Investigating Molecular Kinetics by Variationally Optimized Diffusion Maps

Lorenzo Boninsegna,[†] Gianpaolo Gobbo,[‡] Frank Noé,*[,§] and Cecilia Clementi*[,†]

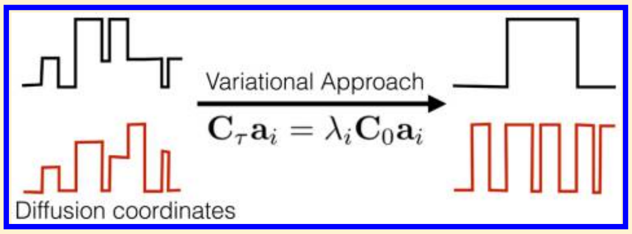
[†]Center for Theoretical Biological Physics and Department of Chemistry, Rice University, 6100 Main Street, Houston, Texas 77005, United States

[‡]Maxwell Institute for Mathematical Sciences and School of Mathematics, The University of Edinburgh, Peter Guthrie Tait Road, Edinburgh EH9 3FD, United Kingdom

[§]Department of Mathematics, Computer Science and Bioinformatics, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany

Ⓢ *Supporting Information*

ABSTRACT: Identification of the collective coordinates that describe rare events in complex molecular transitions such as protein folding has been a key challenge in the theoretical molecular sciences. In the Diffusion Map approach, one assumes that the molecular configurations sampled have been generated by a diffusion process, and one uses the eigenfunctions of the corresponding diffusion operator as reaction coordinates. While diffusion coordinates (DCs) appear to provide a good approximation to the true dynamical reaction coordinates, they are not parametrized using dynamical information. Thus, their approximation quality could not, as yet, be validated, nor could the diffusion map eigenvalues be used to compute relaxation rate constants of the system. Here we combine the Diffusion Map approach with the recently proposed Variational Approach for Conformation Dynamics (VAC). Diffusion Map coordinates are used as a basis set, and their optimal linear combination is sought using the VAC, which employs time-correlation information on the molecular dynamics (MD) trajectories. We have applied this approach to ultra-long MD simulations of the Fip35 WW domain and found that the first DCs are indeed a good approximation to the true reaction coordinates of the system, but they could be further improved using the VAC. Using the Diffusion Map basis, excellent approximations to the relaxation rates of the system are obtained. Finally, we evaluate the quality of different metric spaces and find that pairwise minimal root-mean-square deviation performs poorly, while operating in the recently introduced kinetic maps based on the time-lagged independent component analysis gives the best performance.


Variational Approach
$\mathbf{C}_\tau \mathbf{a}_i = \lambda_i \mathbf{C}_0 \mathbf{a}_i$
Diffusion coordinates

INTRODUCTION

Molecular dynamics (MD) simulations have now reached considerable maturity. A few years ago, extensive sampling of protein systems was still unfeasible—except in a few exceptional projects such as folding@home[1] or the Anton supercomputer.[2] Now, it is commonly possible to achieve hundreds of microseconds of cumulative simulation time by harvesting the computational power of graphics processing units,[3−5] thus enabling extensive sampling of many biomolecular processes at moderate cost.

With the ability to generate vast amounts of MD data on a broad scale, analyzing these data and interpreting them in physicochemically relevant models has become a bottleneck. Consequently, the past years have seen a surge of interest in kinetic models that describe both the equilibrium behavior and the transition dynamics among a set of discrete conformational states. Popular examples include Markov models or Markov state models (MSMs),[6−12] hidden Markov models,[13,14] Diffusion Maps,[15,16] transition networks,[17−19] and Langevin models.[20] All of these models aim at achieving a simplified and interpretable yet accurate picture of the dynamics observed in the available MD trajectories. There are several questions of interest: Which are the most relevant long-lived states or structures of the molecular system?[9,14,21−24] What are their probabilities, transition rates, and relaxation time scales?[11,25−27] What are the transition pathways or mechanisms leading from reactants such as the unfolded or dissociated state to products such as the folded or associated states?[28−33]

It has been realized[6] that the most interesting quantities containing information on both the equilibrium and the slow kinetic properties of a molecular system are the dominant eigenvalues and eigenfunctions of the Markov operator. The eigenfunctions are typically nearly constant on the long-lived (metastable) states, but they change their sign between those regions. In this way, they encode parts of the state space where the system generally remains for a long time, and it rarely transitions between them. This concept is known as metastability and is a typical feature of biomolecules. The metastable regions of state space are frequently associated with biological function of the molecule, e.g., the ability/inability to bind to a binding partner. Therefore, it is precisely these regions that we are most interested in. The eigenvalues, on the

Received: August 5, 2015
Published: November 2, 2015

DOI: 10.1021/acs.jctc.5b00749

other hand, contain the information about the time it takes until such a rare transition might occur.

Indeed, many quantities of interest can be computed given an approximation of the eigenvalues and eigenfunctions, including metastable states,[6,9,14,21] coarse-grained Markov models,[34] and experimentally measurable observables.[27,35] Once the eigenfunctions are obtained, they can serve to define optimal reaction coordinates[33,36,37] and even simulation methods that are efficient in sampling the rare events.[38,39] Most of the molecular kinetics models above directly aim at reconstructing or approximating these eigenvalues and eigenfunctions, or can be shown to effectively attempt to do so.[15,16,40]

Since it was realized that the quality of the molecular kinetics model crucially depends on its ability to approximate the dominant eigenfunctions,[15,25,40] substantial research has gone into attempting to improve this approximation. Two developments are particularly noteworthy: First, the time-lagged (or time-structured) independent component analysis (TICA)[41−43] conducts a linear transformation of the molecular coordinates (e.g., internal coordinates such as distances or angles) onto a maximally slow subspace. It can be shown[42] that this method provides the optimal linear subspace for representing the eigenfunctions of the dynamical system, and thus a very good starting point for constructing a Markov model or other kinetics models that further improve the approximation of the eigenvectors. Second, Diffusion Maps[15] and particularly Locally Scaled Diffusion Maps[16] attempt to construct a direct approximation of the same set of eigenfuctions by nonlinearly projecting the high-dimensional configuration space onto a low-dimensional hyperplane spanned by the dominant eigenfunctions to be approximated. In principle, Diffusion Maps allow a much better approximation to the Markov operator eigenfunctions than a cluster discretization of state space, because they can operate on each sample configuration.

Despite these nice properties, a major drawback of diffusion maps has been that they are purely based on the idea that the spatial distribution of sampled configurations has been generated by a diffusion process. The construction of eigenfunction approximations through Diffusion Maps goes thus through the assumption of a specific dynamic model *a priori* (namely, a Fokker−Planck diffusion), and not through actual parametrization of dynamical observables in the data. As a consequence, Diffusion Maps currently have two drawbacks: (1) The validity of diffusion coordinates (DCs) as eigenfunction approximants cannot be self-consistently checked within the Diffusion Map framework, and their approximation quality cannot be improved. (2) The Diffusion Map eigenvalues cannot be directly related to physically meaningful relaxation time scales.

Here we set out to solve these two problems. Recently, one of the authors has contributed to the development of a variational principle and the Variational Approach for Conformation Dynamics (VAC).[44,45] In brief, this theory makes two statements:

1. Many molecular kinetics models, including MSMs, can be understood as attempting to approximate the eigenfunctions of the Markov operator by a linear combination of basis functions. The approximation is exact if the eigenvalue is approximated exactly as well. If the approximation is sub-optimal, the estimated eigenvalue will be too small. This idea can be cast into a rigorous variational formulation.
2. Inspired by the analogy with quantum mechanics, a Method of Linear Variation can be formulated which maximizes the estimation of the dominant eigenvalues in order to systematically exploit the variational bound. This is easily achieved by proposing a set of basis functions, building their linear combinations, and varying the coefficients until the optimal solution is found. The optimized linear combinations are then approximations of the dominant eigenvectors.

Here we exploit the variational principle and the algorithmic idea to validate and variationally optimize DCs. The main idea is that coordinates obtained by the Locally Scaled Diffusion Map method[16] are already a good approximation to the true eigenfunctions of the dynamical operator. We then use the dominant DCs in order to define a basis set that we further optimize using the method of linear variation. As a result, we get an improved approximation of the eigenfunctions, and additionally we get eigenvalues that can be interpreted as physical relaxation time scales. The resulting method, called Variationally Optimized Diffusion Map (VarDM), overcomes both Diffusion Map issues discussed before.

We describe the theory and methodology and apply the VarDM method to analyze two all-atom 100 μs Fip35 WW domain trajectories generated by the special purpose Anton supercomputer.[2] This data has been previously analyzed using a number of other methods,[2,46−50] and it provides an ideal benchmark for our approach.

The article is organized as follows. After briefly introducing the operators implementing the dynamics and discussing why their eigenvectors and eigenvalues are relevant in the context of MD and simulations, we describe how they can be approximated from a simulation data set, specifically focusing on the Principle of Linear Variation and the Locally Scaled Diffusion Map (LSDMap). Particular attention will be devoted to discussing how the concept of distance can play a crucial role in the diffusion map definition. Results for independent VarDM calculations for different setups are compared and contrasted, together with a benchmark MSM calculation. The results show that standard (purely structurally based) protocols to compute distances between molecular configurations are inadequate at approximating higher order dominant eigenprocesses. In contrast, a kinetic distance based on the TICA dimensionality reduction procedure provides an optimal definition of distance for the construction of Diffusion Maps. The molecular mechanisms associated with the data are then interpreted and discussed in comparison with previous studies.

■ THEORY

Conformation Dynamics, Eigenfunctions, and Eigenvalues. A MD simulation can be described as a Markov process in a state space Ω (generally containing both positions and momenta) that samples from an equilibrium probability density π. π is, for our purposes, given by the Boltzmann distribution at a constant temperature T:

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

where $\beta = (k_B T)^{-1}$ is the inverse temperature, $U(\mathbf{x})$ is the potential energy at phase space point $\mathbf{x}$, and Z is the partition function.

Table 1. Overview of Dynamical Operators for Describing Conformation Dynamics and Their Spectral Expansions

	propagator	transfer operator
symbol	$\mathcal{P}(\tau)$	$\mathcal{T}(\tau)$
definition	$\mathcal{P}(\tau)\rho(\mathbf{y}) = \int_{\mathbf{x}\in\Omega} \rho(\mathbf{x})p_\tau(\mathbf{y}\vert\mathbf{x})\,\mathrm{d}\mathbf{x}$	$\mathcal{T}(\tau)u(\mathbf{y}) = \frac{1}{\pi(\mathbf{y})}\int_{\mathbf{x}\in\Omega} \pi(\mathbf{x})u(\mathbf{x})p_\tau(\mathbf{y}\vert\mathbf{x})\,\mathrm{d}\mathbf{x}$
scalar product	$\langle f, g\rangle_{\pi^{-1}} = \int_{\mathbf{x}\in\Omega} \frac{1}{\pi(\mathbf{x})}f(\mathbf{x})g(\mathbf{x})\,\mathrm{d}\mathbf{x}$	$\langle f, g\rangle_{\pi} = \int_{\mathbf{x}\in\Omega} \pi(\mathbf{x})f(\mathbf{x})g(\mathbf{x})\,\mathrm{d}\mathbf{x}$
eigenfunctions	$\phi(\mathbf{x}) = \pi(\mathbf{x})\cdot\psi(\mathbf{x})$	$\psi(\mathbf{x}) = \pi^{-1}(\mathbf{x})\cdot\phi(\mathbf{x})$
spectral expansion	$\mathcal{P}(\tau)\rho(\mathbf{y}) = \sum_{k=0}^{\infty} \mathrm{e}^{-\kappa_k\tau}\langle\phi_k, \rho\rangle_{\pi^{-1}}\phi_k$	$\mathcal{T}(\tau)u(\mathbf{y}) = \sum_{k=0}^{\infty} \mathrm{e}^{-\kappa_k\tau}\langle\psi_k, u\rangle_{\pi}\psi_k$
truncated expansion	$\mathcal{P}(\tau)\rho(\mathbf{y}) \approx \sum_{k=0}^{m} \mathrm{e}^{-\kappa_k\tau}\langle\phi_k, \rho\rangle_{\pi^{-1}}\phi_k$	$\mathcal{T}(\tau)u(\mathbf{y}) \approx \sum_{k=0}^{m} \mathrm{e}^{-\kappa_k\tau}\langle\psi_k, u\rangle_{\pi}\psi_k$
stationary density	$\phi_0(\mathbf{x}) = \pi(\mathbf{x}) = \mathrm{e}^{-\beta U(\mathbf{x})}$ (Boltzmann density)	$\psi_0(\mathbf{x}) = 1$
corresponding generator	forward generator $\mathcal{L}$: $\mathcal{P}(\tau) = \exp(\tau\mathcal{L})$	backward generator $\mathcal{L}^*$: $\mathcal{T}(\tau) = \exp(\tau\mathcal{L}^*)$

Next, a formal way of describing the dynamics has to be introduced. We would like to stay as generic as possible and only state a few general properties that our dynamics must fulfill, without focusing on any specific model. First, the dynamics should be Markovian in full phase space; i.e., there is a probability density $p_\tau(\mathbf{y}|\mathbf{x})$ of making a transition, that is, to find the system at state $\mathbf{y}$ at a later time $t + \tau$ during the MD trajectory, given that it was at state $\mathbf{x}$ at time t. Many of the choices that can be made in a MD program, such as the thermostat and integrator used, will affect $p_\tau(\mathbf{y}|\mathbf{x})$ and will therefore also affect the kinetics, i.e., the transition rates between different conformations. However, independently of the specific choice of the dynamics, we require $p_\tau(\mathbf{y}|\mathbf{x})$ to have the following properties:

(i) The dynamics are ergodic; i.e., if we run them long enough $(\tau \to \infty)$, $p_\tau(\mathbf{y}|\mathbf{x})$ will sample from the Boltzmann density $\pi(\mathbf{y})$.

(ii) The dynamics satisfy detailed balance:

$$\pi(\mathbf{x})p_\tau(\mathbf{y}|\mathbf{x}) = \pi(\mathbf{y})p_\tau(\mathbf{x}|\mathbf{y})\ \forall\ \mathbf{x}, \mathbf{y} \in \Omega \tag{1}$$

Physically, eq 1 is consistent with the second law of thermodynamics in its formulation that work cannot be produced from thermal energy alone. Unfortunately not every implementation of MD obeys eq 1. It turns out that we can still work with dynamics implementations that fulfill a generalized form of detailed balance, where forward and backward probabilities are equal when momenta are reversed.[51] Langevin dynamics is an example of such a dynamics with generalized detailed balance. In this case eq 1 can be restored at the cost of Markovianity if x is in position space and momenta are integrated out. Nonetheless, since typical τ values are far larger than the autocorrelation times of momenta, the dynamics are still very closely Markovian in position space at time scale τ such that the present theoretical picture is conceptually useful. For the sake of simplicity, we will just assume for the subsequent discussion that eq 1 is fulfilled.

We now introduce a propagator formulation of such an ergodic and Markovian dynamics following the discussion given in ref 45. Using the state space function ρ_t, we can formally write the action of the MD in terms of a propagator:

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

where ρ_t quantifies the instantaneous probability density at time t; i.e., ρ_t measures for an ensemble of copies of a molecular system what fraction of its population is in which conformation. The propagator $\mathcal{P}$ propagates this density in time. Given the above requirements, it is obvious that after sufficiently many applications of $\mathcal{P}$, or for a long enough time τ, the resulting density will just be the equilibrium density $\pi(\mathbf{x})$. Table 1 defines $\mathcal{P}$ as a function of $p_\tau(\mathbf{y}|\mathbf{x})$. Note that we will never explicitly compute $\mathcal{P}$ or carry out the integrals in Table 1. However, we know that when running MD simulations, we effectively sample these integrals, and this gives us a way to construct a model for $\mathcal{P}$ and through that for the most interesting kinetic quantities.

Although it may seem like a purely mathematical comment, it is sometimes convenient or even necessary to work with a different operator than $\mathcal{P}$.[45] It is then appropriate to define the relative density u_t by

$$u_t(\mathbf{x}) = \frac{\rho_t(\mathbf{x})}{\pi(\mathbf{x})} \tag{2}$$

which is obtained by comparing the instantaneous density with the stationary density. Instead of working with $\mathcal{P}$, we can work with the transfer operator $\mathcal{T}(\tau)$ that transports u densities in time:

$$u_{t+\tau} = \mathcal{T}(\tau)u_t$$

When $\rho_t(\mathbf{x})$ is identical to the equilibrium density, it follows from the definition in eq 2 that the relative density is identical to 1. Table 1 summarizes the properties of both operators.

It turns out that both the $\mathcal{P}(\tau)$ and $\mathcal{T}(\tau)$ operators are of great interest. The eigenfunctions and eigenvalues of these operators contain the essential information on the molecular kinetics; i.e., what are the long-lived states, and what are the transition rates between them? Understanding the mathematical properties of these operators allows to design computational methods to approximate their spectra and thus chemically interesting quantities. A plethora of different methods such as Markov models, Diffusion Maps, VAC, and many more are based on this idea.

Both $\mathcal{P}$ and $\mathcal{T}$ operators share the same eigenvalues. The largest eigenvalue is $\lambda_0 = 1$, whereas all remaining eigenvalues are strictly smaller than 1:

$$1 = \lambda_0 > \lambda_1 \geq \cdots$$

The eigenfunction corresponding to the largest eigenvalue $\lambda_0 = 1$ is just the stationary Boltzmann density for the propagator and the constant function for the transfer operator. All eigenvalues except the first decay exponentially with the lag-time τ; i.e., we have

$$\lambda_k(\tau) = \mathrm{e}^{-\kappa_k \tau} \tag{3}$$

with some relaxation rates κ_k. Thus, knowing the eigenvalues allows us to compute the relaxation rates, or the relaxation time scales:

$$t_k = \kappa_k^{-1} = -\frac{\tau}{\log|\lambda_k(\tau)|} \tag{4}$$

These relaxation rates or time scales are important quantities to compare to experiments, as they directly affect the measurement signal in various kinetic experiments such as fluorescence correlation spectroscopy,[27] 2D IR spectroscopy,[52] temperature jump spectroscopy,[30] neutron scattering[53] (a comprehensive review was written by Keller *et al.*[35]).

The eigenfunctions ϕ and ψ are the key quantities for understanding molecular mechanisms. It was proposed in ref 6 that the different metastable (long-lived) states of a molecule could be identified by sorting its microstates according to the signs they have in the eigenfunctions corresponding to the slowest relaxations, i.e., those with the largest eigenvalues. This idea was further refined in ref 54. In Noé *et al.*,[27] it was used in order to construct a systematic way of giving a structural interpretation to ensemble kinetics experiments with the help of MD simulations. Both MSMs and Diffusion Map algorithms attempt to approximate the eigenfunctions ϕ or ψ. It was found that the accuracy by which these eigenfunctions are approximated is crucial for the accuracy of the model.[25,40]

When studying molecular kinetics we are usually interested in the slow processes only. Therefore, for a sufficiently large choice of τ (e.g., in the order of nanoseconds), only a finite number m of the terms in the spectral expansion are still present, and the operator's action can be understood in terms of only finitely many processes, and we obtain the truncated expansions shown in Table 1. We are then interested in methods that allow us to compute (or approximate) a few dominant eigenvalues λ_1, ..., λ_m and their corresponding eigenfunctions ϕ_1, ..., ϕ_m. Alternatively we may also work with ψ_1, ..., ψ_m, provided that some estimate of the equilibrium distribution π is available.

Method of Linear Variation and the Special Variational Principle. A number of useful computational methods have been developed in past years to approximate the eigenvalues and eigenfunctions of the propagator or transfer operator. Recently, a very general approximation principle for such purpose was proposed: VAC[45]. Briefly, if f is some approximation model of the first nontrivial $\mathcal{T}(\tau)$ eigenfunction, $f \approx \psi_1$, then the Rayleigh quotient,

$$\hat{\lambda}_1(\tau) = \frac{\langle f, \mathcal{T}(\tau) f\rangle_\pi}{\langle f, f\rangle_\pi} \tag{5}$$

is an estimate to the corresponding eigenvalue $\hat{\lambda}_1(\tau) \approx \lambda_1(\tau)$ (see Table 1 for the definition of the scalar product). The variational principle states that $\hat{\lambda}_1$ always underestimates the true eigenvalue ($\hat{\lambda}_1(\tau) \leq \lambda_1(\tau)$), equivalence holding only if $f = \psi_1$. If f is an approximation to other eigenfunctions (ψ_2, ...) and constructed such that it is orthogonal to previous exact eigenfunctions ($\psi_0 = 1$, ψ_1, ...), then this variational principle also carries over to subsequent eigenvalue/eigenfunction pairs. Thus, a computational algorithm can approximate eigenvalue/eigenfunction pairs by proposing a model for the eigenfunction and then optimizing it by maximizing the corresponding Rayleigh coefficient.

From a practical perspective, the crucial insight from the VAC is that the Rayleigh coefficients can be easily computed for a given model function f in terms of its autocorrelation function.[45] Furthermore, the search for the optimal model can be performed in a single step if we consider eigenfunctions to be a weighted sum of basis functions $\{\chi_1(x), ..., \chi_m(x)\}$, i.e.,

$$\hat{\psi}_i(\mathbf{x}) = \sum_{j=1}^{m} a_{ij}\chi_j(\mathbf{x}) \tag{6}$$

In this case, the optimal solution consists of finding the optimal coefficients a_{ij} which are just given by applying the generalized Ritz (or Rothaan–Hall) method to conformation dynamics.[45] One computes the correlation matrices $\mathbf{C}(\tau)$ and $\mathbf{C}(0)$ with elements,

$$c_{ij}(\tau) = \langle \chi_i, \mathcal{T}(\tau)\chi_j\rangle_\pi \tag{7}$$

$$c_{ij}(0) = \langle \chi_i, \chi_j\rangle_\pi \tag{8}$$

and obtains the optimal coefficients a_{ij} by solving the generalized eigenvalue problem,[44,45]

$$\mathbf{C}(\tau)\mathbf{a}_i = \hat{\lambda}_i \mathbf{C}(0)\mathbf{a}_i \tag{9}$$

where $\mathbf{a}_i \in R^m$ is the vector of coefficients of the functions χ_i to build the ith optimized linear combination. We refer to the Methods section for a description of how the correlation matrix elements (eqs 7 and 8) are actually computed from a MD trajectory. The generalized eigenvalue problem (eq 9) occurs in other kinetic models and estimators, such as Markov models with fuzzy partitions of unity,[55] core- or milestoning-based MSMs,[26] TICA,[42] and ordinary MSMs that can all be treated as special cases of the VAC with specific choices of basis sets (see ref 45 for a discussion).

When using the Method of Linear Variation to compute eigenvalues and eigenvectors, we can derive a special variational principle that holds for the solutions of the eigenvalue problem (eq 9):

1. If we use the exact eigenvectors 0 through m as basis functions $\chi_i = \psi_i$, we will obtain

$$c_{ij}(0) = \langle \psi_i, \psi_j\rangle_\pi = \delta_{ij} \tag{10}$$

$$c_{ij}(\tau) = \langle \psi_i, \mathcal{T}(\tau)\psi_j\rangle_\pi = \delta_{ij}\lambda_i(\tau) \tag{11}$$

and thus,

$$\mathbf{C}(0) = \mathrm{Id} \tag{12}$$

$$\mathbf{C}(\tau) = \mathbf{\Lambda}(\tau) \tag{13}$$

so the eigenvalue problem becomes trivial, and we recover the exact eigenvalues $\hat{\lambda}_i(\tau) = \lambda_i(\tau)$ for all $i = 1$, ..., m.

2. If we have a basis set error—i.e., the exact eigenvectors *cannot* be represented as a linear combination of basis functions as in eq 6—then we have the following special variational constraint on the eigenvalues:

$$\hat{\lambda}_i(\tau) < \lambda_i(\tau) \qquad \text{for all } i = 1, \ ..., \ m$$

We have the variational principle for all m eigenvalues because our eigenvectors are by construction an orthogonal set of vectors.

3. The accuracy of the VAC strictly depends on the choice of the basis functions.

(Locally Scaled) Diffusion Map and Diffusion Coordinates. The Diffusion Map algorithm is an independent approach to approximate the transfer operator eigenfunctions ψ_i starting from an equilibrium MD data set. The description of the algorithm and its ideas that we present here closely parallels that given by Coifman *et al.*[56]

In the Diffusion Map formalism we assume that the configurations in the sampling were generated by a specific dynamical model being pure diffusion in the potential $U(\mathbf{x})$ (from now on, $\mathbf{x}$ will denote a point in the configuration space). Then, the system dynamics is governed by the Fokker–Planck equation,

$$\frac{\partial \rho}{\partial t} = \mathcal{L}\rho = \frac{\Delta \rho}{\beta} + \nabla(\rho \nabla U) \tag{14}$$

where again $\rho(\mathbf{x})$ is the probability density of state $\mathbf{x}$, $\beta = (kT)^{-1}$ is the inverse temperature, and the gradient ∇U is minus the drift (force) at position $\mathbf{x}$. The operator $\mathcal{L}$ is the diffusion generator, and it is a differential operator that models the instantaneous time change of the state space density. It is yet another way of modeling the system's dynamics, and it is related to the propagator $\mathcal{P}(\tau)$ by an exponentiation, i.e., $\mathcal{P}(\tau) = \exp(\tau\mathcal{L})$, see Table 1. $\mathcal{L}$ has a discrete spectrum with nonpositive eigenvalues $\{-\kappa_j\}_{j=0}^{\infty}$, where $\kappa_0 = 0 < \kappa_1 \leq \kappa_2 \leq ...$, are the system's relaxation rates of eq 3. $\mathcal{L}$ has associated eigenfunctions $\{\phi_j\}_{j=0}^{\infty}$ that are identical to the propagator eigenfunctions (see Table 1). Like for propagators and transfer operators, we can write (at least formally) the solution of eq 14 as

$$\rho_{t+\tau}(\mathbf{y}) = \sum_{i=0}^{\infty} \langle \phi_i(\mathbf{x}) | \rho_t(\mathbf{x}) \rangle_{\pi^{-1}} e^{-\kappa_i \tau} \phi_i(\mathbf{y}) \tag{15}$$

which in the limit of long τ converges to $\phi_0(\mathbf{x}) = \pi(\mathbf{x})$, corresponding to the eigenvalue $\kappa_0 = 0$, which is given by the Boltzmann equilibrium distribution.

A complementary approach is to use the backward Fokker–Planck equation, which implements the time evolution of π-reduced probability densities:

$$\frac{\partial u}{\partial t} = \mathcal{L}^* u = \frac{\Delta u}{\beta} - \nabla u \cdot \nabla U \tag{16}$$

The backward generator $\mathcal{L}^*$ has the same eigenvalues as operator $\mathcal{L}$ but the same eigenfunctions $\psi_i(x)$ as the transfer operator (see Table 1). It is related to the transfer operator $\mathcal{T}$ by $\mathcal{T}(\tau) = \exp(\mathcal{L}^*)$, see Table 1. An infinite decomposition similar to eq 15 can be written for the reduced u probability as well.

For any starting point $\mathbf{x} \in \Omega$, let $\rho_\tau(\mathbf{y}|\mathbf{x})$ be the solution of the forward Fokker–Planck equation with initial condition $\rho_0(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{y})$. The diffusion distance between any of the two points $\mathbf{x}_1,\mathbf{x}_2 \in \Omega$ at time τ is defined as the distance between the corresponding probability densities at time τ, when initialized at $\mathbf{x}_1$ or at $\mathbf{x}_2$ at time 0.[56] The distance is measured in the Hilbert space $L_2(\Omega,w)$ with the weight function $w(\mathbf{x}) = 1/\varphi_0(\mathbf{x}) = \pi^{-1}(\mathbf{x})$. The distance can be written as

$$D_\tau^2(\mathbf{x}_1, \mathbf{x}_2) = \| \rho_{t+\tau}(\mathbf{y}|\mathbf{x}_1) - \rho_{t+\tau}(\mathbf{y}|\mathbf{x}_2) \|^2_{\pi^{-1}} \tag{17}$$

Using eq 15 and its relation with eq 16, we obtain the spectral representation:

$$D_\tau^2(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i \geq 1} e^{-2\kappa_i \tau} \| \psi_i(\mathbf{x}_1) - \psi_i(\mathbf{x}_2) \|^2 \tag{18}$$

Many times, molecular systems are metastable and their dynamics is dominated by a (finite) number m of slow processes encoding the more interesting physics, while all the other processes m+1, ... are much faster over time and encode more subtle, less crucial dynamical features. From a mathematical standpoint, this means that the propagator eigenvalues display a spectral gap at an index m, i.e., $\kappa_m \ll \kappa_{m+1}$. The infinite expansions above, eqs 15 and 18, can then be truncated up to the mth term, if only the slow regime is of interest:

$$\rho_{t+\tau}(\mathbf{y}) = \phi_0(\mathbf{x}) + \sum_{i=1}^{m} \langle \phi_i(\mathbf{x}) | \rho_t(\mathbf{x}) \rangle_{\pi^{-1}} e^{-\kappa_i \tau} \phi_i(\mathbf{y}) \tag{19}$$

The Diffusion Map algorithm can then be seen as the specific nonlinear mapping from the original high dimensional $\mathbf{x}$ configuration space to a mth dimensional Euclidean space that preserves the diffusion distance between each pair of the points in the data set.[15,16,56] The main result to be mentioned here is that the low-dimensional representation space is spanned by the eigenvectors of the generator $\mathcal{L}$, which are automatically computed upon carrying out the projection.

Now consider that we are given a sample of points, $\{\mathbf{x}_i\}_{i=1,...,N}$, generated for example by a long MD trajectory. The elementary steps of the Diffusion Map algorithm are the following:

1. Consider the kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\| \mathbf{x}_i - \mathbf{x}_j \|^2}{2\varepsilon^2} \right) \tag{20}$$

where ε is a scale parameter. $\|\cdots\|$ denotes the distance between configurations $\mathbf{x}_i$ and $\mathbf{x}_j$. The LSDMap formulation[16] differs from the original Diffusion Map[15] in the definition of a local scale parameter $\varepsilon(\mathbf{x})$ that is a function of the configuration $\mathbf{x}$, rather than a constant value ε. For reasons of symmetry, ε^2 in eq 20 is replaced by $\varepsilon(\mathbf{x}_i)\varepsilon(\mathbf{x}_j)$. The local scale has not been derived from a diffusion process in the input coordinate space, but is rather a model parameter that has practically proven to perform significantly better than a constant scale for the definition of reaction coordinates and in adaptive sampling schemes.[16,39] It is worth mentioning that the particular choice for both $\|\cdots\|$ and ε may severely affect the algorithm performance, as we will discuss later.

2. Normalize the kernel as

$$\tilde{K}_{i,j} = \frac{K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{\sum_k K(\mathbf{x}_i, \mathbf{x}_k) \sum_k K(\mathbf{x}_j, \mathbf{x}_k)}} \tag{21}$$

3. Define $D_i = \sum_j \tilde{\tilde{K}}_{i,j}$ and construct the Diffusion Map transition matrix,

$$M_{i,j} = \frac{\tilde{K}_{i,j}}{D_i} \tag{22}$$

4. Calculate the first m eigenvectors of M, $\{\tilde{\psi}_i\}$. Note that the eigenvectors $\tilde{\psi}_i$ are defined on the sampled configurations $\mathbf{x} \in \Omega$ pointwise.

The eigenvectors $\tilde{\psi}_i$ are called *diffusion coordinates* (DCs) and are pointwise approximations of the exact eigenvectors $\psi_i(x)$:

$$\tilde{\psi}_i(\mathbf{x}) = \mathrm{DC}_i(\mathbf{x}) \approx \psi_i(\mathbf{x}) \tag{23}$$

Please note that at this point nothing can really be said or inferred about the relationship between the diffusion map transition matrix eigenvalues and the exact generator or propagator eigenvalues.

Choice of a Distance Metric. The LSDMap algorithm has been tested extensively and has been applied to study the free energy landscape of a number of molecular systems.[16,33,36,38,39] The use of a constant local scale as proposed in the original formulation[56] was found to be inadequate for MD data, as thoroughly discussed in the references above. Another crucial choice in LSDMap is the definition of the distance metric involved in the diffusion kernel (eq 20). Similarly, the choice of a suitable distance is a key ingredient for kinetic clustering protocols in the definition of kinetic models such a MSMs. In the Diffusion Map scenario, the prescription to compute distances influences the definition of diffusion distance preserved upon projection; in the MSMs, it is fundamental to discretize the configuration space, lump similar configurations together and separate dissimilar ones, making sure that slow-interconverting states are distinguished and well separated.

Standard choices for distance metrics for molecular systems, such as the root-mean-square deviation (RMSD), generally use static information from each configuration to define structural similarity. However, it has been shown in a number of publications[57,58] that kinetic similarity (that is, how long it takes for a configuration to evolve into another) is crucial in the definition of distance relationship between the molecular configurations. Distance metrics based on purely geometrical criteria suffer when slow conformational transitions between geometrically proximal states exist, such as register shift dynamics in β-sheet protein topologies, or, in general, geometrically similar states separated by large free energy barriers.

A general and transferrable protocol to design alternative concepts of distance metrics accounting for the interplay between structural and kinetic similarity is lacking. A first step toward this direction would be to compare and contrast the performance of different structural distance metrics. It was shown[50,57,58] that different geometrical metrics capture kinetic properties in a specific, metric-dependent fashion. Here we contribute to this discussion by using the variational principle to decide between different choices of distance metrics, in that it provides us with a quantitative tool to assess the performance of each metric in the context of LSDMap.

■ METHODS

Fip35 Setup and Simulation. WW domains are small, independently folding protein domains that bind to proline-rich sequences. Their topology is characterized by two β-hairpins, which form a three-stranded β-sheet. A number of mutants of the 35 amino acids WW domain of the human protein Pin1 have been engineered which fold in few tens of microseconds,[59] the fastest of them being the Fip35 mutant. The mutants small size and their ultrafast kinetics make them ideal benchmark systems, for which numerical simulations can be compared to a large body of experiments.

Recently, D.E. Shaw Research has generated two 100 μs Fip35 trajectories using the AMBER99SB-ILDN all-atom force field in TIP3P explicit water at 337 K using a Nosé−Hoover thermostat with a relaxation time of 1.0 ps.[2] This data set is used here as the equilibrium sampling onto which multiple VarDM calculations were performed. The setup details are given below.

LSDMap Setup. The LSDMap analysis was repeated for different choices of distance metrics and different choices for the local scale parameter ε in order to assess to what extent the algorithm performance is affected.

Four choices of distance metrics were considered: three of them (RMSD, DD, and CMD) have an immediate structural interpretation in terms of simple physical degrees of freedom; the other metric (KD) is less intuitive and requires a more detailed discussion.

Structural Metrics: RMSD, DD, and CMD. Let $\mathbf{x}^\alpha$ and $\mathbf{x}^\beta$ be any two molecular configurations. Their mutual root-mean-square deviation is defined as

$$\mathrm{RMSD}(\mathbf{x}^\alpha, \mathbf{x}^\beta) = \sqrt{\frac{1}{N}\sum_i (x_i^\alpha - x_i^\beta)^2} \tag{24}$$

where the sum runs over the N heavy atoms. The two configurations have to be preliminarly aligned in an optimal way such that their Euclidean distance is minimized.

Inspired by the discussion in the work of Cossio *et al.*,[57] we considered two additional structural metrics, the dihedral and contact map distances. The dihedral distance is defined as

$$\mathrm{DD}(\mathbf{x}^\alpha, \mathbf{x}^\beta) = \sqrt{\frac{1}{M}\sum_i \frac{1}{2}(1 - \cos(\varphi_i(\mathbf{x}^\alpha) - \varphi_i(\mathbf{x}^\beta)))} \tag{25}$$

where the set of M dihedrals $\{\varphi_i\}$ formed by the backbone atoms N, C_α, C, and N is considered. As a reduced number of degrees of freedom is used to measure distances with the DD metric, it is expected to perform faster than RMSD, and it is an appealing candidate for heavy-duty computations, such as biased MD.[39]

The contact map distance is defined as a standard Euclidean distance in the space of heavy atom contacts:

$$\mathrm{CMD}(\mathbf{x}^\alpha, \mathbf{x}^\beta) = \sqrt{\frac{1}{N^{\alpha\beta}}\sum_{i\neq j} (C_{ij}(\mathbf{x}^\alpha) - C_{ij}(\mathbf{x}^\beta))^2} \tag{26}$$

where

$$C_{ij}(\mathbf{x}) = \frac{1 - (r_{ij}(\mathbf{x})/r_0)^8}{1 - (r_{ij}(\mathbf{x})/r_0)^{12}}$$

is a smooth definition for a contact formation, r_{ij} being the Euclidean distance between heavy atom pair $i-j$ and r_0 = 0.35 nm being an appropriate cutoff, and

$$N^{\alpha\beta} = \sqrt{\left(\sum_{ij} C_{ij}^\alpha\right)\left(\sum_{ij} C_{ij}^\beta\right)}$$

is a normalization constant.[57] CMD is popularly employed in bioinformatics in structure analysis, such as in structure prediction algorithms;[60] hence, it appears to be a relevant alternative to RMSD while approximately retaining the same level of coarse-graining description.

Kinetic Distance (KD). The last distance metric considered in the analysis is KD,[61] based on the time-lagged independent component analysis (TICA).[42,43] TICA is a linear dimensionality reduction technique which builds linear combinations out of a chosen set of molecular coordinates $\{r_i(t)\}$, such as atomic positions, distances, and angles, in such a way that the eigenvectors and eigenvalues of the propagator of the dynamics, $\mathcal{T}$, are approximated. A comprehensive and detailed description of the algorithm can be found in ref 42. In practice, TICA performs a variational optimization as described in the Theory section. Given MD data, a set of input coordinates $\{r_i(t)\}$ is chosen and the mean-free coordinates,

$$y_i(t) = r_i(t) - \langle r_i(t)\rangle_t \tag{27}$$

are defined. The y_i's are then used as input basis functions χ for the variational calculation specified by eqs 7−9. Let us denote as $\hat{\psi}_i$ and $\hat{\lambda}_i$ the approximated eigenvectors and eigenvalues, respectively, from the TICA calculations (at a given lag time reference value τ_{TICA}).

The dominant linear combinations span a low-dimensional space where the system's slowest processes live. The KD is then defined as the Euclidean distance in this space upon scaling the $\hat{\psi}_i$ by the corresponding eigenvalues in the following way:[61]

$$\text{KD}(\mathbf{x}^\alpha, \mathbf{x}^\beta) = \sqrt{\sum_{i=1}^{q} \hat{\lambda}_i^2 (\hat{\psi}_i(\mathbf{x}^\alpha) - \hat{\psi}_i(\mathbf{x}^\beta))^2} \tag{28}$$

It was shown that using KD as a distance metric for clustering yields MSMs with better approximation qualities than when unscaled TICA coordinates are used.[61] Moreover, the eigenvector rescaling in eq 28 weights the TICA coordinates according to their "slowness", thus making the question of how many coordinates need to be taken into account in the clustering obsolete.

In our protocol, the input parameters $\{r_i(t)\}$ to the TICA calculations were chosen to be all the mutual C_α distances (that means each configuration in the data set is labeled by a list of all the mutual distances) at a lag time $\tau_{\text{TICA}} = 0.1$ μs. It is worth pointing out that this distance metric reduces to the standard diffusion distance (eq 18) if the MD data set was generated by a reversible purely diffusive process.

The four distance metrics (eqs 24−26 and 28) were used to run separate and independent LSDMap calculations on the Fip35 data set, by using the algorithm detailed in the Theory section. Two different choices were used for the local scale parameter $\varepsilon(\mathbf{x})$ in the diffusion map kernel (eq 20):

1. a constant value $\varepsilon(\mathbf{x}) = \varepsilon$, where different values of ε were tested (Diffusion Map), or
2. the distance to $N_k(\mathbf{x})$, the kth nearest neighbor configuration to configuration $\mathbf{x}$ in the data set:

$$\varepsilon(\mathbf{x}) = \|\mathbf{x} - N_k(\mathbf{x})\|^2 \tag{29}$$

The DCs are computed as the eigenvectors of the diffusion map transition matrix M in eq 22. Subsequently, these coordinates were orthogonalized and used as basis set $\{\chi_i\}$ for the variational method, as described in the Theory section.

Variational Optimization. Were this LSDMap approximation exact, then the stationary eigenvector $\varphi_0(\mathbf{x})$ of M would be identical to $\pi(\mathbf{x}) = e^{-\beta U(\mathbf{x})}$ and then the eigenvectors would be orthonormal with respect to this density ($\langle\tilde{\psi}_i, \tilde{\psi}_j\rangle_{\varphi 0} = \langle\psi_i\psi_j\rangle_\pi = \delta_{ij}$).

Because we assume to have some finite approximation errors, we perform an orthogonalization procedure in order to obtain self-consistent DCs. We first compute the overlap matrix,

$$c_{ij}(0) = \langle\tilde{\psi}_i, \tilde{\psi}_j\rangle_{\varphi_0} \tag{30}$$

where $\varphi_0^{\text{T}} = \varphi_0^{\text{T}} M$ is the stationary eigenvector obtained from the diagonalization of the diffusion map transition matrix M. We then solve for the eigenvectors of the overlap matrix,

$$\mathbf{C}(0)\hat{\psi}_i = \hat{\lambda}_i \hat{\psi}_i \tag{31}$$

and use the resulting eigenvectors,

$$\chi_i = \hat{\psi}_i \tag{32}$$

as a basis in the Method of Linear Variation described above. Note that the generalized eigenvalue problem (eq 9) does not require the basis set to be orthogonal; however, the orthogonalization in eq 9 is made with respect to the stationary density of the transfer operator, while the orthogonalization in eq 31 is done with respect to the Diffusion Map stationary density. Because of the approximations involved, these orthogonalization procedures are slightly different, and eqs 30−32 are done in order to start the variational method with a self-consistent basis set.

If the DCs were exact, $\hat{\psi}_i = \tilde{\psi}_i = \psi_i$, then the variational principle correlation matrices would be diagonal (see eqs 10−13). On the other hand, the variational principle can be used as an *a posteriori* validation of the approximation quality of the DCs and to further optimize them.

Matrix elements can be computed on the fly by using a stochastic realization of the trajectory $\{x_1, x_2, ..., x_k, ..., x_T\}$. Upon orthogonalization (eq 31), we compute the variational principle correlation matrices (eqs 7 and 8) as direct time averages:

$$c_{ij}(\tau) \approx \frac{1}{T - \tau} \sum_{k=1}^{T-\tau} \chi_i(x_k)\chi_j(x_{k+\tau}) \tag{33}$$

$$c_{ij}(0) \approx \frac{1}{T} \sum_{k=1}^{T} \chi_i(x_k)\chi_j(x_k) \tag{34}$$

Furthermore, we enforce the time-lagged correlation matrix to be symmetric,

$$C(\tau) = \frac{1}{2}(C(\tau) + C(\tau)^T) \tag{35}$$

and then solve the eigenvalue problem (eq 9) for different values of the lag time τ. The eigenvalues $\hat{\lambda}_i(\tau)$ are used to approximate the kinetic time scales of the eigenprocesses, whereas the generalized eigenvectors $\{\mathbf{a}_i\}$ are the linear combination coefficients to build the approximants $\hat{\psi}_i = \sum_j \mathbf{a}_{ij}\chi_j$ to the eigenvectors of the propagator.

Markov State Model. As a comparison, a MSM was built on the Fip35 data set, using the pyEMMA package (www.pyemma.org).[62] The KD introduced above (eq 28) was used in the definition of the MSM, and 1000 k-mean clusters were used as clustering protocol. It is worth pointing out that other MSM-based Fip35 studies[49,50] use metrics for clustering that are

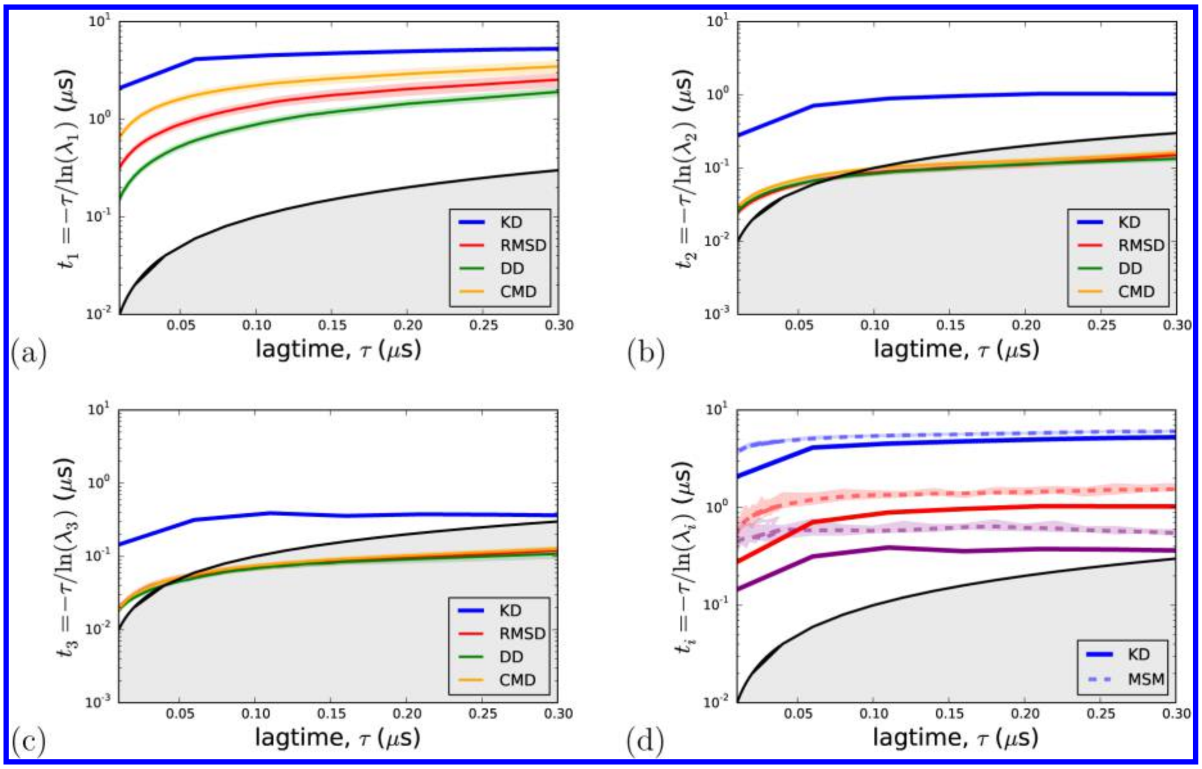


Figure 1. Implied time scales (a) t_1, (b) t_2, and (c) t_3 as from the VarDM calculations ($k = 2000$) using RMSD, DD, CMD, and KD distance metrics. All metrics are equally efficient in targeting $t_1 \approx 5.5$ μs, even though KD estimates converge faster and almost simultaneously. The second and third implied time scales t_2 and t_3 are severely underestimated by all three structural metrics considered and affected by serious convergence issues. (d) Comparison of the time scales resulting from KD VarDM and MSM time scales: t_1 is blue, t_2 red, and t_3 purple.

different from KD. Time scale convergence was studied and compared to VarDM results and literature.

■ RESULTS AND DISCUSSION

Folding of WW Domain Fip35 Mutant. The folding of the Fip35 mutant of the WW domain has been widely investigated. For example, MD simulations[63,64] were performed for a time interval longer than 10 μs, but no folding transitions were observed. Pande and co-workers[65] tackled the problem using a worldwide distributed computing scheme, and found transitions proceeding through a multitude of qualitatively different and nearly equiprobable folding pathways. A very different conclusion was reached by Shaw and collaborators,[2] by analyzing ultralong MD trajectories with multiple unfolding/folding events, obtained using a special-purpose supercomputer. The authors concluded that Fip35 folds predominantly along a pathway in which the first hairpin is fully structured, before the second hairpin begins to fold. On the same line, they concluded that no relevant barriers are present in the folding process; hence, Fip35 is an incipient downhill folder.

Krivov[46] challenged this interpretation and proposed the existence of intermediates revealed by using a novel optimized reaction coordinate on the same data set. Moreover, he proposed that an alternative folding pathway also exists, in which the second hairpin forms before the first hairpin, though it is 5 times less likely than the main pathway.

MSM analysis[49,50] of the same Fip35 data set also suggested that folding mechanism is heterogeneous, though it was realized that the majority of the folding flux flows through the path reported by Shaw.[2] Characteristic time scales were estimated to be 5 μs, and faster processes were also investigated systematically.

Recently, Mori and Saito[48] built on the popular Principal Component Analysis approach to propose a Dynamic Component Analysis using temporal information from the trajectory and concluded that Fip35 is not an incipient downhill folder. They succeeded in identifying two misfolded states where the trajectory happens to be temporarely trapped; hence, explaining the reason for dynamic heterogeneity.

Finally, Berezovska *et al.*[47] presented a network-based analysis of the same data by looking at local fluctuations of conventional order parameters such as RMSD and recapitulated previous results by showing that folding occurs along two preferential pathways, both involving intermediates formation. Their estimate for the mean first-passage folding time is 4.3 μs.

Beccara *et al.*[66,67] generated their own MD data set using the DRP protocol[68] and confirmed that Fip35 folds mostly through two channels which differ in the order in which hairpins are formed, and that folding involves intermediates.

In the following we contribute to the discussion by analyzing the Fip35 data set using the VarDM hybrid approach based on LSDMap and the variational principle, as described above. Our analysis reveals that standard structural metrics used to define configuration similarity are not appropriate, in that they do not fully capture the kinetic properties of the system. We propose a new way of computing distances between configurations which turns out to be extremely effective, kinetically and computationally. Our results show that the Fip35 folding mostly proceeds via on-pathway intermediate(s), in qualitative agreement with previous results based on reaction coordinate optimization.

Results. *Optimal LSDMap Parametrization.* The original data set was downsampled to 100 000 uniformely spaced points, so that two consecutive configurations are separated by

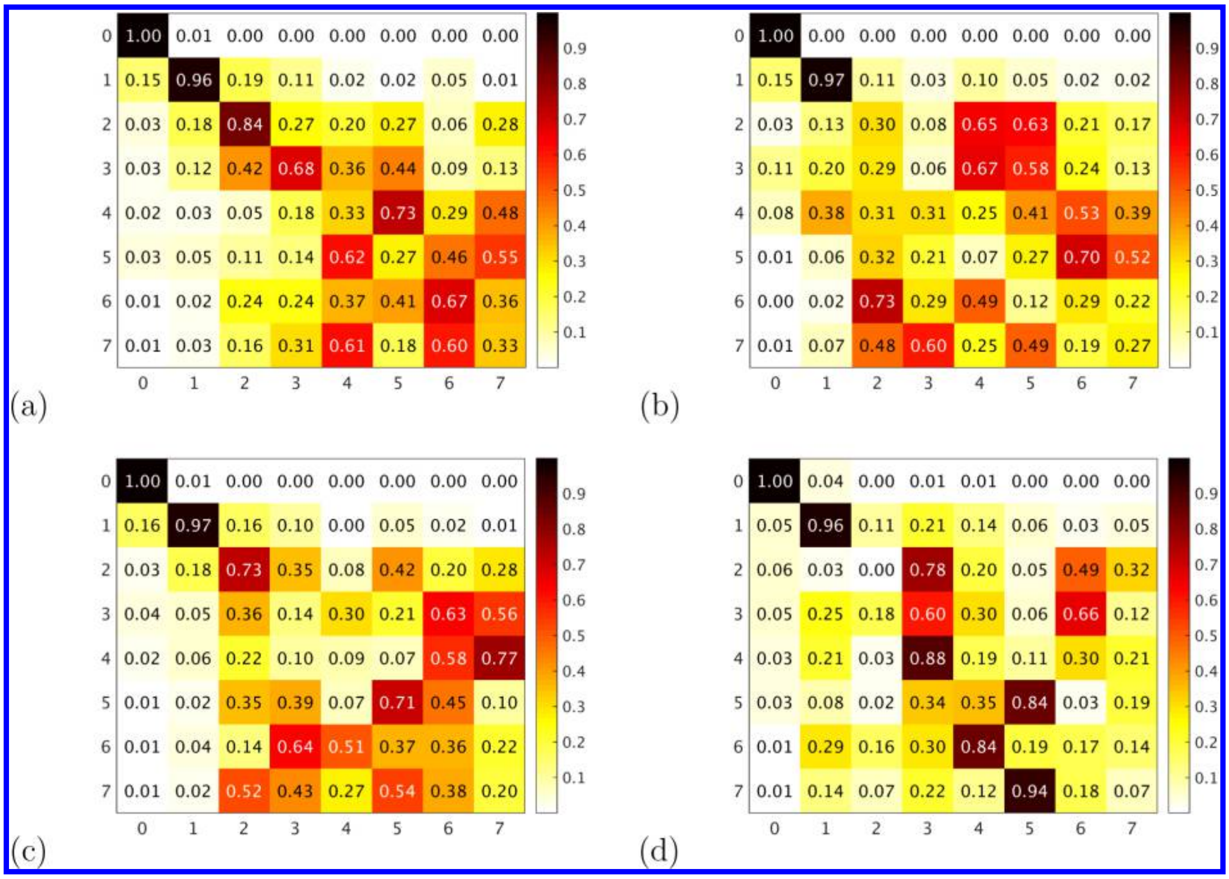


Figure 2. Truncated linear combination coefficient matrix A for (a) RMSD, (b) DD, (c) CMD, and (d) KD. The entry (i,j) is the matrix element a_{ij} and represents the component of the ith optimized eigenvector onto the jth LSDMap diffusion coordinate. Ideally we would expect an identity matrix $a_{ij} = \delta_{ij}$, but these results show that this is not the case; hence, the DCs are not invariant upon optimization.

a 2 ns time interval. Multiple VarDM calculations were performed using eqs 24−26 with both constant and adaptive local scales, as described above, and an input basis set of 100 DCs.

The distribution of the local scale values obtained using the k-nearest neighbor definition provided in eq 29 is bimodal (data not shown), suggesting that the definition of an appropriate constant ε value may be difficult, once again supporting the need of an adaptive, point-dependent local scale.[16,39] Results for different constant local scale values are discussed in the Supporting Information.

Results obtained by using eq 29 are very robust when varying the number of nearest neighbors k over a broad range from 5 to 10 000, as shown in the left panel of Figure S1. This result indicates that appropriate network connectivity is ensured by an extremely small fraction of points, and that the data set manifold shape is preserved.

In the following we will show the results obtained with k = 2000.

Eigenvalues and Implied Time Scales. The variational eigenvalue problem (eq 9) is solved for increasing values of the lag time τ to study the convergence of both eigenvalues and eigenvectors. In principle, both quantities are lag time independent, since the generator (eq 14) does not depend on τ explicitly. However, in practice, some equilibration over short time scale is expected, where time scales and eigenvectors gradually converge to their asymptotic values. This can be interpreted formally by invoking theoretical results about MSM convergence, once it is realized that the variational formulation with characteristic functions as basis functions is equivalent to a MSM.[45] On the other hand, if too long of a lag time is chosen, then decorrelation time and trajectory length start being comparably long, and estimates are no longer reliable. Thus, we expect only a limited range of lag time values where physical convergence of model parameters is reached, as customarily observed.[25,42,44,45] A bootstrapping procedure was employed to estimate statistically meaningful errorbars for the time scales and the linear combination coefficients.

As pointed out in the previous section, we are discussing here optimization results obtained choosing k = 2000 for local-scale computation. Figure 1 shows the convergence of the first three implied time scales associated with Fip35 kinetics,

$$t_i = -\frac{\tau}{\log \hat{\lambda}_i(\tau)} \qquad i = 1,\ 2,\ 3 \tag{36}$$

obtained with the structural metrics and the KD metric (eq 28) (Figure 1a−c). A comparison with the MSM results is shown in Figure 1d. The semi-logarithmic scale helps highlight the improvement of convergence upon using the kinetic metric; indeed, t_1 enters its convergence region around τ = 0.1 μs and τ = 4 μs when KD or structural distance is used, respectively (for visualization purposes, the plots are truncated). The corresponding estimates for the folding characteristic time scale are 6.2 μs and (5.5 ± 2.0) μs, respectively, in agreement with each other, with MSM results and other estimates obtained using alternative analysis procedures.[46−50] By invoking the variational bound, we claim that all four metrics metrics employed in the calculations provide consistent results within the error bars in

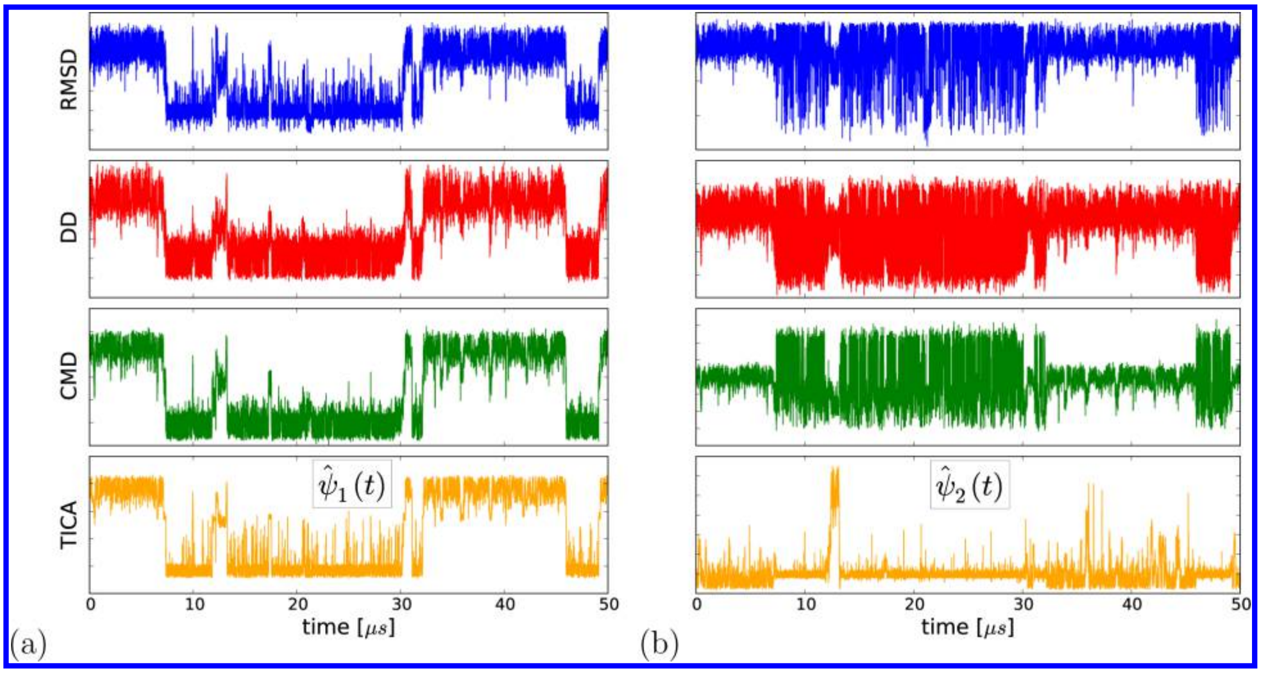


Figure 3. Time sequences for (a) $\hat{\psi}_1$ and (b) $\hat{\psi}_2$ from different metric calculations: RMSD (blue), DD (red), CMD (green), and KD (orange), for the first Fip35 trajectory.[2] The first eigenprocess is correctly captured by all four metric choices consistently, whereas the second is not.

the estimation of the first time scale, though KD and CMD perform slightly better in absolute terms.

In contrast, a different scenario arises when the second time scale t_2 is considered. On one hand, KD calculations converge nicely to $t_2 \approx 1$ μs; on the other hand, structural metric calculations show (see Figure 1) that the unphysical region is entered even before real convergence sets in, and that the process occurs on a time scale $t_2 = (60 \pm 5)$ ns, which is about 2 orders of magnitude smaller than the folding process. According to the results obtained with the structural metrics, Fip35 kinetics displays a marked time scale separation: a microsecond folding process couples with faster nanosecond eigenmotions. In addition, t_1 would converge in such a regime where higher order processes have died off already, and this poses a consistency problem. Indeed, the set of time scales should be the solution of the eigenvalue problem (eq 9) at a well-defined lag time τ, and so it is reasonable to assume that a given range of τ exists where all time scales converge; clearly, this is not the case here.

However, KD results are much more robust, all time scales converges almost simultaneously, and provide a t_2 estimate which is on the microsecond time scale (Figure 1b), mirrowing the benchmark MSM calculations (Figure 1d).

The situation is even more critical for implied time scale t_3, which is degenerate with t_2 in RMSD, DD and CMD calculations (Figure 1c), while is well-separated from t_2 by an order of magnitude, as MSM calculations (Figure 1d) show and the KD setup confirms.

All in all, the time scale analysis suggests that the convergence issues and the large time scale separation $t_1 \gg t_2 \approx t_3$ exhibited in RMSD, CMD, and DD calculations is an artifact of the structural metrics which systematically miss the second slowest process in the system dynamics. The use of purely structural metrics in LSDMap results in the systematic underestimation of the characteristic time scale associated with Fip35 processes faster than plain folding. Now we turn our attention to the optimized eigenvectors.

Optimized Eigenvectors. Discussing how the optimization affects the DCs allows us to assess to what extent LSDMap results are reliable and whether they require any *a posteriori* optimization. As a visual aid to the discussion, we show in Figure 2 the linear combination coefficient matrix $[A]_{ij} = a_{ij}$ (see eq 6) for each calculation setup. We took into account multiple fast processes, to better elucidate that some of those may be surprisingly important in shaping the optimized eigenvectors. Let us now briefly discuss the results qualitatively and their implications. We considered the equilibrium eigenprocess, and the first seven nontrivial eigenprocesses ($m = 7$). It is worth recalling, though, that the first 100 DCs were fed to the optimization procedure; we show in Figure 2 just a portion of the A matrix for the sake of clarity.

We see an optimal 1:1 correspondence between the variationally optimized equilibrium distribution $\hat{\psi}_0(x)$ and LSDMap Boltzmann $\mathrm{DC}_0(x)$ and the first nontrivial eigenvectors $\hat{\psi}_1(x)$ and $\mathrm{DC}_1(x)$, for any metric choice. However, this type of regularity is partially lost if higher orders are considered. Indeed, different patterns arise, which compromise the diagonal structure of the coefficients we would expect. Off-diagonal elements become important, and RMSD and CMD are the only setups where the diagonal structure is approximately preserved even for $\hat{\psi}_2(x)$, despite the corresponding eigenvalue being completely missed (see Figure 1). On the other hand, KD does not display a nice diagonal structure, in spite of the time scales being well approximated; indeed, $\hat{\psi}_2$, $\hat{\psi}_3$, and $\hat{\psi}_4$ all have large components along DC_3.

It is quite unexpected to see that coefficient a_{22} is exactly equal to zero in panel (d). Physically, this shows that the DC_2 direction in the subspace is practically orthogonal to the optimized $\hat{\psi}_2$ direction and indicates that LSDMap further optimization is necessary.

In conclusion, our analysis shows that LSDMap calculations equipped with structural metrics such as RMSD, DD, or CMD project onto a low-dimensional space systematically missing the second nontrivial process, whereas KD works very well, as

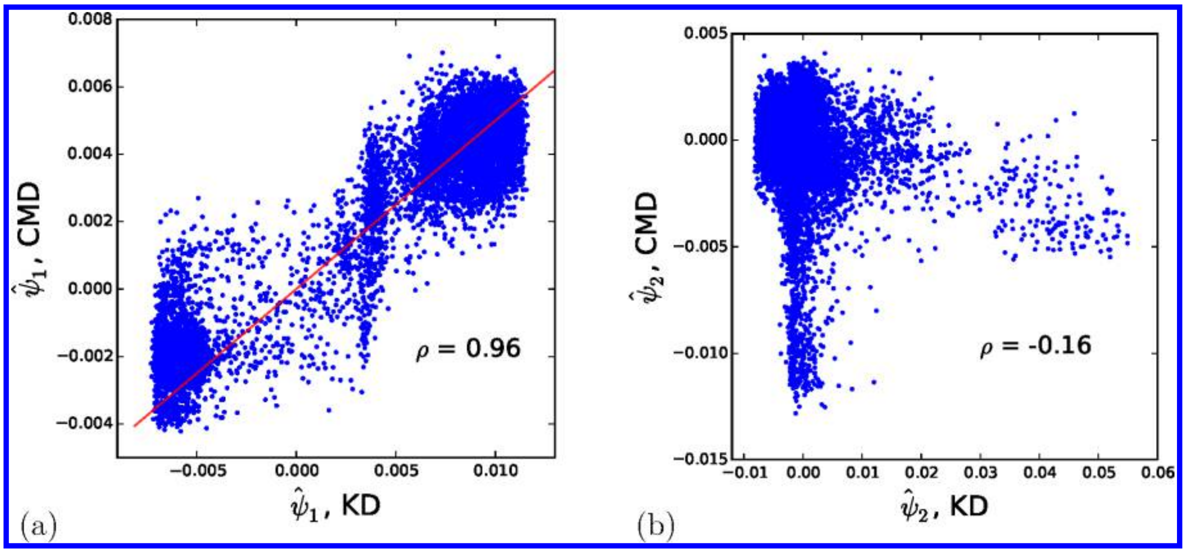


Figure 4. Statistical correlations between KD and CMD optimized eigenvectors (a) $\hat{\psi}_1$ and (b) $\hat{\psi}_2$. The Pearson correlation coefficient ρ is also shown.

comparison with MSM calculations shows. Nevertheless, bare LSDMap results can be further optimized by building optimal DC linear combinations according to the prescription in eq 9.

Eigenvector Correlations. We plotted the time evolution of the estimates $\hat{\psi}_1$,$\hat{\psi}_2$ for the four distance metrics considered and their cross-correlations, to check *a posteriori* to what extent the results correlate with one another. Figures 3 and 4 clearly show how $\hat{\psi}_1$ results from all metric calculations are consistent with one another. Exactly the same transitions between the folded and unfolded state (negative and positive $\hat{\psi}_1$ values, respectively) are found in each setup, and the Pearson coefficient ρ = 0.96 guarantees almost perfect correlation. This is consistent with previous time scale results Figure 1 showing that all distance metric setups nicely approximate the true folding time scale t_1 upon optimization.

The scenario is more complicated if we look at the time evolution of the second DC (Figure 3b). The KD highlights a much slower process than structural metrics do, and transitions shown in the top and bottom panels do not correlate at all, as represented in Figure 4b. This was expected, since structural metrics systematically miss the second eigenprocess, whereas the KD metric captures it correctly.

Free Energy and Structure Analysis. As discussed in previous sections, the KD distance metric provides superior results than popular structural metrics. We now consider the free energy plot as a function of the variationally optimized KD eigenvectors in Figure 5, and extract physical observations about the dynamics of Fip35.

There is a clear separation between the folded and unfolded state (positive and negative ψ_1 respectively), as the structural metrics also encoded. The two of them are separated by a minimum at $\psi_1 \approx 0.004$ which appears to be an intermediate (A) along the folding pathway. The presence of this state shows that folding cannot be downhill incipient, in agreement with other studies.[46−48] In addition, there is an additional basin (C) just outside the unfolded state at $\psi_1 \approx 0.007$, and a cloud of dispersed states (B) at $\psi_1 \approx 0.004$, which elongates orthogonally to the horizontal folding direction. To better investigate the features of these states, heavy atom contact maps were built on sets of structures sampled from the trajectory.

The contact map analysis reveals that the first hairpin is formed in the intermediate (A), but the second is not. This uncovers one of the two folding pathways discussed in the literature.[46,47,66,67] Previous MSM models[49] also identify this pathway in the transition network as that featuring the largest probability flux. In addition, the contact map shows that states falling in B all share a native-like structure, though some key contacts are rearranged. In particular, the first hairpin is out-of-register. Projecting the trajectory onto the $\hat{\psi}_1$−$\hat{\psi}_2$ plane reveals that B is accessed only from the unfolded state, and no directed connections exist to or from the folded state located in the bottom left of Figure 5. Thus, state B represents a misfolded state, similar to those found in previous studies.[47,48] Finally, state C exhibits a random-coil structure, where the second hairpin starts being formed (a number of key contacts are being formed). Projecting productive trajectory portions onto the free energy plot in Figure S2 reveals that state C does uncover an additional transition pathway, where hairpin 2 is formed, followed by hairpin 1 and the immediate transition to the folded state. As a reference, two sample transitions are plotted on the top of the free energy plot in Figure S2, the blue and red trajectories display hairpin 1−hairpin 2 and hairpin 2−hairpin 1 folding events, respectively.

■ CONCLUSIONS

We introduced an innovative technique which combines Locally Scaled Diffusion Map and the Method of Linear Variation to extract physically meaningful collective coordinates from a molecular dynamics data set. This formulation is particularly appealing, since it overcomes issues with both building blocks. It allows us to compute information about the time scales shaping up the kinetics, which is not possible in the original classical LSDMap algorithm.

Diffusion Maps themselves are limited by the fact that they assume an overdamped diffusion process in the metric space used, and this may not allow the eigenfunctions of the Markov backward propagator to be represented perfectly. However, the individual diffusion coordinates do not need to be perfect approximations to the real eigenfunctions to provide a good basis set for the variational procedure. The requirement is only that the first m backward propagator eigenfunctions (of

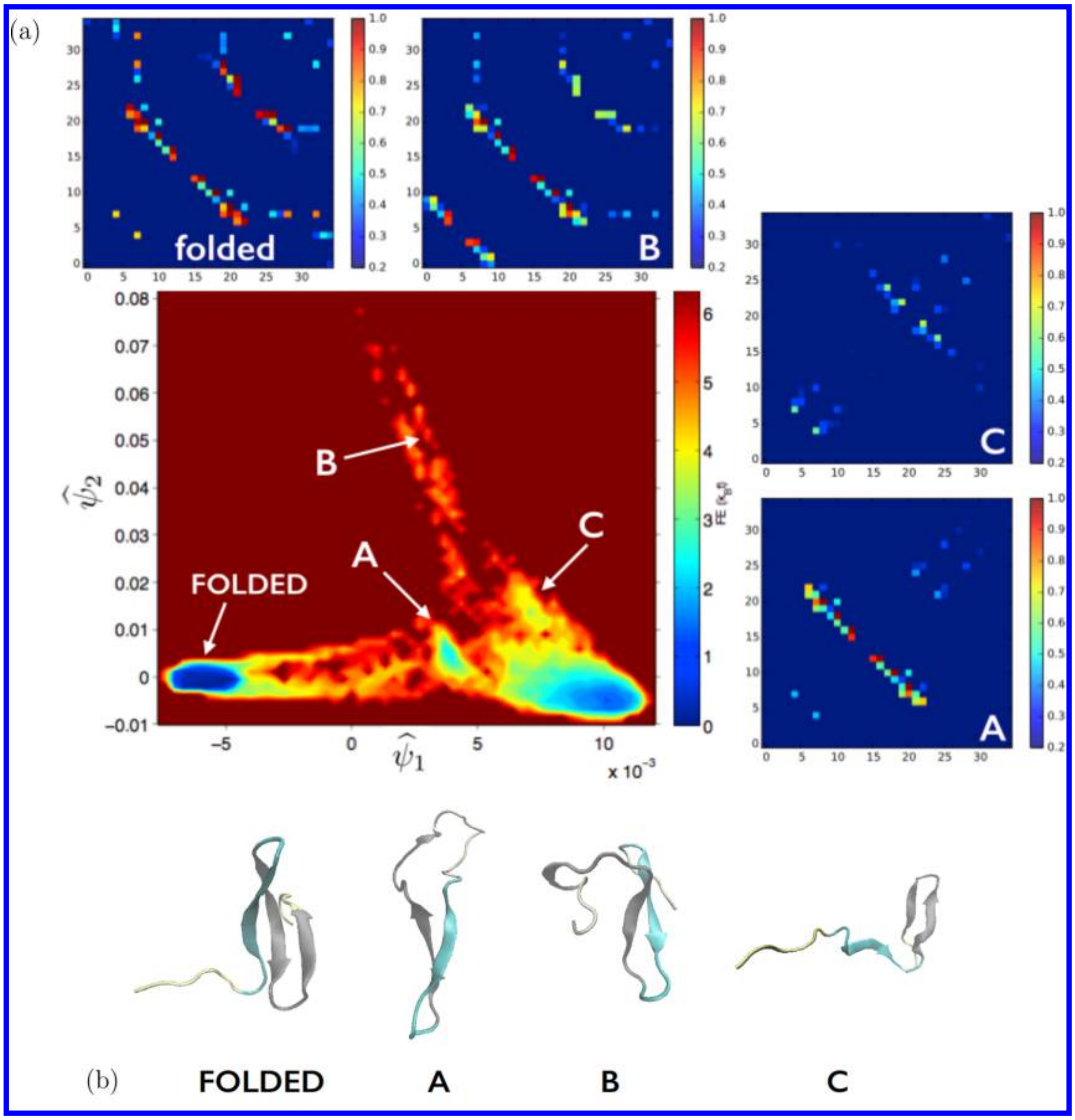


Figure 5. (a) Free energy plot as a function of the optimized KD VarDM coordinates, and sample contact maps. Folded and unfolded states are clearly separated along the $\hat{\psi}_1$ direction; additional states are labeled A, B, and C. See text for complete description. (b) Sample structures from these states.

interest) need to be well representable by the set of n DCs used. If n goes to N (number of samples), then the basis set accuracy improves, and the systematic error is reduced.

In addition, the VAC is not limited to using one type of basis function. DCs can be combined with other basis functions, such as MSM characteristic functions, Gaussians, etc. The motivation for the choice of Diffusion Maps here is that, by experience, they usually provide good approximations to the true backward propagator eigenfunctions,[16,33,36] so they are likely good basis set components in a variational approach.

We applied our VarDM protocol to study the Fip35 protein with two purposes. First of all, we were interested in assessing to what extent changing LSDMap parameters would affect the quality of the results. Hence, we adopted four complementary and independent definitions of distance metrics to formalize the concept of distance in the configuration space. Our calculations showed that standard structural metrics based on RMSD, dihedrals, or contacts are not appropriate, and that introducing an alternative distance metric that can capture the kinetic properties at least approximately, such as TICA, significantly outperforms them. Second, our approach allows us to characterize the slow processes in the folding dynamics of Fip35. The characteristic folding time scale calculated with our protocol is consistent with previously proposed calculations adopting alternative analysis algorithms. The analysis of free energy profiles supports the intermediate folding hyphothesis previously proposed in other studies.[46−48,66,67] Additionally, an orthogonal motion involving a misfolded state was also found, consistent with previous work.[47,48]

The proposed technique is general and easily implemented.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00749.

Results for different values of local scale, and two sample transition pathways (Figures S1 and S2) (PDF)

AUTHOR INFORMATION

Corresponding Authors

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

*E-mail: cecilia@rice.edu.

Funding

We acknowledge funding from the following grants: ERC starting grant 307494-pcCell and Deutsche Forschungsgemeinschaft NO 825/3-1 (F.N.), National Science Foundation CHE-1152344 and CHE-1265929, Welch foundation C-1570 (C.C.), and EPSRC Grant No. EP/K039512/1 (G.G.). Simulations were performed on the following shared resources at Rice University: DAVinCI, supported in part by the Data Analysis and Visualization Cyber Infrastructure, funded by NSF under grant OCI-0959097, and BlueBioU, supported in part by NIH award NCRR S10RR02950 and an IBM Shared University Research (SUR) Award in partnership with CISCO, Qlogic, and Adaptive Computing.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are grateful to D.E. Shaw Research for making their Fip35 trajectory available, and to Giovanni Ciccotti for enlightening discussions. We thank the members of Clementi and Noé groups for discussion and advice.

REFERENCES

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

(2) 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.

(3) Eastman, P.; Pande, V. S. Efficient Nonbonded Interactions for Molecular Dynamics on a Graphics Processing Unit. *J. Comput. Chem.* **2010**, *31*, 1268−1272.

(4) Harvey, M.; Giupponi, G.; Fabritiis, G. D. ACEMD: Accelerated Molecular Dynamics Simulations in the Microseconds Timescale. *J. Chem. Theory Comput.* **2009**, *5*, 1632−1639.

(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) 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.

(7) Swope, W. C.; Pitera, J. W.; Suits, F. Describing Protein Folding Kinetics by Molecular Dynamics Simulations: 1. Theory. *J. Phys. Chem. B* **2004**, *108*, 6571−6581.

(8) Singhal, N.; Pande, V. S. Error Analysis and Efficient Sampling in Markovian State Models for Molecular Dynamics. *J. Chem. Phys.* **2005**, *123*, 204909.

(9) 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.

(10) 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.

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

(12) 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 797; Springer: Heidelberg, 2014.

(13) McGibbon, R. T.; Ramsundar, B.; Sultan, M. M.; Kiss, G.; Pande, V. S. Understanding Protein Dynamics with L1-Regularized Reversible Hidden Markov Models. Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014.

(14) 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.

(15) 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 Deffnition of Data: Diffusion Maps. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 7426−7431.

(16) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. Determination of Reaction Coordinates Via Locally Scaled Diffusion Map. *J. Chem. Phys.* **2011**, *134*, 124116.

(17) Rao, F.; Caflisch, A. the Protein Folding Network. *J. Mol. Biol.* **2004**, *342*, 299−306.

(18) Noé, F.; Krachtus, D.; Smith, J. C.; Fischer, S. Transition Networks for the Comprehensive Characterization of Complex Conformational Change in Proteins. *J. Chem. Theory Comput.* **2006**, *2*, 840−857.

(19) Muff, S.; Caflisch, A. Kinetic Analysis of Molecular Dynamics Simulations Reveals Changes in the Denatured State and Switch of Folding Pathways upon Single-Point Mutation of a -Sheet Miniprotein. *Proteins: Struct., Funct., Genet.* **2008**, *70*, 1185−1195.

(20) Hegger, R.; Stock, G. Multidimensional Langevin Modeling of Biomolecular Dynamics. *J. Chem. Phys.* **2009**, *130*, 034106.

(21) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. Identification of Almost Invariant Aggregates in Reversibly Nearly Uncoupled Markov Chains. *Lin. Alg. Appl.* **2000**, *315*, 39−59.

(22) Bowman, G. R. Improved Coarse-Graining of Markov State Models Via Explicit Consideration of Statistical Uncertainty. *J. Chem. Phys.* **2012**, *137*, 134111.

(23) Yao, Y.; Cui, R. Z.; Bowman, G. R.; Silva, D.-A.; Sun, J.; Huang, X. Hierarchical Nyström Methods for Constructing Markov State Models for Conformational Dynamics. *J. Chem. Phys.* **2013**, *138*, 174106.

(24) Hummer, G.; Szabo, A. Optimal Dimensionality Reduction of Multistate Kinetic and Markov-State Models. *J. Phys. Chem. B* **2015**, *119*, 9029−9037.

(25) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; 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.

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

(27) Noé, F.; Doose, S.; Daidone, I.; Löllmann, M.; Chodera, J. D.; Sauer, M.; Smith, J. C. Dynamical Fingerprints for Probing Individual Relaxation Processes in Biomolecular Dynamics with Simulations and Kinetic Experiments. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 4822−4827.

(28) E, W.; Vanden-Eijnden, E. Towards a Theory of Transition Paths. *J. Stat. Phys.* **2006**, *123*, 503−523.

(29) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. Transition Path Theory for Markov Jump Processes. *Multiscale Model. Simul.* **2009**, *7*, 1192−1219.

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

(31) Held, M.; Metzner, P.; Prinz, J.-H.; Noé, F. Mechanisms of Protein-Ligand Association and Its Modulation by Protein Mutations. *Biophys. J.* **2011**, *100*, 701−710.

(32) Silva, D.-A.; Bowman, G. R.; Sosa-Peinado, A.; Huang, X. A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the LAO Protein. *PLoS Comput. Biol.* **2011**, *7*, e1002054.

(33) Zheng, W.; Qi, B.; Rohrdanz, M. A.; Caflisch, A.; Dinner, A. R.; Clementi, C. Delineation of Folding Pathways of a β-Sheet Miniprotein. *J. Phys. Chem. B* **2011**, *115*, 13065−13074.

(34) Kube, S.; Weber, M. a Coarse Graining Method for the Identification of Transition Rates Between Molecular Conformations. *J. Chem. Phys.* **2007**, *126*, 024103.

(35) Keller, B.; Prinz, J.-H.; Noé, F. Markov Models and Dynamical Fingerprints: Unraveling the Complexity of Molecular Kinetics. *Chem. Phys.* **2012**, *396*, 92−107.

(36) Zheng, W.; Rohrdanz, M. A.; Maggioni, M.; Clementi, C. Polymer Reversal Rate Calculated Via Locally Scaled Diffusion Map. *J. Chem. Phys.* **2011**, *134*, 144109.

(37) Prinz, J.-H.; Chodera, J. D.; Noé, F. Spectral Rate Theory for Two-State Kinetics. *Phys. Rev. X* **2014**, *4*, 011020.

(38) Zheng, W.; Rohrdanz, M. A.; Clementi, C. Rapid Exploration of Configuration Space with Diffusion-Map-Directed Molecular Dynamics. *J. Phys. Chem. B* **2013**, *117*, 12769−76.

(39) Preto, J.; Clementi, C. Fast Recovery of Free Energy Landscapes Via Diffusion-Map-Directed Molecular Dynamics. *Phys. Chem. Chem. Phys.* **2014**, *16*, 19181−19191.

(40) Sarich, M.; Noé, F.; Schütte, C. On the Approximation Quality of Markov State Models. *Multiscale Model. Simul.* **2010**, *8*, 1154−1177.

(41) Molgedey, L.; Schuster, H. G. Separation of a Mixture of Independent Signals Using Time Delayed Correlations. *Phys. Rev. Lett.* **1994**, *72*, 3634−3637.

(42) 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.

(43) 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.

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

(45) Noé, F.; Nüske, F. A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems. *Multiscale Model. Simul.* **2013**, *11*, 635−655.

(46) Krivov, S. V. The Free Energy Landscape Analysis of Protein (FIP35) Folding Dynamics. *J. Phys. Chem. B* **2011**, *115*, 12315−12324.

(47) Berezovska, G.; Prada-Gracia, D.; Rao, F. Consensus for the Fip35 Folding Mechanism? *J. Chem. Phys.* **2013**, *139*, 035102.

(48) Mori, T.; Saito, S. Dynamic Heterogeneity in the Folding/unfolding Transitions of FiP35. *J. Chem. Phys.* **2015**, *142*, 135101.

(49) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. a.; Pande, V. S. Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. *J. Am. Chem. Soc.* **2011**, *133*, 18413−18419.

(50) McGibbon, R. T.; Pande, V. S. Learning Kinetic Distance Metrics for Markov State Models of Protein Conformational Dynamics. *J. Chem. Theory Comput.* **2013**, *9*, 2900−2906.

(51) Gardiner, C. *Stochastic Methods: A Handbook for the Natural and Social Sciences*; Springer Series in Synergetics; Springer: Berlin/Heidelberg, 2009.

(52) Zhuang, W.; Cui, R. Z.; Silva, D.-A.; Huang, X. Simulating the T-Jump-Triggered Unfolding Dynamics of Trpzip2 Peptide and Its Time-Resolved IR and Two-Dimensional IR Signals Using the Markov State Model Approach. *J. Phys. Chem. B* **2011**, *115*, 5415−5424.

(53) Lindner, B.; Yi, Z.; Prinz, J.-H.; Smith, J. C.; Noé, F. Dynamic Neutron Scattering from Conformational Dynamics. I. Theory and Markov Models. *J. Chem. Phys.* **2013**, *139*, 175101.

(54) Deuflhard, P.; Weber, M. *Linear Algebra Appl.* **2005**, *398*, 161−184.

(55) Weber, M. Meshless Methods in Conformation Dynamics. Ph.D. thesis, Freie Universität Berlin, 2006.

(56) Coifman, R. R.; Kevrekidis, I. G.; Lafon, S.; Maggioni, M.; Nadler, B. Diffusion Maps, Reduction Coordinates, and Low Dimensional Representation of Stochastic Processes. *Multiscale Model. Simul.* **2008**, *7*, 842−864.

(57) Cossio, P.; Laio, A.; Pietrucci, F. Which Similarity Measure Is Better for Analyzing Protein Structures in a Molecular Dynamics Trajectory? *Phys. Chem. Chem. Phys.* **2011**, *13*, 10421−10425.

(58) Zhou, T.; Caflisch, A. Distribution of Reciprocal of Interatomic Distances: A Fast Structural Metric. *J. Chem. Theory Comput.* **2012**, *8*, 2930−2937.

(59) Jäger, M.; Nguyen, H.; Crane, J. C.; Kelly, J. W.; Gruebele, M. The Folding Mechanism of a Beta-Sheet: The WW Domain. *J. Mol. Biol.* **2001**, *311*, 373−393.

(60) Caprara, A.; Carr, R.; Istrail, S.; Lancia, G.; Walenz, B. 1001 Optimal PDB Structure Alignments: Integer Programming Methods for Finding the Maximum Contact Map Overlap. *J. Comput. Biol.* **2004**, *11*, 27−52.

(61) Noé, F.; Clementi, C. Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation. *J. Chem. Theory Comput.* **2015**, *11*, 5002−5011.

(62) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Perez-Hernandez, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, DOI: 10.1021/acs.jctc.5b00743.

(63) Freddolino, P. L.; Park, S.; Roux, B.; Schulten, K. Force Field Bias in Protein Folding Simulations. *Biophys. J.* **2009**, *96*, 3772−3780.

(64) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. Ten-Microsecond Molecular Dynamics Simulation of a Fast-Folding WW Domain. *Biophys. J.* **2008**, *94*, L75−L77.

(65) Ensign, D. L.; Pande, V. S. The Fip35 WW Domain Folds with Structural and Mechanistic Heterogeneity in Molecular Dynamics Simulations. *Biophys. J.* **2009**, *96*, L53−L55.

(66) a Beccara, S.; Škrbić, T.; Covino, R.; Faccioli, P. Dominant Folding Pathways of a WW Domain. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 2330−2335.

(67) a Beccara, S.; Fant, L.; Faccioli, P. Variational Scheme to Compute Protein Reaction Pathways Using Atomistic Force Fields with Explicit Solvent. *Phys. Rev. Lett.* **2015**, *114*, 098103.

(68) Faccioli, P.; Sega, M.; Pederiva, F.; Orland, H. Dominant Pathways in Protein Folding. *Phys. Rev. Lett.* **2006**, *97*, 108101.