Adaptive spectral clustering with application to tripeptide conformation analysis

Fiete Haack,[1,*] Konstantin Fackeldey,[2,†] Susanna Röblitz,[2,‡]
Olga Scharkoi,[2,§] Marcus Weber,[2,¶] and Burkhard Schmidt[3,**]

[1]*Institut für Informatik, Universität Rostock*
Albert-Einstein-Str. 21, D-18059 Rostock, Germany
[2]*Zuse Institut Berlin, Takustraße 7, D-14195 Berlin, Germany*
[3]*Institut für Mathematik, Freie Universität Berlin,*
Arnimallee 6, D-14195 Berlin, Germany

(Dated: October 10, 2013)

Abstract

A decomposition of a molecular conformational space into sets or functions (states) allows for a reduced description of the dynamical behavior in terms of transition probabilities between these states. Spectral clustering of the corresponding transition probability matrix can then reveal metastabilities. The more states are used for the decomposition, the smaller the risk to cover multiple conformations with one state, which would make these conformations indistinguishable. However, since the computational complexity of the clustering algorithm increases quadratically with the number of states, it is desirable to have as few states as possible. To balance these two contradictory goals, we present an algorithm for an adaptive decomposition of the position space starting from a very coarse decomposition. The algorithm is applied to small data classification problems where it was shown to be superior to commonly used algorithms, e. g., k-means. We also applied this algorithm to the conformation analysis of a tripeptide molecule where six-dimensional time series are successfully analyzed.

[*] fiete.haack@uni-rostock.de
[†] fackeldey@zib.de
[‡] susanna.roeblitz@zib.de
[§] scharkoi@zib.de
[¶] weber@zib.de
[**] burkhard.schmidt@fu-berlin.de

I. INTRODUCTION

One challenging aspect in the simulation of biomolecules is the high dimensionality of the corresponding conformation space. The position states of a molecular system as individual consecutive snapshots from a trajectory can be represented as a set of points in the conformational space. Typically this conformational space is high-dimensional, which renders a rigorous analysis in terms of individual states impossible. Under the assumption, that the potential energy surface is separated by well defined energy barriers, collections of similar states (metastabilities) can be defined. In the conformational space these metastabilites are characterized as subsets, where the dynamical system spends a long time before it switches to another metastability. Within each metastable set the dynamics is fast mixing (cf. Fig. 1)

This set based point of view of metastabilities differs from the classical definition of conformations as minima of the free energy landscape because it also takes into account entropic barriers. Usually, there exist many more energy minima than metastabilities. Multiple minima can well belong to one metastability if there are frequent transitions between these minima. The identification of metastabilities together with their life times and transition patterns is essential for the analysis of a system's long term behavior. Initiated by the pio-

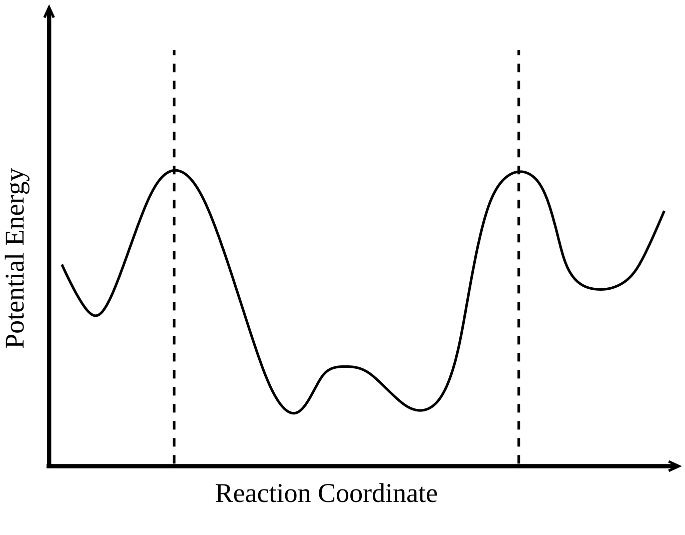

FIG. 1. Sketch of a potential energy along some reaction coordinate. The potential has four local minima but only three metastable states for moderately high temperature separated by the vertical dashed lines.

neering work of Dellnitz, Deuflhard, and Schütte, a multi-scale method, called *conformation dynamics*, has been developed [1–4]. Its main objective is the identification of metastabilities together with their life times and transition patterns. This approach of partitioning the state space and interpreting transition between these sets as a realization of a Markov Chain (Markov State Models), has been quite successful [5–14]. In this mixed deterministic/stochastic approach, the dynamics is modeled as a Markov process in a discretized finite state space, which results in a nearly decomposable transition probability matrix. By considering the transition probabilities as similarities between the states, the application of a cluster algorithm reveals the metastabilities. The aggregation of single molecular configurations into a small number of states in the molecule's position space is necessary for a large amount of configurational data obtained, e. g., from molecular dynamics simulations where intuitive point-wise clustering becomes impossible due to high complexity. If the states are chosen in a naïve way, it might happen that one state covers two or more metastabilities. When applying a cluster algorithm relying on the transition probabilities between the states only these conformations cannot be detected, since the transition behavior within the states is disregarded. Thus, the more states we use for the decomposition, the smaller the risk to cover multiple conformations with one state. However, since the computational complexity of most clustering algorithm increases quadratically with the number of states, it is desirable to have a small number of states. Moreover, the estimated transition probabilities might become statistically unreliable the smaller the states and the fewer configurations per state are available. In the last few years this problem has been addressed by many authors combined with a strategy to find the best trade off between accuracy and complexity [15–19].

Based on a coarse decomposition of the state space, we propose an adaptive scheme, which accounts for geometric as well as for dynamical aspects of the states in each portion of the decomposition. Our idea is to decompose the object space Ω by a Voronoi tessellation, to build the transition probability matrix based on these sets, and to apply the robust Perron Cluster Cluster Analysis (PCCA+) [6] in order to identify the metastabilities. PCCA+ is the successor of the PCCA method ([3]), where the primal version only allowed for a "hard" clustering and the latter allows for a fuzzy clustering [20]. At this point our procedure is similar to the automatic state decomposition algorithm proposed by Chodera et al. [15]. In contrast to Ref. [15], we use an adaptive refinement scheme to detect and refine exclusively those partitions that contain metastabilities. This refinement is not only based on the geometric similarity between objects in one cell, but also relies on intracell transition probabilities. Thus, only partitions that actually contain more than one metastabilities will be refined. Thereby, we avoid the risk of missing conformations that are covered by the same state, while having a minimal set of partitions at the same time. In the following we will first explain, how the metastable clusters are derived from a given partitioning, and subsequently describe the adaptive partitioning scheme in detail.

II. STATE SPACE DECOMPOSITION BY MEMBERSHIP BASIS FUNCTIONS

We seek for a clustering method that combines geometric and dynamic aspects. To do so a suitable decomposition of the position space Ω is needed. In the literature (e. g. Ref. [3]), a discretization of Ω into Voronoi cells is used to compute transition probabilities between different subsets of the position space. However, for our purposes such a discretization is not sufficient, since only the dynamic aspects are mirrored, whereas the geometric aspects are unaccounted. This is possible, if the discretization of the position space Ω is not based on sets but on membership functions having values between zero and one and thus allowing for the computation of an overlap matrix providing the geometric information.

Let us consider a canonical ensemble (constant number of particles, constant volume and constant temperature), where, the positions q and the momenta p of all atoms are given according to the Boltzmann distribution:

$$\pi(q,p) \propto \exp(-\beta H(q,p)).$$

Here $\beta = 1/k_B T$ is the inverse temperature T multiplied with the Boltzmann constant k_B and H denotes the Hamiltonian function which is given by $H(q,p) = V(q) + K(p)$, where $V(q)$ is the potential and $K(p)$ is the dynamic energy. The canonical density can be split into a distribution of momenta $\pi(q)$ and positions $\eta(p)$ where $\pi(q) \propto \exp(-\beta V(q))$ and $\eta(p) \propto \exp(-\beta K(p))$. In the forthcoming we assume, that the states $\{q_i\}_i$ stem from a molecular dynamics simulation (trajectory) being π distributed.

For the discretization step, we use n radial basis functions with nodes $\{\hat{q}_1, \ldots, \hat{q}_n\}$ with the Gaussian similarity measure $\exp(-\alpha\, d(i,j)^2)$ where the parameter α controls the width of the neighborhoods and $d(i,j) = \|q_i - q_j\|_{l_2} = \sqrt{\sum_{k=1}^{d}(q_{i_k} - q_{j_k})^2}$. Following the partition of unity method of Shepard [21] we obtain:

$$\varphi_i(q_k) = \frac{\exp(-\alpha\, d(q_k, \hat{q}_i)^2)}{\sum_{j=1}^{n} \exp(-\alpha\, d(q_k, \hat{q}_j)^2)}, \quad i = 1, \dots n. \tag{1}$$

The basis functions can be interpreted as membership functions since they are non-negative

$$\varphi_i(q) > 0, \quad \forall q \in \Omega, \quad i = 1, \dots, n \tag{2}$$

and form a partition of unity

$$\sum_{i=1}^{n} \varphi_i(q) = 1, \quad \forall q \in \Omega. \tag{3}$$

The basis function φ_i can be interpreted as a relaxation of a Voronoi cell with center at $\hat{q}_i$. In the limit case as $\alpha \to \infty$ the Voronoi discretization is recovered. The shape parameter α determines the overlap M_{ij} between two basis functions φ_i and φ_j defined as

$$M_{ij} := \frac{\int_\Omega \varphi_i(x)\varphi_j(x)\pi(x)dx}{\int_\Omega \varphi_i(x)\pi(x)dx} \approx \frac{\sum_{k=1}^{N} \varphi_i(q_k)\varphi_j(q_k)}{\sum_{k=1}^{N} \varphi_i(q_k)} =: K_{ij}^{(0)}. \tag{4}$$

The larger α, the smaller is the overlap, as illustrated in Fig. 2. The example is based on a small, artificial 2-dimensional (2D) data set that is partitioned by two basis functions depending on different α values. The left part shows how points in-between the two partitions share their membership and thus create an overlap between the two soft partitions, indicated by the orange color. In the middle we show how the partition with large α becomes almost characteristic (Voronoi cells). According to the colorization of the data points, we have a very distinct separation and thus only a very small overlap. In the right panel of the figure one can see the partitioning with small α–value. Consequently, all data points have almost the same membership values for both clusters, indicated by the same color orange.

In the context of geometric clustering the membership values of the basis functions represent the similarity of the given data point q_k to the current representative node $\hat{q}_i$ with respect to the similarity to the rest of the nodes, i.e. for each state $q \in \Omega$ we calculate the relative similarity to all nodes $\{\hat{q}_1, ..., \hat{q}_n\}$ (see Sec. 1). Instead of seeking for a dynamic similarity only, by using a Voronoi discretization, we introduce the dynamic similarity which accounts for the geometric as well as dynamic aspects between the different states. Before we explain our similarity indicator, we need to introduce a time discretization parameter τ. Now and in the forthcoming we assume that the states q_i are given by a classical molecular

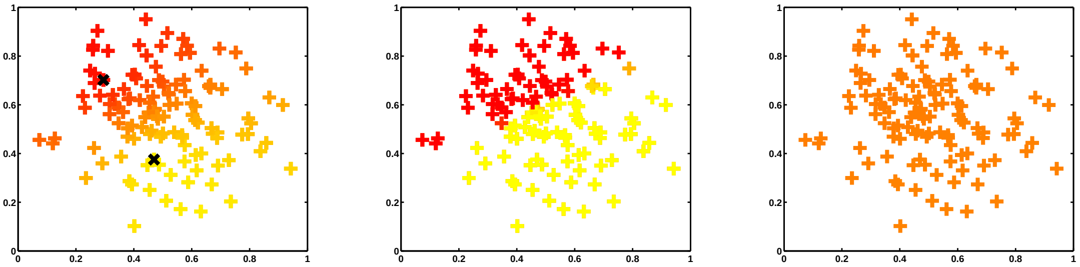

FIG. 2. Soft partitioning of a small, artificial data set by two membership functions. The center nodes $\hat{q}_1$ and $\hat{q}_2$ of the corresponding membership functions φ_1 and φ_2 are depicted only in the left picture as black crosses, but have the same position in all three data sets. In each picture the data points are colored by their membership to a respective partition. Red for the upper left partition and yellow for the lower right partition, orange for intermediate cases. From left to right: $\alpha = 2, 100, 0.1$

dynamics trajectory of a system, that is a sequence of points in the phase space which are connected in time with a time step h (typically in the order of femtoseconds). By choosing $\tau = \tilde{n}h \quad, \tilde{n} \gg 1$ we do not consider each state of the trajectory but only every $\tilde{n}$th step. Analogously to (4) we now can define the dynamic similarity as $K_{ij}^{(L\tau)}$ between two basis functions φ_i and φ_j for a time lag $L\tau$ as

$$K_{ij}^{(L\tau)} := \frac{\sum_{k=1}^{N} \varphi_i(q_k)\varphi_j(q_{k+L})}{\sum_{k=1}^{N} \varphi_i(q_k)}, \tag{5}$$

where q_k is the k-th state of the system and q_{k+L} is the $(k+L)$-th state of the molecular system. This indicator considers the similarity of basis functions. More precisely it is an estimate of the overlap between two basis functions. By normalization the matrix $K^{(L\tau)}$ is stochastic and thus the entries $K_{ij}^{(L\tau)}$ are bounded by 1. We now employ $K^{(L\tau)}$ as a refinement indicator in the following adaptive scheme.

III. ADAPTIVE ALGORITHM AND TRANSITION MATRIX

Since we use global basis functions, any initial partitioning covers the complete state space Ω. In order to use as few basis functions as possible, the nodes should be located only in the relevant parts of the object space, i.e., parts where many data objects are located. However, it is not possible to separate two different metastable sets in the process of clustering if they are covered by only one basis function. Therefore all relevant parts of the object space (i.e. all clusters/metastabilities) must be covered sufficiently. That means, we want to avoid partitions that are

- redundant: strongly overlapping basis functions, since they share the same substructure,
- uninformative outsiders: small separated basis functions, which contain only a very small amount of data points and have a poor overlap with other partial densities.

With the following locally adaptive partitioning algorithm we aim to improve the initial selection of nodes and thus find an optimal soft partition of the object space. The main idea is to check each local basis function for the existence of further metastabilities and, if found, to refine the basis function by adding a user-defined number of s nodes that represent the metastable sets.

For one specific basis function φ_i, the algorithm has the following structure.

1. Select all states q_j with $\varphi_i(q_j) > \varphi_t(q_j)\,\forall t \neq i$.
2. Perform the k-means algorithm with s clusters on the selected objects: Choose k cluster $\mathsf{C} = \{\mathsf{C}_1, \ldots, \mathsf{C}_\mathsf{k}\}$ which minimize
$$\sum_{j=1}^{k} \min_{\bar{q}\in\Omega} \sum_{q_a \in \mathsf{C}_j} \|q_a - \bar{q}_j\|.$$
Select the states nearest to the k computed centroids $\{\bar{q}_1, ..., \bar{q}_k\}$ as new temporal nodes $\{\widetilde{q}_{i1}, \ldots, \widetilde{q}_{is}\}$ (Trials for the center of new basis functions.).
3. Compute the dynamic similarity matrix $K^{(L\tau)}$ (5) based on the temporal set of basis functions $\{\widetilde{\varphi}_{i1}, \ldots, \widetilde{\varphi}_{is}\}$ with
$$\widetilde{\varphi}_{il}(q_k) = \frac{\exp(-\alpha\, d(\tilde{q}_{il}, q_k)^2)}{\sum_{j=1}^{s} \exp(-\alpha\, d(\tilde{q}_{ij}, q_k)^2)}, \quad l = 1, \ldots k.$$
4. Select from the trial nodes the ones for which
$$K_{ii}^{(L\tau)} > \rho, \quad 1 - \varepsilon \sim \rho < 1, \quad i = 1, \ldots, k.$$
where $\varepsilon > 0$ close to zero.

5. Replace the primal center node q_i of the basis function φ_i, by all accepted trial nodes.

After a successful iteration, the complete partitioning is recomputed based on the updated list of nodes, and the above algorithm is applied again to all newly added basis functions. The iteration continues until no new basis functions are added. The resulting transition probability matrix $K^{(L\tau)}$, can now be used to detect metastabilities in the set of basis functions, i.e. calculating a coarse grained transition probability matrix P_c by applying spectral clustering (PCCA+) as described in Section IV. We would like to give some detailed comments on the proposed algorithm. If there are metastabilities within the basis function, k-means will probably deliver center points in different metastable regions. It might happen that the points selected by the k-means routine represent molecular configurations with low statistical weights. Therefore, the objects closest to the selected k-means center points are selected as nodes for the temporal basis functions. Therefore it remains to be checked whether the clusters proposed by the k-means algorithm really separate different metastable sets. The k-means algorithm will always deliver a local partitioning into s clusters independent of the actual amount of metastabilities covered by the basis function. Only in this case the basis function will be refined. For this purpose, we consider the dynamic similarities between the temporal basis functions. A new node q_{il} generated by the k-means algorithm will be only accepted if it's temporal basis function has a self similarity larger than a certain threshold ρ.
To show the influence of the threshold ρ on the number of basis functions, we performed simulations on another synthetic 2D data set with three different thresholds (Fig 3). The closer ρ approximates one, the fewer basis functions are needed and the increase of the number of basis functions is smaller than for lower thresholds.

In order to interpret the entries of matrix $K^{L\tau}$ as transition probabilities we have to minimize the overlaps between the basis functions (Voronoi tessellation). This can be accomplished by setting $\alpha \to \infty$ such that each basis function φ_i becomes an indicator function ,i.e. $\varphi_i = \mathbf{1}_{A_i}$ where

$$\mathbf{1}_{A_i}(q) := \begin{cases} 1 & \text{if } q \in A_i \\ 0 & \text{otherwise} \end{cases} .$$

Thus the set A_i corresponds to basis function φ_i. This allows us now to compute the transition probabilities independently of the shape parameter α which leads to the (set

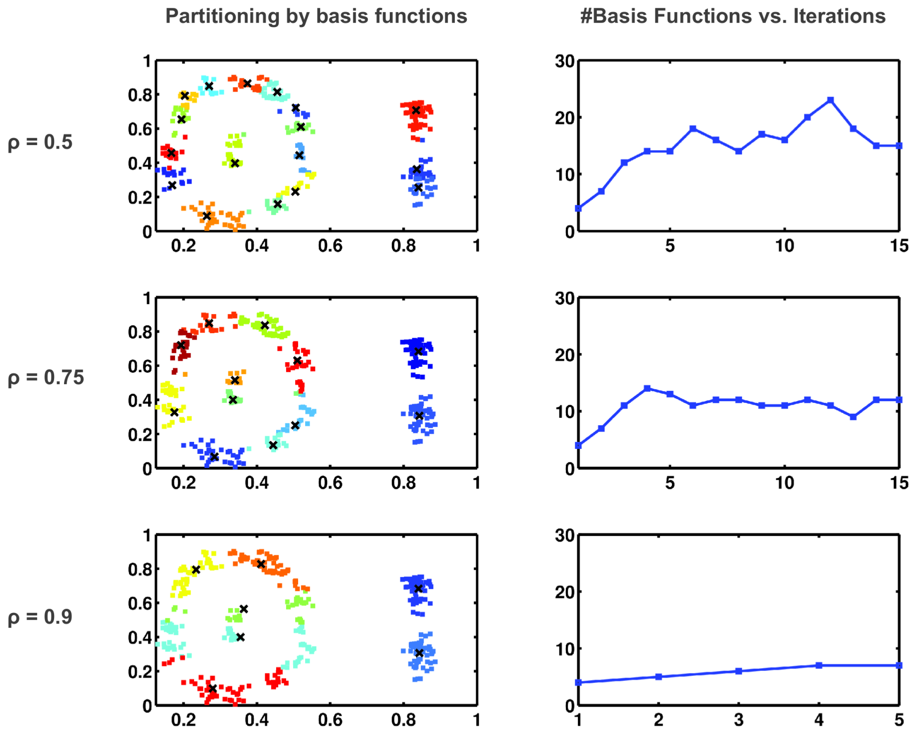

FIG. 3. *Left:* Data set partitioned with three different thresholds. Black crosses are the center nodes of the basis functions. For threshold $\rho = 0,5/0.75/0.9$ we obtained 17/11/7 partitions. *Right:* Number of basis functions in dependence of the number of iterations.

based) transition probability matrix P_{ij}^{τ} :

$$P_{ij}^{\tau} \approx \frac{\#[q_k \in A_i, q_{k+1} \in A_j]}{\#[q_k \in A_i]}, \quad i,j = 1, \ldots N. \tag{6}$$

On the basis of matrix P^{τ} we are now in a position to describe the metastabilities as linear combination of the basis functions $\{\varphi_i\}_i$, i.e. each metastability C_J as a linear combination of the basis functions $\{\varphi_i\}_{i=1}^{n}$:

$$C_J(q) = \sum_{i=1}^{n} G_{iJ}\phi_i(q), \qquad J = 1, ..., n_c. \tag{7}$$

The matrix G relates the basis functions $\{\varphi_i\}_i$ to the conformations $(C_J)_J$, i.e. we seek for a linear combination of the coefficients $g_J = [G_{1J}, G_{2J}, ..., G_{nJ}]$ such that the dynamics of the system shows a metastable behavior. More precisely, the metastability criterion can be given by

$$P^{\tau}\mathbf{g}_J \approx \mathbf{g}_J. \tag{8}$$

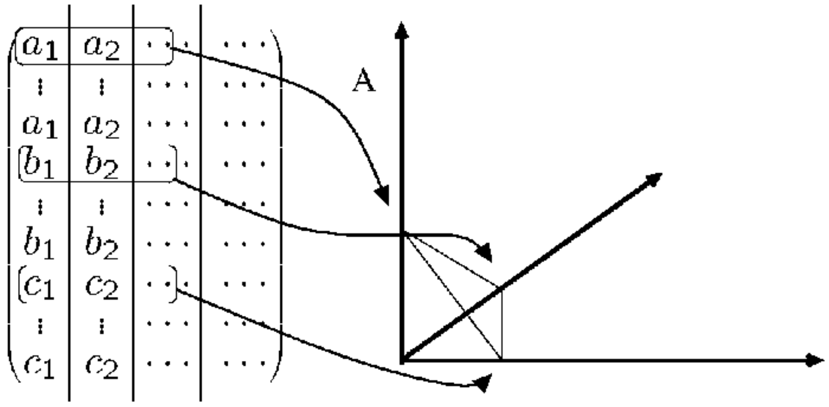

FIG. 4. Eigenvectors $(a_1, ..., a_1, b_1, ..., b_1, ...)^T, (a_2, ..., a_2, b_2, ..., b_2, ...)^T ...$ of the transition matrix P. The rows of these eigenvectors $(a_1, a_2, a_3, ...), (b_1, b_2, b_3, ...)$ are piecewise constant and can be interpreted as vertices of a simplex.

IV. SPECTRAL CLUSTERING BY PCCA+

Having introduced the description of the metastable sets as linear combinations of the sets $\{A_i\}_i$ by (7) and the metastability criterion by (8), we now aim at a coarse grained matrix P_C giving the transition probabilities between the metastable sets, which can be described as a linear combination of the $(A_i)_i$. In earlier works, e.g. [22], the degree of membership of each set (A_i) to a metastable state was confined to either one (membership) or zero (no membership). This condition could be relaxed [6] and is briefly presented in the following.

In case of a decomposable Markov chain or, equivalently, a disconnected similarity graph, an appropriate permutation of objects according to their connectedness results in a block-diagonal matrix P^τ with n_C blocks. This matrix has an n_C-fold eigenvalue $\lambda = 1$. The corresponding eigenvectors $X = [x_1, \ldots, x_{n_C}]$ are piecewise constant on the blocks and can thus be used to identify the clusters. In fact, the rows of X can be considered as vertices of an $(n_C - 1)$-dimensional simplex. Every object can be assigned to one of the n_C vertices and thus to one of the n_C clusters (cf. [6]). Generally, the matrix P^τ constructed from practical data is not decomposable. However, if there are n_C hidden clusters, P has a cluster of eigenvalues $1 = \lambda_1 > \lambda_2 > \ldots > \lambda_{n_C} > 1 - \varepsilon$ near the Perron eigenvalue $\lambda_1 = 1$ [6, 23].

The rows y_i of the corresponding eigenvectors still nearly form a simplex. Since the first eigenvector is always constant, the rows can be considered as vertices of a $(n_C - 1)$-

dimensional simplex, cf. Fig. 4.

The goal of PCCA+ is to identify the vertices of a simplex σ_{n_C-1} such that all points y_i are located within the simplex. Then every point y_i can be assigned to one of the n_C vertices and thus to one of the n_C clusters by a certain membership vector $\mathbf{g}_J = [G_{1J}, ..., G_{nJ}]$.

The identification of such a simplex is equivalent to finding a non-singular transformation matrix $\mathcal{A}$ such that

$$G = X\mathcal{A}$$

and

(1a) $G_{ij} \geq 0 \quad \forall i \in \{1, \ldots, n\}, j \in \{1, \ldots, n_C\}$
(positivity),

(1b) $\sum_{J=1}^{n_C} G_{iJ} = 1 \quad \forall i \in \{1, \ldots, n\}$
(partition of unity).

Among the feasible transformation matrices we search for a matrix $\mathcal{A}$ such that the resulting membership vectors $\mathbf{g}_I$ are as metastable as possible. Metastability is expressed by the fact that the diagonal elements of P_C are as close as possible to 1 (the probability to leave a metastable set, given by the sum of the off-diagonal elements, is as low as possible). It has been shown that instead of maximizing the metastability a maximization of the crispness of the membership functions is also possible [24]. This aims at a clustering which also allows for an interpretation of P_C as a Markov Chain. Crispness means that the columns $\{G_{:j}\}_{j=1,\ldots,n}$ should be as close to indicator vectors as possible (crispness). We can measure this crispness by

$$I(\mathcal{A}; X, \pi) = \frac{1}{n_C} \sum_{I=1}^{n_C} \frac{\langle \mathbf{g}_I, \mathbf{g}_I \rangle_\pi}{\langle \mathbf{g}_I, e \rangle_\pi} \leq 1, \tag{9}$$

where e denotes the vector with all entries equal to 1. The closer $I(\mathcal{A}; X, \pi)$ to one the more crisp is the decomposition into metastabilities. In the PCCA+ algorithm this is achieved by maximizing the objective function $I(\mathcal{A}; X, \pi)$. One has to maximize a convex function with linear constraints, which is not a trivial task. However, the optimization problem can be solved by the Nelder-Mead [25] algorithm provided that a good initial guess for $\mathcal{A}$ is available. This starting guess is obtained by the *inner simplex algorithm* as described in [26]. Once the membership functions $\mathbf{g}_i$ have been computed, one can compute a coarse grained transition probability matrix P_c by projecting the original matrix P^τ onto the metastabilities [27],

$$P_c = (G^\top \pi_D G)^{-1} G^\top \pi_D P G = \mathcal{A}^{-1} \Lambda \mathcal{A}, \tag{10}$$

where π_D denotes a diagonal matrix with the vector π on the diagonal and Λ denotes a diagonal matrix with the eigenvalues $\lambda_1, \ldots, \lambda_{n_C}$ on the diagonal. The matrix P_c is not necessarily a stochastic matrix because it can get negative entries when the membership functions χ_i are far from being characteristic. However, P_c has row sum one and is the correct propagator for densities restricted to the metastabilities [27]. In contrast to P^τ, a set-based transition matrix P_c preserves the Markov property in a "better way": Under the assumption that the trajectory reaches equilibrium within a conformation (metastable subset) before exiting from it, the probabilities of transitions to any other conformation is independent of all but the previous conformation, i. e., there is no memory of earlier conformations. Only if this condition is met, the dynamics can be modeled by a Markov chain which allows for long time simulations [15, 28–31]. For critical remarks on the validity of such models see Ref. [32].

Since the number of clusters n_C is unknown in advance, it is recommended to run the cluster algorithm several times with different input values for n_C and to choose the "best" solution. Since $I(\mathcal{A}; X, \pi) \leq 1$, we choose the number n_C for which $I(\mathcal{A}; X, \pi)$ is maximal.

V. EXAMPLE

A. Geometric Clustering of Simple 2D Examples

A common practice to characterize newly derived clustering algorithms is to use simple two-dimensional data sets. In contrast to complex high-dimensional data sets, artificial 2D examples can be directly represented in terms of two dimensional scatter plots, which is particularly useful for the comparison of different cluster algorithms. To evaluate the presented adaptive spectral clustering (ASC) algorithm we applied it to several classification problems and compared the results with the k-means (KM) [33] and the common-nearest-neighbor-cluster (CNN) [34] algorithm, which is a modified variant of the Jarvis-Patrick algorithm [35]. It is based on the local data-point density around a certain point i. In contrast to the original Jarvis-Patrick algorithm, the local density is measured by the number of common nearest neighbors within a certain cut-off distance from that point i. All three clustering algorithms have been applied with varying parameters to each of the synthetic data sets to gain optimal results for any of the algorithms. For the CNN, both parameters, the *nearest-neighbor-number cutoff* and the nearest-neighbor-distance cutoff, had to be defined by the user prior to clustering. With regard to the adaptive spectral clustering, the threshold ρ was kept fixed at 0.9.

We created five synthetic 2D data sets, that represent common geometrical classification problems. The data sets have been initially seeded with 5 nodes, and extended to 15 - 25 basis functions by adaptive partitioning. Based on the soft partitions, the data sets were clustered by PCCA+. Note, that the classification of these examples is solely based on geometric similarity (4). A combination of dynamic and geometric similarity will be presented in the following section when applying our algorithm to a conformational analysis of a tripeptide molecule. The results of each clustering algorithm applied to the test data sets are shown in Fig. 5. All three cluster algorithms successfully clustered the first data set. For the remaining ones, CNN and ASC gave similar results, whereas k-means could not resolve the underlying clusters.

Obviously our adaptive spectral clustering algorithm is capable of handling typical geometrical classification problems, like spherical shapes as well as elongated structures. More importantly, for all test cases our adaptive partitioning scheme decomposed the state space, such that all hidden clusters could by successfully identified by the subsequent clustering algorithm regardless of shape and structure. We thus receive a soft partition of the state space, that sufficiently covers all clusters/metastabilities with a minimal set of membership basis functions. The obtained set of basis functions can be subsequently used for geometric clustering, as done here with simple 2D examples; or for dynamic clustering as discussed in Sec. III and exemplified in the next section. Thereby, the clustering is performed on $n << N$ basis functions, instead of a complete similarity matrix $N \times N$, that is typically needed for spectral clustering, with N and n being the number of states and basis functions, respectively, c.f. Eqs. (1) - (4). Hence, the adaptive spectral clustering has a significantly reduced computational complexity, while obtaining the same or even slightly better results compared to other established clustering methods, like k-means or CNN.

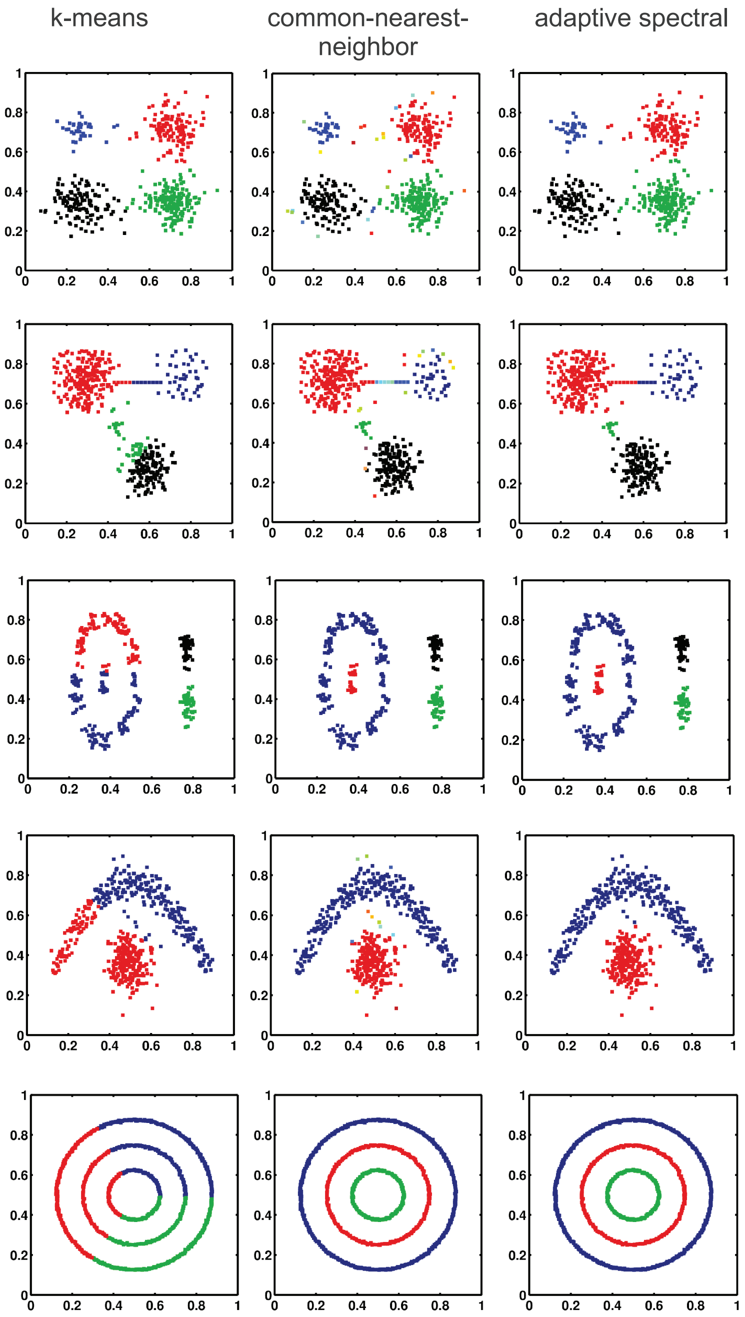

FIG. 5. Results of three different clustering algorithms on various sets of 2D test data. Left: k-means (KM). Middle: common-nearest-neighbor (CNN). Right: adaptive spectral clustering algorithm (ASC) based on geometric similarity. The color of the data points indicates the assigned cluster memberships. Note that for ASC the color of the points indicates the cluster with the highest degree of membership.

B. Application: Conformations of model tripeptide

a. Choice of model system. As another application of the clustering algorithm, we study the conformational dynamics of ZAibProNHMe (benzyloxycarbonyl–aminoisobutyryl-L-prolyl-methylamide) tripeptide molecule (see Fig. 6) as a model system for the adaptive algorithms introduced above. For this molecule a relation between the conformational structures and their mid–IR spectra has been established previously by means of density functional theory (DFT) and normal mode calculations [36, 37] and also preliminary work on adaptive spectral clustering has been published in [38]. In addition, the reason for this choice is that sequences of the rare amino acid Aib (α aminoisobutyric acid) and Pro (proline) are of considerable pharmaceutic interest as β sheet breakers in antibiotic peptides [39, 40]. For example, an Aib-Pro sequence occurs at the amino terminal of alamethicin, an antibiotic produced by trichoderma fungui, which can act as a voltage-dependent ionophore in cell membranes. Aib-Pro sequences are also found in other peptaibols which are used to reduce bacteria and fungal plant pathogens in the soil [41]. The combination of the two methyl groups (in Aib) and the steric restrictions introduced by the pyrrolidine ring (in Pro) cause a strong competition between γ (C_7 ring) and β (C_{10} ring) turn structures in ZAibProNHMe [39], see Fig. 6, which are found at similar energies [36–38].

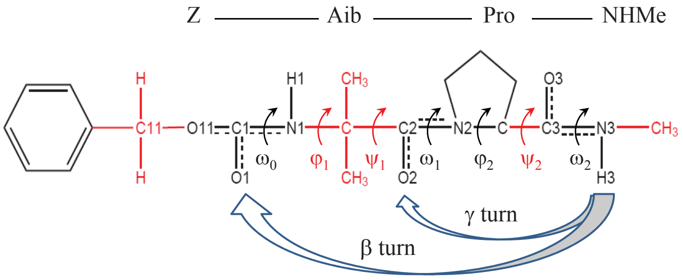

FIG. 6. Primary structure of ZAibProNHMe model peptide including definition of dihedral angles. [37]

b. Minimum energy structures In the present work the ZAibProNHMe peptide *in vacuo* is modeled in terms of the MMFF force field [42, 43]. The parameterization is achieved with the help of the tool EPOS, which is a part of the amiraMol libraries [44]. As a first step to characterize the conformational landscape, minimum energy structures have been obtained using the conjugate gradient method [45], starting from the minimum energy conformations of our previous DFT calculations [37]. Our results are given in Tab. I where we use the notation of Refs. [36, 37]. The γ (A) and β (I, II') turn structures differ mainly in the ψ_2 angle while up- and down-puckering of the pyrrolidine ring (U, D) can be distinguished from the values of the angles χ_1, χ_2. Furthermore, the various A structures (A1, A2, ...) differ essentially in their Aib orientations specified by torsion angles ϕ_1, ψ_1 where primed and unprimed structures denote sign changes of those angles. All major classes of conformations found in our previous quantum chemical DFT based calculations (see Tab. I and supplementary material of Ref. [37]) also represent local minima of the MMFF model, with very similar values of the dihedral angles. Even the relative energies are in most cases within a few kJ/mol from the previous DFT results, as indicated in the last column of Tab. I. Notable exceptions are the II'bU and the A3bU, A5bD, A5'bD, and A6'bD structures where the MMFF energies are more than 10 kJ/mol higher than the corresponding DFT values. Of particular importance is a rather broad basin encompassing A1bD, A2'bD, IbU, and IbD conformations, the first and last of which represent the global minima of the DFT and MMFF potential energy surface. Within that basin, interconversion between γ and β turn structures is expected to accessible at relatively low energies. In addition to the A-type conformers with all peptide bonds being in trans position, also D-type conformers with the three ω angles being cis-trans-cis are found in our $T = 900$ K trajectory discussed below. They are not included in our Tab. I because all ϕ, ψ Ramachandran angles deviate by no more than a few degrees from those of the corresponding A-type structures.

c. Clustering of trajectory data The clustering methods introduced in the previous sections shall be illustrated here for a molecular dynamics simulation of the ZAibProNHMe model peptide. To this end, a 39 ns trajectory is generated using the ZIBgridfree software package [46]. The size of the time step is $h = 1.3$ fs and we use every 60th step of the trajectories in the further data processing, i.e. $\tau = \tilde{n}h = 78$fs. We employ the Nose-Hoover algorithm [47, 48] to approximately sample a canonical (NVT) ensemble for $T =$ 900 K. While we are aware that this temperature is not realistic in peptide chemistry we have chosen this rather high value because the barriers between different conformations of a peptide are typically much higher in the gas phase than in aqueous solutions. In addition, when comparing with our 600K trajectory (not shown here) we found that the 900K simulation displays not only more conformational freedom but also a richer hierarchy of conformations which renders this case more challenging for our clustering algorithms.

Conformation	ϕ_1	ψ_1	ϕ_2	ψ_2	χ_1	χ_2	ΔE_{MMFF}	ΔE_{DFT}
A1bD	176	174	-80	75	30	-38	2.2	0.0
A2bD	60	46	-81	76	30	-38	3.9	2.9
A2'bD	-58	-44	-80	72	34	-36	0.1	1.9
A2'bU	-60	-41	-74	77	-11	30	10.5	1.7
A3bD	70	-167	-81	72	34	-36	12.4	9.9
A3bU	70	-167	-75	76	-13	31	24.3	13.1
A3b'D	-73	169	-81	75	31	-37	10.1	11.8
A4bD	-172	52	-79	73	31	-39	6.3	11.0
A4b'D	173	-52	-79	76	30	-38	7.0	10.2
A5bD	77	-101	-83	68	34	-38	28.5	13.1
A5b'D	-75	110	-81	76	32	-37	26.7	11.5
A6bD	-139	53	-79	77	30	-38	12.7	10.9
A6b'D	118	-50	-81	75	33	-36	22.8	11.8
IbD	-56	-40	-81	-9	32	-37	0.0	1.9
IbU	-57	-34	-69	-24	-22	36	2.9	2.2
II'bD	72	-169	-83	51	36	-37	12.9	11.7
II'bU	65	-149	-69	-18	-28	37	24.3	15.5

TABLE I. Minimum energy structures for ZAibProNHMe model tripeptide from MMFF force field: Dihedral angles (ϕ, ψ, χ in degrees) and relative energies (ΔE in kJ/mol). A$_n$ ($n = 1, 2, \ldots$, indicating different Aib orientations) are all–trans ($\omega_{0,1,2} \approx 180$) conformations of γ turn structures, where the prime denotes an inversion of the signs of ϕ_1 and ψ_1. Classes I and II' are β turn structures. U,D indicate up- and down puckering of the Pro ring. In all cases, the Z–cap is in b orientation. For comparison, DFT results from Ref. [37] are shown in the last column.

The clustering techniques are applied to the time series of the six torsional coordinates $\omega_0, \phi_1, \psi_1, \omega_1, \psi_2, \omega_2$, see Fig 6. We omit here the Z–cap orientation as well as the ring puckering, are chemically relevant partly to keep our model calculations not unnecessarily complicated, but also because these degrees of freedom are essentially independent of the other backbone torsional angles, see our previous work [37]. Furthermore, it is noted that

ϕ_2 is essentially blocked inside the Pro ring, see Fig. 6.

The extraction of torsional angles from the molecular trajectory and the subsequent analysis by means of adaptive, spectral clustering has been carried out by our software package ”MetaStable” which is available via the SourceForge web site [49]. In the first step we examined the influence of the threshold ρ (section III) on the number of basis functions for the 900 K trajectory by setting the maximum number of iterations to three and observing the number of basis functions (Fig 7). We started with 40 seed nodes in a Voronoi tessellation and used a time lag of $L\tau = 20 \times 78$ fs. As expected, the lower threshold leads to more basis functions which is in good agreement with the results from section III. As can be seen in the Figure 7 altering the threshold from 0.4 to 0.5 reduces the number of basis functions drastically, since with larger threshold the criterion for generating a new function becomes more demanding. We also computed the second largest eigenvalue of the transition matrix P_c, as an indicator for the inherent slowest time scale in the dynamics. No clear trend of increasing or decreasing of λ_2 in dependence of the threshold or number of basis functions could be observed. However, the second largest eigenvalue decreases with longer lag time $L\tau$ as shown in Figure 7. This result is in good agreement with theory since for a lag time which equals the original time step τ of the trajectory, each state would be a metastable state and thus $\lambda_2 = 1$.

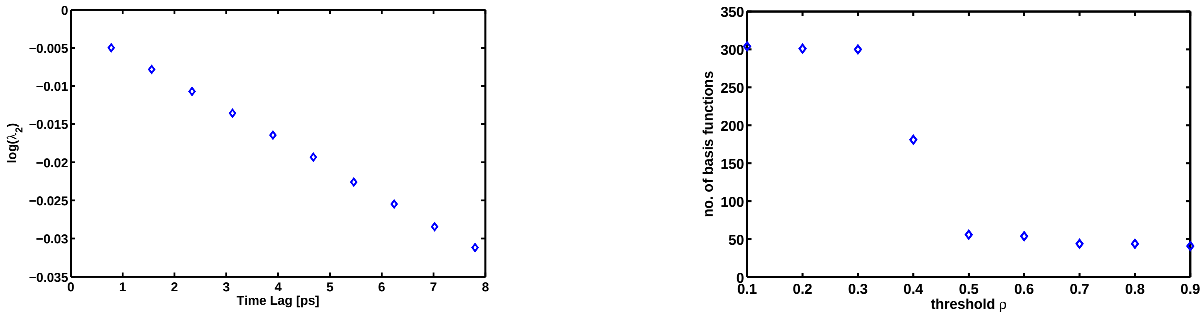

FIG. 7. *Left:* Log of the second largest eigenvalue of the transition matrix of ZAibProNHMe according to the time lag $L\tau$. The slope of this graph is the implied time scale of the system, which is about 230 ps. *Right:* Comparing the threshold ρ with the number of basis functions for ZAibProNHMe at 900 K with 40 initial seeds and a time lag $L\tau = 20 \times 0.078$ps.

In the second step, the spectral clustering technique is applied to approximately determine the eigenvectors of the transfer operator and hence detect the metastable regions of the conformational space spanned by the ZAibProNHMe torsion angles. Here we use a partitioning of the peptide's conformational space generated for a time lag of $L\tau = 64 \times 0.078 \approx 5$ ps. Starting from 42 initial Voronoi seed functions, the basis is adaptively refined leading to 153 basis functions after 3 iteration steps. Subsequently, we perform the metastability analysis by means of the PCCA+ technique. The spectrum of the corresponding transition matrix $P\tau$ is shown in the upper part of Fig. 8. The second eigenvalue, $\lambda_2 = 0.97887$, implying a time scale of 234 ps, characterizes the slowest dynamics. Separated by a small spectral gap, the following eigenvalues $\lambda_3 \dots \lambda_9$ are found between 0.92 and 0.72, with time scales between 60 and 15 ps. After another small gap, the eigenvalues $\lambda_{10} \dots \lambda_{17}$ are lying between 0.67 and 0.56, with time scales between 12 and 9 ps. After yet another, very pronounced gap, the remaining eigenvalues are below 0.4, with time scales of 5 ps and below.

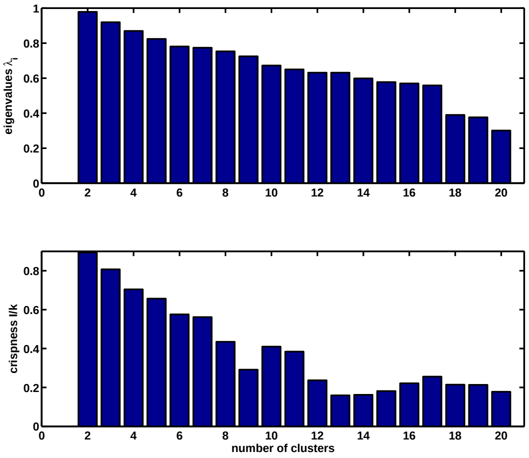

FIG. 8. Spectral clustering of 900K / 39 ns trajectory for ZAibProNHMe peptide by PCCA+ technique. Upper part: Spectrum of transition matrix P_C. Lower part: Objective function / crispness of decomposition into metastable sets. For clarity, only the first twenty states are shown.

Next, we consider the metastability criterion i.e. the objective function $I(\mathcal{A}; X, \pi)$ versus number of clusters, cf. (9). It can be seen that this indicator has a decreasing tendency. The two local maxima at $n_C = 10$ and $n_C = 17$ are based on the fact, that a decomposition into $n_C = 9$ or $n_C = 13, 14, 15$ appears to be unfavorable. Note that the maxima of the objective function are approximately (but not exactly) coinciding with the spectral gaps mentioned above. First, let us consider the case of two clusters which represents the most metastable decomposition. Inspection of the time series of the torsional coordinates reveals that in the major cluster all three peptide bonds are in *trans* position, $\omega = \pm 180$ deg, corresponding to the all–*trans* structures of class A and I listed in Tab. I. The minor cluster contains class D conformations where the first and third peptide bonds are in *cis* position, $\omega_0 \approx \omega_2 \approx 0$, $\omega_1 \approx \pm 180$ deg. Note that these conformations do not play a role for peptides at room temperature but are found here due to the rather high temperature ($T = 900$ K) of our test calculations. The weights of the two clusters are 0.887 and 0.113 which corresponds to a free energy difference of about 15 kJ/mol (by simple Boltzmann inversion). The lifetimes of the type D structures are on the order of a few 100 ps, thus qualitatively agreeing with the implicit time scale inferred from the second eigenvalue of the transition matrix. When choosing a decomposition into three clusters, the type D cluster splits up into two clusters with weights 0.084 and 0.029. While the former one still encompasses several, unresolved D structures, the latter one is essentially centered around the D2 local minimum energy structure (numbering of Aib orientations in analogy to that of the A structures as given in Tab. I). When choosing four clusters, the former D-cluster spawns off a cluster around the D2' minimum, with a statistical weight of 0.012 only. When further increasing the number of clusters, also the major cluster encompassing the all–*trans* structures decays into sub-clusters.

A typical case is the result for 10 clusters given in Fig. 9 where histograms of the most important dihedral angles $(\omega_0, \phi_1, \psi_1, \psi_2)$ of the peptidic backbone are shown. As can be seen from the distribution of ω_0 angles in the upper part of the figure, the five leading (and the tenth) clusters have all their peptide bonds in *trans* positions while for the remaining ones ω_0 (as well as ω_2, not shown) are in *cis* position. Their weights sum up to 0.889 and 0.111, almost in coincidence with the results for only two clusters, which again confirms that the *trans*(A)–*cis*(D) flipping of the planar peptide bonds arrangements gives rise to the main metastability, i. e., the one with the longest implicit time scale. The lower part of Fig. 9 reveals that all ten clusters exhibit rather broad distribution of ψ_2 angles, encompassing both the regimes around $\psi_2 \approx 80$ (γ turn, type A, D) and $\psi_2 \approx 0$ (β turn, type I) so that these classes cannot be uniquely resolved on the basis of the present clustering of the torsional degrees of freedom. However, cluster #5, preferentially (but not exclusively) located in the type I regime, presents the only exception. In contrast, the assignment of the cluster memberships based on the Ramachandran angles ϕ_1, ψ_1 is essentially clear. While the leading all–*trans* cluster #1 is still delocalized, clusters #2, #3, #4 can be assigned to A2', A4' and A2 structures, respectively, where #2 appears to contain also type I structures. This observation that A2' and I cannot be clearly distinguished in our clustering procedures is in agreement with the broad basin and low barriers in the potential energy surface [37]. A similar picture arises for the *cis* structures (type D). While cluster #6 corresponds to an unresolved mixture of several D structures, it is straightforward to assign clusters #7, #8, #9 to minimum energy structures D2, D4', and D2', respectively.

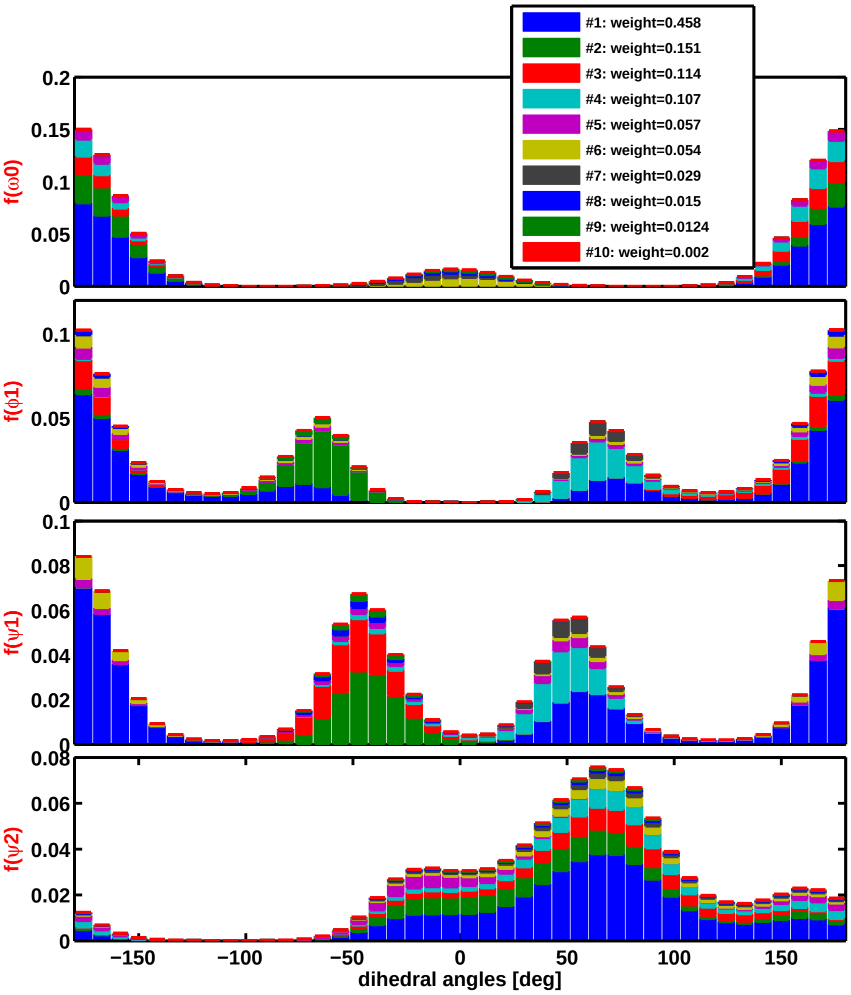

FIG. 9. Histograms of selected dihedral angles from 900K / 39 ns trajectory for ZAibProNHMe peptide, decomposed into 10 clusters. From bottom to top: $\omega_0, \phi_1, \psi_1, \psi_2$. Coloring indicates membership to the clusters, with weights given in the legend.

Although not explicitly included in our metastability decomposition of the (torsional!) state space, it is also instructive to look at the histogram of the O–H distances characterizing the formation of γ or β turns through hydrogen bonds by closing 7– or 10–membered rings, respectively. Fig. 10 shows that most of the ten conformations detected in the metastability analysis of the torsional angles display wide distributions, encompassing both H–bonded (≈ 0.2 nm) as well as non–bonded situations. Nevertheless, a few tendencies can be seen in the upper part of that figure: Out of the all–*trans* clusters, #2 can form β turns, while #3, #4 as well as the type D (*cis*) conformations (#6...#9) are incompatible with this secondary structure element. The situation for the formation of γ turns is even less clear, see lower part of Fig. 10. While clusters #1...#8 do not exhibit clear preferences, only clusters #5 and #10 are found at rather large O...H distances of 0.6 nm where H–bonding can be safely ruled out.

Finally, it is mentioned that a further refinement of the decomposition beyond the case of ten clusters displayed in Fig. 9 and Fig. 10 does not necessarily lead to more detailed information. We investigated the situation for 17 clusters (local maximum of objective function in lower part of Fig. 8) and found an essentially unchanged picture. The important conformations are centered at the same potential minima as in the 10 cluster analysis, with the only exceptions of two new clusters centered in the A4 and D4 regions. All additional clusters bear statistical weights below 0.001 and are thus of no statistical significance.

In summery, our scheme clearly reveals relations between the identified metastable clusters and minimum energy structures of the molecular system. Moreover, by changing the number of clusters , n_C, a hierarchy of clusters has been identified. A coarse clustering only shows basins separated by high free energy barriers, while a fine clustering resolves more and more local minima of the PES.

VI. CONCLUSION

For high-dimensional data sets containing many single data points an adaptive clustering approach is proposed. This means, that the high-dimensional space is decomposed into subsets and these subsets are assigned to different clusters. The decomposition has to be fine enough to resolve the barriers between the clusters and coarse enough to provide locally enough statistical data to discriminate between densely populated and sparsely populated regions.

The main idea of our adaptive approach is to decide, whether a given subset of the data

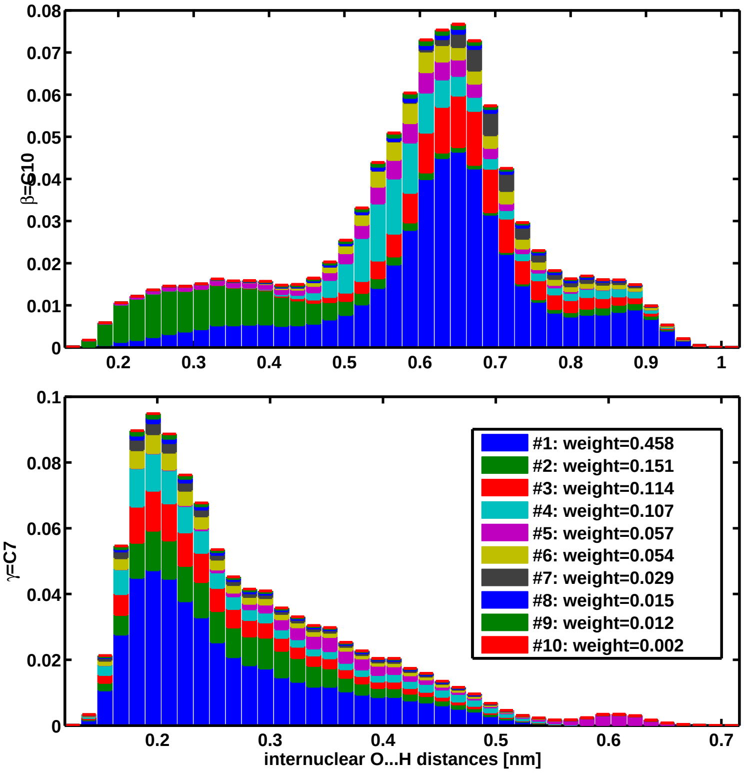

FIG. 10. Histograms of selected internuclear distances from 900K / 39 ns trajectory of ZA-ibProNHMe peptide, decomposed into 10 clusters. O...H distance corresponding to formation of a β turn (upper) and a γ turn (lower part). Coloring indicates membership to the clusters, with weights given in the legend.

space has to be refined or not. Our approach is thus based on a discretization of the state space the main problem of which is the curse of dimensionality. While a method which, e. g., is based on a systematical splitting of the space along its coordinates, would suffer from this, our method circumvents a coordinate based splitting by using internal distances only. Solely the total number of sampled states, their pairwise distances, and the number of metastabilities determine its run time.

ACKNOWLEDGMENTS

Financial support by the Deutsche Forschungsgemeinschaft (DFG) through project A19 of the DFG research center MATHEON is acknowlegded.

[1] M. Dellnitz and O. Junge, SIAM J. Numer. Anal. **36**, 491 (1999).

[2] C. Schütte, Habilitation thesis, Department of Mathematics and Computer Science, Freie Universität Berlin (1999).

[3] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte, Lin. Alg. Appl. **315**, 39 (2000).

[4] P. Deuflhard, in *Trends in Nonlinear Analysis*, edited by M. Kirkilionis, S. Krömker, R. Rannacher, and F. Tomi (Springer Berlin, 2003), pp. 269–287.

[5] M. Karpen, D. J. Tobias, and C. Brooks, Biochemistry **32**, 412 (1993).

[6] P. Deuflhard and M. Weber, in *Lin. Alg. Appl. – Special Issue on Matrices and Mathematical Biology*, edited by M. Dellnitz, S. Kirkland, M. Neumann, and C. Schütte (Elsevier, 2005), vol. 398, pp. 161 – 184.

[7] W. Ren, E. Vanden-Eijnden, P. Maragakis, and W. E, J. Chem. Phys. **123**, 134109 (2005).

[8] A. K. Faradjian and R. Elber, J. Chem. Phys. **120**, 10880 (2004).

[9] V. J. Bhute and A. Chatterjee, The Journal of Chemical Physics **138**, 084103 (2013).

[10] G. R. Bowman, J. Chem. Phys. **137**, 134111 (2012).

[11] C. Schütte, F. Noe, J. Lu, M. Sarich, and E. Vanden-Eijnden, J. Chem. Phys. **134**, 204105 (2011).

[12] E. Vanden-Eijnden and F. A. Tal, J. Chem. Phys. **123**, 184103 (2005).

[13] E. Vanden-Eijnden, M. Venturoli, G. Ciccotti, and R. Elber, J. Chem. Phys. **129**, 174102 (2008).

[14] K. Fackeldey, A. Bujotzek, and M. Weber, in *Meshfree Methods for Partial Differential Equations VI* (Springer Berlin, 2012), vol. 89 of *Lecture Notes in Computational Science and Engineering*, pp. 141 – 154.

[15] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, J. Chem. Phys. **126**, 155101 (2007).

[16] M. Lin, J. Zhang, H.-M. Lu, R. Chen, and J. Liang, J. Chem. Phys. **134**, 075103 (2011).

[17] A. T. Hawk, J. Chem. Phys. **138**, 154105 (2013).

[18] Y. Yao, R. Z. Cui, G. R. Bowman, D.-A. Silva, J. Sun, and X. Huang, J. Chem. Phys. **138**, 174106 (2013).

[19] B. Keller, X. Daura, and W. F. van Gunsteren, J. Chem. Phys. **132**, 074110 (2010).

[20] K. Fackeldey, S. Röblitz, O. Scharkoi, and M. Weber, Tech. Rep. 11-27, ZIB, Takustr.7, 14195 Berlin (2011).

[21] D. Shepard, in *Proc. 23rd ACM Nat. Conf.* (Brandon/Systems Press, Princeton, 1968), pp. 517 – 524.

[22] P. Deuflhard, M. Dellnitz, O. Junge, and C. Schütte, in *Computational Molecular Dynamics: Challenges, Methods, Ideas*, edited by P. Deuflhard, J. Hermans, B. Leimkuhler, A. E. Mark, S. Reich, and R. D. Skeel (Springer Berlin, 1999), vol. 4 of *Lecture Notes in Computational Science and Engineering*, pp. 98–115.

[23] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, J. Comp. Phys. **151**, 146 (1999).

[24] S. Röblitz and M. Weber, Advances in Data Analysis and Classification **7**, 147 (2013), ISSN 1862-5347.

[25] J. Nelder and R. Mead, Computer Journal **7**, 308 (1965).

[26] M. Weber and T. Galliat, ZIB-Report 02-12, Zuse Institute Berlin (2002).

[27] S. Kube and M. Weber, J. Chem. Phys. **126**, 0241203 (2007).

[28] S. P. Elmer, S. Park, and V. S. Pande, J. Chem. Phys. **123**, 114902 (2005).

[29] S. P. Elmer, S. Park, and V. S. Pande, J. Chem. Phys. **123**, 114903 (2005).

[30] N. Singhal, C. D. Snow, and V. S. Pande, J. Chem. Phys. **121**, 415 (2004).

[31] F. Noe, I. Horenko, C. Schütte, and J. C. Smith, J. Chem. Phys. **126**, 155101 (2007).

[32] D. Nerukh, C. H. Jensen, and R. C. Glen, J. Chem. Phys. **132**, 084104 (2010).

[33] J. B. MacQuee, in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press, 1967), pp. 281–297.

[34] M. Goswami, R. Sarmah, and D. K. Bhattacharyya, Int. J. Comput. Vision Robot. **2**, 115 (2011).

[35] R. Jarvis and E. Patrick, IEEE Transactions on Computers **C**, 1025 (1973).

[36] I. Compagnon, J. Oomens, G. Meijer, and G. von Helden, J. Am. Chem. Soc. **128**, 3592 (2006).

[37] H. Zhu, M. Blom, I. Compagnon, A. M. Rijs, S. Roy, G. von Helden, and B. Schmidt, Phys.

Chem. Chem. Phys. **12**, 3415 (2010).

[38] F. Haack, S. Röblitz, O. Scharkoi, B. Schmidt, and M. Weber, in *AIP Conference Proceedings* (2010), pp. 1585–1588.

[39] B. V. V. Prasad, N. Shamala, R. Nagaraj, R. Chandrasekaran, and P. Balaram, Biopolymers **18**, 1635 (1979).

[40] A. Aubry, D. Bayeul, H. Brückner, N. Schiemann, and E. Benedetti, J. Pept. Sci. **4**, 502 (1998).

[41] C. Chutrakul, M. Alcocer, K. Bailey, and J. F. Peberdy, Chem. Biodiv. **5**, 1694 (2008).

[42] T. A. Halgren, J. Comput. Chem. **17**, 490 (1998).

[43] T. A. Halgren, J. Comput. Chem. **20**, 730 (1999).

[44] J. Schmidt-Ehrenberg, D. Baum, and H.-C. Hege, in *Proceedings of IEEE Visualization 2002*, edited by R. J. Moorhead, M. Gross, and K. I. Joy (2002), pp. 235 – 242.

[45] R. Fletcher and C. M. Reeves, Computer Journal **7**, 149 (1964).

[46] M. Weber and H. Meyer, Tech. Rep. 05-17, ZIB, Takustr.7, 14195 Berlin (2005).

[47] S. Nosé, J. Chem. Phys. **81**, 511 (1984).

[48] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).

[49] F. Haack, S. Röblitz, B. Schmidt, and M. Weber, *MetaStable: A MATLAB software package for metastability analysis of molecular conformations.*, Available via http://sourceforge.net/p/trajlab/metastable (2012).