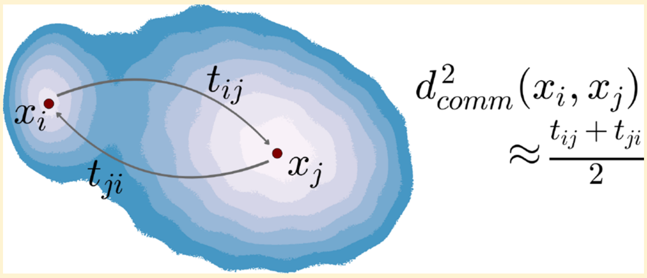
Commute Maps: Separating Slowly Mixing Molecular Configurations for Kinetic Modeling

Frank Noé,*[,†] Ralf Banisch,[†] and Cecilia Clementi*[,‡]

[†]Department of Mathematics, Computer Science and Bioinformatics, FU Berlin, Arnimallee 6, 14195 Berlin, Germany

[‡]Center for Theoretical Biological Physics, and Department of Chemistry, Rice University, 6100 Main Street, Houston, Texas 77005, United States

ABSTRACT: Identification of the main reaction coordinates and building of kinetic models of macromolecular systems require a way to measure distances between molecular configurations that can distinguish slowly interconverting states. Here we define the commute distance that can be shown to be closely related to the expected commute time needed to go from one configuration to the other, and back. A practical merit of this quantity is that it can be easily approximated from molecular dynamics data sets when an approximation of the Markov operator eigenfunctions is available, which can be achieved by the variational approach to approximate eigenfunctions of Markov operators, also called variational approach of conformation dynamics (VAC) or the time-lagged independent component analysis (TICA). The VAC or TICA components can be scaled such that a so-called commute map is obtained in which Euclidean distance corresponds to the commute distance, and thus kinetic models such as Markov state models can be computed based on Euclidean operations, such as standard clustering. In addition, the distance metric gives rise to a quantity we call total kinetic content, which is an excellent score to rank input feature sets and kinetic model quality.



1. INTRODUCTION

Molecular dynamics (MD) simulations are increasingly easy to generate in a high-throughput manner, using high-performance computers, distributed computing networks, or clusters of graphical processing units (GPUs).[1−12] This technological advance has opened up the possibility to sample biomolecular processes that involve rare events, including peptide and protein folding, protein conformational changes, and protein−ligand binding extensively.[9,13−16] As a result of this development, mass trajectory data of aggregate milliseconds can now be routinely generated, shifting the problem more and more from simulation to data analysis. Consequently, automation-capable analysis and machine learning methods to extract key mechanistic and kinetic information have received great attention recently.[17−39]

Essentially all these methods use some sort of distance metric by which they measure the similarity between molecular configurations, before proceeding to further analyses. Markov state models and core-based Markov models are based on state space discretization that employ grids or clustering in some metric space.[34,40] In kernel-based methods, the kernel function translates distances in some metric space into a similarity measure.[31,41,42] In the variational approach of conformation dynamics (VAC), the distance metric is defined by means of a basis set that nonlinearly transforms molecular coordinates or order parameters.[29,43,44] The key concern of these methods is to estimate mechanisms and kinetics from molecular dynamics, and in order to do that successfully they require a distance metric that is able to distinguish between molecular configurations when they are kinetically distant. An ideal distance metric for this purpose is one that assigns large distances between pairs of configurations when transitions between them are rare, and small distances when they are rapidly mixing.

For a complex macromolecular system it is nearly impossible to design a suitable distance metric *a priori*, and it is often misleading to use metrics based on structural similarities because large-scale motions such as loop motions can be fast, while small-scale motions such as isomerization of packed rings or dissociation of salt bridges can be slow. In Markov state modeling, adaptive or iterative approaches have been suggested to design discretizations that would give rise to a good resolution of the slow processes, without using an excessive number of clusters.[22,45−47] Recently, we have proposed the construction of a kinetic map based on a transformation of the input coordinate space into a new space in which Euclidean distances approximately correspond to *kinetic distances*.[38] While it was shown to outperform existing metrics in distinguishing slow and fast events, this approach still suffered the limitation that the kinetic map—and the corresponding distance metric—critically depended on the choice of a delay or lag time parameter used. In addition, the distance metric did not lend itself to a clear physical interpretation.

In the present paper we extend the previous approach[38] by integrating over the lag time parameter. We call the resulting

Received: July 29, 2016

DOI: 10.1021/acs.jctc.6b00762

distance "commute distance", $d_{\text{comm}}(\mathbf{x},\mathbf{y})$. The d_{comm} is completely independent of parameter choices in toy systems and robust to parameter changes when estimated from molecular dynamics data. Moreover, twice the squared commute distance, $2d^2_{\text{comm}}$, is shown to approximate the commute time, that is, the mean time needed to transition from one molecular configuration to the other, and back. The relevance of the commute time as a metric has has been pointed out in graph theory.[48] We demonstrate that commute distance can be practically computed from sampled MD data by employing the VAC or TICA.[35,36,49] We present an algorithm that allows the transformation of the molecular coordinates or order parameters used as an input to a new space in which the Euclidean distance corresponds to the commute distance. We call this transformed space the "commute map". Clustering operations in the commute map are shown to provide high-quality partitioning into metastable sets, thus allowing the robust definition of Markov state models.

2. THEORY

2.1. Commute Distance. We consider a dynamical system with a state space Ω, and an associated propagator, $\mathcal{P}_\tau$, that is, a Markov operator that propagates a probability density of states $\rho_t(\mathbf{x})$, $\forall \mathbf{x} \in \Omega$ in time as

$$\rho_{t+\tau}(\mathbf{y}) = \int_{\mathbf{x}\in\Omega} \rho_t(\mathbf{x})\, p_\tau(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{x} \tag{1}$$

$$= \mathcal{P}_\tau \rho_t(\mathbf{x}) \tag{2}$$

In the definition above, $p_\tau(\mathbf{y}|\mathbf{x})$ is the transition density, that is, the probability that the dynamics process, when located at configuration $\mathbf{x}$ at time t, will be found at configuration $\mathbf{y}$ at time $t + \tau$. We assume that the system has a unique equilibrium density (the Boltzmann distribution) defined by

$$\pi(\mathbf{x}) = \mathcal{P}_\tau \pi(\mathbf{x}) \tag{3}$$

We define a distance metric that assigns to each pair of molecular configurations $\mathbf{x}_1$, $\mathbf{x}_2$ a distance that is related to the time needed to travel between these configurations. Following ref 50, we start with the definition of the squared kinetic distance at lag time τ:

$$D^2_\tau(\mathbf{x}_1, \mathbf{x}_2) = \left\| p_\tau(\mathbf{y}|\mathbf{x}_1) - p_\tau(\mathbf{y}|\mathbf{x}_2) \right\|^2_{\pi^{-1}} \tag{4}$$

$$= \int_{\mathbf{y}\in\Omega} \frac{|p_\tau(\mathbf{y}|\mathbf{x}_1) - p_\tau(\mathbf{y}|\mathbf{x}_2)|^2}{\pi(\mathbf{y})}\, \mathrm{d}\mathbf{y} \tag{5}$$

This definition has the following interpretation: We start an ensemble at initial state $\mathbf{x}_1$ and another ensemble at initial state $\mathbf{x}_2$, let these ensembles evolve for time τ and then compute the distance between their probability distributions. If the starting points were in the same metastable state, or only separated by small barriers that could be overcome within time τ, then their distributions will have become similar after time τ, and their kinetic distance is small. On the other hand, if the two states were separated by large barriers that could not be overcome within time τ, their distributions will still be very different after time τ, and their kinetic distance is large.

At this point, eq 4 is purely a mathematical definition, but it has been shown that it can be approximated from simulation data.[38] The kinetic distance (eq 4) was originally introduced in ref 51 as the natural diffusion distance for diffusion processes, and when expressed in scaled diffusion coordinates one obtains the well-known diffusion map.[31,50] In ref 38 we used the same idea and extended its definition and application to other stochastic processes which have a unique stationary distribution, in particular to MD simulations.

Unfortunately, the metric (eq 4), and thus all algorithms using it, depend very strongly on the lag time τ. It is evident from its definition that this kinetic distance can separate fast and slow processes if there is a clear time scale separation and an appropriate value for the lag time τ is chosen in between fast and slow time scales. Unfortunately these time scales are often unknown at the beginning of the analysis, and one would like a metric that can robustly detect them without requiring any a priori knowledge of the system under investigation.

To avoid this dependency, we here introduce the following integrated version of the kinetic distance and call it commute distance, or short d_{comm}, for reasons that will become clear later. Its square is defined by

$$d^2_{\text{comm}}(\mathbf{x}_1, \mathbf{x}_2) = \int_{\tau=0}^{\infty} \left\| p_\tau(\mathbf{y}|\mathbf{x}_1) - p_\tau(\mathbf{y}|\mathbf{x}_2) \right\|^2_{\pi^{-1}} \mathrm{d}\tau \tag{6}$$

and we will examine its properties in the following sections. This distance metric measures the time-integrated difference between the two distributions when starting in two states $\mathbf{x}_1$ and $\mathbf{x}_2$. It is obvious that when the starting points are equal, $\mathbf{x}_1 = \mathbf{x}_2$, then IKD^2 = 0. Since we assume the dynamics to possess a unique equilibrium distribution, the integrand of eq 6 tends to 0 for $\tau \to \infty$, and the commute distance $d_{\text{comm}}(\mathbf{x}_1, \mathbf{x}_2)$ will have a finite value when this convergence is sufficiently fast. In practice this is the case when the dynamics have a maximum relaxation time scale, as we expect it to have in molecular dynamics. As a counterexample, $d_{\text{comm}}(\mathbf{x}_1, \mathbf{x}_2)$ is not a meaningful quantity for free diffusion on an infinite domain, as there is no finite upper bound to the relaxation time scales in this process.

In the following, in order to explicitly distinguish it from the d_{comm}, we indicate the previously defined kinetic distance $D_\tau(\mathbf{x}_1, \mathbf{x}_2)$ (defined by eq 4) as τ-kinetic distance.

2.2. Spectral Form of the Commute Distance and the Commute Map. The mathematical definition eq 6 is only useful for very simple systems for which the transition density $p_\tau(\mathbf{y}|\mathbf{x}_1)$ can be analytically written and integrated. However, it is straightforward to obtain an expression for the commute distance that is useful to approximate it in practice, and does not require execution of the integration over τ numerically. Following ref 38, we note that for every metastable Markov process that fulfills the detailed balance with respect to the equilibrium distribution $\pi(\mathbf{x})$, that is, $\pi(\mathbf{x})\, p_\tau(\mathbf{y}|\mathbf{x}) = \pi(\mathbf{y})\, p_\tau(\mathbf{x}|\mathbf{y})$, the transition density can be approximated by the spectral decomposition:

$$p_\tau(\mathbf{y}|\mathbf{x}) = \sum_{j=1}^{n} \lambda_j(\tau)\psi_j(\mathbf{x})\pi(\mathbf{y})\psi_j(\mathbf{y}) + \mathcal{P}^{\text{fast}}(\tau)\rho_t(\mathbf{x}) \tag{7}$$

where $\lambda_j(\tau)$ and $\psi_j(\mathbf{x})$ are the eigenvalues and eigenfunctions of the backward propagator, $\mathcal{T}$, that is, the operator adjoint to the propagator $\mathcal{P}$. The contribution $\mathcal{P}^{\text{fast}}(\tau)\rho_t(\mathbf{x})$ contains the fast processes and vanishes at lag times larger than the slowest relaxation time scale contained therein, such that $p_\tau(\mathbf{y}|\mathbf{x}) \approx \sum_{j=1}^{n} \lambda_j(\tau)\, \psi_j(\mathbf{x})\, \pi(\mathbf{y})\, \psi_j(\mathbf{y})$. When a realization of the dynamical process is available, $\lambda_j(\tau)$ and $\psi_j(\mathbf{x})$ can be approximated with methods discussed in section 2.4. We use the convention to normalize all eigenfunctions as $\langle \psi_i(\mathbf{x}) | \psi_j(\mathbf{x}) \rangle_\pi = \delta_{ij}$ and sort the eigenvalues by norm

DOI: 10.1021/acs.jctc.6b00762
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

$$\lambda_1 = 1 \geq |\lambda_2| \geq ... \geq |\lambda_n|$$

These eigenvalues are exponentially decaying functions of the lag time τ:

$$\lambda_j(\tau) = e^{-\tau/t_j} \tag{8}$$

where t_j is the relaxation time scale of the jth process.

When eq 7 is inserted into eq 4, it can be easily shown that the kinetic distance at lag time τ, $D_\tau(\mathbf{x}_1, \mathbf{x}_2)$ can be estimated as follows:

$$D_\tau^2(\mathbf{x}_1, \mathbf{x}_2) \simeq \sum_{j=2}^{n} (\lambda_j(\tau)\psi_j(\mathbf{x}_1) - \lambda_j(\tau)\psi_j(\mathbf{x}_2))^2 \tag{9}$$

Note that this expression does not depend on the stationary process with eigenvalue one, as $\psi_1(\mathbf{x})$ is a constant function, and thus $\psi_1(\mathbf{x}_1) - \psi_1(\mathbf{x}_2) \equiv 0$. Inserting this result into eq 6, and using eq 8 for the eigenvalue dependence on τ, we can compute an expression of the commute distance:

$$d_{\text{comm}}^2(\mathbf{x}_1, \mathbf{x}_2) = \int_0^\infty D_\tau^2(\mathbf{x}_1, \mathbf{x}_2)\, d\tau \tag{10}$$

$$= \sum_{j=2}^{n} (\psi_j(\mathbf{x}_1) - \psi_j(\mathbf{x}_2))^2 \int_0^\infty \lambda_j^2(\tau)\, d\tau \tag{11}$$

$$= \frac{1}{2}\sum_{j=2}^{n} t_j(\psi_j(\mathbf{x}_1) - \psi_j(\mathbf{x}_2))^2 \tag{12}$$

$$= \sum_{j=2}^{n} \left(\sqrt{\frac{t_j}{2}}\psi_j(\mathbf{x}_1) - \sqrt{\frac{t_j}{2}}\psi_j(\mathbf{x}_2)\right)^2 \tag{13}$$

In the above expression, t_j is the relaxation time scale of the jth process, as defined in eq 8.

Proceeding analogously as in ref 38, a similar but practically less useful result can be obtained for nonreversible dynamics.

Equation 10 shows that the commute distance equals the Euclidean distance in the space of coordinates:

$$\tilde{\psi}_j = \sqrt{t_j/2}\,\psi_j(\mathbf{x}_1)$$

These coordinates thus form the natural metric space with respect to the presently defined commute distance, and a suitable set of coordinates for subsequent operations using the Euclidean distance, such as clustering approaches. We call this metric space the commute map, in analogy to the terms diffusion map or kinetic map used before.[38,51] The form eq 12 makes it apparent that the squared commute distance has the unit of time. The relationship of square distance and time is reminiscent of a diffusion process. As discussed in the next section, it can be shown that the squared commute distance $d_{\text{comm}}^2(\mathbf{x}_i, \mathbf{x}_j)$ between two configurations $\mathbf{x}_i$ and $\mathbf{x}_j$ is equivalent to half the commute time $t_{\text{comm}}(\mathbf{x}_i, \mathbf{x}_j)$ between $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$t_{\text{comm}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{t_{ij} + t_{ji}}{2} \tag{14}$$

where t_{ij} is the mean first passage time from $\mathbf{x}_i$ to $\mathbf{x}_j$, that is, the expected time for the dynamics governed by (1) to reach $\mathbf{x}_j$ when starting in $\mathbf{x}_i$. Therefore, $d_{\text{comm}}^2(\mathbf{x}_i, \mathbf{x}_j)$ provides a direct measure of the time needed to connect the two configurations $\mathbf{x}_i$ and $\mathbf{x}_j$ kinetically. This insight makes an interesting connection to graph theory, for which the commute time has been found to be an excellent embedding metric for the graph.[48]

Finally, the metric gives rise to a kinetic variance or kinetic content of each eigenvalue/eigenvector pair. Analogously to ref 38, we compute the variance of the commute distance along each coordinate, giving rise to the kinetic content of that process:

$$K_i = \langle \tilde{\psi}_i | \tilde{\psi}_i \rangle_\pi = \frac{t_i}{2}$$

and the total kinetic content of the system:

$$K = \sum_{i=1}^{n} \frac{t_i}{2}$$

Note that the kinetic content has the physical unit of a time, and it equals half the sum of all relaxation times in the system. Clearly, the total kinetic content is only finite if the distribution of relaxation time scales decays sufficiently fast, as we expect for systems investigated by molecular dynamics, but may in practice require some regularization of the estimated eigenvalues (see section 2.4). Note that the kinetic content is maximized by the variational approach (section 2.4) and can be used as a cross-validation score similar to the sum of eigenvalues used in ref 47.

The fraction of kinetic content captured by the first m dimensions of the kinetic map is given by

$$c_m = \frac{\sum_{i=1}^{m} K_i}{K} = \frac{\sum_{i=1}^{m} t_i}{\sum_{i=1}^{n} t_i}$$

This expression can be used as a practical dimensionality reduction strategy as it provides the m dimensions (containing the m slowest processes) needed to explain a desired percentage (e.g., 95%) of the total kinetic content of the system.

2.3. Example: Commute Distance of Simple Markov Jump Processes.

To get a feeling for what the d_{comm} measures, let us first consider a simple two-state, discrete-time system described by the rate matrix of a Markov jump process, or Master equation model:[52]

$$\mathbf{K} = \begin{bmatrix} -k_{12} & k_{12} \\ k_{21} & -k_{21} \end{bmatrix} \tag{15}$$

with stationary distribution:

$$\boldsymbol{\pi} = \frac{1}{k_{12} + k_{21}} \begin{pmatrix} k_{21} \\ k_{12} \end{pmatrix}$$

and eigenvalues $\kappa_{1,2}$ and eigenvectors $\psi_{1,2}$:

$$\kappa_1 = 0 \qquad \boldsymbol{\psi}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\kappa_2 = -k_{12} - k_{21} \quad \boldsymbol{\psi}_2 = \frac{1}{\sqrt{k_{12}k_{21}}} \begin{pmatrix} -k_{12} \\ k_{21} \end{pmatrix}$$

Note that the rate matrix has the same eigenvectors as the corresponding transition matrices $\mathbf{P}(\tau) = \exp(\tau\mathbf{K})$, where τ is the lagtime. Here we use a rate matrix rather than a transition matrix, because the time-integral in the definition of the commute distance (eq 6) cannot be directly evaluated in a scenario where the true dynamics occurs in (macroscopic) discrete time steps. The rate matrix above has only one relaxation time scale (the exchange process between the two states):

$$t_2 = -\frac{1}{\kappa_2} = \frac{1}{k_{12} + k_{21}} \tag{16}$$

DOI: 10.1021/acs.jctc.6b00762

Inserting these results into eq 10 gives the following expression for the commute distance between the two substates:

$$\begin{aligned} d_{\mathrm{comm}}^2(1, 2) &= \frac{t_2}{2}\frac{(k_{21} + k_{12})^2}{k_{12}k_{21}} \\ &= \frac{1}{2}\frac{k_{12} + k_{21}}{k_{12}k_{21}} \\ &= \frac{1}{2}\left(\frac{1}{k_{12}} + \frac{1}{k_{21}}\right) \end{aligned}$$

We compare this quantity with the mean first-passage time t_{12}, from state 1 to state 2, and t_{21}, from state 2 to 1. The general definition of mean first passage time t_{ij} between two states i and j for a Markov jump process with rate matrix $\mathbf{K} = [k_{ij}]$:

$$t_{ij} = \begin{cases} 0 & i = j \\ -\dfrac{1}{k_{ii}} - \displaystyle\sum_{k \neq i} \frac{k_{ik}}{k_{ii}} t_{kj} & i \neq j \end{cases} \tag{17}$$

in the case of the two-state system above trivially reduces to

$$t_{12} = \frac{1}{k_{12}}$$

$$t_{21} = \frac{1}{k_{21}}$$

and

$$t_{\mathrm{comm}}(1, 2) = t_{12} + t_{21} = \frac{1}{k_{12}} + \frac{1}{k_{21}} = 2d_{\mathrm{comm}}^2(1, 2)$$

Thus, for the simple example, it is straightforward to show that twice the square of the commute distance is equal to the commute time, that is, the sum of the first passage times from state 1 to state 2 and return. We note that when drawing an analogy to a one-dimensional diffusion process, d_{comm} has the role of a diffusion constant and t_{comm} is the time needed to diffuse a unit length.

How is d_{comm}^2 related to the round-trip time more generally? We show in the Appendix that if the full set of exact eigenfunctions ψ_i and associated time scales t_i are used in eq 10, the square of the commute distance is always equivalent to the half commute time:

$$d_{\mathrm{comm}}^2(\mathbf{x}_i, \mathbf{x}_j) = \frac{t_{ij} + t_{ji}}{2} = \frac{t_{\mathrm{comm}}(\mathbf{x}_i, \mathbf{x}_j)}{2} \tag{18}$$

As each term of eq 12 makes a nonnegative contribution to d_{comm}^2, it is obvious that when we only include the first $n < \infty$ terms in the sum, the correspondingly truncated commute distance $d_{\mathrm{comm},n}^2$ will be a lower bound to the half round-trip time:

$$d_{\mathrm{comm},n}^2(\mathbf{x}_i, \mathbf{x}_j) \leq \frac{t_{ij} + t_{ji}}{2} \tag{19}$$

2.4. Estimating Commute Distances and Commute Maps from Data.

The variational approach to approximate eigenfunctions of Markov operators, also called variation approach of conformation dynamics (VAC[29,43]) and the time-lagged independent component analysis (TICA[35,36]) as a special case of the VAC, can be used to approximate relaxation time scales t_j and eigenfunctions $\psi_j(\mathbf{x})$, and thus used to compute commute distances and commute maps according to eq 10.

We briefly restate here the algorithm resulting from refs 29 and 43 in matrix notation. Suppose we are given a simulation trajectory measured in some set of order parameters that are assumed to be able to trace the slow kinetics (e.g., interatomic or inter-residue distances, torsions, and Cartesian coordinates in some reference frame, arbitrary nonlinear functions of molecular coordinates, or any combination of the listed options). Denote these order parameters $\tilde{x}_{t,i}$, with time index $t = 1, ..., T$ (trajectory length T) and dimension index $i = 1, ..., n$. To compute commute distances, we will always strip the mean from the coordinates, in particular we will use:

$$\mu_i = \sum_{t=1}^{T-\tau} \tilde{x}_{t,i}$$

where τ is the lag time at which we will be performing the calculation, and we remove the mean by $x_{t,i} = \tilde{x}_{t,i} - \mu_i$. We then define the two matrices

$$\mathbf{X}_0 = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & & \vdots \\ x_{T-\tau,1} & \cdots & x_{T-\tau,n} \end{bmatrix} \quad \mathbf{X}_\tau = \begin{bmatrix} x_{\tau,1} & \cdots & x_{\tau,n} \\ \vdots & & \vdots \\ x_{T,1} & \cdots & x_{T,n} \end{bmatrix}$$

and compute the moment matrices

$$\begin{aligned} \mathbf{C}_{00} &= \mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 \\ \mathbf{C}_{0\tau} &= \mathbf{X}_0^{\mathrm{T}}\mathbf{X}_\tau \end{aligned}$$

Optionally we can use the symmetrized version with $\mu_i = \sum_{t=1}^{T-\tau} \tilde{x}_{t,i} + \sum_{t=\tau}^{T} \tilde{x}_{t,i}$, $\mathbf{C}_{00} = \mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \mathbf{X}_\tau^{\mathrm{T}}\mathbf{X}_\tau$, and $\mathbf{C}_{0\tau} = \mathbf{X}_0^{\mathrm{T}}\mathbf{X}_\tau + \mathbf{X}_\tau^{\mathrm{T}}\mathbf{X}_0$ in order to enforce a real-valued solution, and then solve the generalized eigenvalue problem:

$$\mathbf{C}_{0\tau}\mathbf{R} = \mathbf{C}_{00}\mathbf{R}\hat{\mathbf{\Lambda}}$$

with eigenvector matrices $\mathbf{R}$ and the diagonal matrix of eigenvalues, $\hat{\mathbf{\Lambda}}$ and where we use the convention that the eigenvectors are normalized to fulfill $\mathbf{r}_i\mathbf{C}_{00}\mathbf{r}_i^{\mathrm{T}} = 1$ for all i. According to the variational principle of conformation dynamics, the eigenvalues resulting from this solution are approximations to the true eigenvalues from below $\hat{\lambda}_i(\tau) \leq \lambda_i(\tau)$ for all i and τ.[29] The eigenvectors of this solution approximate the transfer operator eigenfunctions $\psi_i(\mathbf{x}_t)$ evaluated on the sampled configurations $\mathbf{x}_t$ as

$$\mathbf{\Psi} = [\psi_{ti}] \approx \mathbf{X}_0\mathbf{R}$$

With the use of the normalization condition above, the resulting eigenfunction trajectories have variance 1: $(T - \tau)^{-1} \sum_{t=1}^{T-\tau} \psi_{\mathrm{ti}}^2 = 1$. To use this approach to estimate the commute map, we simply compute the estimated relaxation time scales

$$\hat{t}_i(\tau) = -\frac{\tau}{\ln|\hat{\lambda}_i(\tau)|}$$

Given the estimated time scales $\hat{t}_i(\tau)$, we can compute the commute map as

$$\mathbf{\Psi}' = \mathbf{X}_0\mathbf{R}\,\mathrm{diag}\left(\sqrt{\hat{t}_i/2}\right) \tag{20}$$

Euclidean distances in these commute map coordinates approximate the commute distance, that is, for any two data points $\mathbf{x}_1$ and $\mathbf{x}_2$ sampled along the trajectory, we have

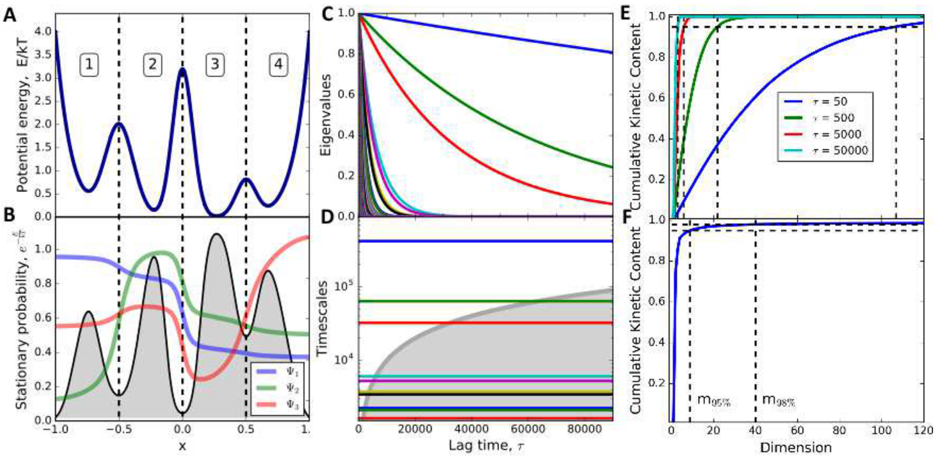


Figure 1. τ-Kinetic distance versus commute distance in a four-state kinetic model. (A) Potential energy of the model system (see ref 40 for details). Dashed black lines mark the four metastable basins of the model. (B) Equilibrium distribution is shown in gray, and the eigenfunctions corresponding to the three slowest processes, Ψ_1 (blue), Ψ_2(green), Ψ_3 (red) are shown. (C) Eigenvalues as a function of the lag time. (D) Relaxation time scales. The region where $t_i < \tau$ is shaded in gray. (E, F) Cumulative kinetic content for the τ-kinetic distance metric used in ref 38 with different choices of lag time τ, and the commute time metric used here, respectively. Dashed black vertical lines indicate number of dimensions required to explain 95% of the kinetic content for the τ-kinetic distance (E) and both 95% or 98% for the d_{comm} (F).

$$\hat{d}^2_{\text{comm}}(\mathbf{x}_1, \mathbf{x}_2) = \left\|\boldsymbol{\psi}'_{t1} - \boldsymbol{\psi}'_{t2}\right\|_2^2$$

Note that by removing the mean, we have removed the stationary eigenpair ($\lambda_1 = 0$, $\psi_1 = 1$) from the solution. Thus, the relaxation time scales obtained from the algorithm above will be variational approximations of the dynamic relaxation time scales, that is, $\hat{t}_1 \approx t_2, ..., \hat{t}_n \approx t_{n+1}$. The stationary eigenpair does not contribute to the commute distance, which motivates us to use the mean-free coordinates for the present application of the variational approach.

In the algorithm described above, we can only approximate n (number of the chosen molecular order parameters or basis functions) eigenfunctions. According to eq 19, the Euclidean distances in $\boldsymbol{\Psi}'$ are thus expected to be lower bounds to the half round-trip times between the corresponding configurations. Furthermore, since the variational principle implies $\hat{t}_i(\tau) \leq t_i(\tau)$ in absence of statistical error, this systematic approximation error involved in the variational approach above will lead to a further underestimation of the half round-trip times. The point-wise approximation error of the eigenfunctions, $\hat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x})$ is more difficult to quantify, and the statistical error due to finite data size in both $\hat{t}_i$ and $\hat{\psi}_i$ can go in both directions. Nonetheless we expect and practically observe the approximate bound:

$$\hat{d}^2_{\text{comm}}(\mathbf{x}_1, \mathbf{x}_2) \lesssim \frac{t_{12} + t_{21}}{2}$$

It is also worth noting that in this paper, we do not, in practice, execute the integral over τ to compute eq 10 algorithmically, but rather use the integrated expression eq 12 with an estimate $\hat{t}_i(\tau)$ obtained at a single estimation lag time τ. The key advantage of the commute distance compared with the previous τ-kinetic distance is that the choice of τ is no longer arbitrary, because t_j is a property of the system rather than a model parameter, and it is well-known that t_j can be approximated by monitoring the convergence of $t_j(\tau)$ as a function of τ. However, the new kinetic distance is in practice also robust if we use a τ for which $t_j(\tau)$ is not converged (see results discussed below in section 4).

2.5. Regularization of Time Scales. In practice, the small eigenvalues and corresponding time scales $t_i < \tau$ are often both spurious from a mathematical perspective and not very relevant from a physicochemical perspective. Processes that are fast compared with the lag time cannot be reliably estimated as their contribution to the signal has decayed according to eq 7. To avoid contamination of the commute distance by these spurious estimates, we can employ a regularization that dampens small time scales. Here we use

$$t'_i(\tau) = \hat{t}_i(\tau)\, \phi(\hat{t}_i, \tau)$$

with

$$\phi(\hat{t}_i, \tau) = \frac{1}{2}\tanh\left(\pi \frac{\hat{t}_i(\tau) - \tau}{\tau} + 1\right)$$

to modify time scales $\hat{t}_i(\tau)$ estimated from simulation data, which has the effect of damping estimated time scales smaller than τ, and then employ t'_i in (20) to construct the commute map.

3. FOUR-WELL POTENTIAL

We compare the performance of the commute map with that of the τ-kinetic map approach proposed in ref 38 in a diffusion process on the metastable example system introduced in ref 40. In particular, we compare the ability of the different metrics to distinguish states that are kinetically well separated. The system's potential energy contains four wells (Figure 1A), and the dynamics are defined by a Metropolis jump process between neighbors in a 1000-state grid on the domain $x \in [0, 1]$. This system has four metastable states that contain most of the equilibrium density (Figure 1B), and three slow relaxation time scales corresponding to the transition processes between the wells that are well separated from the faster relaxation time scales (Figure 1D). As this system has an exact microstate transition matrix $\mathbf{P} = [p_{ij}] \in \mathbb{R}^{1000\times1000}$, we can compute arbitrary long-time expectations by using appropriate powers of that matrix, without requiring sampling and introducing statistical error. The eigenvalues of $\mathbf{P}(\tau) = \mathbf{P}^{\tau}$ as a function of τ are shown in Figure 1C, with the corresponding relaxation time scales (independent of τ) in Figure 1D.

Figure 1 panels E and F show the cumulative kinetic content of the τ-kinetic distance defined previously[38] in comparison with the commute distance defined here. In the τ-kinetic distance, the

DOI: 10.1021/acs.jctc.6b00762
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

cumulative kinetic content, thus the number of dimensions $m_{95\%}$ needed to explain 95% of the total kinetic content, vary strongly with the lag time τ used to compute the metric, with $m_{95\%}$ taking values of 107, 22, 6, and 3 for τ equal to 50, 500, 5000, and 50000, respectively. In the integrated version of the metric, by definition the cumulative kinetic content is independent of the choice of τ, because the relaxation time scales do not depend on τ. In this case, 95% of the kinetic content is obtained with $m_{95\%} = 9$, and 98% with $m_{98\%} = 40$.

Figures 2 and 3 compare the τ-kinetic distance and the commute distance between each pair of points $(\mathbf{x}_i, \mathbf{x}_j)$. The τ-kinetic distance performs poorly for small values of the lag time τ, and states separated by large barriers are only clearly distinguished for large values of τ (>5000), likely because at small τ there are many eigenvalues close to 1 contaminating the distance metric (Figure 2). Figure 3 presents the results obtained when $m_{95\%}$ (top panels) or $m_{98\%}$ (bottom panels) eigenfunctions are used to compute the d_{comm} (see eq 12). Although small differences between the $m_{95\%}$ and $m_{98\%}$ results can be noticed in the values of d_{comm} connecting the faster interconverting regions (left panels), in both cases the commute distance correctly distinguishes quickly and slowly interconverting pairs of states independently of τ. The right panels of Figure 3 show the ratio between $2d^2_{\text{comm}}(\mathbf{x}_i, \mathbf{x}_j)$ and the commute time, $t_{\text{comm}} = (t_{ij} + t_{ji})/2$, between all pairs of states $(\mathbf{x}_i, \mathbf{x}_j)$. Consistent with the results discussed above, this ratio approaches 1 for all pairs of states when more eigenfunctions are used in the d_{comm} estimation. However, even when only $m_{95\%} = 9$ eigenfunctions are used, the ratio $2d^2_{\text{comm}}(\mathbf{x}_i, \mathbf{x}_j)/t_{\text{comm}}$ is very close to 1 when the states $(\mathbf{x}_i, \mathbf{x}_j)$ are well separated by a barrier, and significantly deviates from 1 only for pairs of states within the same metastable basin. In practical applications, in the analysis of trajectories of much more complex systems, the calculation of distances or mean first passage times between states that are kinetically very close (with respect to the lagtime used in the estimations) is not relevant as these states are usually aggregated in the same macrostate. Thus, d_{comm} can in practice be estimated with a limited number of slow coordinates.

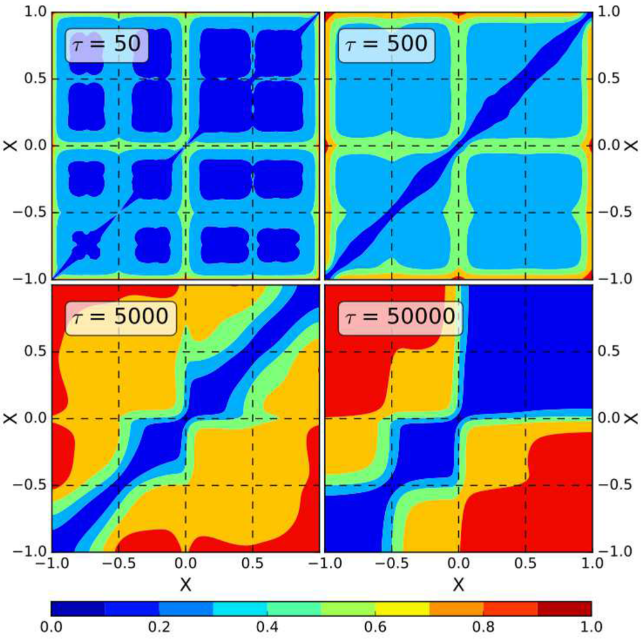


Figure 2. Kinetic distance in a four-well potential model. τ-Kinetic distance $D_\tau(\mathbf{x}_i, \mathbf{x}_j)$ as proposed in ref 38 between all pairs of states $(\mathbf{x}_i, \mathbf{x}_j)$ in the four-well system, computed for four different choices of lag time, τ. For illustrative purposes, for each choice of τ, the distances are normalized to the corresponding maximum values, $\max_{i,j} D_\tau(\mathbf{x}_i, \mathbf{x}_j)$, that are equal to 51.1 ($\tau = 50$), 17.9 ($\tau = 500$), 4.1 ($\tau = 5000$), and 2.3 ($\tau = 50000$). Different colors map to values as illustrated in the color bar at the bottom. Dashed black lines mark the four metastable states as defined in Figure 1.

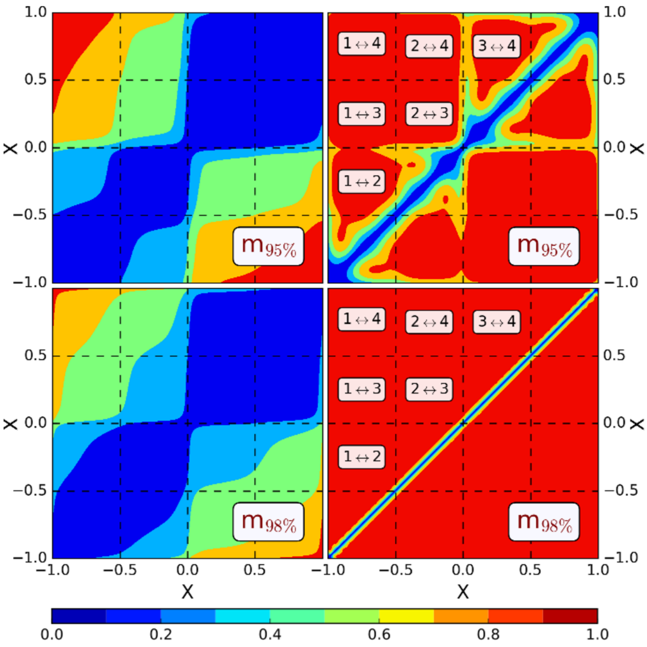


Figure 3. Robustness and interpretation of commute distance in a four-well potential model. The square of the commute distance proposed here, $d^2_{\text{comm}}(\mathbf{x}_i, \mathbf{x}_j)$, between all pairs of states $(\mathbf{x}_i, \mathbf{x}_j)$ in the four-well system is shown on the left panels, for two different choices of dimensions in eq 12: using the first $m_{95\%}$ (top) or $m_{98\%}$ (bottom) coordinates needed to explain 95% or 98% of the total kinetic content. In both cases the commute distance clearly distinguishes between the slowly interconverting metastable states. As in Figure 2, the values are normalized to their corresponding maximum, $\max_{i,j} d^2_{\text{comm},95\%}(\mathbf{x}_i, \mathbf{x}_j)\tau \approx 1.7 \times 10^6\tau$ and $\max_{i,j} d^2_{\text{comm},98\%}(\mathbf{x}_i, \mathbf{x}_j)\tau \approx 2.2 \times 10^6\tau$. The right panels show the ratio $2d^2_{\text{comm}}(\mathbf{x}_i, \mathbf{x}_j)/t_{\text{comm}}(\mathbf{x}_i, \mathbf{x}_j)$ of twice the squared commute distance to the commute time, $t_{\text{comm}}(\mathbf{x}_i, \mathbf{x}_j) = t_{ij} + t_{ji}$, for all pairs of states $(\mathbf{x}_i, \mathbf{x}_j)$, for $m_{95\%}$ (top) or $m_{98\%}$ (bottom). For all figures, different colors map to values as illustrated in the color bar at the bottom. Dashed black lines mark the four metastable states as defined in Figure 1.

To compare the practical performance of d_{comm} and the τ-kinetic distance as metrics for Markov state modeling, we test their ability to find the four metastable basins by simple k-means clustering with four states. Representative results are shown in Figure 4 by using the first $m_{95\%}$ eigenfunctions (using $m_{98\%}$ produces indistinguishable results). When the τ-kinetic map $\tilde{\psi}_j = \lambda_j(\tau)\psi_j(\mathbf{x})$ is used, the clustering result depends strongly on the choice of the lag time τ in its definition (Figure 4A). The four metastable basins are properly distinguished for $\tau = 5000$, but small lag times fail completely in obtaining a metastable partition using four clusters, and will consequently result in a very poor Markov state model. For the longest tested lag time $\tau = 50000$, the separation between basins three and four is lost. This can be explained by the fact that at this lag time, the eigenvalue λ_3 has become negligible and the corresponding eigenvector ψ_3 that

DOI: 10.1021/acs.jctc.6b00762

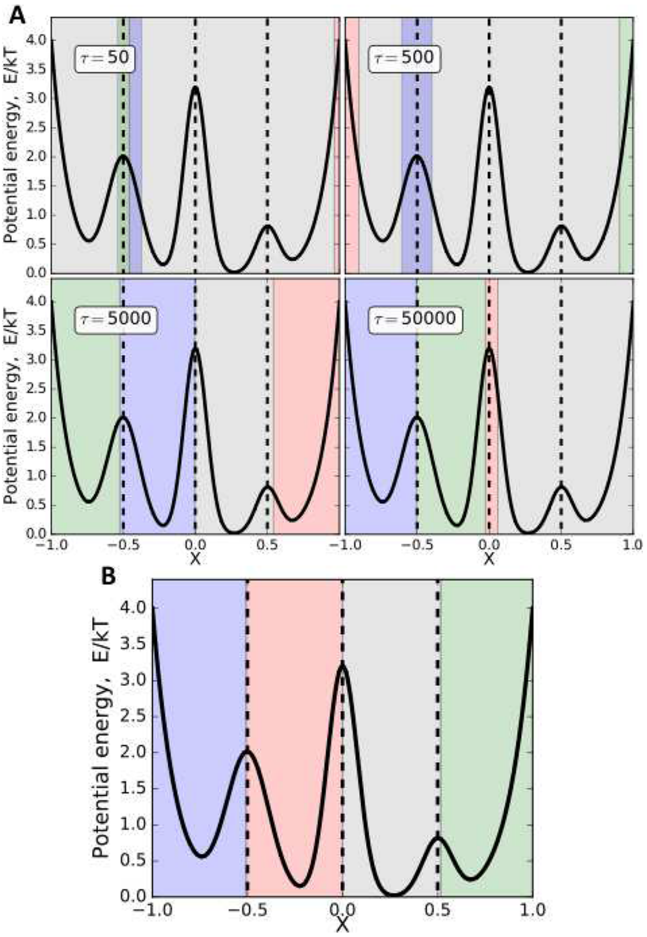


Figure 4. Quality of cluster discretization for different kinetic distances. k-Means clustering with $k = 4$ is used in different kinetic maps to probe their ability to separate different metastable basins with few clusters. The potential energy function is plotted in each figure, and the four metastable states are marked by black dashed lines, as in Figure 1A. (A) With the use of the τ-kinetic map, $\tilde{\psi}_j = \lambda_j(\tau)\,\psi_j(\mathbf{x})$, as a metric produces results that are strongly dependent on the lag-time chosen. Only intermediate lag times (here $\tau = 5000$) provide accurate distinction of the four metastable basins. (B) With the use of the commute map, $\tilde{\psi}_j = \sqrt{t_j/2}\,\psi_j(\mathbf{x})$, the four metastable basins are accurately and reliably separated by a four-state k-means clustering.

distinguishes between basins three and four is no longer contributing to the metric (see Figure 1).

In contrast, using k-means with four clusters on the commute map, $\tilde{\psi}_j = \sqrt{t_j/2}\,\psi_j(\mathbf{x})$, accurately and reliably separates the four metastable basins of the system. In practice, Markov state models are built with more clusters than metastable states, in order to obtain a good approximation quality.[40,45] Nonetheless, the ability of a metric to provide a good separation of the slow processes with relatively few states and without prior knowledge of system-specific parameters (e.g., the choice of the lag time τ) is key in obtaining reliable Markov state models.

To further illustrate this point, we use the different cluster discretizations obtained above and construct Markov state models. As the exact transition matrix $\mathbf{P}$ is given, the coarse-grained Markov state model at lag time T can be obtained by

$$p_{IJ}(T) = \frac{\sum_{i\in I, j\in J} \pi_i p_{ij}(T)}{\sum_{i\in I} \pi_i} \tag{21}$$

where I and J are the sets of microstates in the corresponding clusters, $\pi_i = \mathrm{e}^{-E(\mathbf{x}_i)/k_\mathrm{B}T}$ is the equilibrium probability of microstate i, and $p_{ij}(T) = [\mathbf{P}^T]_{ij}$. We can compare the performance of these Markov models by virtue of the variational approach of conformation dynamics (VAC),[29,43] which states that the approximated eigenvalues, as obtained for instance in a Markov model, always underestimate the true eigenvalues: $\lambda_i^{\mathrm{MSM}}(T) \le \lambda_i^{\mathrm{exact}}(T)$, and that equality is only obtained when the approximated eigenfunction equals the exact one. As a result, the estimated relaxation time scales $t_i(T) = -T/\ln|\lambda_i(T)|$ are also underestimated, $t_i^{\mathrm{MSM}}(T) \le t_i^{\mathrm{exact}}$, and this principle provides a practical criterion to evaluate the performance of a metric distance, as the choice yielding larger estimates for the relaxation time scales should be preferred.

Figure 5 shows the ratio of the Markov state model time scales to the exact time scales, $t_i^{\mathrm{MSM}}(T)/t_i^{\mathrm{exact}}$. Values of 1 indicate the perfect approximation, and values much smaller than 1 indicate a poor approximation. Different numbers of clusters, ranging from 4 to 30, and different choices of lag time T are used in the construction of the Markov model for each choice of the distance metric. In addition to the lag time T used in the construction of the Markov model (eq 21), the kinetic distance, $D_\tau(\mathbf{x}_i, \mathbf{x}_j)$, also depends on the lag time τ in its definition (see eq 9), and results are presented for values of τ = 50, 500, 5000, 50000 (different rows in Figure 5A).

The three slowest time scales are reported in different columns in the figure. As expected, the estimated time scales t_i^{MSM} are closer to the corresponding exact time scales t_i^{exact} if a larger number of clusters and longer lag times T are used in the definition of the Markov model.[40,45] However, the time scales depend strongly also on the lag time τ chosen in the definition of the kinetic distance metric itself, and are severely underestimated when small values of τ are used. On the other hand, the estimated time scales obtained when d_{comm} is used in the discretization step of the Markov model construction are generally much more accurate than the corresponding results obtained with the τ-kinetic distance, comparable only when the optimal value of lag time τ = 5000 is used in the τ-kinetic distance (see also Figure 4A).

4. APPLICATION TO MOLECULAR DYNAMICS SIMULATION

We compare the robustness of the commute map with that of the previously proposed τ-kinetic map approach also in the analysis of molecular dynamics simulations of a biomolecular system. We used a one-millisecond simulation of a 58-residue protein, bovine pancreatic trypsin inhibitor (BPTI), produced by D. E. Shaw research using the Anton supercomputer (see ref 3 for simulation details). As in our previous work,[38] the BPTI trajectory was subsampled to 100 000 frames, and the Cartesian coordinates of the protein C_α-atoms were used as an input data set.

We again first inspect the robustness of the metric itself as a function of lag time. In contrast to the four-well example in which the exact time scales could be computed from the knowledge of the exact transition matrix, in this example the commute distance will depend on the lag time τ, because the time scales t_i are estimated from the data. However, the hope is that the effect of τ on d_{comm} is less strong than on the τ-kinetic distance, or that a suitable value of τ can be easily chosen by inspecting the results.

In kinetic maps based on the τ-kinetic distance (eq 4) the eigenvalues $\lambda_i(\tau)$ are used as scaling factors. As the eigenvalues decay exponentially in τ with different time scales, the scaling

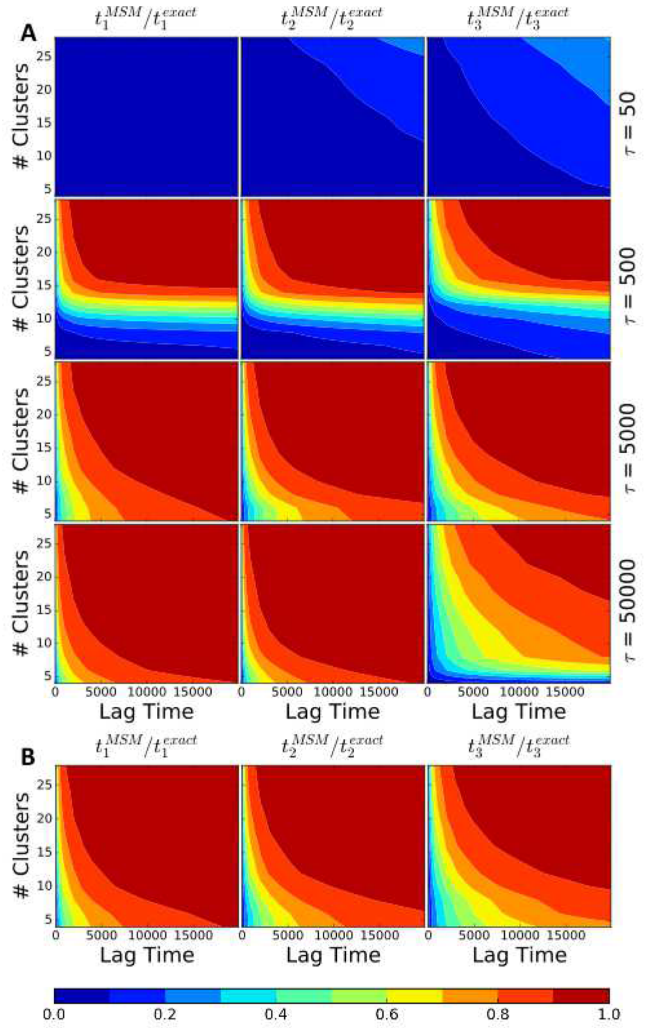


Figure 5. Quality of Markov state model built with different kinetic distances. Ratios of the three slowest Markov state model time scales to the exact time scales, $t_i^{\mathrm{MSM}}(\tau)/t_i^{\mathrm{exact}}$ are shown. Values of 1 (red) indicate the perfect approximation, and values close to 0 (blue) a poor approximation. Color scheme as in Figure 2. (A) MSMs using a cluster discretization of the previously proposed τ-kinetic map.[38] Each row reports results obtained with a different choice of the lag time τ used to compute the metric, while the lag time on the horizontal axis is the Markov state model lag time, T. Consistent with the results reported in Figure 4A, good results for all slow time scales are only obtained when using a suitable choice of lag time τ (here τ = 5000) to define the metric. (B) Same as in panels A, but using the commute distance d_{comm} proposed here.

factors depend strongly on the choice of the lag time τ (Figure 6A,B). In the present approach using d_{comm} (eq 6), the scaling factors are the square roots of the regularized time scales, $\sqrt{t_i'}$. This choice has the attractive advantage that $\sqrt{t_i'}$ is an invariant of the system that we try to approximate and is thus a more objective choice than using eigenvalues $\lambda_i(\tau)$. Indeed, $\sqrt{t_i'}$ values tend to converge at sufficiently long lag times (Figure 6C), thus defining a range of lag times τ that are reasonable choices to estimate them.

Interestingly, for this system, the relative scaling factors $\sqrt{t_i'(\tau)/t_1'(\tau)}$ are approximately constant even for short lag times τ at which the absolute values $\sqrt{t_i'(\tau)}$ are not converged yet (Figure 6D). Since many geometric analysis approaches such as the k-means clustering methods are sensitive to relative, but not absolute distances, this result implies that the present commute map is much more robust with respect to the choice of the lag time τ than previous approaches.

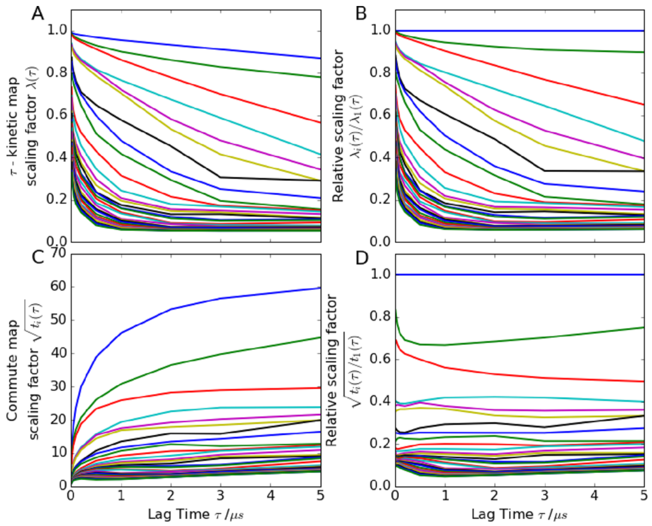


Figure 6. Relative scaling factors for kinetic map and commute map for BPTI. (A) Scaling factors $\lambda_i(\tau)$ for the τ-kinetic map proposed in ref 38. (B) Relative scaling factors $\lambda_i(\tau)/\lambda_1(\tau)$. (C) Scaling factors $\sqrt{t_i(\tau)}$ used for the commute map proposed here. (D) Relative scaling factors $\sqrt{t_i(\tau)/t_1(\tau)}$. The commute map is robust with respect to different choices of lag time τ. In each figure, different colors indicate different eigenvalue numbers, $i \in [1, 30]$.

Results from a Markov model analysis based on different metrics are reported in Figure 7. TICA is performed using a lag time τ = 10 ns. We consider five metrics: Euclidean distance with (i) projections onto the first two, or (ii) first six TICA coordinates, and (iii) in full TICA space, (iv) τ-kinetic map, and (v) commute map of all scaled TICA-coordinates. For the given setting, the commute map provides a better compression than the τ-kinetic map (Figure 7A,B), requiring 29 in contrast to 38 dimensions to explain 95% of the kinetic content.

We then constructed reversible Markov state models from the different projections of the data into the five different metric spaces. In all cases, k-means clustering with 200 clusters was used. Implied time scales estimates of Markov state models are shown in Figure 7C. While low-dimensional TICA projections provide well-converged estimates of the three slowest relaxation time scales in the Markov state model, they are still significantly underestimated when compared to both the kinetic map and the commute map results. Using all (unscaled) TICA coordinates performs worse than low-dimensional TICA projections, as the time scales after the slowest are significantly underestimated. Overall, the present commute map performs better than the τ-kinetic map, and significantly better than low-dimensional TICA projections. This is most apparent in the comparison of the total kinetic content (Figure 7D). The commute map also reveals that the slowest relaxation time scales and the total kinetic content is not perfectly converged. This is an indicator of either incomplete sampling as has been noted for this data set before[3,53] or that the features and discretization used here still exhibit significant projection error.[40,45]

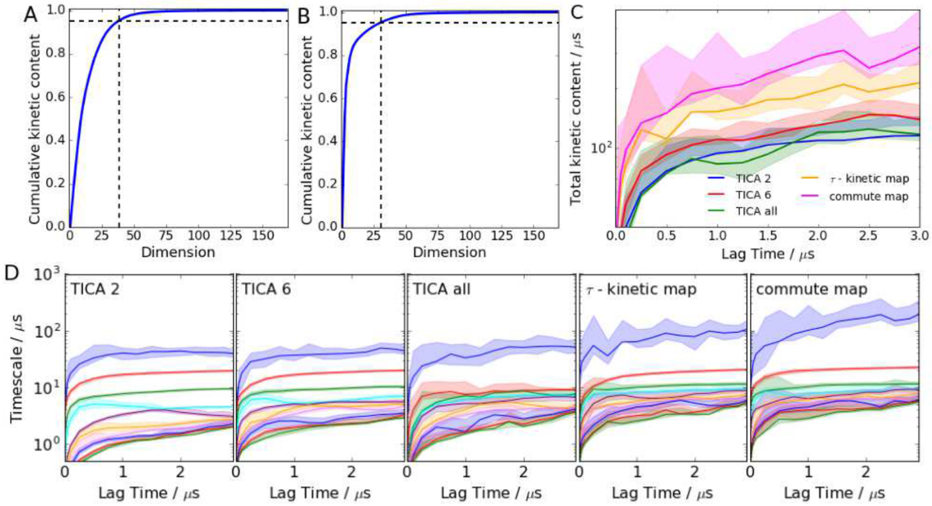


Figure 7. Markov model analysis of BPTI molecular dynamics trajectories by using different TICA projections and kinetic maps. (A) Cumulative kinetic content of the τ-kinetic map approach proposed in ref 38 with a lag time $\tau = 10$ ns. (B) Cumulative kinetic content of the d_{comm} approach proposed here. (C) Total kinetic content $K = \sum_i t_i$ with 95% confidence intervals for different choices of the subspace used. The present commute map approach performs best. (D) Implied time scales with 95% confidence intervals of the Markov state model built on different projections: TICA 2, TICA 6, all TICA dimensions, the τ-kinetic map, and the commute map were compared. The gray area denotes the numerically unreliable regime time scale < τ.

5. CONCLUSIONS

We propose a new distance metric, the commute distance (d_{comm}), which can distinguish between slowly interconverting states in dynamical systems, and thus serve as a basis for finding good reaction coordinates and building kinetic models. The d_{comm} is defined as the lagtime-integral of the τ-kinetic distance metric proposed in ref 38. In practice, we use the d_{comm} concept to construct a kinetic map as follows:

1. Obtain estimates of the slowest relaxation time scales and approximations to the eigenfunctions of the Markov backward propagator by means of the time-lagged independent component analysis (TICA), or the variational approach of conformation dynamics (VAC). Algorithmically, this step involves the computation of two covariance matrices from simulation data, the solution of an eigenvalue problem, and the projection of the simulation data onto the eigenvector matrix.
2. The coordinates of the transformed data are now ordered from the slowest to the fastest relaxation process. Now scale each coordinate by $\sqrt{t_i/2}$, where t_i is the estimated relaxation time scale of the corresponding process. The resulting coordinates define the commute map, in which the Euclidean distance between any two points corresponds to d_{comm}.

We have shown that by definition, $2d_{\mathrm{comm}}^2$ is equal to the commute time, or round-trip time between pairs of states. This equivalence is only obtained when all the eigenfunctions are used in the estimation, but a good approximation to the round-trip time between states in different metastable states can be obtained even if only a small set of slow eigenfunctions are used. Note that a good metric for constructing kinetic models needs to be able to separate slowly mixing states, while it is less important to accurately represent the commute times between fast-mixing states.

In contrast to the previous τ-kinetic distance, the commute distance is well-defined as the scaling factors applied to the eigenfunction trajectories are using t_i, which is a physical quantity of the simulated system, rather than τ which is a parameter. In practice, t_i is obtained by monitoring the convergence of $t_j(\tau)$ as a function of τ. It has been empirically observed in the presently studied molecular example that the relative scaling factors $\sqrt{t_i(\tau)/t_1(\tau)}$ are rather constant even in regimes of τ where the absolute scaling factors $t_i(\tau)$ are not converged. This indicates that when geometric operations are used for postprocessing which are insensitive to the absolute scaling (length unit) of the data, it may be possible to rely on d_{comm} even when using very short lag times. We can currently not prove whether this observation is generally valid, and this aspect will require further investigation.

■ APPENDIX A: COMMUTE DISTANCE AS AN ESTIMATE OF THE ROUND TRIP TIME

The definition of twice the commute distance $2d_{\mathrm{comm}}^2(\mathbf{x}_i, \mathbf{x}_j)$ (eq 10) between two states $\mathbf{x}_i$ and $\mathbf{x}_j$ is equivalent to the definition of the commute time, $t_{\mathrm{comm}}(\mathbf{x}_i, \mathbf{x}_j) = t_{ij} + t_{ji}$, that is, the mean time to go from $\mathbf{x}_i$ to $\mathbf{x}_j$ and back (eq 14). This equivalence can be proven by expressing the mean first passage time t_{ij} from $\mathbf{x}_i$ to $\mathbf{x}_j$ in terms of the eigenfunctions $\psi_k(\mathbf{x})$ in eq 7 and corresponding time scales t_k. We assume that the fast term in the spectral decomposition (eq 7 above) can be ignored at lagtimes of interest.

It is useful to introduce a function $h_{\mathbf{x}_i}(\mathbf{x})$ which takes the value of the mean first passage time from any state $\mathbf{x} \in \Omega$ to the target state $\mathbf{x}_i$. The function $h_{\mathbf{x}_i}(\mathbf{x})$ can be decomposed in the basis set defined by the eigenfunctions $\{\psi_k(\mathbf{x})\}$ of the backward operator $\mathcal{T}_\tau$ as

$$h_{\mathbf{x}_i}(\mathbf{x}) = \sum_{k>1} (\psi_k(\mathbf{x}_i) - \psi_k(\mathbf{x}))\, \psi_k(\mathbf{x}_i)\, t_k \tag{22}$$

where $t_i = -\tau \ln\lambda_i(\tau)$ are the time scales associated with the eigenvalues $\lambda_i(\tau)$. The first eigenfunction $\psi_1(\mathbf{x})$ does not contribute to the sum in eq 22 as $\psi_1(\mathbf{x}_i) - \psi_1(\mathbf{x}) \equiv 0$. This decomposition can be proven by using the defining equation for $h_{\mathbf{x}_i}(\mathbf{x})$. The generalization of definition of mean first passage time $t_{ji} = h_{\mathbf{x}_i}(\mathbf{x}_j)$ given in eq 17 for $h_{\mathbf{x}_i}(\mathbf{x})$ becomes

$$\mathcal{L}h_{\mathbf{x}_i}(\mathbf{x}) = -1 \quad \forall\ \mathbf{x} \neq \mathbf{x}_i$$
$$h_{\mathbf{x}_i}(\mathbf{x}_i) = 0 \tag{23}$$

The operator $\mathcal{L}$ in eq 23 is the generator of the backward operator $\mathcal{T}_\tau$, that is, $\mathcal{T}_\tau = e^{\mathcal{L}\tau}$. $\mathcal{L}$ has the same eigenfunctions $\psi_i(\mathbf{x})$ of $\mathcal{T}_\tau$, with associated eigenvalues $\kappa_i = -\tau^{-1}\ln\lambda_i(\tau) = -t_i^{-1}$, where $\lambda_i(\tau)$ are the eigenvalues of $\mathcal{T}_\tau$, and t_i are the corresponding time scales. This expression trivially satisfies the requirement $h_{\mathbf{x}_i}(\mathbf{x}_i) = 0$. It is also straightforward to show that it solves eq 23, by directly evaluating the action of the generator $\mathcal{L}$ on it:

$$\begin{aligned}
&\mathcal{L}\sum_{k>1}(\psi_k(\mathbf{x}_i) - \psi_k(\mathbf{x}))\,\psi_k(\mathbf{x}_i)t_k \\
&= \sum_{k>1}\mathcal{L}\psi_k(\mathbf{x}_i)\,\psi_k(\mathbf{x}_i)t_k - \mathcal{L}\psi_k(\mathbf{x})\,\psi_k(\mathbf{x}_i)t_k \\
&= \sum_{k>1}\psi_k(\mathbf{x}_i)\,\psi_k(\mathbf{x})
\end{aligned} \tag{24}$$

where we have used $\mathcal{L}\psi_i(\mathbf{x}) = -\psi_i(\mathbf{x})/t_i$ and $\mathcal{L}(\text{constant}) \propto \mathcal{L}\psi_1(\mathbf{x}) = \kappa_1 = 0$.

Because the eigenfunctions $\{\psi_k(\mathbf{x})\}$ form a π-orthonormal basis set (where $\pi(\mathbf{x})$ is the equilibrium distribution), they can be used to decompose any function, including the delta function:

$$\frac{\delta(\mathbf{x} - \mathbf{x}_i)}{\pi(\mathbf{x})} = \sum_{k=1}^{\infty}\psi_k(\mathbf{x})\int\psi_k(\mathbf{x}')\delta(\mathbf{x}' - \mathbf{x}_i)\,\mathrm{d}\mathbf{x}'$$

Thus, $\forall \mathbf{x} \neq \mathbf{x}_i$:

$$\begin{aligned}
0 &= \frac{\delta(\mathbf{x}_i - \mathbf{x})}{\pi(\mathbf{x})} \\
&= \sum_{k=1}^{\infty}\psi_k(\mathbf{x}_i)\int\psi_k(\mathbf{x}')\,\delta(\mathbf{x}' - \mathbf{x})\,\mathrm{d}\mathbf{x}' \\
&= \sum_{k=1}^{\infty}\psi_k(\mathbf{x}_i)\,\psi_k(\mathbf{x}) \\
&= \psi_1(\mathbf{x}_i)\,\psi_1(\mathbf{x}) + \sum_{k>1}\psi_k(\mathbf{x}_i)\,\psi_k(\mathbf{x})
\end{aligned}$$

and, using the fact that $\psi_1(\mathbf{x}) = 1$:

$$\sum_{k>1}\psi_k(\mathbf{x}_i)\,\psi_k(\mathbf{x}) = -1 \tag{25}$$

When eq 24 and eq 25 are combined, it is clear that eq 22 is the solution of eq 23.

By using the decomposition 22, we can now express the commute time t_{comm} in terms of the eigenfunctions and time scales:

$$\begin{aligned}
t_{\mathrm{comm}}(\mathbf{x}_i, \mathbf{x}_j) &= t_{ij} + t_{ji} = h_{\mathbf{x}_i}(\mathbf{x}_j) + h_{\mathbf{x}_j}(\mathbf{x}_i) \\
&= \sum_k[(\psi_k(\mathbf{x}_i) - \psi_k(\mathbf{x}_j))\psi_k(\mathbf{x}_i) + (\psi_k(\mathbf{x}_j) - \psi_k(\mathbf{x}_i))\psi_k(\mathbf{x}_j)]t_k \\
&= \sum_k t_k(\psi(\mathbf{x}_i) - \psi_k(\mathbf{x}_j))^2 = 2d^2_{\mathrm{comm}}(\mathbf{x}_i, \mathbf{x}_j).
\end{aligned}$$

In the case of a discrete-state system (as the example discussed in section 2.3), the set of eigenfunctions is finite and the sum can be evaluated. In more complex applications, in practice the sum is truncated to the first $n < \infty$ terms, and $2t^2_{\mathrm{comm}}(\mathbf{x}_i, \mathbf{x}_j)$ is a lower bound to the commute $t_{\mathrm{comm}}(\mathbf{x}_i, \mathbf{x}_j)$ (see also discussion in section 2.4).

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: frank.noe@fu-berlin.de.
*E-mail: cecilia@rice.edu.

Funding

C.C. is supported by National Science Foundation (Grants CHE-1152344 and CHE-1265929) and the Welch Foundation (Grant C-1570). F.N. is supported by European Commission ERC starting Grant 307494-pcCell and Deutsche Forschungsgemeinschaft (Grant NO 825/3-1).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to D. E. Shaw Research for sharing the BPTI data.

■ REFERENCES

(1) Shirts, M.; Pande, V. S. Screen Savers of the World Unite! *Science* **2000**, *290*, 1903–1904.

(2) Buch, I.; Harvey, M. J.; Giorgino, T.; Anderson, D. P.; De Fabritiis, G. High-throughput all-atom molecular dynamics simulations using distributed computing. *J. Chem. Inf. Model.* **2010**, *50*, 397–403.

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

(4) Mei, C.; Sun, Y.; Zheng, G.; Bohm, E. J.; Laxmikant, K.; Phillips, J. C.; Harrison, C. Enabling and Scaling Biomolecular Simulations of 100 Million Atoms on Petascale Machines with a Multicore-optimized Message-driven Runtime. Proceedings of the 2011 ACM/IEEE conference on Supercomputing. Seattle, WA, 2011.

(5) Le Grand, S.; Goetz, A. W.; Walker, R. C. SPFP: Speed without compromise - a mixed precision model for GPU accelerated molecular dynamics simulations. *Comput. Phys. Commun.* **2013**, *184*, 374–380.

(6) Doerr, S.; De Fabritiis, G. On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations. *J. Chem. Theory Comput.* **2014**, *10*, 2064–2069.

(7) Preto, J.; Clementi, C. Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. *Phys. Chem. Chem. Phys.* **2014**, *16*, 19181–19191.

(8) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854.

(9) Plattner, N.; Noé, F. Protein conformational plasticity and complex ligand binding kinetics explored by atomistic simulations and Markov models. *Nat. Commun.* **2015**, *6*, 7653.

(10) Pronk, S.; Pouya, I.; Lundborg, M.; Rotskoff, G.; Wesén, B.; Kasson, P. M.; Lindahl, E. Molecular Simulation Workflows as Parallel Algorithms: The Execution Engine of Copernicus, a Distributed High-Performance Computing Platform. *J. Chem. Theory Comput.* **2015**, *11*, 2600–2608.

(11) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **2016**, *12*, 1845–1852.

(12) Zimmerman, M. I.; Bowman, G. R. FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *J. Chem. Theory Comput.* **2015**, *11*, 5747–5757.

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

DOI: 10.1021/acs.jctc.6b00762

(14) Buch, I.; Giorgino, T.; De Fabritiis, G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 10184−10189.

(15) Bowman, G. R.; Geissler, P. L. Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 11681−11686.

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

(17) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. A Direct Approach to Conformational Dynamics based on Hybrid Monte Carlo. *J. Comput. Phys.* **1999**, *151*, 146−168.

(18) Deuflhard, P., Weber, M. In *Linear Algebra Applications*; Dellnitz, M., Kirkland, S., Neumann, M., Schütte, C., Eds.; Elsevier: New York, 2005; Vol. *398C*; pp 161−184.

(19) Peters, B.; Trout, B. L. Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys.* **2006**, *125*, 054108.

(20) Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 9885−9890.

(21) 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.

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

(23) Buchete, N. V.; Hummer, G. Coarse Master Equations for Peptide Folding Dynamics. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.

(24) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chem. Phys.* **2007**, *126*, 244111.

(25) Stamati, H.; Clementi, C.; Kavraki, L. E. Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 223−235.

(26) Noé, F.; Wu, H.; Prinz, J.-H.; Plattner, N. Projected and Hidden Markov Models for calculating kinetics and metastable states of complex molecules. *J. Chem. Phys.* **2013**, *139*, 184114.

(27) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMBuilder2: Modeling Conformational Dynamics at the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7*, 3412−3419.

(28) Senne, M.; Trendelkamp-Schroer, B.; Mey, A. S. J. S.; Schütte, C.; Noé, F. EMMA - A software package for Markov model building and analysis. *J. Chem. Theory Comput.* **2012**, *8*, 2223−2238.

(29) Noé, F.; Nüske, F. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Model. Simul.* **2013**, *11*, 635−655.

(30) Rohrdanz, M. A.; Zheng, W.; Clementi, C. Discovering mountain passes via torchlight: methods for the definition of reaction coordinates and pathways in complex macromolecular reactions. *Annu. Rev. Phys. Chem.* **2013**, *64*, 295−316.

(31) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.* **2011**, *134*, 124116.

(32) Bowman, G. R., Pande, V. S., Noé, F., Eds. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation.*; Advances in Experimental Medicine and Biology; Springer: Heidelberg, 2014; Vol. 797.

(33) Ledbetter, P. J.; Clementi, C. A new perspective on transition states: χ^1: separatrix. *J. Chem. Phys.* **2011**, *135*, 044116.

(34) Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. Markov State Models Based on Milestoning. *J. Chem. Phys.* **2011**, *134*, 204105.

(35) Perez-Hernandez, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*, 015102.

(36) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000−2009.

(37) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Perez-Hernandez, G.; Hoffmann, M.; Plattner, N.; Prinz, J.-H.; Noé, F.; Wehmeyer, C. PyEMMA 2: A software package for estimation, validation and analysis of Markov models. *J. Chem. Theory Comput.* **2015**, *11*, 5525−5542.

(38) Noé, F.; Clementi, C. Kinetic distance and kinetic maps from molecular dynamics simulation. *J. Chem. Theory Comput.* **2015**, *22*, 5002−5011.

(39) Boninsegna, L.; Gobbo, G.; Noé, F.; Clementi, C. Investigating Molecular Kinetics by Variationally Optimized Diffusion Maps. *J. Chem. Theory Comput.* **2015**, *11*, 5947−5960.

(40) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B. G.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov models of molecular kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134*, 174105.

(41) Schwantes, C. R.; Pande, V. S. Modeling Molecular Kinetics with tICA and the Kernel Trick. *J. Chem. Theory Comput.* **2015**, *11*, 600−608.

(42) Schölkopf, B.; Smola, A.; Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* **1998**, *10*, 1299−1319.

(43) Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noé, F. Variational Approach to Molecular Kinetics. *J. Chem. Theory Comput.* **2014**, *10*, 1739−1752.

(44) Nüske, F.; Schneider, R.; Vitalini, F.; Noé, F. Variational Tensor Approach for Approximating the Rare-Event Kinetics of Macromolecular Systems. *J. Chem. Phys.* **2015**, *144*, 054105.

(45) Sarich, M.; Noé, F.; Schütte, C. On the approximation quality of Markov state models. *Multiscale Model. Simul.* **2010**, *8*, 1154−1177.

(46) Sheong, F. K.; Silva, D.-A.; Meng, L.; Zhao, Y.; Huang, X. Automatic State Partitioning for Multibody Systems (APM): An Efficient Algorithm for Constructing Markov State Models To Elucidate Conformational Dynamics of Multibody Systems. *J. Chem. Theory Comput.* **2015**, *11*, 17−27.

(47) McGibbon, R. T.; Pande, V. S. Variational cross-validation of slow dynamical modes in molecular kinetics. *J. Chem. Phys.* **2015**, *142*, 124105.

(48) Doyle, P. G.; Steiner, J. Commuting time geometry of ergodic Markov chains. 2011, arXiv:1107.2612 (accessed 2016).

(49) Molgedey, L.; Schuster, H. G. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.* **1994**, 72, 3634−3637.

(50) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 7426−7431.

(51) Nadler, B.; Lafon, S.; Coifman, R. R.; Kevrekidis, I. G. Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators. *Adv. Neural Inf. Process. Syst. (NIPS)* **2005**, 955−962.

(52) Sriraman, S.; Kevrekidis, I. G.; Hummer, G. Coarse Master Equation from Bayesian Analysis of Replica Molecular Dynamics Simulations. *J. Phys. Chem. B* **2005**, *109*, 6479−6484.

(53) Suárez, D.; Díaz, N. Sampling Assessment for Molecular Simulations Using Conformational Entropy Calculations. *J. Chem. Theory Comput.* **2014**, *10*, 4718−4729.

DOI: 10.1021/acs.jctc.6b00762