Identification of slow molecular order parameters for Markov model construction

Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé

Citation: J. Chem. Phys. **139**, 015102 (2013); doi: 10.1063/1.4811489
View online: http://dx.doi.org/10.1063/1.4811489
View Table of Contents: http://jcp.aip.org/resource/1/JCPSA6/v139/i1
Published by the AIP Publishing LLC.

Identification of slow molecular order parameters for Markov model construction

Guillermo Pérez-Hernández,[1] Fabian Paul,[2,a),b)] Toni Giorgino,[3,a)] Gianni De Fabritiis,[4,c)] and Frank Noé[1,c)]

[1]*Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany*
[2]*Max-Planck-Institut for Colloids and Interfaces, Division Theory and Bio-Systems, Science Park Potsdam-Golm, 14424 Potsdam, Germany*
[3]*Institute of Biomedical Engineering (ISIB), National Research Council of Italy (CNR), Corso Stati Uniti 4, I-35127 Padua, Italy*
[4]*GRIB, Barcelona Biomedical Research Park (PRBB), C/Dr. Aiguader 88, 08003 Barcelona, Spain*

(Received 11 December 2012; accepted 5 June 2013; published online 3 July 2013)

A goal in the kinetic characterization of a macromolecular system is the description of its slow relaxation processes via (i) identification of the structural changes involved in these processes and (ii) estimation of the rates or timescales at which these slow processes occur. Most of the approaches to this task, including Markov models, master-equation models, and kinetic network models, start by discretizing the high-dimensional state space and then characterize relaxation processes in terms of the eigenvectors and eigenvalues of a discrete transition matrix. The practical success of such an approach depends very much on the ability to finely discretize the slow order parameters. How can this task be achieved in a high-dimensional configuration space without relying on subjective guesses of the slow order parameters? In this paper, we use the variational principle of conformation dynamics to derive an optimal way of identifying the "slow subspace" of a large set of prior order parameters – either generic internal coordinates or a user-defined set of parameters. Using a variational formulation of conformational dynamics, it is shown that an existing method—the time-lagged independent component analysis—provides the optional solution to this problem. In addition, optimal indicators—order parameters indicating the progress of the slow transitions and thus may serve as reaction coordinates—are readily identified. We demonstrate that the slow subspace is well suited to construct accurate kinetic models of two sets of molecular dynamics simulations, the 6-residue fluorescent peptide MR121-GSGSW and the 30-residue intrinsically disordered peptide kinase inducible domain (KID). The identified optimal indicators reveal the structural changes associated with the slow processes of the molecular system under analysis. © *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4811489]

I. INTRODUCTION

Conformational transitions between long-lived or "metastable" states are essential to the function of biomolecules.[1–7] These rare transitions are ubiquitously found in biomolecular processes including folding,[8,9] complex conformational rearrangements between native protein substates[10,11] and ligand binding.[12] Rare conformational transitions can be explicitly traced by either single-molecule experiments[9,13–15] or by high-throughput molecular dynamics (MD) simulations, either realized with few long trajectories[16,17] or with many shorter trajectories.[18–23] MD simulations are unique in their ability to resolve the dynamics and all structural features of a biomolecule simultaneously. When the sampling problem can be overcome and the appropriateness of the force field parameters used is confirmed by accompanying experimental evidence, MD simulations are amongst the most powerful tools to investigate conformational transitions in biomolecules.

A current challenge with high-throughput MD simulations is to extract meaningful information from vast trajectory data in an objective way. To achieve this goal, the last few years have seen vast activity in the development of computational methods that extract kinetic models from the MD data. Kinetic models usually first partition the conformation space into discrete states.[24–34] Subsequently, transition rates or probabilities can be estimated.[31,33–38] The resulting models are often called transition networks,[30,36,39] diffusion maps,[40,41] master equation models,[29,42] Markov models[43] or Markov state models[37,44] (MSM), where "Markovianity" means that the kinetics are modeled by a memoryless jump process between states.

The recent integration of classical statistical mechanics with modern molecular kinetics highlights the crucial role of the eigenvectors and eigenvalues of the Markov model transition matrix or master equation rate matrix. This is because they approximate the exact eigenfunctions and eigenvalues

[a)]F. Paul and T. Giorgino contributed equally to this work.
[b)]Also at Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany.
[c)]Authors to whom correspondence should be addressed. Electronic addresses: gianni.defabritiis@upf.edu and frank.noe@fu-berlin.de

of the propagator of the continuous dynamics.[45] The following eigenvalue equation is fundamental to conformation dynamics:

$$\mathcal{P}\phi_i = \lambda_i \phi_i. \quad (1)$$

Here, $\mathcal{P}$ is the transfer operator that propagates probability densities of molecular configurations,[46,47] ϕ_i are its eigenfunctions, and λ_i are the associated eigenvalues. Equivalent expressions are obtained by expressing the eigenfunctions in different weighted spaces, leading to the transfer operator formulation[46] or the symmetrized propagator formulation.[29] Equation (1) is fundamental because when solving it for the largest eigenvalues and associated eigenfunctions, all stationary and kinetic quantities are defined by them. For example:

- $\mathcal{P}$ is guaranteed to have a unitary stationary eigenvalue and the associated stationary distribution $\mu(\mathbf{x}) = \phi_1(\mathbf{x})$. The ensemble average of an observable o can be calculated from $o(\mathbf{x})$ and $\mu(\mathbf{x})$.
- Experimentally measurable relaxation rates of the system can be computed from the eigenvalues as $\kappa_i = -\tau^{-1}\ln\lambda_i$, or from the corresponding timescales as $t_i = \kappa_i^{-1}$.
- The metastable states (often referred to as "free energy basins"—although we will avoid this term as it would imply the projection onto some pre-defined coordinate set) can be computed from the sign structure of the leading eigenfunctions.[46,48]
- The structural transition associated to each relaxation timescale is defined by the corresponding eigenfunction[49] and corresponds to a transition between metastable sets. This fact can be used to assign structural changes to experimentally measurable timescales.[50]
- Experimentally measurable correlation functions (e.g., fluorescence correlation, intermediate scattering function in dynamic neutron or X-ray scattering) can be computed as a sum of single-exponential relaxations with timescales computed from λ_i and amplitudes from the ϕ_i and the experimental observable.[50–52]
- From the largest m eigenvalues and their associated eigenfunctions, a rank-m propagator can be assembled that can describe the dynamics slower than timescale t_m.[53] From this propagator, many properties can be calculated, such as transition pathways between two sets of configurations.[20,54,55]

The approximation error of all of the above quantities can be cast in terms of the approximation error of the eigenvalues and eigenfunctions.[45,49,56,57] Vice versa, all of the above quantities are easily and precisely computable when the eigenvalues and eigenfunctions of $\mathcal{P}$ have been approximated with high precision. Consequently, any modeling method that attempts to compute the above quantities must aim at approximating the eigenvalues and eigenfunctions of $\mathcal{P}$ – either explicitly or implicitly.

Markov models and most of the other aforementioned kinetic models require a discretization of configuration space to be made. This is typically done by choosing representative configurations by some data clustering method, and then partitioning the configuration space by a Voronoi tesselation. In contrast to other fields of data analysis, the purpose of clusters is not a classification of configurations, but rather a sufficiently fine discretization of configuration space such that the eigenfunctions can be well approximated in terms of step functions on the Voronoi cells.[49] In order to achieve this, the metric must be chosen as a fine partition of the relevant "slow" order parameters, i.e., those which are good indicators of the slow eigenfunctions ϕ_i.

How can the slow order parameters be identified without already having a high-precision Markov model? It has been noted that *a priori* order parameters such as the root mean square distance (RMSD) to a single reference structure, the radius of gyration, or pre-selected distances or angles are often not good indicators of the slow eigenfunctions, and thus bear the danger of disguising the slow kinetics.[25,31,35,58] In order to avoid this, Markov model construction has focused in the last years on the other extreme – using general metrics that are capable of describing every sort of configurational change. Most notable is the minimal RMSD metric, which assigns to each pair of configurations their minimal Euclidean distance subject to rigid-body translation and rotation.[59] Minimal RMSD has been used successfully in many examples, especially protein folding (see Refs. 49 and 60 and references therein). Recent applications include folding of MR121-GSGS-W peptide,[49] folding of FiP35 WW domain, GTT, NTL9, and protein G,[61] and discovery of cryptic allosteric sites in β-lactamase, interleukin-2, and RNase H.[62] However, minimal RMSD tends to fail in situation where the largest-amplitude motions are not the slowest (an example of this is the intrinsically disordered kinase inducible domain (KID) peptide analyzed below). Principal component analysis (PCA) is a frequently used method to reduce the dimension of an order parameter space by projecting it on its linear subspace of the largest-amplitude motions.[63] PCA has also been used successfully in Markov model construction,[20,50] however, it suffers from the similar limitations as minimal RMSD, as there is no general guarantee that large-amplitude motions are associated with slow transitions.

It is an important challenge to find a metric that provides a good indicator of the slow processes, such that a good approximation of the eigenfunctions ϕ_i is feasible with a moderate number of clusters. The aim of this paper is to identify such a method. To be more precise, let $r_1, \ldots, r_d \in \mathbb{R}$ be a possibly large set of d order parameters of a molecular system, which are *a priori* specified by the user. Typical examples of order parameter include intramolecular distances and torsion angles. However, complex order parameters like the instantaneous dipole moment of a molecule, or an experimentally measurable quantity such as a Förster resonance energy transfer (FRET) efficiency may also be included. Given this set of order parameters, we aim to

1. Find the linear combination of order parameters, which optimally approximates the dominant eigenvalues and eigenfunctions, such that a high-precision Markov model can be built in these order parameters with direct clustering.

2. Identify the m order parameters that are best and least redundant indicators for the m dominant eigenfunctions, thus providing the user a direct physical interpretation whose structural changes are associated with the slowest relaxation timescales (feature selection).

Here we use the variational principle of conformation dynamics[47] to derive an optimal solution for problem 1, and show that an existing extension to PCA solves this problem: time-lagged independent component analysis (TICA) combines information from the covariance matrix and a time-lagged covariance matrix of the data.[64] See Ref. 65 for a detailed description of the method. TICA has recently been applied in the analysis of MD data. Naritomi and Fuchigami[66] used TICA to investigate domain motion of the Lysine-, arginine-, ornithine-binding (LAO) protein and compared it to PCA. Mitsutake *et al.*[67] used relaxation mode analysis, a related technique, to analyze the dynamics of Met-enkephalin. Both studies showed that the slow modes were not necessarily associated with large amplitudes, and time-lagged mode analyses were thus better suited to detect them than PCA. While revising this manuscript, another successful application of TICA to Markov model building by the group of Pande and Schwantes[68] has appeared.

Here, we demonstrate the usefulness of TICA coordinates for constructing Markov models for two rather different molecular processes: the conformational dynamics of (i) the small fluorescent peptide MR121-GSGSW, for which good Markov models can be built using a variety of methods, and (ii) the intrinsically disordered 30-residue peptide KID modeled through a large ensemble of explicit-solvent MD simulations.

We also propose a way to approach problem 2 (see above) identifying the optimal indicators of the slowest processes. These indicators inform the user of the structural process that is governing the slow relaxations of the macromolecule. Optimal indicators help in understanding what comprises the slow kinetics, and dramatically the user time to "search" for a structural character of the slow processes from a Markov model.

II. THEORY

We first summarize the variational principle of conformation dynamics stating that the true eigenfunctions are best approximated by a Markov model when the estimated timescales are maximized. We then derive a way to optimally approximate the true eigenfunctions in terms of a linear combination of the original order parameters. It is shown that this method is identical to the TICA that is an established method in statistics. This establishes a new connection between TICA and the optimal approximation of the molecule's relaxation timescales. As a result, TICA provides the optimal linear way of projecting simulation data in order to build Markov models. The TICA problem can easily be solved by subsequently solving two simple eigenvalue problems.

A. Exact dynamics in full configuration space

We start by providing an expression for the propagator of exact continuous molecular dynamics, and show that in order to approximate its long-time behavior, its largest eigenvalues and associated eigenfunctions must be well approximated. The following paragraphs are a short and educative summary of results from Ref. 46. See also Ref. 49 for more details.

We use $\mathbf{x}_t$ to denote the full molecular configuration at time t (if velocities are available, $\mathbf{x}_t$ denotes a point in full phase space) in state or phase space Ω. We assume that the molecular dynamics implementation is Markovian in Ω (i.e., the time step to $\mathbf{x}_{t+\tau}$ is computed based on the current value of $\mathbf{x}_t$ only), and gives rise to a unique stationary density $\mu(\mathbf{x})$, usually the Boltzmann density:

$$\mu(\mathbf{x}) = Z^{-1}\mathrm{e}^{-\beta H(\mathbf{x})},$$

where H is the Hamiltonian, Z is the partition function, and $\beta = (k_B T)^{-1}$ is the inverse temperature. We also assume that the dynamics are statistically reversible, i.e., that the molecular system is simulated in thermal equilibrium. Let us denote a probability density of molecular configurations as ρ_t, and let us subsume the action of the molecular dynamics implementation into the propagator $\mathcal{P}(\tau)$. The propagator describes the probability that a trajectory that is at configuration $\mathbf{x}_t$ at time t will be found at a configuration $\mathbf{x}_{t+\tau}$ a time τ later. In an ensemble view, the propagator takes a probability density of configurations, ρ_t, and predicts the probability density of configurations at later time, $\rho_{t+\tau}$:

$$\rho_{t+\tau} = \mathcal{P}(\tau)\rho_t.$$

We can write the propagator by expanding it in terms of its eigenvalues,

$$\lambda_i(\tau) = \mathrm{e}^{-\frac{\tau}{t_i}},$$

and its eigenfunctions ϕ_i as

$$\rho_{t+\tau}(\mathbf{y}) = \mathcal{P}(\tau)\rho_t(\mathbf{x}) = \sum_{i=1}^{\infty} \mathrm{e}^{-\frac{\tau}{t_i}} \langle \psi_i, \rho_t \rangle \phi_i, \qquad (2)$$

where the eigenfunctions $\phi_i(\mathbf{x})$ take the role of basis functions with which probability densities ρ can be constructed. The first eigenvalue is $\lambda_1 = 1$ and the remaining eigenvalues have a norm strictly smaller than 1. Thus, the first timescale is $t_1 = \infty$ and corresponds to the stationary distribution, while all other timescales t_i are finite relaxation timescales. $\psi_i(\mathbf{x}) = \mu^{-1}(\mathbf{x})\phi_i(\mathbf{x})$ are the eigenfunctions weighted by the inverse of the stationary density. Equation (2) has a straightforward physical interpretation: the scalar product $\langle \psi_i, \rho_t \rangle$ measures the overlap of the starting density ρ_t with the ith eigenfunction and thus determines the amplitude by which this eigenfunction contributes to the dynamics. At any time τ, the new probability density $\rho_{t+\tau}$ is composed of a set of basis functions ϕ_i. With increasing time, the contributions of all basis functions ϕ_i with $i > 1$ vanish exponentially with a timescale given by t_i. After infinite time $\tau \to \infty$, only the first term with $t_1 = \infty$ (and hence, $\exp(-\tau/t_i) = 1$) is left, and the stationary density is reached: $\lim_{\tau\to\infty} \mathcal{P}(\tau)\rho_t = \phi_1 = \mu$. Stationarity implies that μ will not be changed under the action of the propagator:

$$\mathcal{P}(\tau)\mu = \mu.$$

Suppose we are interested in slow timescales $\tau \gg t_{m+1}$. At such large times, the dynamics is governed by the m largest

timescales t_i and eigenfunctions of the propagator:

$$\rho_{t+\tau} = \mathcal{P}(\tau)\rho_t \approx \sum_{i=1}^{m} \mathrm{e}^{-\frac{\tau}{t_i}} \langle \psi_i, \rho_t \rangle \phi_i .$$

All kinetic properties at this timescale and all stationary properties can be accurately computed when the dominant m eigenvalues and eigenfunctions are approximated. This is our goal.

B. Approximation of slowest timescales and the related eigenfunctions

We can make a few general statements on how to approximate the true timescales t_i and eigenfunctions following Ref. 47. These general properties can be used to derive a general method that achieves the aim of this paper: the identification of the slowest order parameters in a molecule. Since ϕ_i and ψ_i are interchangeable using the weights μ, the approximation problem can be described using either kind of eigenfunction. Subsequently, we will always refer to the problem of approximating the weighted eigenfunctions ψ_i.

Consider some function of the molecular configuration, $f(\mathbf{x})$. From Eq. (2), we can express the time-autocorrelation function of f as a function of τ as

$$\langle f(\mathbf{x}_t) f(\mathbf{x}_{t+\tau}) \rangle_t = \sum_{i=1}^{\infty} \mathrm{e}^{-\frac{\tau}{t_i}} \langle \phi_i, f \rangle^2 . \tag{3}$$

Suppose we would know the true eigenfunction $\psi_i(\mathbf{x})$. It is now easy to show[47] that the time-autocorrelation function of $\psi_i(\mathbf{x})$ yields the exact ith eigenvalue, and thus the model timescale $t_i^{\ddagger}$ is identical to the exact ith timescale t_i:

$$\lambda_i^{\ddagger}(\tau) = \langle \psi_i(\mathbf{x}_t)\psi_i(\mathbf{x}_{t+\tau}) \rangle_t = \mathrm{e}^{-\frac{\tau}{t_i}},$$
$$t_i^{\ddagger} = -\frac{\tau}{\ln|\lambda_i^{\ddagger}(\tau)|} = t_i,$$

where model variables are denoted with a double dagger (‡) (see Fig. 1 for illustration).

However, in reality we will not know the exact eigenfunction ψ_i. Suppose that we would *guess* a model function $\psi_2^{\ddagger}$ that is supposed to be similar to ψ_2. When we make sure that $\psi_2^{\ddagger}$ is appropriately normalized, the variational principle of conformation dynamics[47] shows that the time-autocorrelation function of $\psi_2^{\ddagger}$ approximates the true eigenvalue, and the true timescale from below:

$$\langle \psi_2^{\ddagger}(\mathbf{x}_t)\psi_2^{\ddagger}(\mathbf{x}_{t+\tau}) \rangle_t \leq \mathrm{e}^{-\frac{\tau}{t_2}},$$
$$t_2^{\ddagger} \leq t_2, \tag{4}$$

where equality only holds for $\psi_2^{\ddagger} = \psi_2$. Thus, we have a recipe for finding an optimal approximation to the second timescale and its associated eigenfunction. We must seek a function $\psi_2^{\ddagger}$ that has the maximum timescale $t_2^{\ddagger}$.

Similar inequalities can be shown for the other eigenvalues and timescales $t_3, \ldots, t_m$. We can show that if one proposes a model function $\psi_i^{\ddagger}$ that is orthogonal to the exact eigenfunctions 1 through $i-1$, we also have

$$t_i^{\ddagger} \leq t_i . \tag{5}$$

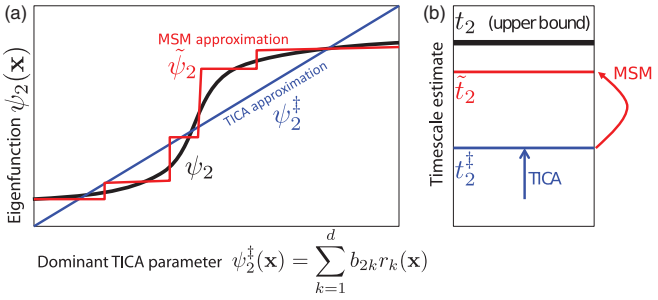

FIG. 1. Scheme illustrating different approximations to the dominant eigenfunction ψ_2 of the molecular dynamics propagator, and the associated approximations to the slowest relaxation timescale t_2. TICA ($\psi_2^{\ddagger}$, blue) approximates the eigenfunction ψ_2 (black) as a linear combination of molecular observables, and the TICA timescale $t_2^{\ddagger}$ associated to the TICA eigenvalue $\lambda_2^{\ddagger}$ underestimates (usually strongly) the true timescale t_2. The estimate is then improved by building a Markov model in TICA space, which approximates the eigenfunction ψ_2 by a step function ($\tilde{\psi}_2$, red) that is constant on the Markov model clusters. The corresponding Markov model estimate of the relaxation timescale, $\tilde{t}_2$, is thus typically larger than the TICA timescale $t_2^{\ddagger}$ and a better estimate of the true timescale t_2.

This variational principle of conformation dynamics is analogous to the variational principle in quantum mechanics.

C. Best approximation of the eigenfunctions

What is the relation of the variational principle above to Markov models? Since the eigenfunctions ψ_i are initially unknown and difficult to guess, it is reasonable to approximate them by functions $\psi_i^{\ddagger}$ that are assembled from a linear combination of basis functions,

$$\psi_i^{\ddagger}(\mathbf{x}) = \sum_{k=1}^{n} b_{ik} \chi_k(\mathbf{x}), \tag{6}$$

which must be defined *a priori*, and the optimization problem then consists of finding the optimal parameters b_{ik} that we will denote by vectors $\mathbf{b}_i \in \mathbb{R}^n$, where we have chosen the dimension of the basis set, n, to be equal to the number of basis functions. The Ritz method[69] provides the optimal set of coefficients for an orthonormal basis set. Formally, if we define the covariance matrix between Ansatz functions as

$$c_{ij}^{\chi}(\tau) = \langle \chi_i(\mathbf{x}_t)\chi_j(\mathbf{x}_{t+\tau}) \rangle_t ,$$

and we require that the basis functions are orthogonal—which is equivalent to them being uncorrelated at lag time 0:

$$\langle \chi_i, \chi_j \rangle_\mu = \langle \chi_i(\mathbf{x}_t)\chi_j(\mathbf{x}_t) \rangle_t = c_{ij}^{\chi}(0) = \delta_{ij}, \tag{7}$$

then the optimal set of coefficients is then given by the eigenvectors $\mathbf{b}_i$ of the following eigenvalue problem:

$$\mathbf{C}^{\chi}(\tau)\mathbf{b}_i = \mathbf{b}_i \lambda_i^{\ddagger}(\tau). \tag{8}$$

Let us now consider the more general case that the Ansatz functions are not orthonormal, i.e., $\langle \chi_i, \chi_j \rangle_\mu \neq \delta_{ij}$. In this situation, we must first orthonormalize the basis coordinates before. This is done via a generalization to Eq. (8). For a non-orthonormal basis set, the optimal approximation to the true eigenvalues and eigenfunctions is obtained by solving the

generalized eigenvalue problem:

$$\mathbf{C}^{\chi}(\tau)\mathbf{b}_i = \mathbf{C}^{\chi}(0)\mathbf{b}_i\lambda_i^{\ddagger}(\tau). \tag{9}$$

One may formally rewrite Eq. (9) as $\mathbf{D}(\tau)\mathbf{b}_i = \lambda_i^{\ddagger}\mathbf{b}_i$, where $\mathbf{D}(\tau) = (\mathbf{C}^{\chi}(0))^{-1}\mathbf{C}^{\chi}(\tau)$ is an orthonormal basis set. Numerically, the matrix inversion of $\mathbf{C}^{\chi}(0)$ is often poorly conditioned and should therefore be avoided. The results (8) and (9) are well known from variational calculus. The supplementary material[70] contains an illustrative derivation of Eq. (9) relevant to the special choice of basis set used in this paper.

D. Optimal linear combination of input order parameters

We are now ready to make the main theoretical step of the present contribution. Based on the above results, we can now formulate a method to find a linear combination of molecular order parameters $\mathbf{r} = (r_1(\mathbf{x}), \ldots, r_d(\mathbf{x}))$, which best resolves the slow relaxation processes. This is done by finding the optimal coefficients for Eq. (6). For this, we define the basis function χ_i to be identical to the mean-free coordinate $r_i(\mathbf{x})$ (if the original order parameters $r_i'(\mathbf{x})$ are not mean-free, then we simply subtract the mean, $r_i(\mathbf{x}) = r_i'(\mathbf{x}) - \langle r_i(\mathbf{x})\rangle_t$):

$$\chi_i(\mathbf{x}) = r_i(\mathbf{x}). \tag{10}$$

Thus, our basis set has $n = d$ dimensions. Now let us compute the correlation matrix of normalized order parameters as

$$c_{ij}^{r}(\tau) = \langle r_i(\mathbf{x}_t)r_j(\mathbf{x}_{t+\tau})\rangle_t = c_{ij}^{\chi}(\tau).$$

Then solving Eq. (9) with the correlation matrix for lag times 0 and τ will provide us with the linear combination of input order parameters that optimally approximates the exact propagator eigenfunctions. See the supplementary material[70] for a sketch of the usual derivation of Eq. (9) for the case of TICA. It so happens that Eq. (9) with the choice of coordinates (10) is known as the TICA in statistics.[64,66] This insight establishes that TICA is an optimal approach (amongst the linear projection methods) to approximate molecular relaxation timescales, and therefore ideally suited to construct Markov models. A robust algorithm to solve Eq. (9) is known as AMUSE algorithm[71] and will be given below.

The eigenfunction approximations via Eq. (6) using the coefficients $\mathbf{b}_i$ are the optimal approximation to the true eigenfunctions and will give an optimal approximation of the timescales. As a result of the variational principle, $\lambda_2^{\ddagger}(\tau) \leq \lambda_2(\tau)$ and

$$t_2^{\ddagger}(\tau) = -\frac{\tau}{\ln\lambda_2^{\ddagger}(\tau)} \leq t_2$$

according to Eq. (4). However, since the true eigenfunctions are generally nonlinear functions of the original order parameters, and the basis set used in Eq. (10) is linear in the original order parameters, it cannot be expected that $\psi_2 \approx \psi_2^{\ddagger}$ is true, and therefore the variational principle can at this point not be extended to further timescales than t_2. In other words, the TICA timescales $t_3^{\ddagger}, \ldots, t_m^{\ddagger}$ may be both under- or overestimated.

E. Markov models and implied timescales

We do not intend to use the TICA timescales directly, but rather use the TICA subspace in order to construct a Markov model by finely discretizing this space. What can be said about the timescales of the resulting Markov model? We can use the variational principle summarized above to bound the timescales of the Markov model. Classical Markov models operate by assigning a configuration $\mathbf{x}$ uniquely to one of the n-geometric clusters used to construct them. It can be shown[45] that this operation is equivalent to use the basis functions

$$\chi_i(\mathbf{x}) = \frac{\mathbf{1}_i(\mathbf{x})}{\sqrt{\pi_i}},$$

i.e., each basis function i is a step function with has a constant value on the configurations belonging to the ith cluster and is zero elsewhere. This basis is an orthonormal basis set: $\langle\chi_i, \chi_j\rangle_\mu = \pi_i^{-1}\int_{\mathbf{x}\in S_i}\mu(\mathbf{x})\,d\mathbf{x} = \delta_{ij}$. Thus, the direct Ritz method applies and as shown in Ref. 47, Eq. (8) becomes

$$\mathbf{T}(\tau)\tilde{\boldsymbol{\psi}}_i = \tilde{\boldsymbol{\psi}}_i\tilde{\lambda}_i(\tau), \tag{11}$$

where $\mathbf{T}(\tau)$ is the Markov model row-stochastic transition matrix, $\boldsymbol{\Psi} = [\tilde{\boldsymbol{\psi}}_1, \ldots, \tilde{\boldsymbol{\psi}}_n]$ are its right eigenvectors, and $\tilde{\lambda}_i(\tau)$ is the eigenvalue estimated form the Markov model (from now on variables with tilde denote quantities estimated via the Markov model). To relate Eqs. (8) and (11), we have used the definition $\mathbf{C}^{\chi}(\tau) = \sqrt{\frac{\pi_i}{\pi_j}}T_{ij}(\tau)$, i.e., the covariance matrix between Ansatz functions χ is the symmetrized transition matrix as given in Ref. 29.

Thus, a Markov model is the Ritz method for the choice of a step-function basis on the clusters used to build it, and thus gives an optimal step-function approximation to the eigenfunctions and maximal eigenvalues amongst all choices of functions that can be supported by the clustering. It follows from Eq. (4) that at least the second timescale will then be underestimated. When the Markov model is sufficiently good in approximating the slowest processes, all of the first m timescales will be underestimated as given by Eq. (5). It was shown[56] that this estimation error becomes smaller when τ is increased. Prinz *et al.*[57] showed that it decreases with τ^{-1}. As a result, when plotting the estimated timescales $\tilde{t}_i(\tau)$ as a function of τ, one obtains the well-known implied timescale plots shown in Figs. 2 and 4, where the estimated timescales $\tilde{t}_i(\tau)$ slowly converge to the true timescale when τ is increased.

We have now seen that both the TICA eigenvalue $\lambda_2^{\ddagger}$ and the corresponding timescale $t_2^{\ddagger}$ are underestimated, as well as the Markov model eigenvalue $\tilde{\lambda}_2$ and the corresponding timescale $\tilde{t}_2$. Unfortunately, we cannot make a rigorous statement of how $t_2^{\ddagger}$ and $\tilde{t}_2$ are related to each other. However, we can make the *ad hoc* statement that we intend to cluster the dominant TICA subspace "sufficiently fine." Thereby the Markov model step functions of the dominant TICA component allow the nonlinear eigenfunction $\psi_2(\mathbf{x})$ to be approximated better than by the linear combination of order parameters (10) directly. For example, it is typical that the eigenfunction $\psi_2(\mathbf{x})$ stays almost constant over a large part of configuration space and then changes abruptly to a different level in the transition state.[46,49] Such a behavior can be much

better described by a step function than by a linear fit. Therefore, we shall here assume that the estimates of the dominant timescale as $t_2^{\ddagger} < \tilde{t}_2 < t_2$: the dominant TICA timescale $t_2^{\ddagger}$ is a lower bound to the true timescale t_2, but typically a poor lower bound. The Markov model timescale $\tilde{t}_2$ is typically larger, and thus a better estimate of the true timescale t_2. This concept is illustrated in Fig. 1.

III. METHODS

A. Principal component analysis and time-lagged independent component analysis

In the present paper, we use PCA in two ways: (1) as a direct dimension reduction tool to yield a subspace for clustering and subsequent Markov model construction in that space, and (2) to transform the original data into the full set of principal components, thus arriving at a decorrelated coordinate set as an input for the subsequent transform into time-lagged independent components.

Like PCA, TICA[64] uses a linear transform to map the original order parameters $\mathbf{r}(t)$ to a new set of order parameters $\mathbf{z}(t)$—the independent components (ICs). Unlike PCs, ICs have to fulfill *two* properties:

1. They are uncorrelated.
2. Their autocovariances at a fixed lag time τ are maximal.

The time-lagged covariance matrix $\mathbf{C}^r_\tau(\tau)$ is defined by

$$c^r_{ij}(\tau) = \langle r_i(t) r_j(t+\tau) \rangle_t,$$

and the estimator for trajectory data containing N time steps is given by

$$c^r_{ij}(\tau) = \frac{1}{N-\tau-1} \sum_{t=1}^{N-\tau} r_i(t) r_j(t+\tau).$$

The elements of $\mathbf{C}^r(\tau)$ are time-lagged autocovariances if $i = j$, and time-lagged cross covariances if $i \neq j$. As shown in the supplementary material,[70] this matrix is symmetric under the assumption of reversible dynamics and in the limit of good statistics. For a finite dataset, symmetricity must be enforced.

We seek a transformation matrix $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_d]$ that diagonalizes $\mathbf{C}^r(0)$ (to fulfill property 1), and maximizes the autocorrelations $c^z_{ii}(\tau) = \mathbf{u}_i^T \mathbf{C}^r(\tau) \mathbf{u}_i$ for every column $\mathbf{u}_i$ of $\mathbf{U}$ (to fulfill property 2). As described in Sec. II D, this is accomplished by solving

$$\mathbf{C}^r(\tau)\mathbf{u}_i = \mathbf{C}^r(0)\mathbf{u}_i \lambda_i^{\ddagger}(\tau). \tag{12}$$

Equation (12) is equivalent to Eq. (9). See the supplementary material[70] for an illustrative derivation of (12). As described in Sec. II D, the second-largest estimated eigenvalue is a lower bound for the real second-largest propagator eigenvalue: $\lambda_2^{\ddagger}(\tau) < \lambda_2(\tau)$.

ICs are now ordered according to the magnitude of the autocovariance $\lambda_i^{\ddagger}(\tau)$, and the IC's with the largest autocovariances $\lambda_i^{\ddagger}(\tau)$ will be called *dominant*. Since the dominant m IC's yield the linear subspace in which most of the slow processes are contained, it is reasonable to now perform a direct clustering in this subspace, thus aiming at approximating the nonlinear behavior of the slowest m eigenfunctions with step functions. This will yield a better approximation to the m slowest timescales. Rewriting Eq. (12) in matrix form, and with the matrix of autocorrelations, $\mathbf{\Lambda}^{\ddagger}(\tau) = \mathrm{diag}(\lambda_1^{\ddagger}(\tau), \ldots, \lambda_d^{\ddagger}(\tau))$ yields

$$\mathbf{C}^r(\tau)\mathbf{U} = \mathbf{C}^r(0)\mathbf{U}\mathbf{\Lambda}^{\ddagger}(\tau). \tag{13}$$

In order to transform an original coordinate vector $\mathbf{r}$ into independent components, we perform

$$\mathbf{z}^T = \mathbf{r}^T\mathbf{U}. \tag{14}$$

How can (13) be solved? If $\mathbf{C}^r(0)$ or $\mathbf{C}^r(\tau)$ were invertible, the generalized eigenvalue problem could be transformed into a normal eigenvalue problem. But as we expect some of our original order parameters to be highly correlated, the determinants of $\mathbf{C}^r$ and $\mathbf{C}^r(\tau)$ will be nearly zero, prohibiting this option. Alternatively, one can seek the solution of (12) via generalized eigensolvers.

However, there is a simple and efficient alternative to this: Problem (13) can also be solved by solving two simple eigenvalue problems using the AMUSE algorithm.[71] It consists of the following steps:

1. Use PCA to transform mean-free data $\mathbf{r}(t)$ into principal components $\mathbf{y}(t)$.
2. Normalize principal components: $\mathbf{y}'(t) = \mathbf{\Sigma}^{-1}\mathbf{y}(t)$.
3. Compute the symmetrized time-lagged covariance matrix $\mathbf{C'}^{y'}_\tau = \frac{1}{2}[\mathbf{C}^{y'}_\tau + (\mathbf{C}^{y'}_\tau)^\dagger]$ of the normalized PCs.
4. Compute an eigenvalue decomposition of $\mathbf{C'}^{y'}_\tau$ obtaining eigenvector matrix $\mathbf{V}$, and project the trajectory $\mathbf{y}'(t)$ onto the dominant eigenvectors to obtain $\mathbf{z}(t)$.

This only works when the eigenvectors of $\mathbf{C'}^{y'}_\tau$ are uniquely defined, i.e., if the eigenvalues are not degenerated.[65] The main idea of this algorithm is that properties 1 and 2 can be fulfilled one after the other. First, steps 1 and 2 use PCA to produce decorrelated and normalized trajectories $\mathbf{y}'(t)$, also known as whitening the data. Then steps 3 and 4 maximize the time lagged autocovariances. Because the matrix $\mathbf{V}$, which is used in step 4, is unitary, it preserves scalar products between the vectors $\mathbf{y}'(t)$. Now if $\mathbf{y}'(t)$ are chosen to be uncorrelated (and properly normalized), then also $\mathbf{z}(t)$ will be uncorrelated.

In summary, the transformation equation (14) can be written as a concatenation of three linear transforms:

$$\mathbf{z}^T(t) = \mathbf{r}^T(t)\mathbf{U} = \mathbf{r}^T(t)\mathbf{W}\mathbf{\Sigma}^{-1}\mathbf{V}. \tag{15}$$

TICA will be used as a dimension reduction technique. Only the dominant TICA components will be used to construct a Markov model.

B. Markov model construction

Having identified the "slow" linear combinations of input order parameters, the hope is that clustering in a low-dimensional linear subspace will provide a useful clustering metric for the accurate and efficient construction of Markov

models with a moderate number of clusters. For clustering and Markov model construction of molecular dynamics data, the packages EMMA,[72] MSMbuilder,[73] Wordom,[74] and METAGUI[75] are currently available. Here, we use the EMMA package.

Markov models are constructed by first performing a data clustering and subsequently converting the trajectory files into discrete trajectory files containing the sequence of cluster indexes visited. For the sake of the current paper, the main analysis is the behavior of the relaxation timescales that are implied by the estimated Markov model.

The k-means algorithm[76] has been used with 1000 centers for clustering in the different coordinate spaces that spanned by the input order parameters $\mathbf{r}$, a set of normalized PCA sub-spaces ($\mathbf{y}'$), and a set of TICA sub-spaces ($\mathbf{z}$). For comparison, the data have also been clustered using the widely used minimal normalized Euclidean distance (short: minimal RMSD-metric[59]) in the full Cartesian space ($\mathbf{x}$) using a distance cut-off that produces an equivalent number of clustercenters. Both algortihms are available via the EMMA command mm_cluster.

Subsequent to the identification of cluster centers, the state space is partitioned by assigning each trajectory frame to its closest cluster center according to the same metric used for clustering. The discretization obtained this way is a Voronoi tessellation of the observed coordinate space, so that the Voronoi cells form a complete partition of the conformation space. This is carried out via the EMMA command mm_assign.

Ultimately, Markov model estimation is done as proposed in Ref. 49 using the maximum probability estimator of reversible transition matrices with a weak neighbor prior count matrix (EMMA default). Both the transition matrix estimation and its diagonalaization are performed by the command mm_timmescales.

The transition matrix $\mathbf{T}(\tau)$ has the right eigenvectors $\tilde{\boldsymbol{\psi}}_i$, the left eigenvectors $\tilde{\boldsymbol{\phi}}_i$, and the eigenvalues $\tilde{\lambda}_i$ according to the following eigenvalue equations:

$$\mathbf{T}(\tau)\tilde{\boldsymbol{\psi}}_i = \tilde{\lambda}_i(\tau)\tilde{\boldsymbol{\psi}}_i,$$
$$\tilde{\boldsymbol{\phi}}_i^T\mathbf{T}(\tau) = \tilde{\lambda}_i(\tau)\tilde{\boldsymbol{\phi}}_i^T.$$

We order eigenvalues by descending norm. When $\mathbf{T}(\tau)$ is connected (irreducible), it will have a unique eigenvalue of norm 1. The corresponding eigenvector can be normalized to yield the stationary distribution $\boldsymbol{\pi}$:

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

Since $\mathbf{T}(\tau)$ fulfills detailed balance, the left and right eigenvectors are related by

$$\tilde{\boldsymbol{\phi}}_i = \mathrm{diag}(\boldsymbol{\pi})\tilde{\boldsymbol{\psi}}_i.$$

The estimated (implied) relaxation timescales of the Markov model are given by

$$\tilde{t}_i = -\frac{\tau}{\ln\tilde{\lambda}(\tau)},$$

which are – ignoring statistical errors – related to the true relaxation timescales by $\tilde{t}_i < t_i$ (see Sec. II), and are typically larger than the timescales implied by the TICA eigenvalues (see Sec. II and Fig. 1).

C. Optimal indicators

Given the final Markov model transition matrix $\mathbf{T}(\tau)$, we can now establish a simple way to quantify how well each of the order parameters r_k serving as an input serves as an indicator of the slow process described by the eigenvector $\tilde{\boldsymbol{\psi}}_i$: We simply compute the correlation between all pairs of order parameters and eigenvectors, and then, for each eigenvector, choose those order parameters that have a maximum correlation:

$$r_{\mathrm{opt}}(i) = \arg\max_{r_k}\mathrm{Corr}(r_k, \tilde{\psi}_i), \tag{16}$$

where the correlation is in practice computed either via a time or MSM-based ensemble average. For the time-average (used here), let $s(t)$ be the trajectory of microstates, i.e., the trajectory that contains at each time point the microstate number that is assigned to the configuration visited at that time point. Then, we approximate the correlation by

$$\mathrm{Corr}(r_k, \tilde{\psi}_i) = \frac{\langle r_k\tilde{\psi}_{i,s(t)}\rangle_t - \langle r_k\rangle_t\langle\tilde{\psi}_{i,s(t)}\rangle_t}{\sqrt{\langle r_k^2\rangle_t\langle\tilde{\psi}_{i,s(t)}^2\rangle_t}}, \tag{17}$$

while for the MSM-based ensemble average, we compute the average value of r_k for every microstate j, obtaining $\bar{r}_{k,j}$, and obtain

$$\mathrm{Corr}(r_k, \tilde{\psi}_i) = \frac{\sum_{j=1}^n \pi_j\bar{r}_{k,j}\tilde{\psi}_{i,j} - \sum_{j=1}^n \pi_j\bar{r}_{k,j}\sum_{j=1}^n \pi_j\tilde{\psi}_{i,j}}{\sqrt{\sum_{j=1}^n \pi_j\bar{r}_{k,j}^2\sum_{j=1}^n \pi_j\tilde{\psi}_{i,j}^2}}. \tag{18}$$

IV. RESULTS

The proposed methodology is demonstrated on two different peptide systems: the fluorescent peptide MR121-GSGSW and the 30-residue intrinsically disordered peptide KID. See the supplementary material[70] for the statistical uncertainties of both data samples.

MR121-GSGSW is a well-studied fluorescent peptide that has been extensively characterized by experiments,[77] simulations,[78] and also by Markov models.[50,79] Here, a data set of two explicit solvent simulations of 4 μs each is used, a set that is publicly accessible as a benchmark dataset for the EMMA software package (see http://simtk.org/home/emma). The details of the simulation setup are described in Ref. 49.

The slowest relaxation timescale of the MR121-GSGSW data set has been estimated to be between 20 and 30 ns, and it has been found that the slowest processes are dominated by the interaction between MR121 and the tryptophan residue (Trp).[50] The data set is used as a benchmark system to test whether Markov model construction in PCA or TICA coordinates manages to identify the slow parameters, approximates the slow processes, and assigns the correct timescales.

Figure 2(a1) shows a sample structure of MR121-GSGSW. Figure 2(b) shows a benchmark for the relaxation

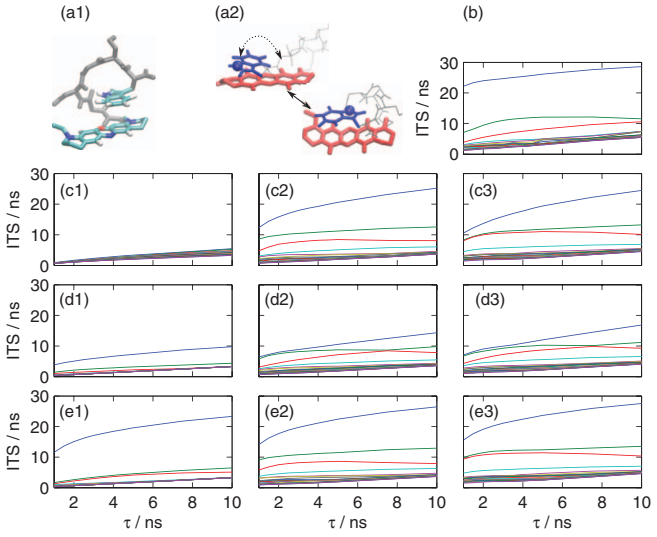

FIG. 2. MR121-GSGSW peptide and its dominant relaxation timescales calculated *via* different Markov model construction methods. (a1) Sample structure of the peptide. (a2) Illustration of the Trp coordinates used. The center position of the Trp and the orientation vectors are given in a coordinate system defined by the MR121 principal axes. (b) Relaxation timescales using regular space RMSD clustering with 1000 clusters. (c)–(e) Relaxation timescales using *k*-means with 1000 clusters and Euclidean metric but operating on different subspaces: (c1) Intramolecular distances between all C_α's and ring centers. (c2) Center position and orientation coordinates of the Trp moiety in the MR121 coordinate system. (c3) Combined coordinate set including intramolecular distances and Trp coordinates. (d1)–(d3) Dominant PCA subspace of the combined coordinate set using 1, 4, and 10 dimensions. (e1)–(e3) Dominant TICA subspace of the combined coordinate set using 1, 4, and 10 dimensions.

timescales computed by a regular-space clustering in pairwise minimal RMSD metric using 1000 cluster centers. The slowest processes are found at about 25 ns, 12 ns, and 8 ns, slightly larger—and thus more accurate according to the variational principle in Eq. (4)—than by the coarser Markov model in Ref. 50. To set up the direct clustering, two internal coordinate sets are considered: (i) the set of 66 distances between 12 coordinates defined by the 5 C_α's and the 7 ring centers involved, and (ii) the center position and the orientation vector coordinates of the Trp sidechain in a coordinate set defined by the MR121 principal axes (see Fig. 2(a2) for an illustration). We refrain from using dihedral angles as input coordinates given that the tertiary interactions between MR121 and Trp (and not the flexible linker) are responsible for the metastability. Figures 2(c1)–2(c3) show the results of direct k-means clustering with 1000 cluster centers in the space of 66 intramolecular distances Fig. 2(c1), only the 9 Trp coordinates Fig. 2(c2), and the combined set Fig. 2(c3). It is clearly seen that the intramolecular distances are not suited to resolve the slowest processes, while the Trp coordinates resolve them very well. This can be understood from the structural arrangements shown in Fig. 3, which are dominated by the relative orientation of the Trp sidechain with respect to the MR121 ring system. Especially the slowest process, the stacking-order exchange of the two ring systems, cannot be well described by the intramolecular distances that are similar when the Trp is "above" or "below" the MR121. Figure 2(c3) shows that discretizing the combined coordinate set resolves the slowest processes with similar timescales as in the 9 Trp-coordinate set alone. This is not always expected, as increasing the dimensionality of the space to be clustered, while keeping the number of clusters constant will often reduce the resolution.

In the subsequent PCA and TICA analysis different linear subspaces of the combined coordinate set were considered. Interestingly, clustering the principal components reduces the quality of the Markov model significantly. This is explained by the fact that the largest-amplitude motion in the present system is the transition between structures in which Trp and MR121 are in contact, and open structures. However, open structures have a very low population giving rise to a rather fast timescale of the opening/closing process. The slowest processes, involving different arrangements and orientations of the Trp and MR121 while being in contact, give rise to comparatively small amplitude motions. Using one and four PCA components (Figs. 2(d1) and 2(d2)), the three slowest processes are not found. Using ten PCA components, the two slowest processes are found, although slightly underestimated, while the third-slowest process is not found.

Figures 2(e1)–2(e3) show that the TICA coordinates perform indeed very well. Using only the single slowest TICA coordinate does resolve the slowest process well and gives rise to a timescale of 20–25 ns, close to the expected value. Using the four slowest TICA coordinates resolves the two slowest processes well, while somewhat underestimating the third process. With ten TICA coordinates all slow processes are well resolved, and the timescales are found to be 27 ns, 13 ns, and 10 ns at a lagtime of $\tau = 10$ ns—slightly larger than in any of the other choices of metrics.

Figure 3(a) illustrates the structural transition involved in the two slowest processes occurring at around 27 and 13 ns computed from the ten-dimensional TICA Markov model. We display the 1000 microstates in a visualization that we shall call *kinetic map*, where the coordinates are given by the two slowest left eigenvectors $\tilde{\boldsymbol{\phi}}_2$, $\tilde{\boldsymbol{\phi}}_3$. For example, a cluster i is drawn at a position $(\tilde{\phi}_{2,i}, \tilde{\phi}_{3,i})$ with its area proportional to its stationary probability π_i. The map is termed "kinetic" because similar positions in eigenvector spaces mean that the states can relatively quickly reach one another, while distant positions only exchange on timescales t_2 on the horizontal and on timescale t_3 on the vertical axis. The left eigenvectors are chosen instead of the right eigenvectors because the left eigenvectors are weighted by the stationary distribution: $\tilde{\phi}_{k,i} = \pi_i \tilde{\psi}_{k,i}$. Thus, points on the border of the map tend to have larger stationary probability. Therefore, the extremal points are at the same time populous and kinetically distant, and can roughly be associated with the most stable "free energy minima," while the smaller clusters connecting them correspond to transition states. The structures, shown for the most populous and kinetically distinct clusters, indicate that the slowest relaxations are associated with a stacking-order exchange of the MR121 and Trp groups, and a rotation of the Trp group with respect to the MR121 group (see "marker" atom shown as a blue sphere).

Figure 3(b) illustrates the optimal indicators of the slowest processes, i.e., the input order parameters that have the largest correlation with the individual right Markov model

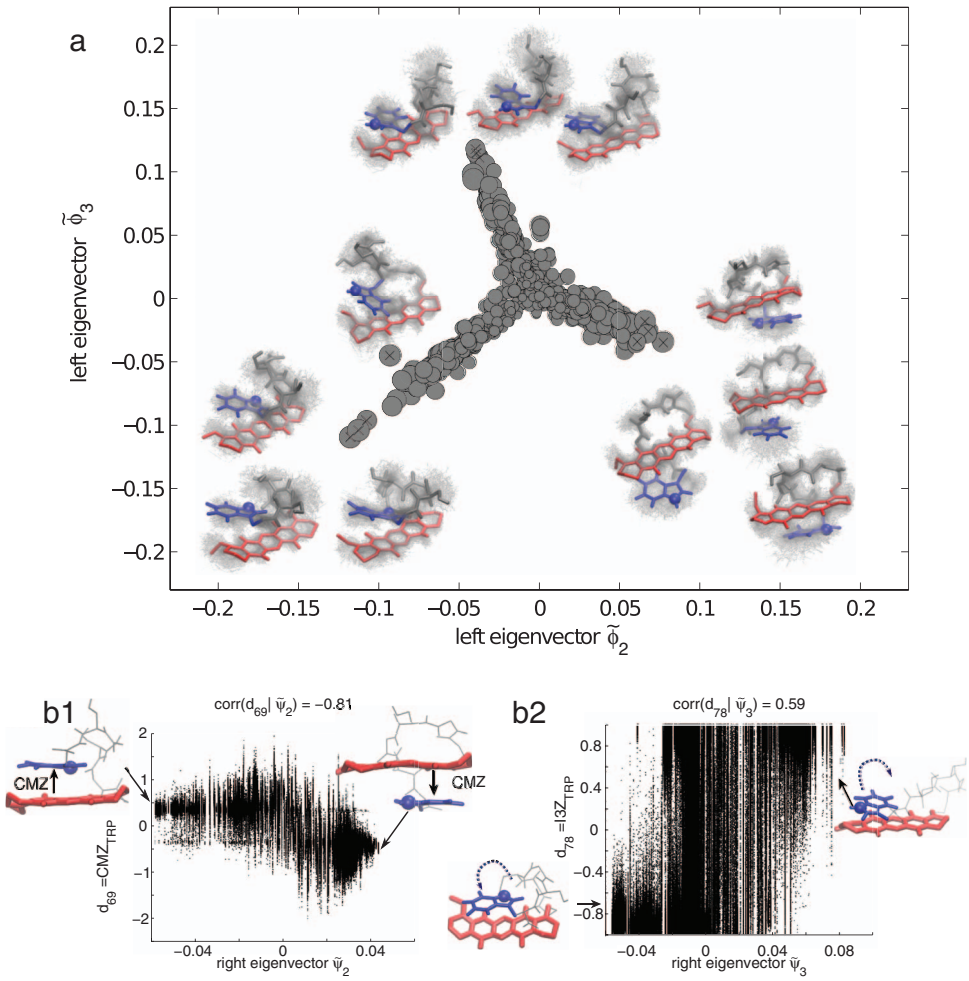

FIG. 3. (a) *Kinetic map* of the two slowest relaxation processes of MR121-GSGSW (around 27 ns and 13 ns) calculated from the Markov model shown in Fig. 2(e3). The grey discs mark the coordinates of the 1000 microstates in the space of the left eigenvectors $\tilde{\boldsymbol{\phi}}_2$, $\tilde{\boldsymbol{\phi}}_3$. The slowest relaxation of the system thus takes place on the horizontal axis, the second-slowest one on the vertical axis, and distances are associated with kinetic separation. The area of a disc is proportional to the stationary probability of the corresponding microstate. Some representative (kinetically distant and populous) microstates are shown as molecular structures. (b) *Optimal indicators* of the slow processes. The scatter plots show the correlation between the second and third right Markov model eigenvectors $\tilde{\boldsymbol{\psi}}_2$, $\tilde{\boldsymbol{\psi}}_3$ and the order parameters most correlated with them. The arrows in the structures show the optimal indicators. (b1) The Trp z-position mediates the stacking order exchange and has a correlation coefficient of 0.81 with the second eigenvector $\tilde{\boldsymbol{\psi}}_2$ (timescale 27 ns). (b2) The smallest Trp axis of inertia mediates the rotation of the side-chain and has a correlation coefficient of 0.59 with the third eigenvector $\tilde{\boldsymbol{\psi}}_3$ (timescale 13 ns).

eigenvectors $\tilde{\boldsymbol{\psi}}_2$ and $\tilde{\boldsymbol{\psi}}_3$. The correlation plots show that the respective order parameters attain clearly different values at the end-states of the transition, i.e., for the minimal and maximal values of the respective eigenvector. At intermediate values of the eigenvectors, i.e., transition states, the order parameter can access many different values. This is easily seen in the slowest process (Fig. 3(b1)), where the best indicator is the Trp z-position that mediates the stacking order exchange (correlation coefficient −0.81 with the second eigenvector $\tilde{\boldsymbol{\psi}}_2$). While the value of the Trp z-position is clearly defined in the transition end-states, where the Trp is located "above" and "below" the MR121 moiety, the transition states include open configurations where the Trp and the MR121 are not in contact at all, and therefore all values of the z-position are accessible in these states. A similar behavior is seen for the second-slowest process (Trp sidechain rotation).

We now turn to the other molecular system. Here, an extensive set of simulations of the KID in explicit solvent was investigated. KID is part of the cAMP response element-binding protein (CREB). CREB is a transcription factor involved in processes as important as glucose regulation and memory, and it binds the CREB-binding protein (CBP), a well-known cancer-related molecular hub with around 300 interacting protein partners.[80] KID belongs to a large and important class of *intrinsically disordered* peptides, encompassing many hormones, domains, and even whole proteins.[81] Unstructured regions perform their function even though they lack a well-defined secondary or tertiary structure in solution. Although standardized algorithms exist to detect unstructured regions on the basis of the primary amino acid sequence, the structural details of how disordered regions exert their function are still elusive. For example, some unstructured domains, including KID, become folded upon binding;[82] it is therefore of much interest (e.g., for the druggability of protein-protein interactions) to investigate whether the presence of pre-formed elements causes folded conformations to be selected from the ensemble (conformational selection),[83]

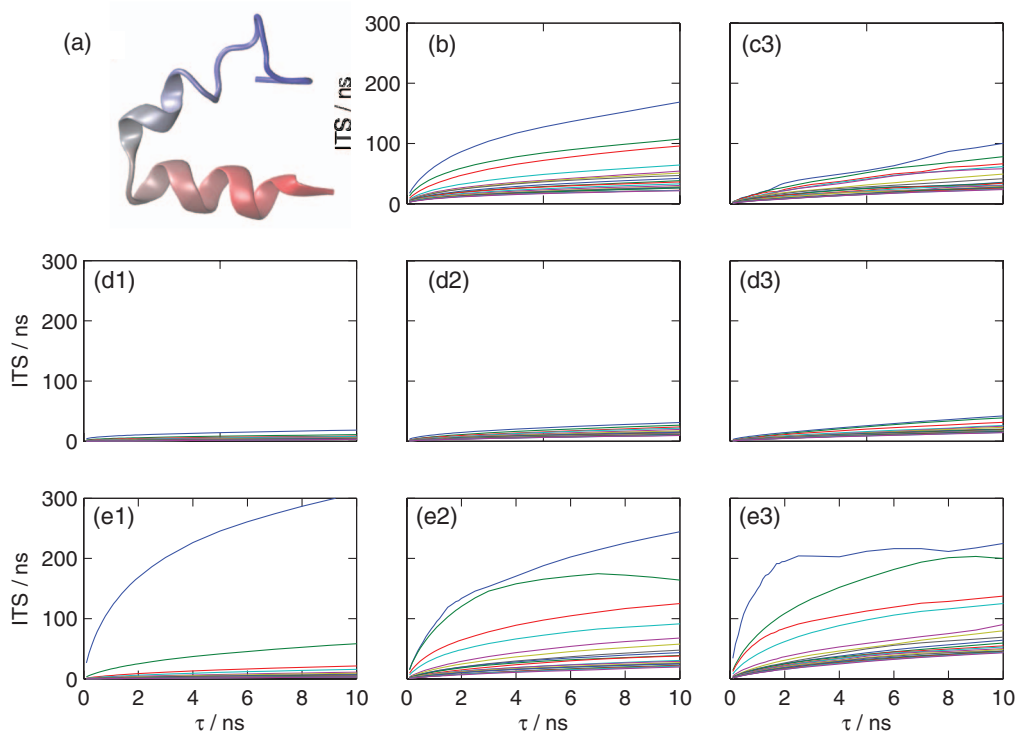

FIG. 4. KID peptide and its estimated dominant relaxation timescales using different Markov model construction methods. (a) Sample structure of KID. (b) Relaxation timescales using regular space RMSD clustering with 1000 clusters. (c)–(e) Relaxation timescales using *k*-means with 1000 clusters and Euclidean metric but operating on different subspaces. (c) All $C_\alpha - C_\alpha$ distances. (d1)–(d3) Dominant PCA subspace of $C_\alpha - C_\alpha$ distances using 1, 4, and 10 dimensions. (e1)–(e3) Dominant TICA subspace of $C_\alpha - C_\alpha$ distances using 1, 4, and 10 dimensions.

or whether the binding rather occurs through induced-fit mechanics.[84]

To shed light on this problem, we set up an ensemble of all-atom simulations of the phosphorylated KID (pKID) domain. We have performed 7706 all-atom explicit-solvent simulations of 24 ns each using the ACEMD software[85] on the GPUGRID distributed computing platform,[23] yielding a total of 185 μs simulation data. However, due to the short simulations, only short lagtimes could be used presenting a challenge to the Markov model construction. The detailed simulation setup is described in the supplementary material.[70]

Figure 4 shows the performance of different metrics in their ability to resolve the slowest processes of KID. KID is a more difficult case than the MR121-GSGSW peptide because its natively unstructured nature gives rise to many fast large-amplitude motions, which will conceal the slow processes in most *ad hoc* metrics. Figures 4(b) and 4(c) show that neither regular-space clustering in minimal pairwise RMSD metric nor direct clustering in $C_\alpha - C_\alpha$ distances yield a converged estimated of the timescales up to lagtimes of 10 ns. Between these two, regular-space RMSD is better, reaching a timescale of about 170 ns at $\tau = 10$ ns, while the direct clustering produces a timescale below 100 ns at $\tau = 10$ ns. Higher choices of lagtimes were avoided as they lead to a severe reduction of the usable data because the connected set of clusters drops significantly below 100% after that point. Figures 4(d1)–4(d3) show that the performance of principal components is even worse than direct clustering giving rise to timescale estimates below 20 ns for one principal component and below 50 ns for ten principal components. This confirms that the largest-amplitude motions are not the slowest in KID.

Figures 4(e1)–4(e3) show the performance of the TICA coordinates using one, four, and ten dimensions. Using only the slowest TICA coordinate, a slow process of >200 ns is found, that has not been resolved by the clustering in any of the other metrics, however, this timescale does not converge for lagtimes up to 10 ns. Using only the four slowest TICA coordinates, there are already three processes resolved that are above 100 ns, and the convergence behavior improves. Using the ten slowest TICA coordinates, five processes slower than 100 ns are resolved. The slowest process converges to a timescale around 220 ns and does so already at a lagtime τ of 2–5 ns. Thus, the lagtime needed is a factor of 50-100 smaller than the timescale of the process, indicating a very good discretization of the corresponding process.

Figure 5(a) illustrates the structural transitions associated with the two slowest relaxation processes of KID as identified by the Markov model using ten-dimensional TICA model. We have decided to focus on the two slowest processes around 200 and 220 ns relaxation time, because they are somewhat separated from the next-slowest processes occurring at around 100 ns. As the peptide has great structural variability, it is of little value to plot all relevant structures. Therefore, we have plotted the positions of the microstates again in a kinetic map using the coordinates of the two dominant left eigenvectors $\tilde{\boldsymbol{\phi}}_2$, $\tilde{\boldsymbol{\phi}}_3$. It is seen that at the slowest timescales, the system rearranges mostly between open and disordered structures (left), structures with one helix folded or partially

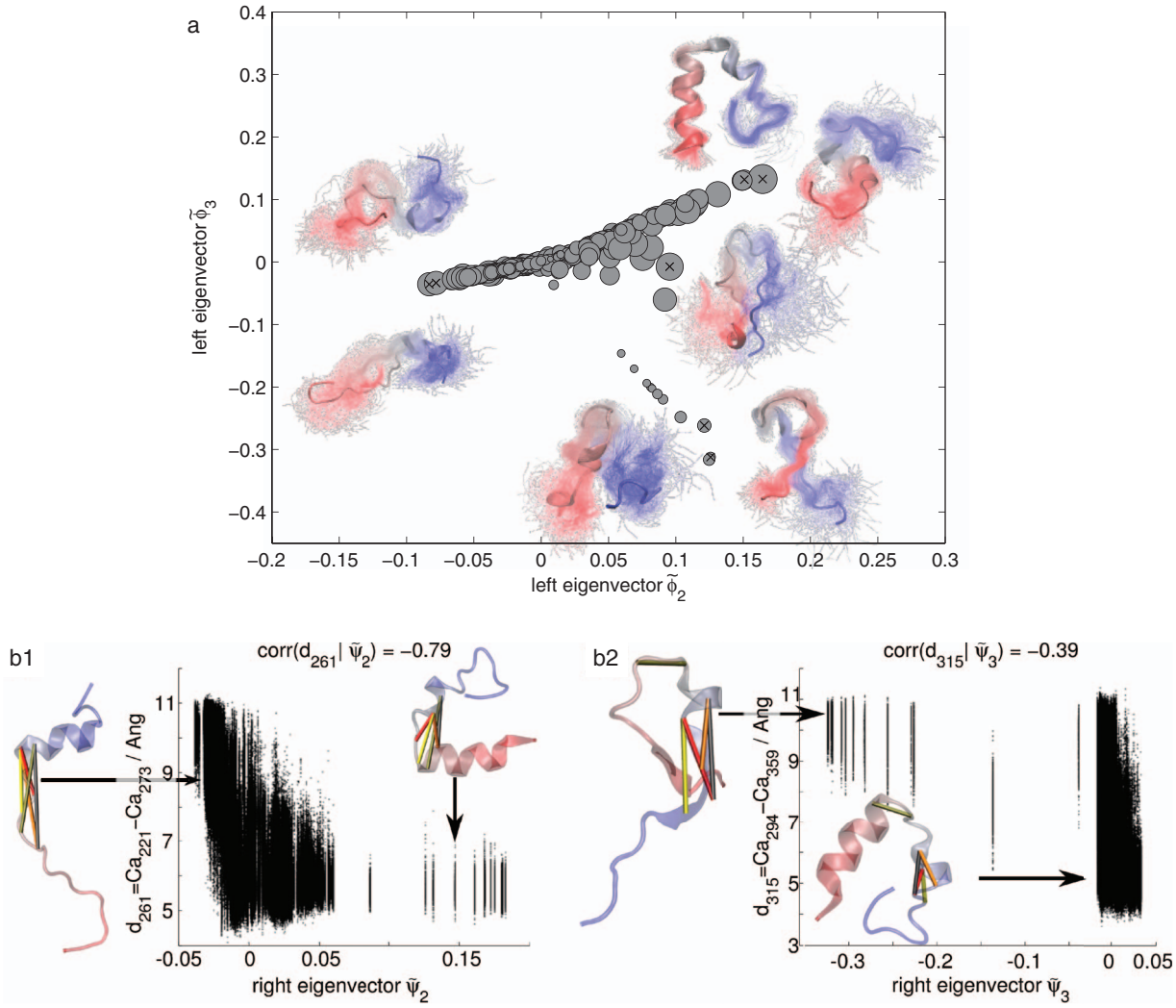

FIG. 5. (a) *Kinetic map* of the two slowest relaxation processes of the KID peptide (around 200 ns and 220 ns) calculated from the Markov model shown in Fig. 4(e3). The grey discs mark the coordinates of the 1000 microstates in the space of the left eigenvectors $\tilde{\boldsymbol{\phi}}_2$, $\tilde{\boldsymbol{\phi}}_3$. The slowest relaxation of the system thus takes place on the horizontal axis, the second-slowest one on the vertical axis, and distances are associated with kinetic separation. The area of a disc is proportional to the stationary probability of the corresponding microstate. Some representative (kinetically distant and populous) microstates are shown as molecular structures. (b) *Optimal indicators* of the slow processes. The scatter plots show the correlation between the second and third right Markov model eigenvectors $\tilde{\boldsymbol{\psi}}_2$, $\tilde{\boldsymbol{\psi}}_3$ and the order parameters most correlated with them. The colored lines show all five best indicators. The slowest process may thus be described as opening/closing of the hinge between the two helical domains of KID (timescale 220 ns), while hinge-closing is associated with at least partial N-terminal helix formation (red). The second-slowest process may be described as partial helix formation in the "blue" region (timescale 200 ns).

folded (top right), and hairpin-like structures (bottom right). Thus, the system has some residual helical structure, although it is not very stable in absence of a stabilizing binding partner.

Figure 5(b) illustrates the optimal indicators of the slowest processes, i.e., the input order parameters that have the largest correlation with the resulting right Markov model eigenvectors $\tilde{\boldsymbol{\psi}}_2$ and $\tilde{\boldsymbol{\psi}}_3$. Like for MR121-GSGSW, the correlation plots show that the respective order parameters are mainly able to distinguish the end-states of the transition, but unlike for MR121-GSGSW, multiple $C_\alpha - C_\alpha$ distances are almost equally good indicators for the same process. Figure 5(b) shows correlation plots of the best indicators of $\tilde{\boldsymbol{\psi}}_2$ and $\tilde{\boldsymbol{\psi}}_3$, but indicates the five best correlations in the structure. It is seen that the slowest process (timescale 220 ns) is best described by a hinge opening and closing, where the closed hinge appears to induce at least partial formation of the N-terminal helix (red, see Fig. 5(b2)). This is consistent with nuclear magnetic resonance (NMR) experiments that have shown the N-terminal region to be approximately 50% helical in the apo-form.[86] The second-slowest process (timescale 200 ns) is best described by partial helix formation of the C-terminal part (blue) of the protein.

V. DISCUSSION

In the present study we have derived a method to find the optimal linear combination of input coordinates for approximating the slowest relaxation processes in complex conformational rearrangements of molecules. It is shown that an implementation for this method is already known in statistics as the TICA method, which is combined here with Markov modeling in order to construct models of the slow relaxation processes and precise estimates of the related relaxation timescales. It is shown that this approach of constructing

Markov models yields slower timescales, and thus a more precise approximation to the true relaxation processes than previous approaches. This is also achieved for the intrinsically disordered peptide KID where established approaches such as direct clustering in distance space, minimal-RMSD-based clustering, or clustering in PCA space did not perform well because the largest-amplitude motions were not good indicators of the slowest relaxation processes.

Beyond having an approach to construct quantitatively accurate Markov models in a way that is more robust than most previous approaches, we readily obtain a way to find *best indicators* of the slowest transitions. Best indicators are those molecular order parameters that are best correlated with the Markov model eigenvectors describing the slowest processes, and thus serve as candidates for good reaction coordinates. Being able to point out such indicators provides a way to make the sometimes complex structural rearrangements readily understandable.

ACKNOWLEDGMENTS

We thank all the volunteers of GPUGRID who donated GPU computing time to the project. We are grateful to Thomas Weikl (MPI Potsdam) for advice and support. G.P.-H. acknowledges support from German Science Foundation DFG fund NO 825-3. F.P. acknowledges funding from the Max Planck Society. T.G. gratefully acknowledges former support from the "Beatriu de Pinós" scheme of the Agència de Gestió d'Ajuts Universitaris i de Recerca (Generalitat de Catalunya). G.D.F. acknowledges support from the Ramón y Cajal scheme and support by the Spanish Ministry of Science and Innovation (Ref. BIO2011-27450). F.N. acknowledges funding from DFG center Matheon and ERC starting grant 307494 pcCell.

[1] A. Gansen, A. Valeri, F. Hauger, S. Felekyan, S. Kalinin, K. Tóth, J. Langowski, and C. A. M. Seidel, Proc. Natl. Acad. Sci. U.S.A. **106**, 15308 (2009).
[2] H. Neubauer, N. Gaiko, S. Berger, J. Schaffer, C. Eggeling, J. Tuma, L. Verdier, C. A. Seidel, C. Griesinger, and A. Volkmer, J. Am. Chem. Soc. **129**, 12746 (2007).
[3] W. Min, G. Luo, B. J. Cherayil, S. C. Kou, and X. S. Xie, Phys. Rev. Lett. **94**, 198302 (2005).
[4] E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern, Nature (London) **438**, 117 (2005).
[5] Y. Santoso, C. M. Joyce, O. Potapova, L. Le Reste, J. Hohlbein, J. P. Torella, N. D. F. Grindley, and A. N. Kapanidis, Proc. Natl. Acad. Sci. U.S.A. **107**, 715 (2010).
[6] J. C. M. Gebhardt, T. Bornschlögl, and M. Rief, Proc. Natl. Acad. Sci. U.S.A. **107**, 2013 (2010).
[7] B. G. Wensley, S. Batey, F. A. C. Bone, Z. M. Chan, N. R. Tumelty, A. Steward, L. G. Kwa, A. Borgia, and J. Clarke, Nature (London) **463**, 685 (2010).
[8] M. Jäger, Y. Zhang, J. Bieschke, H. Nguyen, M. Dendle, M. E. Bowman, J. P. Noel, M. Gruebele, and J. W. Kelly, Proc. Natl. Acad. Sci. U.S.A. **103**, 10648 (2006).
[9] A. Y. Kobitski, A. Nierth, M. Helm, A. Jäschke, and G. U. Nienhaus, Nucleic Acids Res. **35**, 2047 (2007).
[10] S. Fischer, B. Windshuegel, D. Horak, K. C. Holmes, and J. C. Smith, Proc. Natl. Acad. Sci. U.S.A. **102**, 6873 (2005).
[11] F. Noé, D. Krachtus, J. C. Smith, and S. Fischer, J. Chem. Theory Comput. **2**, 840 (2006).
[12] A. Ostermann, R. Waschipky, F. G. Parak, and U. G. Nienhaus, Nature (London) **404**, 205 (2000).
[13] M. Pirchi, G. Ziv, I. Riven, S. S. Cohen, N. Zohar, Y. Barak, and G. Haran, Nat. Commun. **2**, 493 (2011).
[14] J. Stigler, F. Ziegler, A. Gieseke, J. C. M. Gebhardt, and M. Rief, Science **334**, 512 (2011).
[15] H. Chung, J. Louis, and W. Eaton, Science **335**, 981 (2012).
[16] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, Science **330**, 341 (2010).
[17] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science **334**, 517 (2011).
[18] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, J. Am. Chem. Soc. **132**, 1526 (2010).
[19] G. R. Bowman, V. A. Voelz, and V. S. Pande, J. Am. Chem. Soc. **133**, 664 (2011).
[20] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. U.S.A. **106**, 19011 (2009).
[21] I. Buch, T. Giorgino, and G. De Fabritiis, Proc. Natl. Acad. Sci. U.S.A. **108**, 10184 (2011).
[22] S. K. Sadiq, F. Noé, and G. De Fabritiis, Proc. Natl. Acad. Sci. U.S.A. **109**, 20449 (2012).
[23] I. Buch, M. J. Harvey, T. Giorgino, D. P. Anderson, and G. De Fabritiis, J. Chem. Inf. Model. **50**, 397 (2010).
[24] D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).
[25] F. Noé and S. Fischer, Curr. Opin. Struct. Biol. **18**, 154 (2008).
[26] M. E. Karpen, D. J. Tobias, and C. L. Brooks, Biochemistry **32**, 412 (1993).
[27] I. A. Hubner, E. J. Deeds, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **103**, 17747 (2006).
[28] M. Weber, ZIB Report No. 03–04, 2003.
[29] N. V. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057 (2008).
[30] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).
[31] S. Muff and A. Caflisch, Proteins **70**, 1185 (2007).
[32] B. de Groot, X. Daura, A. Mark, and H. Grubmüller, J. Mol. Biol. **309**, 299 (2001).
[33] V. Schultheis, T. Hirschberger, H. Carstens, and P. Tavan, J. Chem. Theory Comput. **1**, 515 (2005).
[34] A. C. Pan and B. Roux, J. Chem. Phys. **129**, 064107 (2008).
[35] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **101**, 14766 (2004).
[36] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, J. Chem. Phys. **126**, 155102 (2007).
[37] J. D. Chodera, K. A. Dill, N. Singhal, V. S. Pande, W. C. Swope, and J. W. Pitera, J. Chem. Phys. **126**, 155101 (2007).
[38] W. C. Swope, J. W. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).
[39] D. Huang and A. Caflisch, PLOS Comput. Biol. **7**, e1002002 (2011).
[40] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proc. Natl. Acad. Sci. U.S.A. **102**, 7426 (2005).
[41] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**, 124116 (2011).
[42] S. Sriraman, I. G. Kevrekidis, and G. Hummer, J. Phys. Chem. B **109**, 6479 (2005).
[43] F. Noé, J. Chem. Phys. **128**, 244103 (2008).
[44] N. Singhal and V. S. Pande, J. Chem. Phys. **123**, 204909 (2005).
[45] M. Sarich, F. Noé, and C. Schütte, Multiscale Model. Simul. **8**, 1154 (2010).
[46] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, J. Comput. Phys. **151**, 146 (1999).
[47] F. Noé and F. Nüske, "A variational approach to modeling slow processes in stochastic dynamical systems," Multiscale Model. Simul. (in press), http://arxiv.org/abs/1211.7103v1.
[48] P. Deuflhard and M. Weber, ZIB Report No. 03–09, 2003.
[49] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Fischbach, M. Held, J. D. Chodera, C. Schütte, and F. Noé, J. Chem. Phys. **134**, 174105 (2011).
[50] F. Noé, S. Doose, I. Daidone, M. Löllmann, J. D. Chodera, M. Sauer, and J. C. Smith, Proc. Natl. Acad. Sci. U.S.A. **108**, 4822 (2011).
[51] B. Keller, J.-H. Prinz, and F. Noé, Chem. Phys. **396**, 92 (2012).
[52] G. S. Buchner, R. D. Murphy, N.-V. Buchete, and J. Kubelka, Biochim. Biophys. Acta **1814**, 1001 (2011).
[53] S. Kube and M. Weber, J. Chem. Phys. **126**, 024103 (2007).
[54] P. Metzner, C. Schütte, and E. V. Eijnden, Multiscale Model. Simul. **7**, 1192 (2009).
[55] A. Berezhkovskii, G. Hummer, and A. Szabo, J. Chem. Phys. **130**, 205102 (2009).

[56]N. Djurdjevac, M. Sarich, and C. Schütte, Multiscale Model. Simul. **10**, 61 (2012).
[57]J.-H. Prinz, J. D. Chodera, and F. Noé, "Spectral rate theory for two-state kinetics," Phys. Rev. (submitted).
[58]G. Berezovska, D. Prada-Gracia, S. Mostarda, and F. Rao, J. Chem. Phys. **137**, 194101 (2012).
[59]W. Kabsch, Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **32**, 922 (1976).
[60]G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, J. Chem. Phys. **131**, 124101 (2009).
[61]K. A. Beauchamp, R. McGibbon, Y.-S. Lin, and V. S. Pande, Proc. Natl. Acad. Sci. U.S.A. **109**, 17807 (2012).
[62]G. R. Bowman and P. L. Geissler, Proc. Natl. Acad. Sci. U.S.A. **109**, 11681 (2012).
[63]A. Amadei, A. B. Linssen, and H. J. C. Berendsen, Proteins **17**, 412 (1993).
[64]L. Molgedey and H. G. Schuster, Phys. Rev. Lett. **72**, 3634 (1994).
[65]A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis* (John Wiley & Sons, 2001) Chap. 18, p. 344.
[66]Y. Naritomi and S. Fuchigami, J. Chem. Phys. **134**, 065101 (2011).
[67]A. Mitsutake, H. Iijima, and H. Takano, J. Chem. Phys. **135**, 164102 (2011).
[68]C. R. Schwantes and V. S. Pande, J. Chem. Theory Comput. **9**, 2000 (2013).
[69]W. Ritz, J. Reine Angew. Math. **135**, 1 (1909).
[70]See supplementary material at http://dx.doi.org/10.1063/1.4811489 for the mathematical derivation of TICA using variational calculus, the symmetrization of the time-lagged covariance matrix, the simulation setup for KID, and statistical uncertainties in the estimated implied timescales.
[71]L. Tong, V. C. Soon, Y. F. Huang, and R. Liu, Circuits Syst. **3**, 1784 (1990).
[72]M. Senne, B. Trendelkamp-Schroer, A. Mey, C. Schütte, and F. Noé, J. Chem. Theory Comput. **8**, 2223 (2012).
[73]K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, J. Chem. Theory Comput. **7**, 3412 (2011).
[74]M. Seeber, A. Felline, F. Raimondi, S. Muff, R. Friedman, F. Rao, A. Caflisch, and F. Fanelli, J. Comput. Chem. **32**, 1183 (2011).
[75]X. Biarnés, F. Pietrucci, F. Marinelli, and A. Laio, Comput. Phys. Commun. **183**, 203 (2012).
[76]S. Lloyd, IEEE Trans. Inf. Theory **28**, 129 (1982).
[77]H. Neuweiler, M. Löllmann, S. Doose, and M. Sauer, J. Mol. Biol. **365**, 856 (2007).
[78]I. Daidone, H. Neuweiler, S. Doose, M. Sauer, and J. C. Smith, PLOS Comput. Biol. **6**, e1000645 (2010).
[79]F. Noé, I. Daidone, J. C. Smith, A. di Nola, and A. Amadei, J. Phys. Chem. B **112**, 11155 (2008).
[80]L. H. Kasper, T. Fukuyama, M. A. Biesen, F. Boussouar, C. Tong, A. de Pauw, P. J. Murray, J. M. A. van Deursen, and P. K. Brindle, Mol. Cell. Biol. **26**, 789 (2006).
[81]V. N. Uversky, Int. J. Biochem. Cell Biol. **43**, 1090 (2011).
[82]K. Sugase, H. J. Dyson, and P. E. Wright, Nature (London) **447**, 1021 (2007).
[83]A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker, and V. N. Uversky, J. Mol. Biol. **362**, 1043 (2006).
[84]B. A. Shoemaker, J. J. Portman, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **97**, 8868 (2000).
[85]M. J. Harvey, G. Giupponi, and G. D. Fabritiis, J. Chem. Theory Comput. **5**, 1632 (2009).
[86]I. Radhakrishnan, G. C. Pérez-Alvarado, D. Parker, H. Dyson, M. R. Montminy, and P. E. Wright, Cell **91**, 741 (1997).