# Diffusion maps, clustering and fuzzy Markov modeling in peptide folding transitions

Lilia V. Nedialkova,[1] Miguel A. Amat,[1,a)] Ioannis G. Kevrekidis,[2,b)] and Gerhard Hummer[3,b)]

[1]*Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey 08544, USA*
[2]*Department of Chemical and Biological Engineering and Program in Applied and Computational Mathematics, Princeton University, Princeton, New Jersey 08544, USA*
[3]*Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Max-von-Laue-Str. 3, 60438 Frankfurt am Main, Germany*

(Received 14 June 2014; accepted 4 August 2014; published online 15 September 2014)

Using the helix-coil transitions of alanine pentapeptide as an illustrative example, we demonstrate the use of diffusion maps in the analysis of molecular dynamics simulation trajectories. Diffusion maps and other nonlinear data-mining techniques provide powerful tools to visualize the distribution of structures in conformation space. The resulting low-dimensional representations help in partitioning conformation space, and in constructing Markov state models that capture the conformational dynamics. In an initial step, we use diffusion maps to reduce the dimensionality of the conformational dynamics of Ala5. The resulting pretreated data are then used in a clustering step. The identified clusters show excellent overlap with clusters obtained previously by using the backbone dihedral angles as input, with small—but nontrivial—differences reflecting torsional degrees of freedom ignored in the earlier approach. We then construct a Markov state model describing the conformational dynamics in terms of a discrete-time random walk between the clusters. We show that by combining fuzzy C-means clustering with a transition-based assignment of states, we can construct robust Markov state models. This state-assignment procedure suppresses short-time memory effects that result from the non-Markovianity of the dynamics projected onto the space of clusters. In a comparison with previous work, we demonstrate how manifold learning techniques may complement and enhance informed intuition commonly used to construct reduced descriptions of the dynamics in molecular conformation space.  [http://dx.doi.org/10.1063/1.4893963]

## I. INTRODUCTION

In theoretical and computational descriptions of the motion of macromolecular systems, the high dimensionality of the underlying conformation space entails a major challenge with respect to the sampling of the essential dynamics, and to the analysis of the resulting trajectories in terms of relevant degrees of freedom. Earlier studies using simple dimensionality reduction techniques, in particular variations of principal component analysis,[1–3] showed that projections onto low-dimensional manifolds can capture even the complex motions of a protein with a manageably small number of degrees of freedom. However, the energy landscape of (bio)polymers and the corresponding structure of populated regions in conformation space are often quite complex. As a result, we expect that linear projection methods will not necessarily provide us with the optimal (lowest-dimensional) representations that reliably separate conformational basins in a clear and useful form.[4,5] A long-standing goal has thus been to develop *nonlinear* projection/embedding methods for the parsimonious description of macromolecular conformation space.

Low-dimensional representations of a high-dimensional conformation space constitute powerful visualization and analysis tools, and can be used to increase the efficiency of sampling in atomistic simulations. The competing requirements to include large numbers of degrees of freedom, to resolve even the finest time scales of covalent bond motions and atomic collisions, and to overcome high enthalpic barriers between large conformational basins stand in the way of proper sampling with direct molecular simulations. With good low-dimensional representations the sampling problem can be addressed using a host of powerful biasing techniques ranging from umbrella sampling[6] to metadynamics[7] and beyond. In poor representations we expect large hysteresis effects because distant, slowly equilibrating regions of conformation space are erroneously not separated in projection.[8] By contrast, in a good representation the projection provides us with suitable "reaction coordinates"[9] along which the motion has minimal memory effects;[10] even driven motion in such coordinates is expected to stay close to equilibrium, thus minimizing dissipation effects because of coupling to unrelaxed motions transverse to the chosen reaction coordinate. Examples of work motivated by these ideas and making use of biased sampling in reduced coordinates include accelerating

a)Currently at Department of Biochemistry and Molecular Pharmacology, University of Massachusetts, Worcester, Massachusetts 01655, USA.
b)Authors to whom correspondence should be addressed. Electronic addresses: yannis@princeton.edu and gerhard.hummer@biophys.mpg.de.

stochastic simulations[11,12] and systematically identifying low dimensional parametrizations of a biomolecule's free energy surface.[13]

An alternative approach to obtaining reduced sets of good global reduction coordinates is to group states locally (as cells in conformation space) and to summarize the dynamics in terms of transitions between such groups. This approach has a long tradition in chemistry, and, in particular, in the concise description of population changes in chemical reaction kinetics. Such projections onto discrete sets of states naturally lead to descriptions of the dynamics in terms of Markov state models[14] or coarse master equations[15,16] with discrete time-stepping or continuous dynamics, respectively. Coarse master equations and Markov state models have attracted much attention because they can be constructed directly from molecular dynamics simulations,[15–21] with the aim to capture the dynamics of most interest, occurring over long time scales. The success of such models, however, strongly depends on the ability to decompose conformation space into a set of meaningful metastable states, associated with low free energy (meta)basins.

In this work, we illustrate the construction of a link between the global and local approaches to dimensionality reduction. Following an initial, data-mining based dimensionality reduction step, we identify possible metastable states within the new, reduced space.

The first step in this strategy relates, conceptually, to data pretreatment ideas used in time series prediction. In the context of a molecular simulation, we think of a trajectory as a set of $N$ points $x_i$ in the $3S$-dimensional conformation space, where $S$ denotes the number of atoms in the molecule. The assumption is that there exists an underlying $l$-dimensional manifold ($l \ll 3S$) close to which the physically important system dynamics takes place. The initial goal is to embed the high-dimensional simulation data in an $l$-dimensional space in a way that consistently groups related conformations together, and separates different ones. Beyond serving as a visualization aid, a successful embedding should also provide a basis for the identification of metastable states for the system of interest. The embedding procedure we use aims to avoid common problems resulting from molecular distance metrics, such as the root-mean-square deviation (RMSD), that work well for short distances between highly similar structures, but lack discriminating power for larger distances. This issue may also be mitigated by initially dividing up the conformation space very finely and subsequently lumping states together.[19,22] In the limit of a very fine partitioning, the Markov state model approaches a diffusion model.[23]

A number of nonlinear dimensionality reduction techniques have been developed in recent years.[24–26] In this paper, we use the diffusion map (DM) approach,[27–31] which has previously been applied to molecular simulation data.[30,32–34] The central idea of DM is to take high-dimensional data that lie on, or close to, a nonlinear low-dimensional manifold and embed them in a low-dimensional space in a way that preserves the intrinsic geometry of the data. In a second step we then construct a Markov chain model formulated in this reduced-dimensional embedding space.

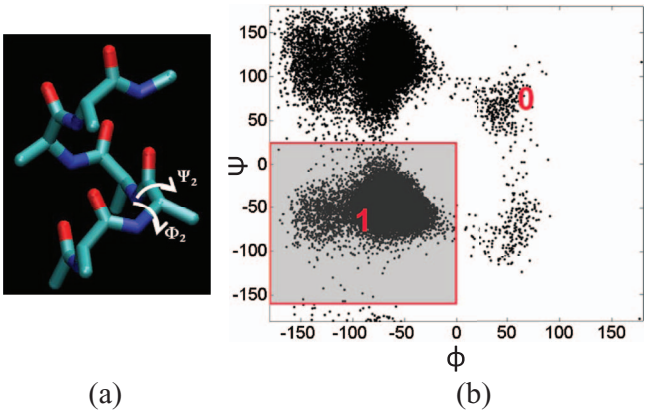



FIG. 1. (a) Schematic representation of Ala5. Individual atoms are marked as follows: red - oxygen; blue - nitrogen; green - carbon; (hydrogens omitted). The dihedral angles of the second alanine residue $\phi_2$ and $\psi_2$ are indicated. (b) Ramachandran map for a single Ala5 residue. The gray shaded area indicates the helical region, and its complement indicates the coil region (1 and 0, respectively, in BH nomenclature).

To illustrate and test this two-step approach we studied the conformational changes of the small molecule alanine pentapeptide (Ala5) (see Fig. 1(a)) in explicit water. We deliberately chose a force field[35] that over-emphasizes the helix content of Ala5 to obtain a minimal model of the helix-coil transition. In this system, the folding transitions between extended coil structures and the folded helix occur on time scales short enough to be accessible to unbiased molecular dynamics simulations. The system thus provides a useful benchmark sharing much of the complexity of larger-scale molecular simulation problems. We also chose this system because it has been studied in detail by Buchete and Hummer (BH),[16] who used informed intuition to partition conformation space and to formulate a master-equation model that describes its dynamics.

The paper is structured as follows: In Sec. II, we outline the DM technique. The details of the molecular dynamics simulation are given in Sec. III, which also summarizes the master-equation model of BH as a basis for comparison. Section IV presents the results of applying the methodology to Ala5; the paper concludes with a brief summary and discussion of these results and of important features of the approach.

## II. THEORY

### A. Diffusion map

Applied to molecular systems, DM[27–32] allows us to perform nonlinear dimensionality reduction in the space of molecular conformations, based on (large, hopefully representative) ensembles of such conformations from simulation data. Consider such an ensemble of $N$ sampled conformations. We construct the $N \times N$ matrix $W$ whose elements are given by

$$w_{ij} = k(x_i, x_j) = \exp\left(-\frac{d^2(x_i, x_j)}{\epsilon}\right), \qquad (1)$$

where $k$ is a kernel function (here the diffusion kernel), $d$ is a pairwise similarity measure (a distance metric) and $\epsilon$ is the kernel width (a system-dependent parameter whose choice

is discussed in Appendix A). There are different possible candidates for $d$—the Euclidean distance being the obvious one—and the analysis would be better served if one were to select a metric which reflected well the underlying system dynamics. In this sense, a more informed choice is expected to produce more revealing results. The matrix $W$ can be interpreted as the adjacency matrix of a graph, each of whose nodes represents one of the data points. We then construct the row-stochastic matrix

$$A = D^{-1}W, \tag{2}$$

where $D$ is a diagonal matrix with elements

$$d_{ii} = \sum_{j=1}^{N} w_{ij}. \tag{3}$$

This in effect defines a random walk on the data graph, represented by $W$, and the matrix $A$ is the corresponding Markov transition matrix. (Note that for the DM we have a row-stochastic matrix, to be consistent with the original work; later, in the construction of Markov state models, we use a column-stochastic matrix, again for consistency with the relevant literature). The matrix $A$ has right eigenvectors $v_0, \ldots, v_{N-1}$, left eigenvectors $u_0, \ldots, u_{N-1}$ and eigenvalues $\lambda_0, \ldots, \lambda_{N-1}$, where $\lambda_0 = 1 \geq |\lambda_1| \geq \ldots \geq |\lambda_{N-1}|$ and $v_0 = (1, \ldots, 1)^T$.

Within the DM formulation, the probability of the random walk being at the point $x_j$ after $t$ steps given a starting point of $x_i$ is given by

$$p(t, x_j|x_i) = \sum_{k=0}^{N-1} v_k(i)\lambda_k^t u_k(j). \tag{4}$$

On the basis of this definition and given the context of a random walk on the data graph, it makes sense to think about the similarity between two points in terms of their "dynamic proximity"—the "ease" of transitioning from one conformation to the other. This motivates the definition of the *diffusion distance* between points $x_i$ and $x_j$ as

$$D_t^2(i, j) = \sum_{k=0}^{N-1} (p(t, x_k|x_i) - p(t, x_k|x_j))^2 \frac{1}{u_0(x_k)}. \tag{5}$$

In a "dynamically meaningful" lower-dimensional representation of the data, the Euclidean distance in the new embedding space should correspond to the relative "ease of transitioning" between states. The mapping from the original space into the new *DM space* is

$$x_i \to \left(\lambda_1^t v_1(i), \lambda_2^t v_2(i), \ldots, \lambda_{N-1}^t v_{N-1}(i)\right). \tag{6}$$

It can be readily shown[29] that the *diffusion distance* between two points is indeed equivalent to their Euclidean distance in DM space,

$$D_t^2(i, j) = \sum_{k=0}^{N-1} \lambda_k^{2t} (v_k(i) - v_k(j))^2. \tag{7}$$

In many practical applications one observes a gap in the eigenvalue spectrum of the matrix $A$, and the diffusion distance may be well approximated by using just the $l \ll N$ eigenvectors corresponding to the leading $l$ eigenvalues (not counting $\lambda_0$). This then achieves the dimensionality reduction, and the $l$-dimensional DM embedding at $t = 0$ is defined as

$$x_i \to (v_1(i), v_2(i), \ldots, v_l(i)). \tag{8}$$

Depending on the time-horizon over which transitions are examined, one may also use the "time-t embedding" (Eq. (6)), e.g., see Ref. 36.

## III. MOLECULAR DYNAMICS

A molecular dynamics simulation of Ala5 in explicit solvent was performed using the GROMACS 4.0.7 molecular simulation package[37] with the AMBER-GSS force field[35] ported to GROMACS.[38,39] The simulation box contained 1050 TIP3P water molecules.[40] We used periodic boundary conditions and the particle mesh Ewald method[41] for long-range electrostatic interactions. The simulation was performed in the NPT ensemble, using a stochastic dynamics integrator at 350 K. This temperature is close to the mid-point of the helix-coil equilibrium, with roughly 65% of the population in the helical state. The pressure was maintained isotropically at 1 bar using the Parrinello-Rahman barostat.[42] Bond lengths involving hydrogen atoms were constrained.[43] The simulation used a time step of 2 fs. The production portion of the simulation used in this work lasted 100 ns, with 50 000 structures saved at 2 ps intervals.

### A. Reference master equation

To compare our results to the analysis of BH, we first introduce relevant notation to describe the Ala5 conformation space. To write their master equation model, BH initially partitioned conformation space into 32 ($2^5$) substates based on a "helix" or "coil" designation for each of the molecule's 5 residues. In one approach, this designation is based on the instantaneous values of the $\phi$ [C–N–$C_\alpha$–C] and $\psi$ [N–$C_\alpha$–C–N] backbone dihedral angles—see Fig. 1(b). Each substate is assigned a 5-digit binary label with 1 and 0 indicating residues in the helical and coil regions of the Ramachandran map (Fig. 1(b)), respectively, and residues from the N to the C terminus ordered left to right. As is evident from the figure, the coil designation for any of the five dihedral angle pairs includes a large section of the Ramachandran map, and thus potentially very different structures. BH later coarse-grain the system further, arriving at a four-state model which recovers the slowest relaxation processes in the system. These four states include two folded states (labeled F1 and F2) and two unfolded ones (labeled U1 and U2). As a first goal, we explore the extent to which the DM embedding, without explicit exploitation of the known peptide stereochemistry, recovers these fine and coarse partitionings of the Ala5 conformation space.

## IV. RESULTS

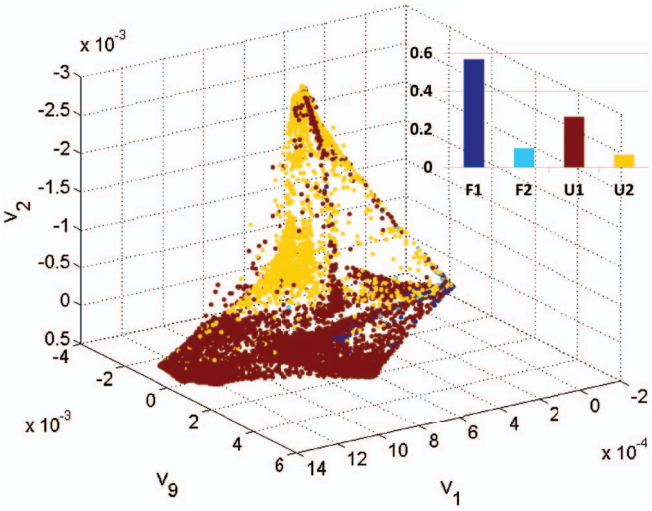
### A. Dimensionality reduction

We use $N = 50\,000$ snapshots from the molecular dynamics trajectory for our analysis. This choice of subsampling the MD simulation attempted to balance the computational

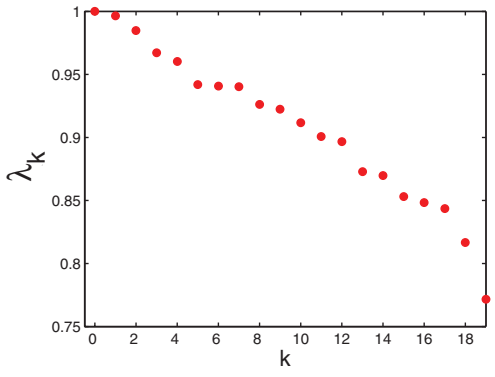
burden of the eigendecomposition of the diffusion map matrix and the need for representative phase space sampling. Each data point is represented by a vector containing the coordinates of all atoms in Ala5 except the hydrogens ($S = 30$). The omission of the hydrogen atoms allows us to discard from the analysis fast modes of the molecule's dynamics. To establish structural similarities between various conformations we pairwise align the structures using the algorithm by Kabsch.[44] This alignment ensures that we remove trivial translational and rotational differences between individual configurations prior to computing a meaningful distance between them. The RMSD was the distance metric $d(x_i, x_j)$ used in our diffusion kernel and the value of the parameter $\epsilon$ was set to 0.1667 Å$^2$. (See Appendix A for details on selecting a suitable value for $\epsilon$.)

In a future publication we plan to detail the outcome when using different metrics, and in particular, a *dihedral angle-based* distance metric, which more directly captures the conformation dynamics of Ala5.

The eigendecomposition of the DM matrix $A$ constructed from the data was performed using the ARPACK[45] numerical library. The eigenvalue spectrum (Fig. 2(b)) exhibits a gap after the 13th eigenvalue ($\lambda_{12}$ in our numbering). A quick computational test (see Ref. 46) shows that each of these eigenvectors corresponds to a different direction on the low-dimensional manifold (no eigenvector appears to be a "higher harmonic" of a previous one). This suggests that the eigenvectors corresponding to the 12 leading eigenvalues (excluding $\lambda_0$) would be sufficient to achieve an adequate global embedding of the trajectory. In practice, it has been observed that different eigenvectors resolve different portions of the data (see also Ref. 13).




(a)




(b)

FIG. 2. (a) Diffusion map embedding of the Ala5 MD simulation data in the space spanned by DM vectors 1, 2, and 9. The inset shows the relative frequency of the four coarse BH states and the scheme for coloring the points. (b) Eigenvalue spectrum of the DM Markov transition matrix $A$.

Figure 2(a) shows the data mapped onto three DM coordinates (the first, second, and ninth eigendirections). The points are colored according to the previously mentioned four coarse BH states (F1, F2, U1, U2). The choice of eigenvectors for this and later visualizations was motivated by an effort to find projections that most clearly separate (selected) BH states/substates in a 3d (or 2d) representation. This embedding demonstrates a certain degree of separation (although not complete) between the coarse states. Conformations in the two unfolded states (yellow and red points, respectively) tend to occupy separate sections of the approximately tetrahedral arrangement of points. The lower right-hand vertex of the tetrahedron is predominantly made up of the two folded states (dark and light blue). Loosely speaking $v_1$ appears to organize points according to a folded/unfolded criterion, $v_2$ distinguishes the points in state U2 and $v_9$ does the same for the points in state U1. Zoom-ins reveal further structure. To explore the clusters naturally arising in the DM embedding, and to quantify the extent to which they relate to the BH substates, we next perform a cluster analysis in DM space.

We have already pointed out that the Euclidean distance between points *in DM space* corresponds to the *diffusion distance* between them. Clustering based on Euclidean distances *in DM space* is, therefore, clustering using the diffusion distance in the original space. If the diffusion distance is representative of the dynamic proximity of points (molecular conformations in our case), one might expect clustering in DM space to be more informative than clustering in physical space.

## B. Clustering

Clustering in the reduced 12-dimensional space was performed using the k-means algorithm[47] as implemented in MATLAB$^{TM}$. The number of clusters, $k$, is an input to the algorithm. We used a range of $k$ values and for each $k$, 10 random initializations were performed. After clustering was completed, the result with the lowest value of the objective function was selected as the clustering result for the given $k$. To choose an optimal $k$, we rated the clusterings according to the degree of successful assignment of the points. This was assessed by using the silhouette score,[48] which is a measure of how well each point fits in the cluster to which it has been assigned. The silhouette score for a point $i$ is given by

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}, \tag{9}$$

where $a(i)$ represents the average distance between point $i$ and all points in its cluster and $b(i)$ is given by the average distance

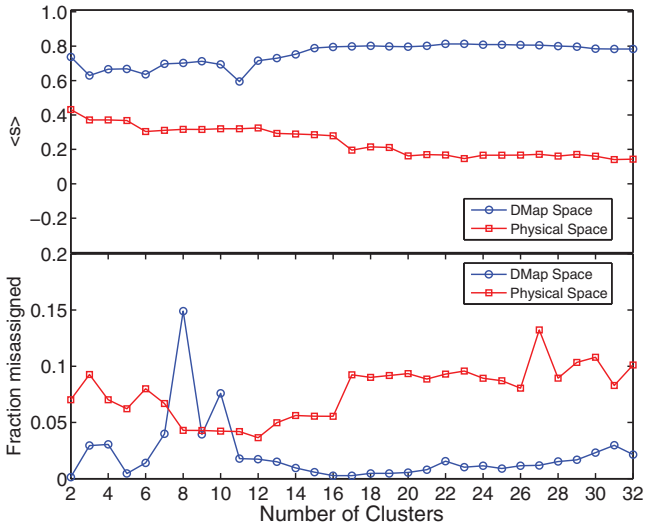



FIG. 3. Average silhouette scores (above) and fraction of misassigned points (below) for the Ala5 data clustering in DM space (circles) and physical space (squares).

between point $i$ and all points in the closest neighboring cluster. Any point with $s(i) \leq 0$ is considered misassigned. We assess the clustering results for different cluster numbers on the basis of the average silhouette score $\langle s \rangle$ over all points as well as the fraction of misassigned points (see Fig. 3), and select to focus on the apparently optimal 16-cluster representation.

### C. DM visualization of clusters

Figure 4 visually compares our DM clusters with the BH substates and states. The top middle panel (Fig. 4(b)) shows (a zoom-in of) the separation of the two folded states (dark and light blue for F1 and F2, respectively). When we look at the same section of the data, but this time in terms of our clusters (Fig. 4(c)), we find two clusters corresponding to the two folded states. The bottom row (Figs. 4(d)–4(f)) shows an analogous result, but for the points in state U1. Comparing the middle and right bottom panels of the figure, we see that clusters naturally found in the data correspond overall consistently with the U1 substates.

The heat map in Fig. 5 summarizes the degree of overlap between the identified DM clusters and the BH substates, which we will now proceed to examine in more detail. Note that cluster indices (as they result from the k-means algorithm) have been systematically reordered for consistent comparisons, based on the conformations they contain, so that the heat map acquires a diagonally dominant appearance.

We start with the following observations: (a) all BH substates in state F1 are grouped into a single cluster; (b) pairs of BH substates which differ just in their 5th residues tend to occupy the same cluster; and (c) the correspondence between DM clusters and BH substates is not perfect (the heat map is not perfectly diagonal).

A specific example of this non-perfect correspondence is provided by a close look at state F2, which predominantly is contained in cluster 12, but also forms a subpopulation of

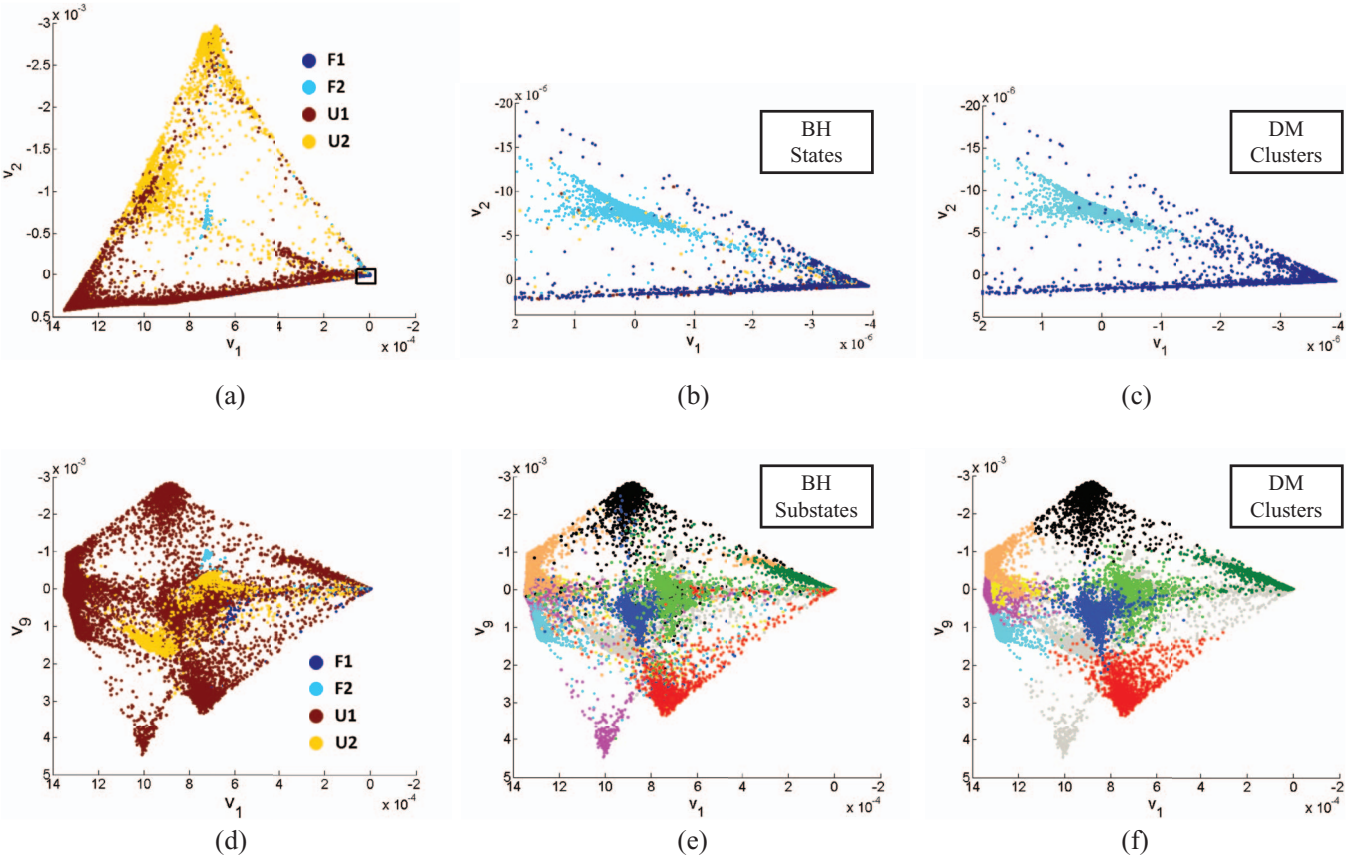



FIG. 4. Visual comparison between BH states/substates and DM clusters. (a), (d) Data colored according to the BH coarse states (the small region outlined in black in (a) is the area shown enlarged in (b) and (c)); (b), (e) data colored according to BH substates; (c), (f) data colored according to DM clusters.

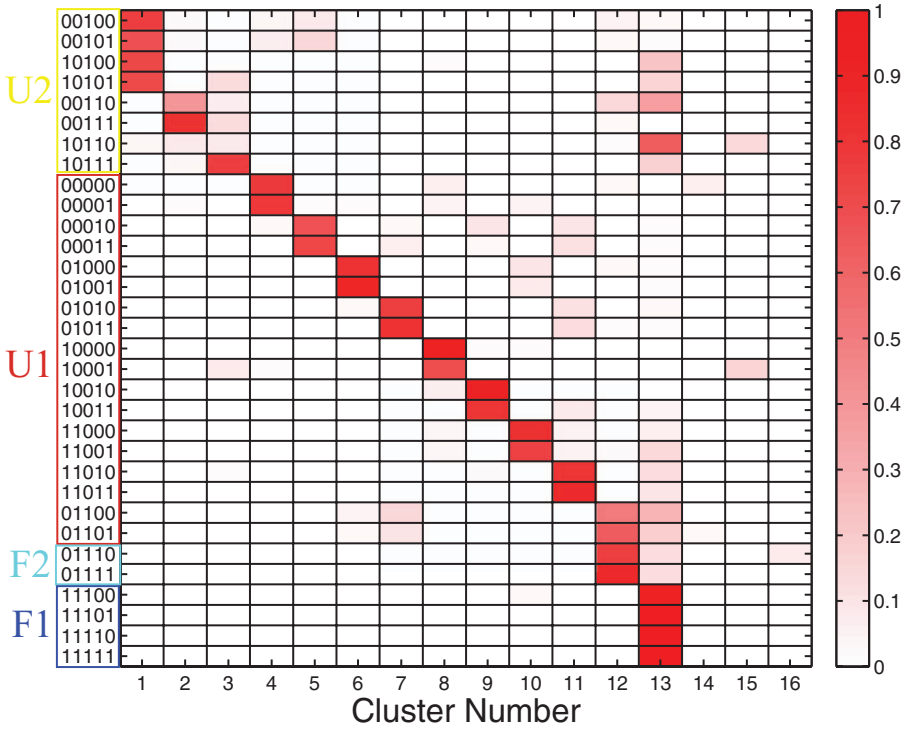



FIG. 5. Heat map showing the level of overlap between BH substates (vertical axis) and DM clusters (horizontal axis; relabeled for clarity). The 1s/0s in the BH labels represent helix/coil configurations of the Ala residues from the N to C terminus ordered left to right. The normalization of the heat map was performed with respect to the rows (BH substates).

cluster 13. To visualize this, Fig. 6(a) highlights in color only the points in state F2, while Fig. 6(b) colors these F2 points according to the DM cluster they occupy. Note that all conformations in state F2 have a first residue in the coil (0) state—a very broad state in the Ramachandran map as already discussed. Remarkably, coloring the conformations in DM space according to the value of the dihedral angle $\psi_1$ of residue 1 reveals a new separation (Fig. 6(c)). Clearly the split of F2 points into different DM clusters is strongly correlated with the value of the dihedral angle $\psi_1$.

Similar observations arise when focusing on other sections of the data. A second example involves the U1 states 01010/01011. According to the heat map in Fig. 5, points from these BH substates are predominantly found in DM cluster 7, with some also present in cluster 11. Figure 6(d) shows just the points in these two BH substates, while Fig. 6(e) shows the same points colored according to their DM assigned cluster. Figure 6(f) and the inset demonstrate, just as in the previous example, that the DM cluster assignment is correlated with a physical variable—the value of the dihedral angle $\psi_1$. Note that for these two BH substates, residue 3 also has the coil (0) designation, but the conformations show limited variation in the value of $\psi_3$ (not shown).

The fine structure in dihedral angle space is, therefore, one of the factors responsible for the not-entirely-perfect correspondence between DM clusters and BH (sub)states. This fine structure was ignored in the BH partitioning of conformation space, where the focus was primarily on the helix-coil transition; it is therefore not surprising that discrepancies arise at this level, and reassuring that they can be clearly rationalized.

Another observation from the heat map in Fig. 5 is that clusters appear to spontaneously "pair up" conformations that differ in the coil-helix designation of *only the fifth* residue. Interestingly, this C-terminal residue was found by BH to relax rapidly (compared to the others); so this grouping into 16 rather than 32 clusters, as obtained entirely from structural data, is consistent with the physical dynamics. In summary, DM analysis leads to a lower-dimensional representation of the data, which clusters points (molecular conformations) in a manner reasonably consistent with the BH description. Moreover, this embedding contains an added level of detail: we see that we are generally able to meaningfully further resolve structure within the Ramachandran map underlying the coil designation.

### D. Clustering in physical space

For comparison purposes, clustering in the original high-dimensional physical space of Cartesian coordinates was also performed by using the original RMSD metric instead of the diffusion distance. Figure 3 shows a comparison of the silhouette scores and the fraction of misassigned points for clustering in *physical space* as well as in DM space. Overall, physical space clustering results in significantly lower silhouette scores and systematically higher fractions of misassigned points, which together suggest less well defined clusters compared to the DM case. We looked in greater detail at the 12- and 16-cluster results (the former, because it shows the

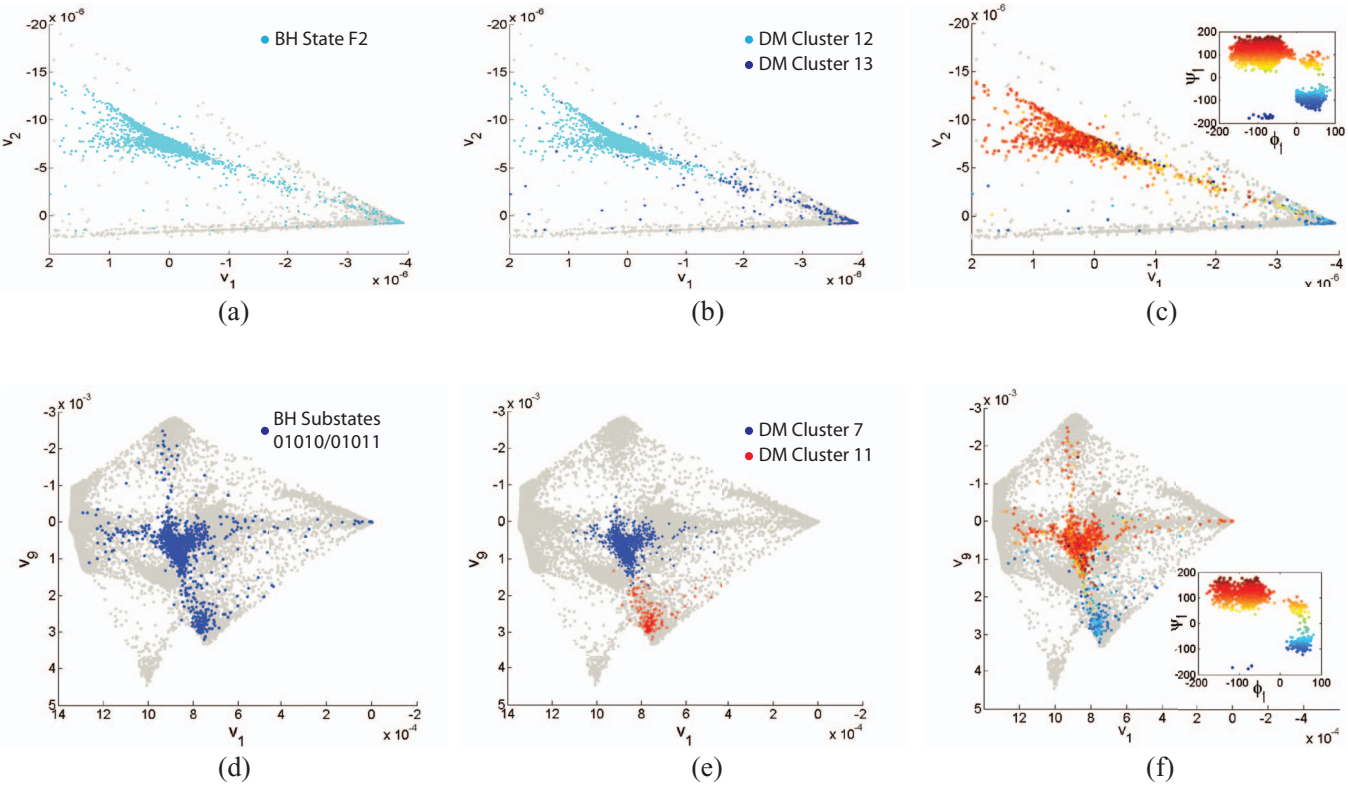



FIG. 6. Illustration and rationalization of the discrepancy between BH states and DM clusters: (a) Points from state F2; (b) points from state F2 colored according to the cluster they are assigned to; (c) points from state F2 colored according to the value of their angle $\psi_1$ (inset: Ramachandran map for residue 1 of points in state F2, color is according to $\psi_1$); (d) points from substates 01010/01011; (e) points from substates 01010/01011 colored according to the cluster they are assigned to; (f) points from substates 01010/01011 colored according to their value of angle $\psi_1$ (inset: Ramachandran map for residue 1 of points in substates 01010/01011, color is according to $\psi_1$).

lowest fraction of misassigned points; the latter, in order to allow direct comparison with the DM case) and found poorer correspondence to the BH substates than in the DM result (see Appendix C). However, we did observe that the $k = 2$ partition is very well matched with BH's coarse folded (F1 and F2) and (U1 and U2) unfolded states (see Table I). We conclude that clustering in the physical space works well on the coarsest scale, but is outperformed by clustering in the DM space when it comes to resolving the more detailed BH substates.

### E. A Markov state model

In order to explore the dynamical relevance of the identified clusters we construct the corresponding discrete-state, discrete-time Markov model. The construction of such Markov models from molecular simulation data has been discussed elsewhere[17,18,21]—here we simply apply some of those ideas. For clarity, we first briefly summarize some key relevant features.

TABLE I. Overlap between clusters resulting from physical clustering and BH coarse states.

| | Cluster 1 | Cluster 2 |
|---|---|---|
| F1 + F2 | 95% | 2% |
| U1 + U2 | 5% | 98% |

In the process of clustering, we have partitioned conformation space into $k = 16$ non-overlapping states. The successive visits to these states, which make up the trajectory, are assumed to constitute a realization of a stochastic process observed at discrete times. The Markov property requires that the conditional probability of future states be independent of the history of the process prior to the current state. If this property holds, then we can describe such a process by a $k \times k$ transition matrix $\Pi(\Delta t)$ whose elements $\Pi_{nm}$ represent the probability of finding the system in state $n$ at time $t + \Delta t$ after observing it in state $m$ at time $t$. (We note that the matrix $\Pi(\Delta t)$ describes Markovian transitions between the states of the reduced model, whereas the matrix $A$ defined in Eq. (2) used in the DM construction describes transitions between the individual structures defining the data graph.) If we consider a vector $P(l\Delta t)$ whose $n$th element is the probability of finding the system in state $n$ at time $l\Delta t$, then we can compute the probabilities at the next discrete time step as

$$P((l+1)\Delta t) = \Pi(\Delta t)P(l\Delta t). \tag{10}$$

The matrix $\Pi$ has left eigenvectors $z_1, \ldots, z_k$ and eigenvalues $\mu_1, \ldots, \mu_k$, where $|\mu_1| \geq |\mu_2| \geq \ldots \geq |\mu_k|$ (here we start indexing at 1 instead of 0, in order to match BH's notation). By construction, $\Pi$ is column-stochastic, which implies that it has an eigenvalue 1. It can be shown that $|\mu_i| \leq 1$, and therefore we have $\mu_1 = 1$. The corresponding right eigenvector (when properly normalized) can be readily seen (just

by observing Eq. (10)) to represent the stationary probability distribution.

The left eigenvectors may be used to group states (as in spectral clustering[49]) and can provide information about the nature of the dynamic processes "captured" by the model. Following the work of BH, we use the $z_i$ normalized to the interval [0, 1] to compute the quantities $\boldsymbol{\sigma}_i(s)$ (defined as the splitting probabilities[50]) for each cluster $s$,

$$\boldsymbol{\sigma}_i(s) = \frac{z_i(s) - \min_k z_i(k)}{\max_k z_i(k) - \min_k z_i(k)}, \tag{11}$$

where $z_i(s)$ denotes the $s$th element of the eigenvector $z_i$. In this way, each eigenvector divides clusters into two groups—one for those with values of $\boldsymbol{\sigma}_i < 0.5$ and one for those with $\boldsymbol{\sigma}_i \geq 0.5$.

We estimate $\Pi(\Delta t)$ from the trajectory as follows. We construct the matrix $T$, whose elements $T_{nm}$ are the number of transitions from state $m$ to state $n$ as observed at discrete time interval $\Delta t$. The condition of detailed balance requires that at equilibrium $T_{nm} = T_{mn}$. To impose this condition, we symmetrize $T$ by setting $T^{sym} = T + T^{\top}$ (the $\top$ in the superscript of the preceding expression signifies the transpose). It should be noted that the imposition of detailed balance may also be seen as supplementing the transition matrix with transitions that would have been observed by running the simulation backwards in time.[51] We then approximate the transition probabilities by

$$\Pi_{nm} \approx \frac{T^{sym}_{nm}}{\sum_n T^{sym}_{nm}}. \tag{12}$$

This expression represents the most likely transition matrix given the observed transitions. However, it should be noted that the finite size of the available data introduces error into this estimate and methods based on Bayesian statistics have been developed to handle this issue.[52,53]

A realization of a Markov process will also appear Markovian if observed on coarser time scales. This is because a large step with a corresponding transition matrix is equivalent to multiple smaller steps using the original transition matrix. That is, $\Pi(l\Delta t) = [\Pi(\Delta t)]^l$, where $l$ is some integer, and we now refer to $l\Delta t$ as the lag time (in our case $\Delta t = 2$ps, because we chose to record MD observations at this interval). This necessarily implies that the corresponding eigenvalues would be similarly related: $\mu_i(l\Delta t) = [\mu_i(\Delta t)]^l$. It follows that the eigenvalues of the transition matrix (except $\mu_1$) decay exponentially with increasing lag time and the corresponding relaxation times are computed this way,

$$\tau_i = -\frac{l\Delta t}{\ln(\mu_i)}. \tag{13}$$

For a true Markov process, the $\tau_i$ must be independent of the lag time and we use this property to test the Markovianity assumption for our data. In general, depending on the time scales of the process we observe, we may expect Markovianity to be applicable only at *long enough* lag times. After filtering out fast recrossings across state partition boundaries, we expect (based on the work of BH) that the lag-time dependence in the current example will be attenuated.

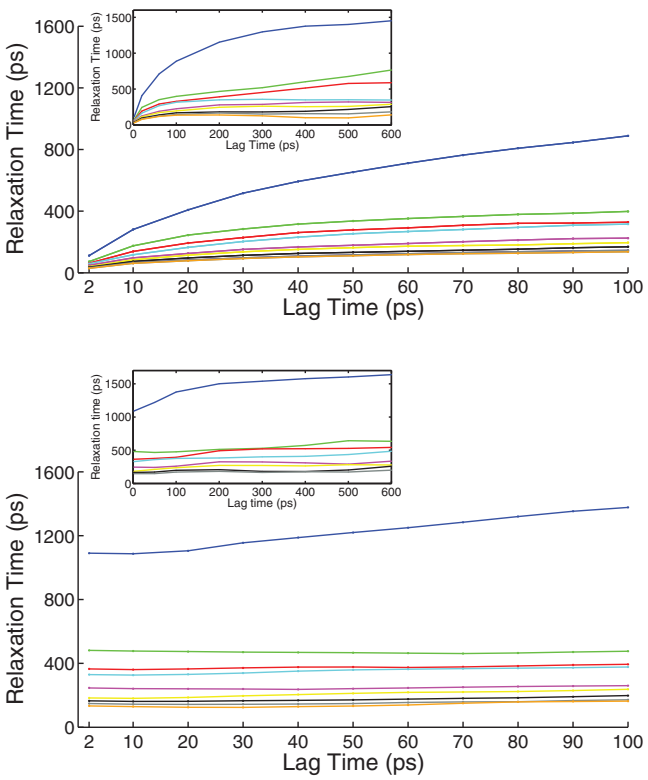



FIG. 7. Convergence of relaxation times as a function of lag time used in the construction of the Markov state models ($\tau_2$ - blue, $\tau_3$ - green, $\tau_4$ - red, $\tau_5$ - cyan, $\tau_6$ - magenta, $\tau_7$ - yellow, $\tau_8$ - black, $\tau_9$ - grey, $\tau_{10}$ - orange). (Top) k-means; (bottom) FCM+TBA. Insets show $\tau_2$ plateauing at larger lag times. (The lowest reported lag time is 2ps—the interval of observation from the MD trajectory.)

Figure 7(top) shows the 9 slowest relaxation times as functions of lag time. Clearly, the dynamics are non-Markovian at short lag times, but then appear to approach Markovian behavior at longer times. At the longest lag times considered, the slowest relaxation time, $\tau_2$, ultimately plateaus at approximately 1500 ps. This is in reasonable agreement with the value reported by BH[16] for this temperature ($\tau_2 \approx 1200$ ps), yet it takes a lag time of ~600 ps to approach this limiting value.

### F. A fuzzy Markov model

Our analysis of Markovianity exhibits the hallmarks of well-known deficiencies in cluster-based Markov state model construction: whereas the relaxation times of the fast processes reach their plateau values relatively quickly, the time scales of slow processes increase slowly with the lag time before finally converging. This is not the result of insufficient sampling: we have sampled for 100 ns, many multiples of the slowest relaxation time. In addition, the BH study was based on trajectories that were twice as long and found qualitatively similar results, in particular with respect to the time scales of relaxation. A very likely reason for non-Markovianity is apparent in Fig. 5: cluster 13 contains not only folded, helical structures, but also a small subpopulation of unfolded, coil structures. The fast escape from this subpopulation will contaminate the convergence of eigenvalues of folding and unfolding, occurring over longer time scales.

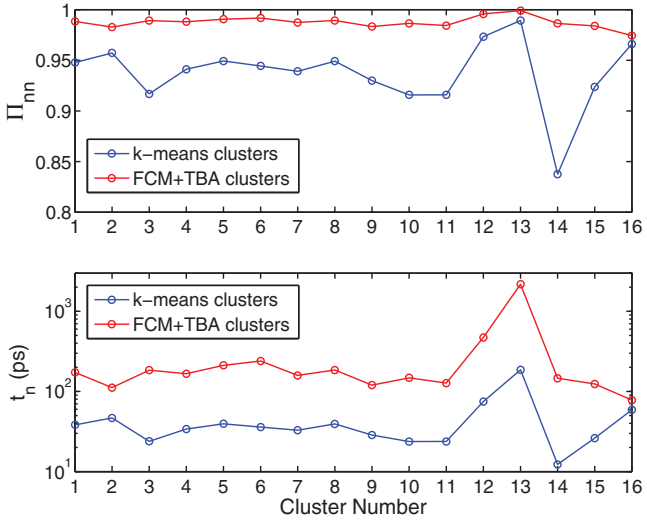



FIG. 8. Self-transition probabilities $\Pi_{nn}$ (top) and average lifetimes $t_n$ on a logarithmic scale (bottom) for state assignments through k-means clustering and through FCM+TBA.

Three possible ways of addressing this general issue are (a) to use a different, more informative metric for the similarity between configurations in the DM construction; (b) to expand the number of discrete states; and (c) to perform clustering using a more thoughtful state assignment procedure. Based on the above observations for cluster 13, we focus here on the latter approach.

We replace the hard clustering algorithm with a fuzzy one—Fuzzy C-means (FCM)[54] as implemented in MATLAB™. (We used a value of 1.4 for the "fuzzifier parameter" (m), because the commonly used value of 2 produced a clustering that was uninformative—all points were assigned equal membership to all clusters.) We should also note that linking fuzzy clustering algorithms with the geometry that diffusion induces on data has also been successfully implemented in the DifFUZZY algorithm of Cominetti *et al.*[57] The fuzzy clustering algorithm allows us to identify those points in a cluster that *unambiguously* belong to it. For every point $i$, the algorithm returns a grade $u_{ij} \in [0, 1]$ indicating the point's degree of membership in cluster $j$. The values 0 and 1 indicate no membership and full membership, respectively. We select all points $i$ with $u_{ij} = 1$ and assign them to the corresponding clusters $j$—these form the *cluster cores*, which we assume to represent true metastable states. The remaining unlabeled points along the trajectory are assigned to one of the cores in a procedure which is inspired by transition path sampling ideas.[9,55] BH call this procedure Transition-Based Assignment (TBA), where the assigned cluster index is To Be Announced(!) only at the end of a transition path, and use it to refine their substate definitions. A similar procedure performed well in numerical tests by Metzner *et al.*[53] The FCM+TBA procedure is applied as follows: if the trajectory leaves one of the cores, and some time later returns to the same core without having entered any other cores, all intermediate unlabeled points are assigned to this core, i.e., no transition took place. If the trajectory leaves a core and some time later visits another core, the first half of the intermediate unlabeled points are assigned to the first core and the second half to the second core. In this way, all of the unlabeled points are assigned to one of the cores. In our example, this procedure results in ≈12% of points changing their assignment.

As one might expect, this FCM+TBA procedure results in higher self-transition probabilities and longer average lifetimes for all clusters (Fig. 8). As a result, we observe less frequent recrossings between clusters. Figure 9 shows such a comparison of recrossings when the trajectory is recorded

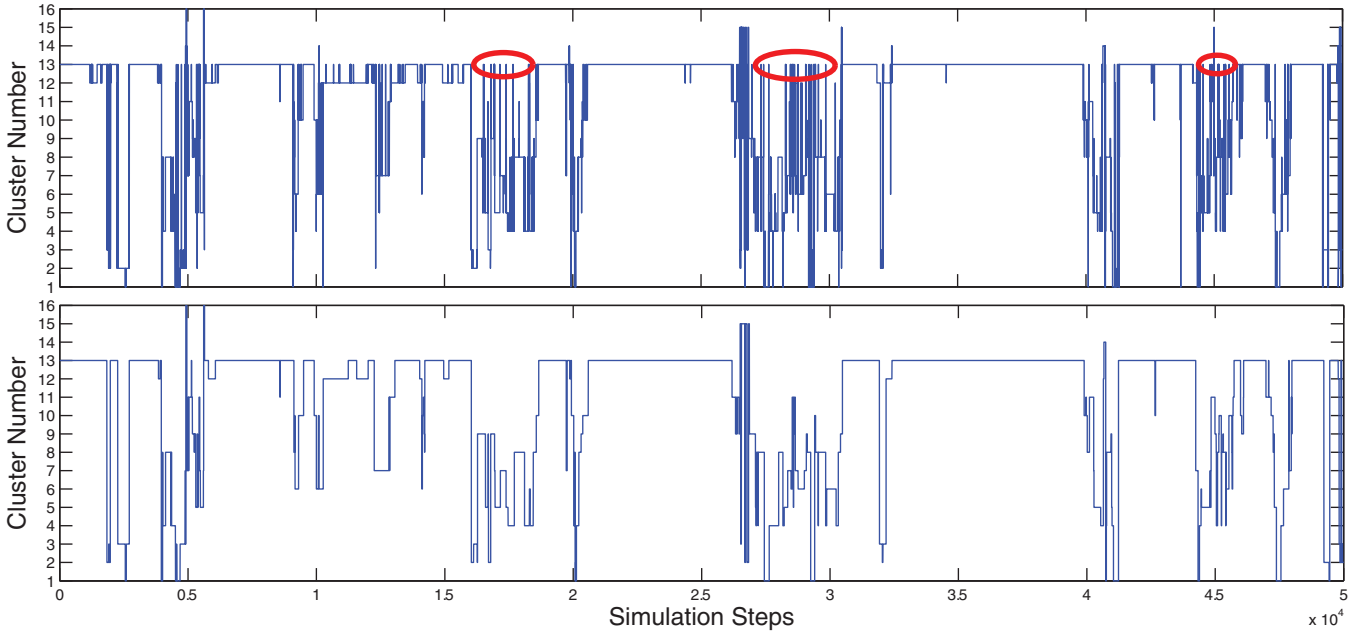



FIG. 9. Trajectory mapped onto the k-means clusters (top) and the FCM+TBA clusters (bottom). Segments with rapid exits from the "folded" cluster 13 are marked by red ovals. In the assignment of BH, the structures visited in these brief entries into cluster 13 tend to fall into the small subset not part of the F1 state, as illustrated in the heat map of Fig. 5. The fuzzy clustering of the FCM+TBA state assignment effectively eliminates these brief entries, explaining the greatly improved convergence of the fuzzy Markov model.

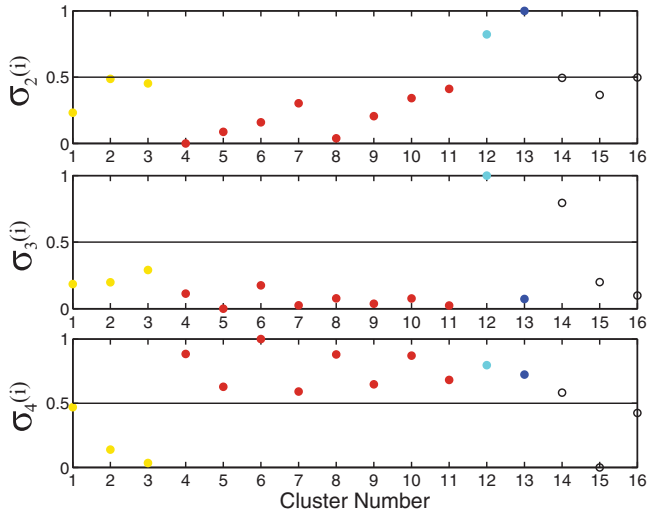



FIG. 10. Lumping of DM clusters based on splitting probabilities. Color indicates the DM cluster's correspondence to the BH coarse states as reflected by the heat map (Fig. 5); blue - F1, cyan - F2, red - U1, yellow - U2, black - cannot be unambiguously assigned.

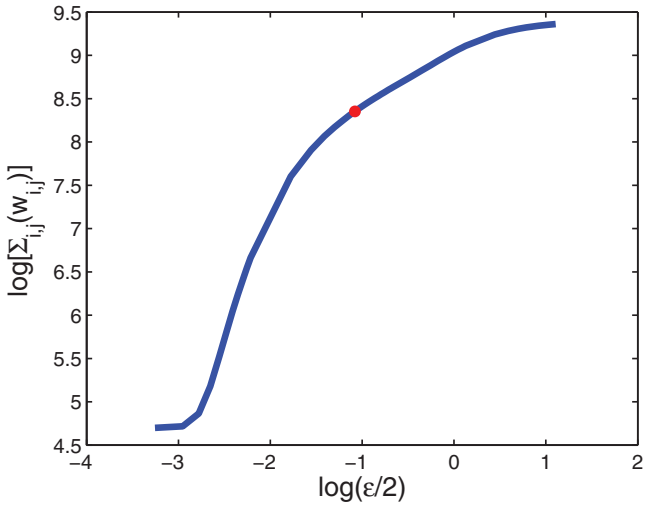



FIG. 11. Plot used to select $\epsilon$. The red dot indicates the value used in this work.

in terms of k-means cluster assignments (top) and in terms of FCM+TBA cluster assignments (bottom). Three particularly problematic regions with frequent short-lived visits to the "bad" subpopulation of cluster 13 are marked in the figure. Figure 7 compares the convergence of relaxation times for the two methods of cluster assignment. Remarkably, even at lag times of only 2 ps FCM+TBA produces a value of 1100 ps; for comparison, the traditional approach gives a value of only 100 ps. Overall, the combination of FCM and TBA substantially attenuated non-Markovianity effects.

One of the goals of Markov state model construction is to identify the slowly relaxing metastable "superstates" (e.g., the F and U states in BH). Based on our transition matrix, we use the leading nontrivial eigenvectors to help lump the clusters into such superstates by computing splitting probabilities (Eq. (11)). Figure 10 shows such a lumping of clusters: the first non-trivial eigenvector $v_2$ separates clusters 12 and 13 from the rest, suggesting a two-state (folded-unfolded) description, consistent with our expectations (as can be seen in Fig. 5, clusters 12 and 13 largely correspond to states F1 and F2, and the rest to the unfolded states). The second non-trivial eigenvector $v_3$ serves to split the folded superstate, separating cluster 12 from cluster 13. It also separates cluster 14, which contains points from both U1 and U2 from all the other unfolded clusters. Finally, the third non-trivial eigenvector $v_4$ splits the unfolded superstate into two groups of clusters corresponding to BH's U1 and U2 states, respectively. We find the consistency of the superstates derived from our model with BH's coarse states encouraging.

## V. CONCLUSIONS

In this paper, we illustrated the use of Diffusion Maps (a nonlinear manifold-learning technique) in processing molecular simulation data representative of folding transitions in alanine pentapeptide. Embedding the data in the lower-dimensional DM space not only provides a basis for visual representations of the high-dimensional peptide conformation space, but also enhances the identification of dynamically important substates using clustering tools. The comparison of DM-space clusters with the clusters identified in Ref. 16 supports the informed intuition of the previous work, in which structures were grouped based on their backbone dihedral angles; furthermore, the DM clustering helps identify and explain certain assignment inconsistencies by resolving dihedral motions ignored in the binary helix-coil scheme of BH. Given an informative embedding, the construction of Markov models by partitioning DM space becomes an obvious next step. Accordingly, we determined Markov state models from the MD dynamics projected onto the DM space clusters. In the construction of these models, tools developed for the fine-tuning of datapoint assignments in the clustering process can naturally be used to good effect in the new embedding. In particular, the combination of fuzzy clustering techniques (here fuzzy C-means clustering) with transition-based assignments of the state (using only core sets) filters out fast recrossings between states. Such recrossings arise commonly in state-assignments based on instantaneous projections. Their elimination in FCM+TBA thus effectively suppresses non-Markovianity effects at short observation times.[16,53]

Ultimately, the success of the MD trajectory analysis process is underpinned by two factors. The first is the (assumed) inherent low-dimensionality of the effective dynamics; if it is not there, it simply cannot be discovered algorithmically. The second factor (if the assumption is correct) is the systematic discovery/quantification of the low-dimensionality. Nonlinear data-mining tools may well deliver more parsimonious parameterizations of such low-dimensional manifolds (effective free energy surfaces); yet the major enabling factor is the use of an *informed metric*: a way to quantify pairwise similarities of nearby conformations. The better the choice of this metric, the easier the entire procedure becomes—from the collection of data through enhanced sampling techniques to the Markov modeling of the dynamics. In a forthcoming publication, we will illustrate the effect of such a "more informed"

TABLE II. Statistics of point silhouette scores for the individual clusters in the optimal 16-cluster representation.

| | Mean | Q1 | Median | Q3 |
|---|---|---|---|---|
| Cluster 1 | 0.687 | 0.657 | 0.768 | 0.786 |
| Cluster 2 | 0.729 | 0.736 | 0.819 | 0.830 |
| Cluster 3 | 0.568 | 0.493 | 0.689 | 0.712 |
| Cluster 4 | 0.631 | 0.588 | 0.731 | 0.759 |
| Cluster 5 | 0.743 | 0.739 | 0.825 | 0.838 |
| Cluster 6 | 0.707 | 0.687 | 0.780 | 0.816 |
| Cluster 7 | 0.724 | 0.719 | 0.799 | 0.815 |
| Cluster 8 | 0.683 | 0.671 | 0.776 | 0.793 |
| Cluster 9 | 0.662 | 0.633 | 0.760 | 0.781 |
| Cluster 10 | 0.609 | 0.565 | 0.687 | 0.741 |
| Cluster 11 | 0.615 | 0.560 | 0.691 | 0.740 |
| Cluster 12 | 0.645 | 0.673 | 0.745 | 0.758 |
| Cluster 13 | 0.884 | 0.921 | 0.922 | 0.922 |
| Cluster 14 | 0.442 | 0.300 | 0.540 | 0.601 |
| Cluster 15 | 0.679 | 0.659 | 0.789 | 0.814 |
| Cluster 16 | 0.868 | 0.895 | 0.915 | 0.923 |

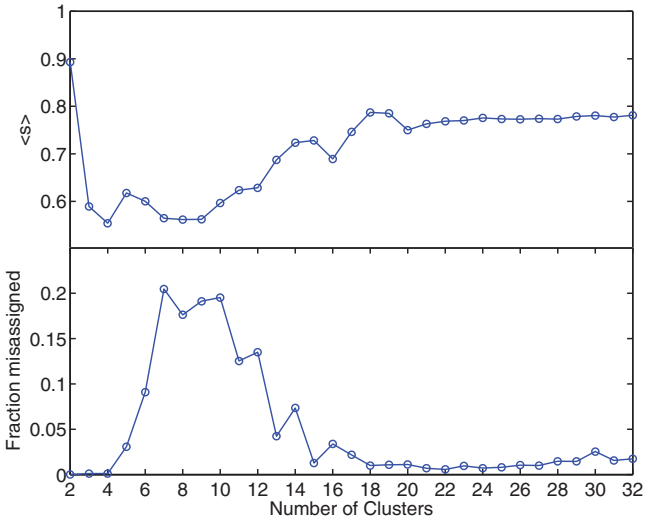



FIG. 12. Average silhouette score (above) and fraction of misassigned points (below) for clustering in 18-d DM space.

(dihedral-angle based) metric on the resulting DM embedding, clustering, and Markov modeling.

## ACKNOWLEDGMENTS

The work of L.V.N., M.A.A., and I.G.K. was partially supported by the U.S. Department of Energy (Grant No. DE-SC0002097); L.V.N. also gratefully acknowledges support from the Princeton University Program in Plasma Science and Technology under U.S. Department of Energy Contract No. DE-AC02-76-CHO-3073. The work of G.H. was supported by the Intramural Research Program of the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health (NIH), and by the Max Planck Society.

## APPENDIX A: CHOICE OF EPSILON

The value of the parameter $\epsilon$ is representative of the size of the neighborhood around a data point over which we consider the RMSD metric to be dynamically informative. As detailed in Ref. 56, there is a range of $\epsilon$ values which result in meaningful embeddings and this range may be identified by constructing a $\log(\sum_{i,j}(w_{ij}))$ vs. $\log(\epsilon/2)$ plot (see Eq. (1)

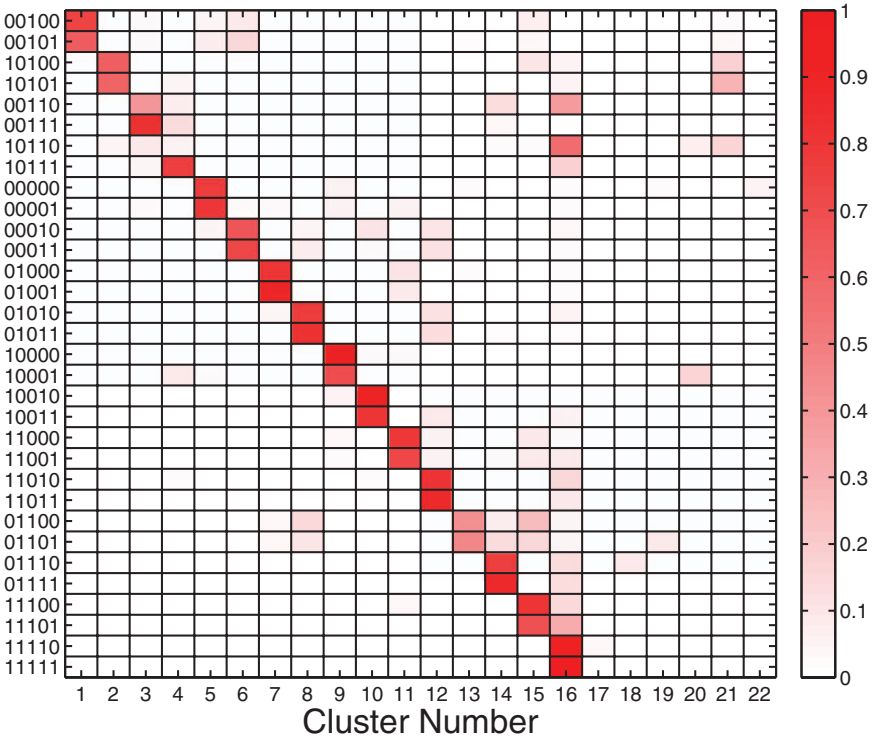



FIG. 13. Heat map showing the level of overlap between BH substates (vertical axis) and DM clusters resulting from clustering in an 18-d DM space.

and Fig. 11). Useful values of $\epsilon$ are those from the linear region(s) of the plot.

In practice, we have found that $\epsilon$ values near the low boundary of the region produce plots with more clearly visible clusters. We refer the reader to the work of Clementi *et al.*[34,58,59] for additional considerations and algorithmic suggestions for the choice of $\epsilon$.

## APPENDIX B: ADDITIONAL CLUSTER STATISTICS

Table II indicates how well formed are the clusters of the optimal ($k = 16$) k-means partitioning. In addition to the mean, we report the three quartile points—Q1, median (Q2), and Q3—of the silhouette scores.

## APPENDIX C: SUPPLEMENTARY RESULTS

### 1. Varying the number of DM eigenvectors

We chose to work in a 12-d reduced DM space on the basis of a gap in the eigenvalue spectrum found after $\lambda_{12}$ (Fig. 2(b)). However, evident in the figure are a number of other gaps, the largest of which appears after $\lambda_{18}$. We repeated the analysis presented in the paper using 18 DM eigenvectors and found qualitatively similar results which we include here.

The k-means cluster scoring in the 18-d space is shown in Fig. 12. Although the number of clusters we selected changed to 22, the correspondence between those clusters and the BH substates was not significantly altered (Fig. 13). One difference is that the state F1 was resolved better than in the 12-d space, where we noted that the 4 substates making up F1 were clustered together. Here substates 11100 and 11101 form part of cluster 15, while 11110 and 11111 are in cluster 16. However, as might be expected, this does not impact the Markovianity of the resulting model. Figure 14 shows the relaxation times obtained from both the k-means and subsequent FCM+TBA constructions. A comparison with Fig. 7 indicates that the behavior is very close to the 12-d result.

### 2. Varying the number of clusters in the 12-d DM space

The procedure for choosing which k-means partitioning to use in our analysis was based on silhouette score metrics (Fig. 3). The apparently optimal 16-cluster assignment only marginally outscored some others, e.g., 2 clusters and 20 clusters. We have investigated the effect of choosing these assignments and include the results here.

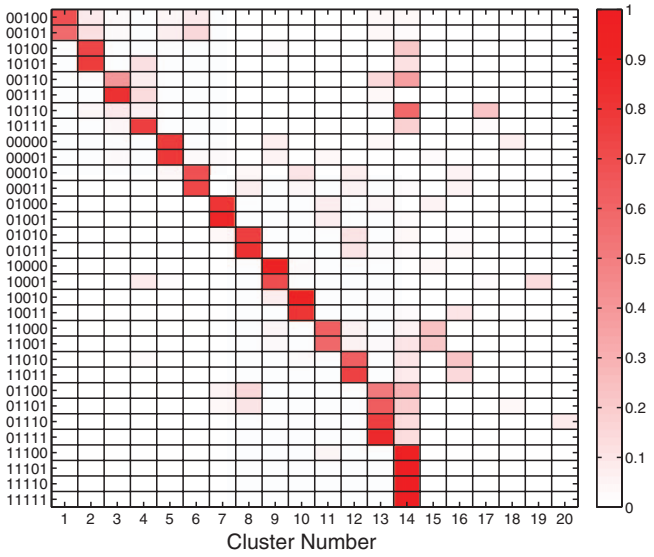



FIG. 15. Heat map showing the level of overlap between BH substates (vertical axis) and DM clusters resulting from clustering in the 12-d DM space, 20 clusters.

Figure 15 shows that the correspondence between BH substates and clusters in the case $k = 20$ is not altogether much different than the 16-cluster result (Fig. 5). One change is that what was cluster 1 in the 16-cluster partitioning has split in two, so that substates 00100/00101 and 10100/10101 now mostly occupy their own clusters (numbered 1 and 2, respectively). Somewhat mirroring the result of adding more eigenvectors, this apparent improvement in detail does not impact the behavior of the model's slow relaxation times. As can be observed in Fig. 16, the k-means assignment is rectified by the FCM+TBA procedure in a way analogous to the case presented in the main body of the paper.

The partition with $k = 2$ performs rather poorly. The clustering scores (Fig. 3) suggest well-formed clusters, however, we see in Fig. 17 that this does not translate to a good correspondence with the BH folded and unfolded states. We believe that this occurs due to the structure of the data in DM space. As was mentioned in the paper, the folded states are

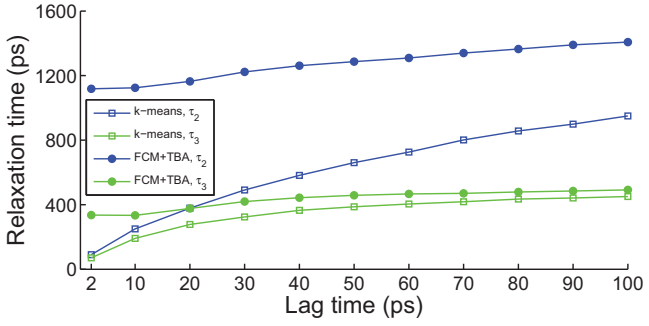



FIG. 14. Convergence of 2 slowest relaxation times as a function of lag time used in the construction of the Markov state model in 18-d DM space.

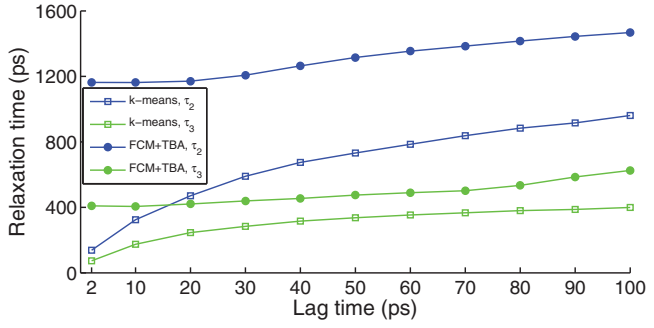



FIG. 16. Convergence of 2 slowest relaxation times as a function of lag time used in the construction of the Markov state model in 12-d DM space, 20 clusters.

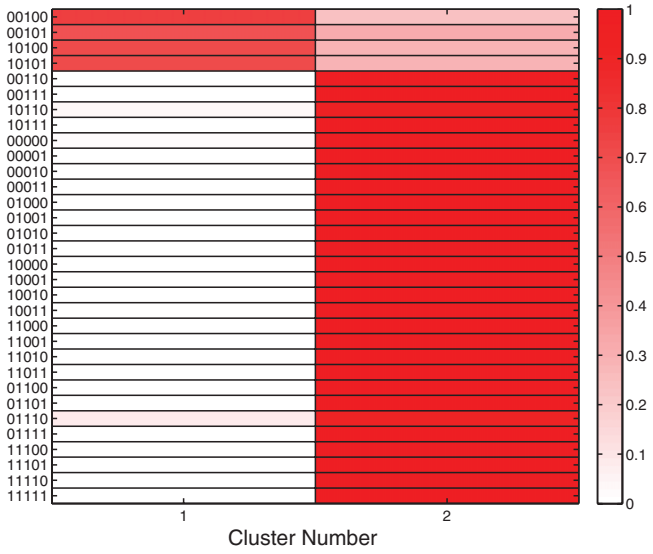



FIG. 17. Heat map showing the level of overlap between BH substates (vertical axis) and DM clusters resulting from clustering in the 12-d DM space, 2 clusters.

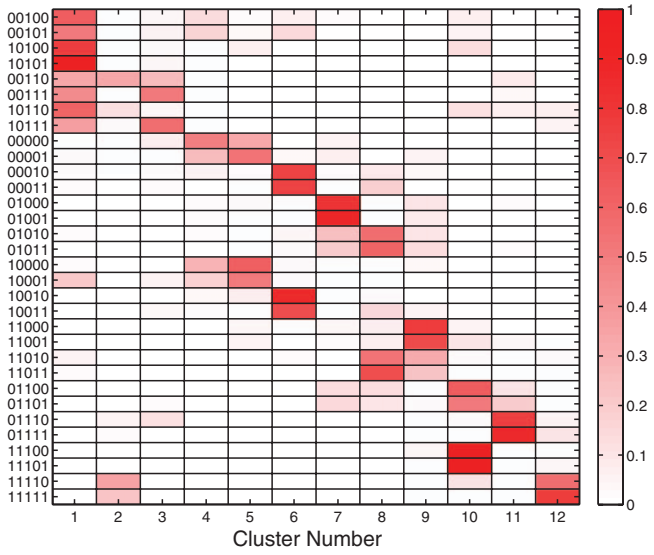



FIG. 19. Heat map showing the level of overlap between BH substates (vertical axis) and DM clusters resulting from clustering into 12 physical space clusters.

concentrated together in a very small section of the space, while the unfolded states are more spread out, occupying a much larger area. The k-means algorithm has a tendency to produce clusters of relatively similar size and we believe this handicaps the $k = 2$ result.

## 3. Physical space clustering

We did a more thorough investigation of the results of physical space clustering. The claim in the main body of the paper was that the "nicest" overlap with the BH classification occurs when $k = 2$. Figures 18 and 19 show that at $k = 12$ and $k = 16$ k-means clustering in the physical space does not result in partitionings which correlate well with the BH substates.

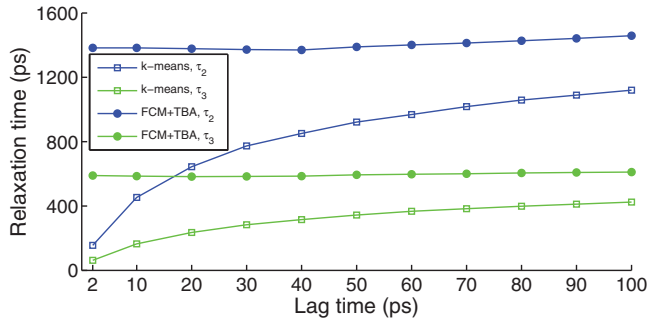



FIG. 20. Convergence of the two slowest relaxation times as a function of lag time used in the construction of the Markov state model in 16 physical space clusters.

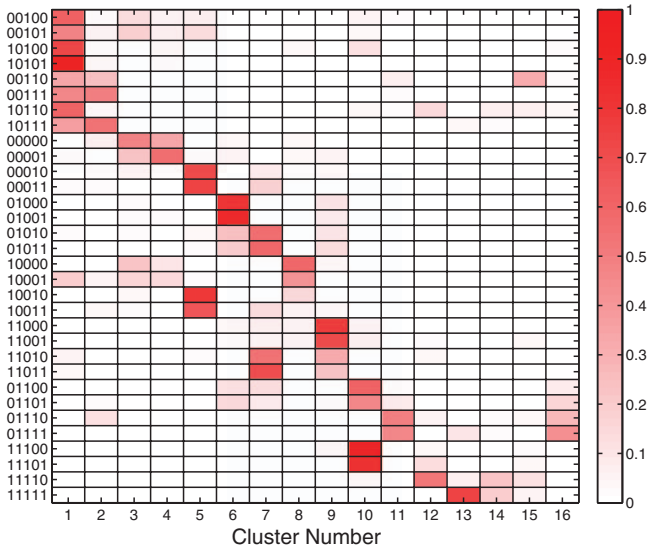



FIG. 18. Heat map showing the level of overlap between BH substates (vertical axis) and DM clusters resulting from clustering into 16 physical space clusters.

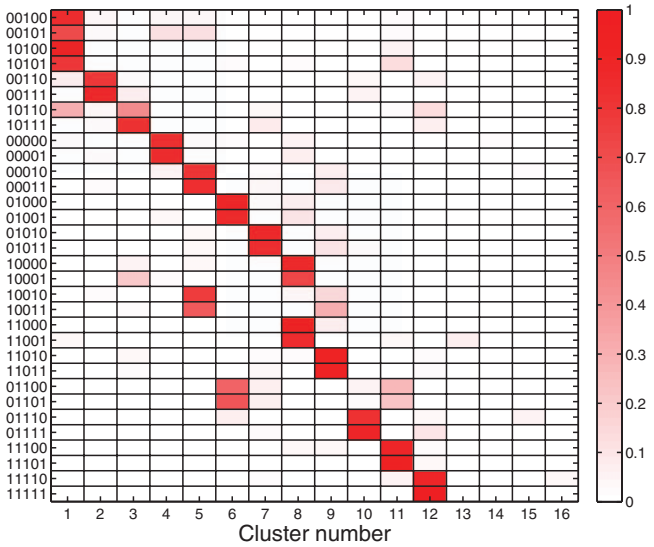



FIG. 21. Heat map showing the level of overlap between BH substates (vertical axis) and macrostates produced by the MSMBuilder Software.

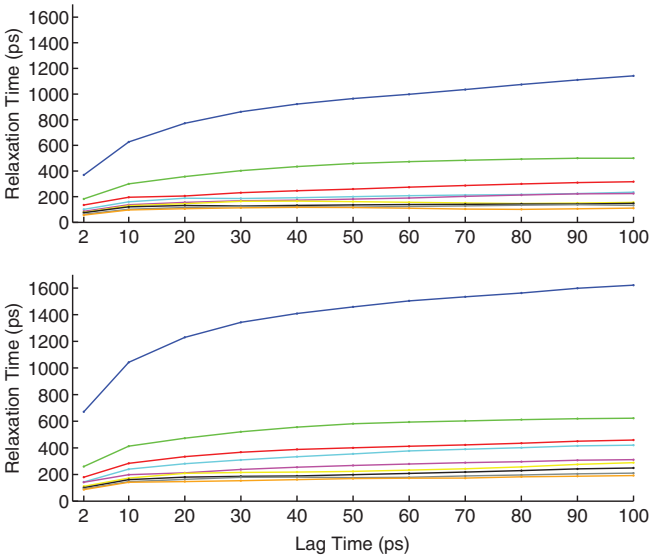



FIG. 22. Convergence of relaxation times as a function of lag time used in the construction of the Markov state models produced by the MSMBuilder Software. (Top) 16 macrostates; (bottom) 200 microstates.

Somewhat surprisingly, however, the FCM+TBA procedure in physical space results in a model with improved Markovianity as can be seen in Fig. 20.

## 4. Comparison with MSMBuilder[19]

As a further means of validation for the work in this paper, we built a Markov State Model using the MSMBuilder Software.[19] This method initially partitions the data into a large number of microstates based on structural similarity, quantified by heavy-atom RMSD. In a subsequent step, it lumps the microstates into macrostates in a way that maximizes the sum of the self-transition probabilities of the macrostates.

We chose a 16-state macrostate model in order to maintain consistency with the results reported so far. Figure 21 shows a level of correspondence to the BH substates which is comparable to ours, although some differences in the lumping of substates are observed. The behavior of the slowest 9 relaxation times for the macrostate model is shown in Figure 22 (top) and is consistent with our DM k-means result.

[1]A. E. García, Phys. Rev. Lett. **68**, 2696 (1992).
[2]A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins: Struct., Funct., Genet. **17**, 412 (1993).
[3]L. S. D. Caves, J. D. Evanseck, and M. Karplus, Protein Sci. **7**, 649 (1998).
[4]P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, Proc. Natl. Acad. Sci. U.S.A. **103**, 9885 (2006).
[5]W. M. Brown, S. Martin, S. N. Pollock, E. A. Coutsias, and J. P. Watson, J. Chem. Phys. **129**, 064118 (2008).
[6]G. M. Torrie and J. P. Valleau, Chem. Phys. Lett. **28**, 578 (1974).
[7]A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A. **99**, 12562 (2002).
[8]E. Rosta, H. L. Woodcock, B. R. Brooks, and G. Hummer, J. Comput. Chem. **30**, 1634 (2009).
[9]P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).
[10]R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. U.S.A. **102**, 6732 (2005).
[11]R. Erban, T. A. Frewen, X. Wang, T. C. Elston, R. Coifman, B. Nadler, and I. G. Kevrekidis, J. Chem. Phys. **126**, 155103 (2007).
[12]C. R. Laing, T. A. Frewen, and I. G. Kevrekidis, Nonlinearity **20**, 2127 (2007).
[13]A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, J. Chem. Phys. **134**, 135103 (2011).
[14]M. Sarich, F. Noé, and C. Schütte, Multiscale Model. Simul. **8**, 1154 (2010).
[15]S. Sriraman, L. G. Kevrekidis, and G. Hummer, J. Phys. Chem. B **109**, 6479 (2005).
[16]N.-V. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057 (2008).
[17]J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, J. Chem. Phys. **126**, 155101 (2007).
[18]N. Singhal, C. D. Snow, and V. S. Pande, J. Chem. Phys. **121**, 415 (2004).
[19]G. R. Bowman, X. H. Huang, and V. S. Pande, Methods **49**, 197 (2009).
[20]F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. U.S.A. **106**, 19011 (2009).
[21]C. Schütte, F. Noé, J. Lu, M. Sarich, and E. Vanden-Eijnden, J. Chem. Phys. **134**, 204105 (2011).
[22]C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, J. Comput. Phys. **151**, 146 (1999).
[23]R. B. Best and G. Hummer, Phys. Chem. Chem. Phys. **13**, 16902 (2011).
[24]J. B. Tenenbaum, V. de Silva, and J. C. Langford, Science **290**, 2319 (2000).
[25]S. T. Roweis and L. K. Saul, Science **290**, 2323 (2000).
[26]D. L. Donoho and C. Grimes, Proc. Natl. Acad. Sci. U.S.A. **100**, 5591 (2003).
[27]R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proc. Natl. Acad. Sci. U.S.A. **102**, 7426 (2005).
[28]R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proc. Natl. Acad. Sci. U.S.A. **102**, 7432 (2005).
[29]S. S. Lafon, "Diffusion maps and geometric harmonics," Ph.D. thesis (Yale University, 2004).
[30]R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, Multiscale Model. Simul. **7**, 842 (2008).
[31]B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis, Appl. Comput. Harmon. Anal. **21**, 113 (2006).
[32]P. Das, T. Frewen, I. Kevrekidis, and C. Clementi, in *Coping with Complexity: Model Reduction and Data Analysis*, Lecture Notes in Computational Science and Engineering Vol. 75 edited by A. Gorban and D. Roose (Springer, Berlin-Heidelberg, 2011), p. 113.
[33]A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti, Chem. Phys. Lett. **509**, 1 (2011).
[34]M. A. Rohrdanz, W. W. Zheng, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**, 124116 (2011).
[35]H. Nymeyer and A. E. García, Proc. Natl. Acad. Sci. U.S.A. **100**, 13934 (2003).
[36]S. Lafon and A. B. Lee, IEEE Trans. Pattern Anal. Mach. Intell. **28**, 1393 (2006).
[37]B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[38]E. J. Sorin and V. S. Pande, Biophys. J. **88**, 2472 (2005).
[39]A. J. DePaul, E. J. Thompson, S. S. Patel, K. Haldeman, and E. J. Sorin, Nucleic Acids Res. **38**, 4856 (2010).
[40]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[41]U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).
[42]M. Parrinello and A. Rahman, J. Appl. Phys. **52**, 7182 (1981).
[43]B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, J. Comput. Chem. **18**, 1463 (1997).
[44]W. Kabsch, Acta Crystallogr., Sect. A **34**, 827 (1978).
[45]R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, (SIAM, Philadelphia, 1998).
[46]A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, Proc. Natl. Acad. Sci. U.S.A. **107**, 13597 (2010).
[47]G. A. F. Seber, *Multivariate Observations* (John Wiley and Sons, Inc., Hoboken, NJ, 1984).
[48]P. J. Rousseeuw, J. Comput. Appl. Math. **20**, 53 (1987).
[49]U. von Luxburg, Stat. Comput. **17**, 395 (2007).
[50]A. Berezhkovskii and A. Szabo, J. Chem. Phys. **121**, 9186 (2004).
[51]G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, J. Chem. Phys. **131**, 124101 (2009).
[52]N. Singhal and V. S. Pande, J. Chem. Phys. **123**, 204909 (2005).

[53]P. Metzner, M. Weber, and C. Schütte, Phys. Rev. E **82**, 031114 (2010).
[54]J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms* (Plenum Press, New York, 1981).
[55]G. Hummer, J. Chem. Phys. **120**, 516 (2004).
[56]R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, IEEE Trans. Image Process. **17**, 1891 (2008).
[57]O. Cominetti, A. Matzavinos, S. Samarasinghe, D. Kulasiri, S. Liu, P. K. Maini, and R. Erban, Int. J. Comput. Intell. Bioinf. Syst. Biol. **1**, 402 (2010).
[58]W. W. Zheng, M. A. Rohrdanz, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**, 144109 (2011).
[59]W. W. Zheng, B. Qi, M. A. Rohrdanz, A. Caflisch, A. R. Dinner, and C. Clementi, J. Phys. Chem. B **115**, 13065 (2011).