained dynamics of flexible molecules. *Mol. Phys.* **1980**, *40*, 1015–1019.

(5) van Gunsteren, W. F.; Karplus, M. Effect of constraints on the dynamics of macromolecules. *Macromolecules* **1982**, *15*, 1528–1544.

(6) van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*, 2nd ed.; Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1992.

(7) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.* **2007**, *126*, 155102.

(8) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* **2007**, *126*, 155101.

(9) Sarich, M; Noé, F; Schütte, C. On the approximation quality of Markov state models. *Multiscale Model. Simul.* **2010**, *8*, 1154.

(10) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Dill, K. A. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model. Simul.* **2006**, *5*, 1214–1226.

(11) Swope, W. C.; Pitera, J. W.; Suits, F. Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.

(12) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and beta-hairpin peptide. *J. Phys. Chem. B* **2004**, *108*, 6582–6594.

(13) Muff, S.; Caflisch, A. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a beta-sheet miniprotein. *Proteins: Struct. Funct. Bioinf.* **2008**, *70*, 1185–1195.

(14) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011–19016.

(15) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* **2009**, *131*, 124101.

(16) Vanden-Eijnden, E.; Venturoli, M. Markovian milestoning with Voronoi tessellations. *J. Chem. Phys.* **2009**, *130*, 194101.

(17) Keller, B.; Daura, X.; van Gunsteren, W. F. Comparing geometric and kinetic cluster algorithms for molecular simulation data. *J. Chem. Phys.* **2010**, *132*, 074110.

(18) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra Appl.* **2000**, *315*, 39–59.

(19) Deuflhard, P.; Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Appl.* **2005**, *389*, 161–184.

(20) Buchete, N. V.; Hummer, G. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B* **2008**, *112*, 6057–6069.

(21) Pan, A. C.; Roux, B. Building Markov state models along pathways to determine free energies and rates of transitions. *J. Chem. Phys.* **2008**, *129*, 064107.

(22) Buchete, N. V.; Hummer, G. Peptide folding kinetics from replica exchange molecular dynamics. *Phys. Rev. E* **2008**, *77*, 030902.

(23) Muff, S.; Caflisch, A. ETNA: Equilibrium transitions network and Arrhenius equation for extracting folding kinetics from REMD simulations. *J. Phys. Chem. B* **2009**, *113*, 3218–3226.

(24) Micheletti, C.; Bussi, G.; Laio, A. Optimal Langevin modeling of out-of-equilibrium molecular dynamics simulations. *J. Chem. Phys.* **2008**, *129*, 074105.

(25) Bowman, G. R.; Ensign, D. L.; Pande, V. S. Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. *J. Chem. Theory Comput.* **2010**, *6*, 787–794.

(26) Jensen, C. H.; Nerukh, D.; Glen, R. C. Sensitivity of peptide conformational dynamics on clustering of a classical molecular dynamics trajectory. *J. Chem. Phys.* **2008**, *128*, 115107.

(27) Singhal, N.; Pande, V. S. Error analysis and efficient sampling in Markovian state models for molecular dynamics. *J. Chem. Phys.* **2005**, *123*, 204909.

(28) Hinrichs, N. S.; Pande, V. S. Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics. *J. Chem. Phys.* **2007**, *126*, 244101.

(29) Noé, F. Probability distributions of molecular observables computed from Markov models. *J. Chem. Phys.* **2008**, *128*, 244103.

(30) Zwanzig, R. Nonlinear generalized Langevin equations. *J. Stat. Phys.* **1973**, *9*, 215–220.

(31) Frenkel, D.; Smit, B. *Understanding Molecular Simulation - From Algorithms to Applications*, 2nd ed.; Elsevier Academic Press: London, United Kingdom, 2002; Vol. 1 of Computational Science Series.

(32) Schwabl, F. *Statistische Mechanik*, 3rd ed.; Springer-Verlag: Berlin Heidelberg New York, 2004.

(33) MacCluer, C. R. The many proofs and applications of Perron's theorem. *SIAM Rev.* **2000**, *42*, 487–498.

(34) Deuflhard, P.; Andreas, P. *Numerical Analysis in Modern Scientific Computing*, 2nd ed.; Springer-Verlag: Berlin Heidelberg New York, 2003; Vol. 1 of Texts in Applied Mathematics.

(35) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.* **2005**, *26*, 1719–1751.

(36) Hockney, R. W. The potential calculation and some applications. *Meth. Comp. Phys.* **1970**, *9*, 136–210.

(37) Schuler, L. D.; Daura, X.; van Gunsteren, W. F. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.* **2001**, *22*, 1205–1218.

(38) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of Cartesian equations of motion of a system with constraints - molecular-dynamics of *n*-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

(39) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.