

Principal component analysis on a torus: Theory and application to protein dynamics

F SCI

Cite as: J. Chem. Phys. **147**, 244101 (2017); <https://doi.org/10.1063/1.4998259>

Submitted: 29 July 2017 • Accepted: 03 November 2017 • Published Online: 22 December 2017

Florian Sittel, Thomas Filk and Gerhard Stock

COLLECTIONS

F This paper was selected as Featured

SCI This paper was selected as Scilight



View Online



Export Citation



CrossMark

ARTICLES YOU MAY BE INTERESTED IN

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

The Journal of Chemical Physics **149**, 150901 (2018); <https://doi.org/10.1063/1.5049637>

Dihedral angle principal component analysis of molecular dynamics simulations

The Journal of Chemical Physics **126**, 244111 (2007); <https://doi.org/10.1063/1.2746330>

Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates

The Journal of Chemical Physics **141**, 014111 (2014); <https://doi.org/10.1063/1.4885338>



Webinar
Quantum Material Characterization
for Streamlined Qubit Development



Register now



Principal component analysis on a torus: Theory and application to protein dynamics

Florian Sittel, Thomas Filk, and Gerhard Stock^{a)}

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

(Received 29 July 2017; accepted 3 November 2017; published online 22 December 2017)

A dimensionality reduction method for high-dimensional circular data is developed, which is based on a principal component analysis (PCA) of data points on a torus. Adopting a geometrical view of PCA, various distance measures on a torus are introduced and the associated problem of projecting data onto the principal subspaces is discussed. The main idea is that the (periodicity-induced) projection error can be minimized by transforming the data such that the maximal gap of the sampling is shifted to the periodic boundary. In a second step, the covariance matrix and its eigendecomposition can be computed in a standard manner. Adopting molecular dynamics simulations of two well-established biomolecular systems (Aib₉ and villin headpiece), the potential of the method to analyze the dynamics of backbone dihedral angles is demonstrated. The new approach allows for a robust and well-defined construction of metastable states and provides low-dimensional reaction coordinates that accurately describe the free energy landscape. Moreover, it offers a direct interpretation of covariances and principal components in terms of the angular variables. Apart from its application to PCA, the method of maximal gap shifting is general and can be applied to any other dimensionality reduction method for circular data. *Published by AIP Publishing.* <https://doi.org/10.1063/1.4998259>

I. INTRODUCTION

The past years have witnessed an explosion of data. To a large part, this trend is driven by advances in computational algorithms and hardware, which have led to a rapidly rising amount of data in fields such as quantum chemistry, fluid mechanics, or molecular dynamics simulation. The interpretation of these computational results usually requires efficient and systematic strategies to reduce the dimensionality of the problem.^{1–6} Although to this end various nonlinear methods have been suggested,^{5,7} it is often most convenient to use a linear transformation. This includes, for example, various versions of independent component analysis^{1,8,9} as well as the commonly used principal component analysis (PCA).^{2,10,11} While these methods typically assume linear input data, in many applications, the motion of the system is best described using circular coordinates. Examples include rotating actors in tracking and control applications¹² or backbone dihedral angles in proteins.¹³ Due to the periodicity of circular data, however, it is not as straightforward to define means, covariances, or linear projections as in standard PCA.¹⁴

To overcome this problem, several approaches have been proposed,^{15–22} including dihedral angle principal component analysis (dPCA)^{16,17} which applies PCA on the sine- and cosine-transformed dihedral angles of proteins or GeoPCA^{18,19} which invokes hyperdimensional spheres to describe the circular motion. However, both methods have their limitations. While dPCA has been shown to represent

the energy landscape of biomolecules in high resolution,^{23–30} the inherent duplication of coordinates and the nonlinearity of the sine and cosine transformations render it difficult to interpret the results in terms of the underlying observables. In the case of GeoPCA, the description of circular motion as grand circles on hyperdimensional spheres renders it equally difficult to interpret the identified principal components. Given that the locus of a set of circular observables is the torus, moreover, it is in general not adequate to limit the dynamics to grand circles on the embedding hypersphere.

To avoid problems associated with transformations of input data, it appears natural to perform a PCA directly on the torus. This requires a generalization of the standard PCA of data points distributed on a vector space to data points distributed on a Riemannian manifold. Such a generalization was developed (termed principal geodesic analysis²⁰) and was also applied to protein dynamics.²¹ Moreover, the so-called torus PCA was proposed,²² which is based on “deforming” the torus into a sphere. As discussed below in Secs. II and III, which briefly review general theory and previous approaches, the existing methods in practice are plagued by various problems, in particular, the problem of projecting the data onto the principal subspaces.

As a main result of this work, Sec. III presents an alternative way to perform a PCA on the torus. To circumvent the inherent projection problem, a specific data structure is assumed, which allows us to introduce a cut of the angular data at some maximal gap. Although this assumption in general may represent a limitation, it is shown to be well fulfilled for the description of backbone dihedral angles of proteins,¹³ which is the main application under consideration. The proposed transformation of the data is linear; hence, no artificial extra

^{a)}Author to whom correspondence should be addressed: stock@physik.uni-freiburg.de

dimensions or deformations of the underlying probability distribution occur. To demonstrate the virtues and shortcomings of the new method termed “dPCA+,” it is applied in Sec. IV to two well-established model problems, namely, conformational transitions of a Aib peptide helix³¹ and the folding of villin headpiece protein.³² In particular, we show that dPCA+ leads to an accurate characterization of the free energy landscape and the underlying metastable conformational states of the system.

II. THEORY

In this section, we first describe the general idea of PCA in terms of geometrical concepts which can then be generalized to Riemannian manifolds. We then focus on the torus by describing several ways to represent a torus geometrically and how these ways can be used to derive different distance measures for data points distributed on a torus. Moreover, we discuss the problem of projecting data onto the principal subspaces.

A. Geometrical interpretation of PCA on a Euclidean space

Given a set of data $\{\mathbf{x}(n)\}$ ($n = 1, \dots, N$) of N points in \mathbb{R}^D , the usual procedure for a PCA on these data points is the following:

1. Determine the arithmetic mean value of the observables,

$$\langle \mathbf{x} \rangle = \frac{1}{N} \sum_{n=1}^N \mathbf{x}(n). \quad (1)$$

2. Obtain the principal axes. To this end, one determines the covariance matrix

$$\mathbf{C}_{ij} = \frac{1}{N-1} \sum_{n=1}^N (\mathbf{x}(n) - \langle \mathbf{x} \rangle)_i \cdot (\mathbf{x}(n) - \langle \mathbf{x} \rangle)_j, \quad (2)$$

which is a symmetric matrix and can be diagonalized. Solving the eigensystem

$$\mathbf{C} \mathbf{e}^{(k)} = \lambda_k \mathbf{e}^{(k)} \quad (3)$$

yields the directions $\mathbf{e}^{(k)}$ and variances λ_k of principal motion ordered by decreasing variance.

3. Projection of the data onto the principal subspace. Each data point \mathbf{x} is projected orthogonally onto the principal subspace which, based on the eigenvalues, has been chosen to be relevant. Let $\mathbf{e}^{(k)}$ ($k = 1, \dots, d$) be the normalized eigenvectors of \mathbf{C}_{ij} corresponding to the d largest eigenvalues, then the k th principal component V_k , i.e., the projection of \mathbf{x} onto the k th eigenvector, is given by

$$V_k(n) = (\mathbf{x}(n) - \langle \mathbf{x} \rangle) \cdot \mathbf{e}^{(k)}. \quad (4)$$

This procedure requires that the data points are elements of a vector space with a scalar product: The mean value [Eq. (1)] is obtained by a particular linear combination of the data points, the formula for the covariance matrix [Eq. (2)] requires the definition of a scalar product, and the projection [Eq. (4)] requires both.

Geometrically, this procedure can also be interpreted in the following way:

1. First, one determines the point in \mathbb{R}^D for which the sum of the squared distances to all data points is minimal (this is the mean value).
2. In a second step, one constructs a linear subspace (a straight line) passing through this point such that the sum of the orthogonal squared distances of all data points to this line is minimized. Successively one can add further orthogonal subspaces such that the sum of the squared orthogonal distances of the data points to these linear subspaces is minimal, etc. The orthogonal variances, which are minimized, correspond to the variances of the data points which remain *unexplained* by the dimensions of the leading PCA subspaces.
3. The point on the linear subspace which is closest to the data point corresponds to the projection of the data point onto this linear subspace.

This geometrical formulation, which was actually the original approach,³³ requires different concepts: the distance between two points and a proper generalization of the notion of a linear subspace.

B. PCA on a Riemannian manifold

The method of PCA on a Euclidean space has been generalized to Riemannian manifolds (principal geodesic analysis).²⁰ In principle, the method can even be generalized to arbitrary metric spaces. However, some of the uniqueness theorems with respect to the construction will only hold under suitable restrictions. This also refers to Riemannian manifolds: Some of the following constructions may only be well-defined locally, i.e., in an open neighborhood of particular points. We will not always emphasize this explicitly, however, we will come back to this point in the context of a PCA on a torus.

The important structure, which allows us to translate the geometrical ideas of a PCA to Riemannian manifolds, is the metric, i.e., a local definition of a distance between two points. Using the metric, one can assign a length to any path on the manifold. The distance $d(\mathbf{x}_0, \mathbf{x}_1)$ between two points is equal to the length of the shortest path joining these two points. A geodesic is a (locally) minimal path, i.e., a path that is a shortest connection between any two sufficiently close points on the path. It can be obtained as the solution of a differential equation: Given a point \mathbf{x}_0 on a Riemannian manifold and a tangent vector \mathbf{v} at this point, we can construct the geodesic through \mathbf{x}_0 which has the vector \mathbf{v} as its tangent vector at \mathbf{x}_0 (i.e., a constant “velocity”). The solution is a path $\mathbf{x}(t)$ such that $\mathbf{x}(t_0) = \mathbf{x}_0$ and $\dot{\mathbf{x}}(t_0) = \mathbf{v}$. In particular, in the context of Lie groups, this path $\mathbf{x}(t)$ is sometimes called the “exponential mapping” of \mathbf{v} .²⁰

For the following, we also will need the notion of a distance $d(\mathbf{x}, M_S)$ of a point $\mathbf{x} \in M$ from a submanifold $M_S \subset M$ of the Riemannian manifold M ,

$$d(\mathbf{x}, M_S) = \min_{\mathbf{y} \in M_S} d(\mathbf{x}, \mathbf{y}). \quad (5)$$

That is, $d(\mathbf{x}, M_S)$ is the minimal length of a geodesic from \mathbf{x} to a point in M_S (if $\mathbf{x} \in M_S$, this distance is zero). One can easily show that this minimal geodesic from \mathbf{x} to the submanifold M_S is orthogonal to M_S in the point where it meets M_S .

Given the set of data points $\{\mathbf{x}(n)\}$ which can be represented as points of a D -dimensional Riemannian manifold, the general idea of a principal geodesic analysis is the following:²⁰

1. Calculate the mean value $\bar{\mathbf{x}} \in M$ as the point on the manifold M for which the sum of the squared distances to all data points becomes minimal. Note that this definition of the mean is in general different to the arithmetic mean $\langle \mathbf{x} \rangle$ as defined in Eq. (1). Let

$$D_0(\mathbf{x}) = \sum_{n=1}^N d(\mathbf{x}, \mathbf{x}(n))^2 \quad (6)$$

be the sum of the squared distances from an arbitrary point $\mathbf{x} \in M$ to all data points, then $\bar{\mathbf{x}}$ is the point in M which minimizes this function. Generically, this point is unique; however, there are notable exceptions: An example is given by data points that are homogeneously distributed on a circle; in this case, any point on the circle is a mean value.

2. Given $\bar{\mathbf{x}}$ and a tangent vector \mathbf{v} at this point, one can construct the geodesic $\mathbf{x}(\bar{\mathbf{x}}, \mathbf{v}; t)$ with $\mathbf{x}(\bar{\mathbf{x}}, \mathbf{v}; t = 0) = \bar{\mathbf{x}}$ and $\dot{\mathbf{x}}(\bar{\mathbf{x}}, \mathbf{v}; t = 0) = \mathbf{v}$. Each such geodesic defines a one-dimensional submanifold $M_1(\bar{\mathbf{x}}, \mathbf{v})$ of the Riemannian space. Now one determines the sum of the squared distances from all data points to this submanifold,

$$D_1(\bar{\mathbf{x}}, \mathbf{v}) = \sum_{n=1}^N d(\mathbf{x}(n), M_1(\bar{\mathbf{x}}, \mathbf{v}))^2. \quad (7)$$

For $\bar{\mathbf{x}}$ fixed, this is a function of \mathbf{v} (and, thereby, a function of the geodesic through $\bar{\mathbf{x}}$). The tangent vector \mathbf{v} for which this function becomes minimal defines the first “principal geodesic.”

One can now add further principal geodesics using the following procedure (which will only be explained for the second principal geodesic because the generalization is straightforward): In addition to \mathbf{v} (the tangent vector for the first principal geodesic), one takes a second tangent vector \mathbf{v}' that can be chosen to be orthogonal to the first one. \mathbf{v} and \mathbf{v}' define a tangent plane at $\bar{\mathbf{x}}$. This tangent plane can be geodesically extended (i.e., one considers the plane spanned by all geodesics defined by the tangent vectors in this tangent plane) to a two-dimensional submanifold. Next, one calculates the sum of the squared distances of data points to this submanifold and minimizes this sum with respect to the vector \mathbf{v}' . The result is the second principal geodesic.

3. For the projection, we map each data point \mathbf{x} on the Riemannian manifold onto the closest point $\mathbf{V} \in M_S$ on the submanifold, i.e., the point that minimizes $d(\mathbf{x}, M_S)$. As we shall see, it is mainly this projection that becomes problematic globally.

Several comments are in order. First, the description of the principal geodesic analysis for a Riemannian manifold uses only local concepts. It does not take the global shape of the manifold into account, in particular whether geodesics can be arbitrarily extended on the manifold or

whether these geodesics “wind around” parts of the manifold. In the case of a sphere, the geodesics are great circles. (If the D -dimensional sphere is embedded into \mathbb{R}^{D+1} , these great circles are the intersections of the sphere with a plane through the origin.)

Second, we wish to point out a peculiar fact concerning the application of these concepts to a PCA on a torus.²² That is, on a torus, one can always find an infinite number of geodesics such that the squared orthogonal distances of the data points to this single geodesic are essentially zero. The reason is that a geodesic with a tangent vector with incommensurable components winds around the torus an infinite number of times and covers the torus densely. This is a special feature of the topological properties of a torus: One single geodesic can be dense on the whole surface due to an infinite winding number. As a consequence, there is no gain in information from such a principal geodesic, even though the orthogonal (or unexplained) variances of the data points from such a single geodesic may be zero and therefore all variances are explained.³⁴ These considerations clearly indicate that a nontrivial PCA on a torus should make strong restrictions on the winding number.

Applying the theory outlined above to the case of a torus, in the following, we consider the definition of suitable distance measures and mean as well as the projection problem.

C. Distance measures on the torus

A point on a D -dimensional torus T^D can be described by D independent angles $\varphi = (\varphi_1, \varphi_2, \dots, \varphi_D)$ that take values in $[-\pi, \pi)$. Being periodic in nature (i.e., $\varphi_i \equiv \varphi_i + 2\pi$), the D -torus is often defined as the manifold $\mathbb{R}^D / (2\pi\mathbb{Z})^D$, i.e., as the quotient space of the Euclidean space \mathbb{R}^D modulo translations by integer multiples of 2π in each variable. This representation shows that the torus differs from (flat) Euclidean space only in its topological properties [Fig. 1(a)] and is therefore also referred to as “flat torus.” The commonly used doughnut-shaped illustration of a two-dimensional torus in \mathbb{R}^3 , on the other hand, is not well suited to define distances because this representation of the torus is intrinsically not flat and the distance in one angular variable depends on the value of the other angular variable.

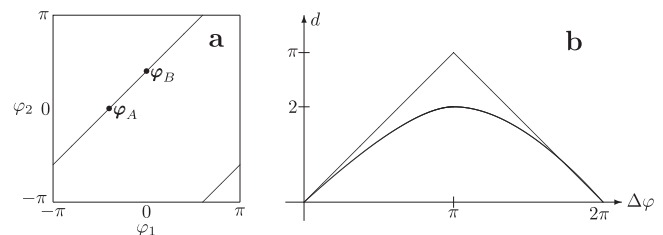


FIG. 1. (a) A two-dimensional flat torus can be visualized as a square with opposite sites topologically identified (i.e., $-\pi$ is identified with $+\pi$). The diagonal lines represent an example for a geodesic that winds around the torus in each direction once. The intrinsic distance between the two indicated points φ_A and φ_B on the torus is directly given by the length of the connecting line. (b) Comparison of the intrinsic distance on the one-dimensional torus (i.e., a circle) [Eq. (13), upper triangular shape] and the corresponding Euclidean distance of the embedding of the circle into a two-dimensional plane [Eq. (15), lower sinus-shaped curve].

Let us now briefly describe two commonly used embeddings of a torus into larger spaces. First, there exists an embedding of the D -dimensional torus into the $2D$ -dimensional real vector space \mathbb{R}^{2D} (Ref. 17)

$$(\varphi_1, \varphi_2, \dots, \varphi_D) \mapsto \mathbf{x}(\boldsymbol{\varphi}) = (\sin \varphi_1, \cos \varphi_1, \sin \varphi_2, \dots, \sin \varphi_D, \cos \varphi_D). \quad (8)$$

For this embedding, the norm of the image vector is

$$|\mathbf{x}(\boldsymbol{\varphi})| = \left(\sum_i (\sin^2 \varphi_i + \cos^2 \varphi_i) \right)^{1/2} = \sqrt{D}. \quad (9)$$

As this norm is independent of the angles, all points lie on a $(2D - 1)$ -dimensional sphere S^{2D-1} of radius \sqrt{D} , which itself is embedded into \mathbb{R}^{2D} . Topologically we have

$$T^D \subset S^{2D-1} \subset \mathbb{R}^{2D}. \quad (10)$$

Given two points $\boldsymbol{\varphi}_A$ and $\boldsymbol{\varphi}_B$ on the torus, we obtain for the scalar product of their images in \mathbb{R}^{2D} ,

$$\begin{aligned} \mathbf{x}(\boldsymbol{\varphi}_A) \cdot \mathbf{x}(\boldsymbol{\varphi}_B) &= \sum_i (\sin \varphi_{Ai} \sin \varphi_{Bi} + \cos \varphi_{Ai} \cos \varphi_{Bi}) \\ &= \sum_i \cos(\varphi_{Bi} - \varphi_{Ai}) \\ &= \sum_i \cos \Delta \varphi_i \end{aligned} \quad (11)$$

with

$$\begin{aligned} \Delta \varphi_i &= \min_T |\varphi_{Bi} - \varphi_{Ai}| \\ &= \min\{|\varphi_{Bi} - \varphi_{Ai}|, 2\pi - |\varphi_{Bi} - \varphi_{Ai}|\}, \end{aligned} \quad (12)$$

where “ \min_T ” means that one has to take the minimal angular distance between two angles (on the torus). This value can always be chosen to be in the interval $[0, \pi]$. The embedding of the D -dimensional torus into \mathbb{R}^{2D} is isometric (i.e., the intrinsic metric on the torus is the same as the induced metric of the embedding or, in other words, the lengths of all paths on the torus are the same as the lengths of the corresponding paths in \mathbb{R}^{2D}).

The different embeddings as given in Eq. (10) give rise to three different notions of distance between two configurations:

- The intrinsic distance on the torus T^D [see Fig. 1(a)] is given by

$$\begin{aligned} d_T(\boldsymbol{\varphi}_A, \boldsymbol{\varphi}_B) &= \sqrt{\sum_i (\min_T |\varphi_{Ai} - \varphi_{Bi}|)^2} \\ &= \sqrt{\sum_i \Delta \varphi_i^2}. \end{aligned} \quad (13)$$

Again, “ \min_T ” and $\Delta \varphi_i$ refer to taking the minimal angular distance on the circle, i.e., a value in the interval $[0, \pi]$. This distance measure assigns the same weight to all angles.

- The geodesic distance on the sphere S^{2D-1} defined by the arcus cosine of the scalar product between two normalized vectors is

$$\begin{aligned} d_S(\boldsymbol{\varphi}_A, \boldsymbol{\varphi}_B) &= \sqrt{D} \arccos \left(\frac{\mathbf{x}(\boldsymbol{\varphi}_A) \cdot \mathbf{x}(\boldsymbol{\varphi}_B)}{|\mathbf{x}(\boldsymbol{\varphi}_A)| |\mathbf{x}(\boldsymbol{\varphi}_B)|} \right) \\ &= \sqrt{D} \arccos \left(\frac{1}{D} \sum_i \cos \Delta \varphi_i \right). \end{aligned} \quad (14)$$

Note that we are dealing with a sphere of radius \sqrt{D} , i.e., the intrinsic circular distance on a unit sphere has to be multiplied by the radius.

- The Euclidean distance in \mathbb{R}^{2D} is

$$\begin{aligned} d_E(\boldsymbol{\varphi}_A, \boldsymbol{\varphi}_B) &= |\mathbf{x}(\boldsymbol{\varphi}_A) - \mathbf{x}(\boldsymbol{\varphi}_B)| \\ &= \min_T \sqrt{|\mathbf{x}(\boldsymbol{\varphi}_A)|^2 + |\mathbf{x}(\boldsymbol{\varphi}_B)|^2 - 2\mathbf{x}(\boldsymbol{\varphi}_A) \cdot \mathbf{x}(\boldsymbol{\varphi}_B)} \\ &= 2\sqrt{\sum_i \left(\sin \frac{\Delta \varphi_i}{2} \right)^2}. \end{aligned} \quad (15)$$

Due to the embeddings [Eq. (10)], we have $d_E \leq d_S \leq d_T$. In particular, for two data points that are far away from each other, these notions of distance can differ significantly. In any case, taking the Euclidean distance [Eq. (15)] or the spherical distance [Eq. (14)] as compared to the torus distance [Eq. (13)] leads to slight deformations of the distance functionals. As a simple illustration, Fig. 1(b) compares the Euclidean distance with the toroidal distance for the case $D = 1$. Note that geodesic and toroidal distances are the same in this case but differ in general for $D > 1$.

D. Circular mean

To compute the mean value $\bar{\varphi}$ of a circular observable φ , we express every observation $\varphi(n)$ as a complex number $\exp(i\varphi(n))$ and compute its average in the complex plane.¹⁴ The associated angle in polar form

$$\begin{aligned} \bar{\varphi} &= \arg \left(\frac{1}{N} \sum_{n=1}^N \exp(i\varphi(n)) \right) \\ &\equiv \arg(x + iy) \end{aligned} \quad (16)$$

is the angle with minimal squared distances on the unit circle [Eq. (12)], which is a convenient choice for a mean value. This approach is computationally equal to averaging the sine and cosine projections of the observations, $x = \langle \cos \varphi \rangle$ and $y = \langle \sin \varphi \rangle$, and computing the resulting mean angle via

$$\bar{\varphi} = \text{atan2}(y, x) = \begin{cases} \arctan \frac{y}{x} & x > 0 \\ \arctan \frac{y}{x} \pm \pi & x < 0, \pm y > 0 \\ \pm \frac{\pi}{2} & x = 0, \pm y > 0 \\ \text{undefined} & x = 0, y = 0 \end{cases} \quad (17)$$

Based on the relation $\tan \varphi = y/x$, the atan2 function uses the correct signs of y and x , corresponding to the given projection on the full circle.

Adopting a double peak distribution of angles on a unit circle as a simple illustration, Fig. S1 of the [supplementary material](#) displays the circular mean $\bar{\varphi}$ of the data together with the averages $\langle \cos \varphi \rangle$ and $\langle \sin \varphi \rangle$. While the circular mean by definition lies on the unit circle (or one-dimensional torus), the mean constructed from the two-dimensional embedding [Eq. (8)] does not because in general $\langle \cos \varphi \rangle^2 + \langle \sin \varphi \rangle^2 \neq 1$.

E. Projection problem on the torus

In the case of periodic data, a fundamental problem exists concerning the projection of data points onto the chosen principal axes. In general, it is not possible to find a projection of the torus onto a geodesic which preserves neighborhoods in the sense that any two points, which are close to each other on the torus, will remain to be close to each other when being projected onto the geodesic. The problem occurs for geodesics with non-vanishing winding numbers in more than one direction and it increases with increasing winding numbers.

To demonstrate the origin of the problem, Fig. 2(a) shows a geodesic on a two-dimensional torus (the diagonal line) which winds around each direction once. The dashed lines indicate points that are diametrical to this geodesic, in the sense that for each point on this line, there exist two different shortest paths of the same length to points on the geodesic. Obviously, the projection problem does not occur if the data are distributed within the region defined by the dashed lines. However, when

the data are also distributed in the vicinity of a dashed line, they are projected onto different regions on the geodesic, depending on whether they are located on the left- or right-hand side of the dashed line. As an illustrative example, two samplings of two-dimensional Gaussian distributions are shown in Fig. 2(a). The blue data points lie near the geodesic; hence, their projection shown in Fig. 2(b) yields a single-peaked distribution. The orange data points in the vicinity of the dashed line, on the other hand, spuriously yield a distribution with two smaller maxima.

When all data points in the sample are projected according to their embedding in the plane (i.e., neglecting the periodicity), neighborhoods in the orange data sample are preserved. The projection error induced by this procedure cannot be avoided but is small as long as the data points are mainly distributed within the region defined by the dashed line. The latter can be achieved, if we shift the circular mean value [Eq. (17)] of the data to the origin of the square representation of the torus. (Of course, this implies choosing a new geodesic onto which the data are to be projected.) This procedure has been suggested in Refs. 21 and 24.

Doing so, however, we may have introduced a second source for a splitting of nearby data points: Data which are distributed at the boundary of the square might be close to each other due to the periodicity but get projected onto different parts of the geodesic because of the choice of the boundary. This might even occur for data points that are mainly distributed around a geodesic. We will refer to this problem as “residual projection error.”

As a demonstration of this effect, Fig. 2(c) shows the sampling of a two-component model (orange points) that consists of a Gaussian centered at the origin and a second Gaussian located at the periodic boundaries of circular space. Although the data are centered with respect to their mean, the projection of the data onto the principal axis in Fig. 2(d) erroneously splits up the second Gaussian at the periodic boundaries.

To minimize this residual projection error, we may determine the optimal position of the origin of the square (or, in general, of the hypercube). This can be achieved by choosing the origin such that the boundary of the hypercube cuts the data along some maximal gap,²² i.e., a minimum of the point density. Proceeding this way, the blue points in Fig. 2(c) show a correct representation of the two Gaussians by two clusters, which also yields a correct projection of the data on the principal axis showing two maxima [Fig. 2(d)]. That is, in cases where the data structure allows one to shift the data such that the boundaries are not crossed, the procedure completely eliminates the residual projection error. In general (i.e., in cases where the data exhibit few crossings over the boundaries), the cut at the maximal gap does not completely eliminate the problem but at least minimizes the projection error and can therefore be considered as an optimal solution to the projection problem.

III. APPROACHES TO CIRCULAR PCA

Before we introduce a new method to construct a PCA on circular data (circular PCA), it is instructive to briefly review previous formulations.

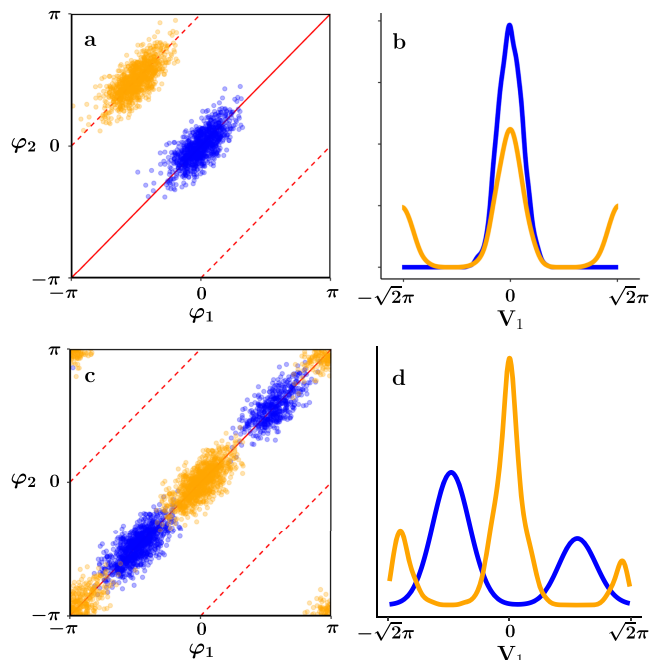


FIG. 2. Illustration of (top) the projection problem and (bottom) the maximal gap approach, using two samplings (orange and blue points) of two-dimensional Gaussian distributions in periodic space. (a) Two-dimensional torus with a geodesic (the red line) onto which we want to project the data as well as two dashed lines which are diametrical to this geodesic. The blue data points lie near the geodesic; hence, their projection is straightforward. Located near the dashed line, on the other hand, the orange data are projected onto different regions on the geodesic, depending on whether they are located on the right-hand side of the dashed line (these points give rise to the peak in the center) or on the left-hand side (these data points give rise to the other peak). (b) Projection of the two data sets onto the geodesic. The centered data yield a single-peaked distribution (blue), the data points from the vicinity of the dashed line are spuriously splitted into two maxima (orange). (c) A two-component model, consisting of a Gaussian centered at the origin and a second Gaussian located at the periodic boundaries of circular space (orange points). Although the data are centered with respect to their mean, the projection of the data onto the principal axis ignoring the periodic nature of the underlying space (d) erroneously splits up the second Gaussian at the periodic boundaries. On the other hand, by cutting the data along their maximal gap, the resulting data (blue points) correctly exhibit two clusters in (c) and two maxima in (d).

A. Angular PCA

Maybe the simplest way to treat periodic data is to first shift the circular mean of the sampled distribution to the center of the space (under consideration of periodic boundaries). Subsequently, one simply ignores the circular nature of the data and performs a standard PCA [Eqs. (1)–(3)] on the shifted data. This approach was introduced in the analysis of dihedral angles of RNA²⁴ and termed “angular PCA.” If the mean-centered data still crosses the periodic boundaries [cf. Fig. 2(c)], however, fast fluctuations of the data (due to jumps from $-\pi$ to π and vice versa) may occur, which can lead to serious errors in the calculation of the covariance matrix.

B. Dihedral angle PCA

The dihedral-angle PCA (dPCA) has been proposed to analyze circular data in the context of protein dynamics.^{16,17} In this case, a data point $\varphi = (\varphi_1, \varphi_2, \dots, \varphi_D)$ is mapped into \mathbb{R}^{2D} according to Eq. (8). Essentially, dPCA consists of a standard PCA, where the data points are now treated as elements of \mathbb{R}^{2D} . One first determines the mean value of the data (which does not necessarily lie on the torus, see Fig. S1 of the [supplementary material](#)) and then calculates the covariance matrix with respect to the Euclidean distance [Eq. (15)] of points as part of the higher dimensional Euclidean space. Finally, one projects the data points onto the linear subspaces (or, to be more precise, the affine subspaces) which, according to the PCA, contain the most relevant variances.

This procedure has been successfully applied to the modeling of biomolecular free energy landscapes.^{23–30} There are several cases in which it can be shown that this method yields good results: (1) when the data points are located in the vicinity of some mean value (in which case the approximating affine space will be close to a tangent space of the torus), (2) if the data points are localized in a region which winds around the torus once. In this case, a linear space that cuts the torus can be a good approximation. The disadvantage is that the actual data points are projected onto linear spaces in \mathbb{R}^{2D} but may still be distributed on a lower dimensional submanifold of the torus, so the dimensional reduction may not be optimal. Furthermore, for large angular variations in the data points, the usage of the Euclidean distance (instead of the toroidal distance) can lead to deformations in the projections of data points to the PCA space^{35,36} (Fig. S1 of the [supplementary material](#)). Simply put, the sine and cosine of the same sampled region may show a high and a low gradient, e.g., close to 0° or 180° where the cosine has a small slope, while the sine is close to the point of steepest descent. This effect puts different weights on the transformed coordinates, which effectively increase and decrease their covariances. Depending on the given sampling (e.g., the chosen origin of the angular data), it is thus difficult to interpret the resulting covariances with respect to the underlying dihedral angles. Representative examples of this effect will be discussed in Sec. IV (Fig. 5 and Fig. S9 of the [supplementary material](#)).

C. GeoPCA

The authors of Refs. 18 and 19 claim that the D angular variables of a torus are essentially the angular variables

of a sphere. While this is true when the torus is embedded in \mathbb{R}^{2D} [Eq. (10)], the sphere considered in Ref. 18 has dimension $2D - 1$. The authors then propose to use the geometry of the sphere and apply the geodesic PCA which essentially leads to great circles as geodesics. However, the proposed algorithm (“GeoPCA”) seems to deal with a D -dimensional sphere embedded in \mathbb{R}^{D+1} instead of a D -dimensional torus. Judging from the article, the authors of Ref. 18 are not aware of the change in topology and the deformation of distances close to the poles of the sphere.³⁷ Already the two-dimensional case shows that for the sphere, only one angular variable (which usually parametrizes the equator) is periodic, while the other angular variable parametrizes a half-circle (from north- to south pole). To illustrate the projection problem for this case, we consider data points that are mainly distributed around the equator of this sphere in \mathbb{R}^3 so that the equator becomes the principal component. For the two poles (north and south pole), the projection onto the equator is not unique because any point on the equator would be a satisfactory solution. As a consequence, points in the immediate vicinity of the poles (and very close to each other on the sphere) may be projected onto points on the equator which are far away.

D. Torus PCA

In Ref. 22, it is proposed to “deform” the D -dimensional torus to a D -dimensional sphere by cutting the torus along a $(D - 1)$ -dimensional hyperplane and contracting the two resulting hyperplanes to the poles of a sphere. This D -dimensional sphere can be parametrized by generalized Euler angles. For the D -dimensional sphere, a nested PCA is performed, i.e., a successive reduction of dimensions according to $S^D \supset S^{D-1} \supset S^{D-2} \dots \supset S^1$. For data points lying on a sphere S^k ($1 < k \leq D$), a subsphere S^{k-1} is chosen in such a way that the orthogonal squared distance of the data points is minimized. The data points are then projected onto the subsphere along great circles orthogonal to the subsphere and passing through the data point. S^{k-1} is not necessarily of maximal radius, i.e., it is not necessarily a “geodesic” subsphere. This is one essential difference as compared to GeoPCA. Furthermore, for the determination of this subsphere as a “best approximation” to the data points of the larger sphere, they use “projected toroidal distances,” i.e., not the metrical structure of the sphere but a distance measure that essentially assigns to the points a distance identical to the one on the torus.

Adopting a practical point of view, we note that the approach cannot be applied to high-dimensional data (in the examples below, $D \approx 10^2$) because it requires clustering in D -dimensional space. Moreover, the interpretation of covariances and principal components is not straightforward.

E. New approach

The discussion above has shown that it is mainly the projection problem (and not the definition of distance and mean) that undermines previous approaches to a PCA on the torus. As worked out in Sec. II E, however, there is an optimal solution to this problem. That is, the residual projection error is minimized by transforming the data such that the maximal gap of the sampling is shifted to the periodic boundary. In

cases where the structure of the data allows us to clearly identify this maximal gap, the problem is basically solved: Once the data are successfully transformed, we can treat the data as linear and compute the covariance matrix and its eigendecomposition in the standard way [Eqs. (1)–(3)]. In this way, we also minimize spurious crossings of the periodic boundaries which may corrupt the calculation of the covariance matrix. Using the transformed data, moreover, the final projection step [Eq. (4)] is well-defined and does not cause any problems.

In practice, there are numerous ways to choose the optimal cut which maximizes the gap between data points. For example, one can (1) maximize the squared distance of the cut to the nearest data points, (2) minimize the data point density in a corridor (with a suitably chosen width) around the cut, or (3) use MD trajectories and choose the cut such that the number of crossings of the cut is minimized. Here we have chosen a procedure based on the second approach: By computing a histogram with a bin width of five degrees, we select the center of the bin with lowest population as the maximal gap. In case of multiple bins of equally low population, we sum over their respective neighboring bins and select the one with lowest overall population in the neighborhood. For the description of maximal gaps in the dihedral distributions of proteins, this approach has proven to be efficient and robust.

Let us discuss virtues and shortcomings of the new approach, henceforth referred to as dPCA+. First, we note that the method is in its practical realization quite similar to the angular PCA²⁴ discussed in Sec. III A. Unlike the latter, however, dPCA+ respects the special topology of the torus by preserving the correct neighborhoods of the data points (and therefore avoids spurious crossings of the periodic boundaries). Second, the transformation of the data is linear; hence, no artificial extra dimensions or deformations of the underlying probability distribution occur (as, e.g., in dPCA). By avoiding nonlinear transformations or deformations of the original topology into spheres (as in Refs. 18–22), moreover, dPCA+ is appealing in its conceptual simplicity and computational efficiency. Finally, dPCA+ yields directly interpretable covariance matrices and eigenvectors, which readily reveal the contributions of the various circular variables. The main assumption underlying dPCA+ is that the data indeed show a significant gap in their distribution. Although this may represent a limitation in general, we note that the opposite case of nearly uniformly sampled variables anyway is in contrast to the very concept of finding a low-dimensional subspace for cluster analysis. As the description of backbone dihedral angles of proteins¹³ represents the main application considered here, in the following, we study the validity of the maximal gap assumption for this case.

To this end, Fig. 3 shows the Ramachandran (ϕ, ψ) plots of several representative amino acids, as obtained from the molecular dynamics trajectory of villin headpiece³² (see below). Due to the steric hindrance of the side chains, these (ϕ, ψ) distributions typically show two main regions that reflect right-handed α -helical and β -extended structures, respectively. Moreover, weak signatures of left-handed structures (i.e., $\phi \geq 0$) may be found. Considering the Ramachandran

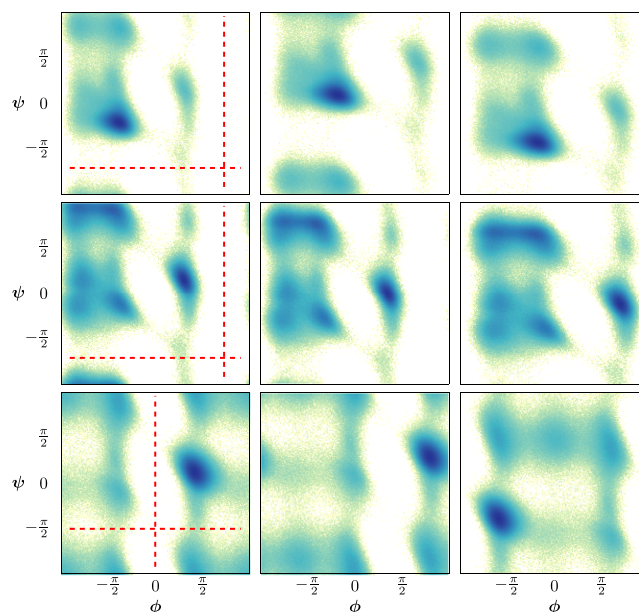


FIG. 3. Ramachandran plots for (from top to bottom) Ala16, Ser2, and Gly33 of HP-35, using (from left to right) original, mean-shifted, and gap-shifted data, respectively. The red dashed lines indicate the maximum gap of the data.

plots as original data, we now consider the effect of (i) shifting the circular mean value of the data to the origin of (ϕ, ψ) space (“mean shifting”) and (ii) shifting the data such that the maximal gap of the sampling is located at the periodic boundary (“gap shifting”).

We first consider residue Ala16, which is part of the second α -helix of villin headpiece and therefore mainly found in α_R conformation. In the unfolded basin, however, the helix may be distorted, which gives rise to β -extended and even some left-handed structures. The original Ramachandran plot represents the data fairly well, with the exception of a few points close to $\psi \geq -180^\circ$ which belong to the β -conformation at $\psi \lesssim 180^\circ$. The mean-shifted data, however, are cut right through the middle of the β -conformation. That is, in the case of Ala16, mean-shifting actually leads to worse results than using the original data. On the other hand, the gap-shifted data correctly place the α -helical and β -extended structures in the middle of the (ϕ, ψ) plane, which virtually eliminates any residual projection error. The second example is Ser2 at the N-terminus, which is very flexible and samples most of the sterically accessible conformational space. Again, we find that the original Ramachandran plot represents the data fairly well but artificially cuts the β -conformation at the periodic boundary $\psi = \pm 180^\circ$. In this case, mean-shifting clearly improves matters by cutting less of the β -conformation off. The gap-shifted data again lead to the best representation of the angular data, although we find residual periodic transitions at $\psi = \pm 180^\circ$ for the rarely sampled left-handed structures. We finally consider Gly33. Being a glycine (with little steric hindrance) located at the C-terminus, Gly33 shows an atypical and widely spread angular space that may be considered as a worst case scenario for the distribution of protein backbone dihedral angles. While mean-shifting does not improve the situation, the gap-shifted data are seen to at least minimize the projection error.

IV. APPLICATION TO PROTEIN DYNAMICS

To demonstrate the potential of the proposed dimensionality reduction technique, we choose two well-established model systems whose conformational dynamics are well described by changes of backbone dihedral angles: Aib₉, a 9-residue achiral peptide showing left- to right-handed transitions of the entire peptide helix,³¹ and villin headpiece (HP-35), a 35-residue protein as a standard example of reversible folding.³² In previous molecular dynamics (MD) simulation work,^{31,38} dPCA studies of both systems have revealed complex free energy landscapes. Comparing results from dPCA and dPCA+, here we show that dPCA+ naturally provides a straightforward interpretation of the systems' covariance and correlation matrices as well as the composition of principal components. Employing a recently developed density-based clustering technique,³⁹ moreover, we find that dPCA+ also yields an unprecedented structural resolution of metastable states in reduced dimensions.

A. Computational methods

1. MD simulations

Aib₉ ($\text{H}_3\text{C} - \text{CO} - (\text{NH} - \text{C}_\alpha(\text{CH}_3)_2 - \text{CO})_9 - \text{CH}_3$) was recently studied by Buchenberg *et al.*,³¹ using the GROMACS program suite⁴⁰ with the GROMOS96 43a1 force field⁴¹ and explicit chloroform solvent.⁴² Here we adopt eight MD trajectories at 300 K of each 2 μs length, using a time step of 4 ps ($4 \cdot 10^6$ frames). The simulation data of HP-35 were kindly provided by the D. E. Shaw research group,³² who performed equilibrium MD simulations at various temperatures for wild-type HP-35 and various mutants, using the Amber ff99SB*-ILDN force field⁴³⁻⁴⁵ and TIP3P explicit water.⁴⁶ Here we adopt a 300 μs trajectory at 360 K that exhibits 61 folding-unfolding transitions, using a time step of 200 ps ($1.5 \cdot 10^6$ frames).

2. Identification of metastable states

On the basis of the dimensionality reduction obtained by dPCA or dPCA+, we used a recently developed density-based clustering technique,³⁹ combined with the most probable path (MPP) algorithm by Jain *et al.*⁴⁷ to identify the metastable states of Aib₉ and HP-35. In brief, the density-based geometric clustering algorithm by Sittel *et al.*³⁹ identifies high-density regions in the given coordinate space, based on the local density of structures measured as a population in a D -dimensional hypersphere. The optimal hypersphere radius is an input parameter that can be found for equilibrium systems using the Boltzmann distribution as a heuristic. The approach has been shown to yield well-defined microstates separated by local free energy barriers. The MPP algorithm represents a dynamic clustering technique that lumps microstates based on their metastabilities and most probable transition pathways.⁴⁷ As the projection on a low-dimensional space may induce spurious transitions in the vicinity of energy barriers, in a final step, we identify core regions of the microstates and count transitions only if the core region of the other state is reached.⁴⁷

3. Ramacolor plots

To efficiently visualize secondary structure differences between metastable states, we use the “Ramacolor” method,³⁹ which assigns a unique color to every point in a (ϕ, ψ) Ramachandran plot, see Fig. 4(b). To assign a color to residue n of a protein in some structural ensemble (e.g., a metastable state), we average over the colors pertaining to all (ϕ_n, ψ_n) frames of the ensemble. As an example, Fig. 4(c) shows the Ramacolor plot of the twenty highest populated metastable states of Aib₉ which assigns a specific color to the (ϕ, ψ) conformations of the seven inner residues of Aib₉. As a consequence of the color code in Fig. 4(b), we find that right-handed structures are drawn in green, left-handed structures are drawn in blue, while right-handed and left-handed excited states are shown in reddish and dark green/purple.

B. Chiral transitions of Aib₉ peptide

Aib₉ is a small helical peptide that nonetheless exhibits complex structural dynamics.³¹ The complexity arises from a hierarchical free energy landscape of the system,⁴⁸⁻⁵⁰ which reflects coupled dynamical processes on several time scales. They correspond to chiral left- to right-handed transitions of the entire peptide helix that happen on a μs time scale, conformational transitions of individual residues which take about

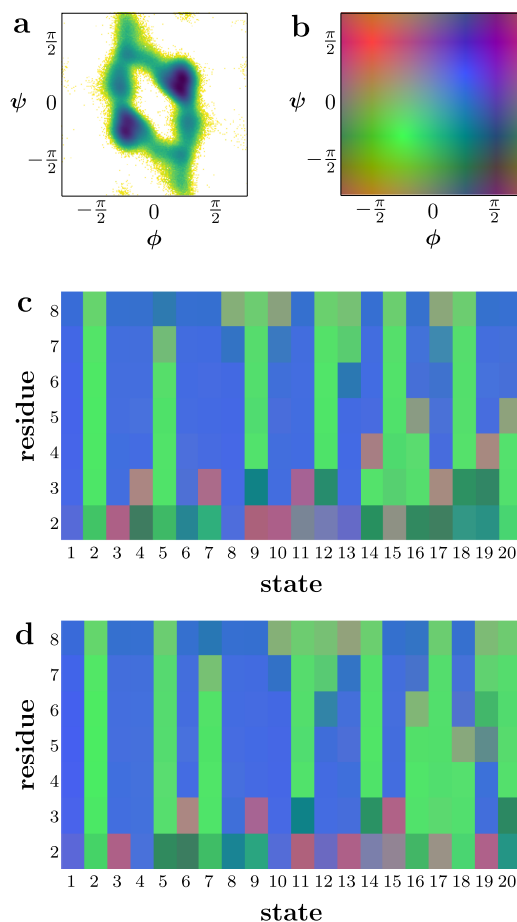


FIG. 4. (a) Ramachandran (ϕ, ψ) density of the inner residues of Aib₉. (b) (ϕ, ψ) -dependent definition of color space.³⁹ [(c) and (d)] Ramacolor plots obtained for the 20 highest populated metastable states of Aib₉ comparing results from (c) dPCA and (d) dPCA+.

1 ns, and the opening and closing of structure-stabilizing hydrogen bonds which occur within tens of ps and are triggered by sub-ps structural fluctuations.³¹ Showing the Ramachandran (ϕ , ψ) plot of Aib₉ (averaged over the inner residues), Fig. 4(a) reveals that the achiral peptide indeed samples both left-handed ($\phi \geq 0$) and right-handed ($\phi \leq 0$) conformations with similar probability. The main conformational states at $\approx(\mp 50^\circ, \mp 45^\circ)$ represent a right- and left-handed helix, respectively. Moreover, for each chirality, we find (at least) one excited conformational state at $\approx(\mp 68^\circ, \pm 45^\circ)$.

Let us first consider the covariance matrix of Aib₉, as obtained directly from the dihedral angles (in the case of dPCA+) and from their cosine and sine transforms (in the case of dPCA). Representing the resulting matrices as color plots, Fig. 5 portrays one of the most obvious advantages of the new approach. In the case of dPCA [Figs. 5(a) and 5(c)], we see a peculiar checker board pattern that hampers a straightforward interpretation in terms of the underlying dihedral angles. The pattern is caused by the sine and cosine transformations that put different weights on the transformed coordinates. (For Aib₉, we find $\langle \cos \phi \rangle, \langle \cos \psi \rangle \approx 0$ and $\langle \sin \phi \rangle, \langle \sin \psi \rangle \geq 0$.) An additional problem occurs for dPCA when we consider the correlation matrix of the transformed variables [Fig. 5(c)], whose diagonal elements are equal to one by definition. This normalization may amplify the effect of variables that have only little contribution to the total variance, such as the cosine projections in the case Aib₉. As shown in Fig. S2 of the [supplementary material](#), this leads to a significant decrease in the amount of variance explained by the first few principal components. Including four principal components, for example, we recover only 52% of the total fluctuations by dPCA (compared to 85% in the case of dPCA+).

Since there are no such issues when we compute the covariances directly from the dihedral angles, the resulting covariance and correlation matrices in Figs. 5(b) and 5(d) are straightforward to interpret. For simplicity, we focus on the latter in the following. Overall, we notice higher correlations of dihedral angles of inner residues, which may be expected

because typically the outer residues have more freedom to move than the inner ones. In particular, we find that the dihedral angles ϕ_1, ψ_1 and ϕ_8, ψ_8 of the terminal ends are largely uncorrelated to the dynamics of the inner dihedral angles. In the subsequent PCA, we discard these coordinates because they do not yield any information on the correlated dynamics of Aib₉. (We note in passing that an uncorrelated coordinate adopting two states results in a trivial doubling of all microstates and therefore unnecessarily complicates the analysis.) Hence, in total, 13 dihedral angles are considered in the PCA.

Another obvious thing to observe is the stripy pattern of the correlation matrix, which shows that the ϕ dihedral angles in general are higher correlated than the ψ dihedral angles. To explain this finding, we recall from Fig. 4(a) that ϕ angles discriminate left- and right-handed ground states, while ψ discriminate ground- and excited states. While the excited states represent short-lived intermediate states, the left- and right-handed ground states are long-lived and define the chirality of a residue. As the latter is important for the propensity of the chirality of the adjacent residues, these residues are strongly correlated via the ϕ angles. Finally we notice the tendency that angles closer to the C-terminal generally show higher correlation than angles closer to the N-terminal. This is caused by the additional CO-group of the N-terminus (as compared to the C-terminus), which facilitates stronger coupling to the solvent due to hydrogen bonding.

Let us now discuss the eigenvectors and principal components obtained from the various approaches. We first note that due to the high symmetry in (ϕ , ψ)-space [Fig. 4(a)], the maximal gap of the data is naturally at the periodic boundaries, i.e., no additional shifting of the data is required to minimize the residual projection error. By diagonalizing the respective covariance matrix, we obtain the eigenvectors of dPCA and dPCA+, the components of which are shown in Fig. S3 of the [supplementary material](#). In the case of the dPCA, the oscillatory pattern of the eigenvector components again reflects the above discussed disparity of sine and cosine transformed variables, which defies a straightforward interpretation. On the other hand, the structure of the dPCA+ eigenvectors directly reveals the contributions of the various dihedral angles. Similar results are also found for the eigenvectors obtained from the correlation matrices.

Figure 6 shows the resulting free energy landscapes $F(V_1, V_2) = -k_B T \ln P(V_1, V_2)$, obtained for the first two principal components V_1 and V_2 of dPCA and dPCA+. As discussed in Ref. 31, the landscape shows two main conformational states *R* and *L*, where all residues are either right- or left-handed, as well as numerous intermediate states accounting for the transition between *R* and *L*. Using the covariance matrix, the results obtained for dPCA and dPCA+ are quite similar [Figs. 6(a) and 6(b)]. In the case of dPCA+, we also obtain similar results when we employ the correlation matrix [Fig. 6(d)]. As mentioned above, however, the dPCA using correlations fails to reproduce the correct free energy landscape due to the artificial overemphasis of cosine projections.

As a simple means to describe the dynamics of the principal components, Fig. S4 of the [supplementary material](#) shows the time evolution of the autocorrelation functions $C_i(t) = \langle \delta V_i(t) \delta V_i(0) \rangle / \langle \delta V_i^2 \rangle$ with $\delta V_i(t) = V_i(t) - \langle V_i \rangle$. As discussed

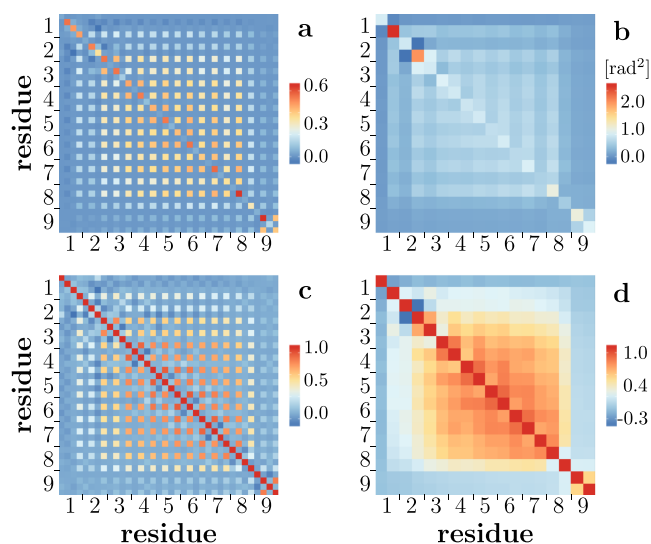


FIG. 5. Covariance (top) and correlation (bottom) matrices of Aib₉ as obtained directly from the dihedral angles (right) and from their cosine and sine transforms (left).

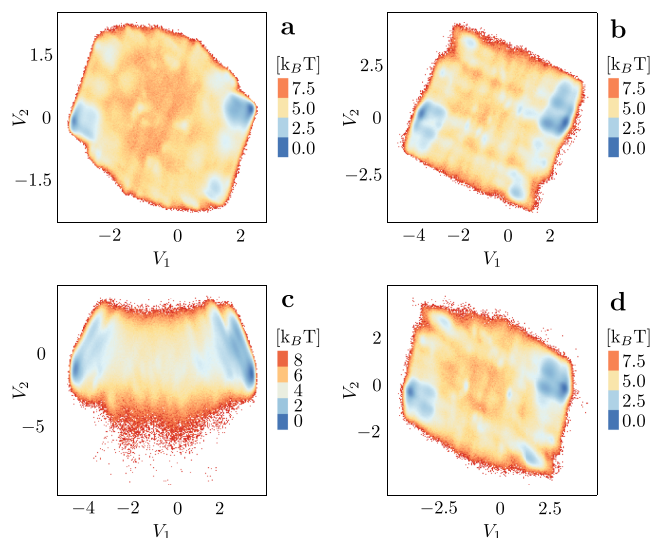


FIG. 6. Projections of the MD data of Aib₉ on the first two principal components obtained from dPCA (left) and dPCA+ (right), based on covariances (top) and on correlations (bottom).

elsewhere,³¹ the decay of the first component reflects chiral left- to right-handed transitions of the entire peptide (the slowest process), while the next four components account for left- to right-handed transitions of individual residues (the next slowest processes). Interestingly, the decay times of the first few principal components are very similar for both methods.

As usual, we selected from the principal components the first few (in this case five) components that show multi-peaked, clustered distributions, and a slow decay of the autocorrelation function.²³ Using these first five principal components, we performed density-based clustering³⁹ of the data obtained from dPCA and dPCA+, which gave 53 and 105 geometrically distinct microstates, respectively. In both cases, a hypersphere radius of 0.2 was found to yield optimal results (see Sec. IV A). Figures 4(c) and 4(d) show the resulting Ramacolor plots of the highest populated microstates for both methods, where the right- or left-handed residues are drawn in green and blue, respectively. Somewhat surprisingly, in the case of dPCA+, certain states seem to share the same geometry (compare, e.g., all-blue states 1 and 4 or all-green states 2 and 5). A closer analysis of the (ϕ, ψ) -distributions reveals, however, that these states in fact correspond to either 3_{10} - or α -helical

structures, which are separated by about 10° in the Ramachandran plot (Fig. S5 of the [supplementary material](#)). This small angular difference, coupled with the overall variance of the states makes them hard to distinguish in the color space of the Ramacolor plots. Nonetheless, the dPCA+ based clustering readily achieves the structural discrimination of the two kinds of helices, which dPCA could not.

C. Folding dynamics of HP-35

As a second example, we choose a 300 μ s trajectory³² of the fast-folding HP-35 protein which was analyzed recently³⁹ using a combination of dPCA and clustering techniques, in order to construct a Markov state model^{6,51,52} of the folding dynamics. In brief, dPCA was found to give 10 relevant principal components, for which density-based clustering obtained 543 microstates and MPP clustering generated 12 macrostates (see Sec. IV A). Using the same protocol, we now perform dPCA+ on the data. Based on the shape of one- and two-dimensional projections of the free energy landscape (Figs. S6 and S7 of the [supplementary material](#)) showing several distinguishable clusters, we selected principal components 1-5 and 7 for further analysis. That is, dPCA+ requires only six components, while dPCA required ten. Nonetheless, we find that the autocorrelation functions of these components decay in a quite similar way for both methods (Fig. S8 of the [supplementary material](#)).

Employing density-based clustering with a hypersphere radius of 0.3, we obtain 76 microstates, that is, less than 15% as obtained for dPCA. This discrepancy is, however, not due to a higher resolution of dPCA but can be traced back to spurious extrema in the dPCA free energy landscape, which are caused by a combination of nonlinear trigonometric transformations and incomplete sampling of the data. That is, more than 500 of these microstates are only very little populated and account for the many structures in the broad unfolded basin. The remaining higher populated microstates, on the other hand, are structurally close to the ones found by dPCA+.

We next compare the 12 macrostates obtained for dPCA and dPCA+, which were constructed using MPP dynamic clustering. Displaying Ramacolor plots pertaining to these states, Figs. 7(a) and 7(c) reveal that both methods exhibit

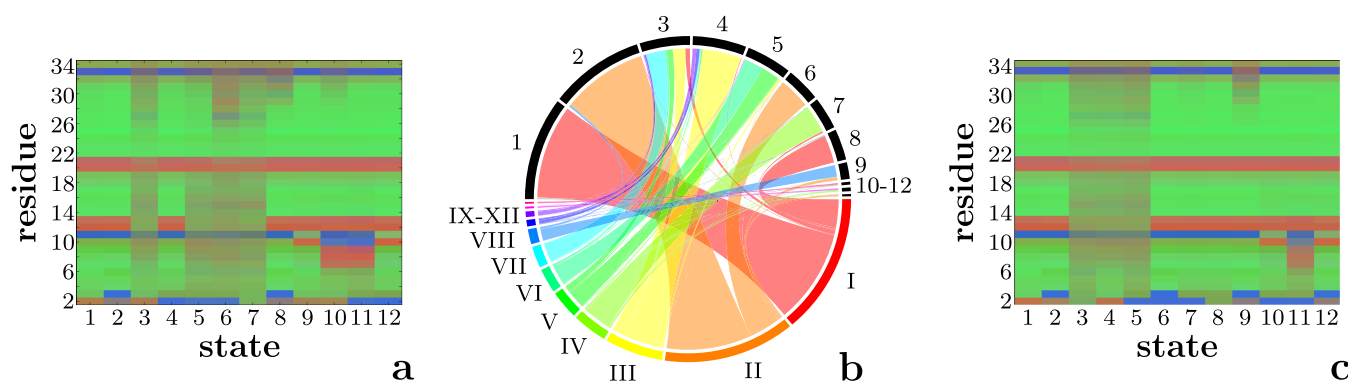


FIG. 7. Ramacolor plots of the 12 macrostates of HP-35 (ordered by decreasing population) as obtained for (a) dPCA and (c) dPCA+. (b) Matching wheel of the correspondence of dPCA states (lower half circle) and dPCA+ states (upper half circle), revealing the different state assignment of the two methods.

structurally similar states, which—roughly speaking—feature the three α -helices of HP-35 connected by some short turn structures. Remarkably, though, dPCA+ is able to provide a clearer separation of metastable native and intermediate states from states assigned to the entropic unfolded basin.^{39,47} To highlight the sometimes subtle differences between the different clusterings, we matched both state trajectories against each other (i.e., by counting the number of times, a certain state i found in dPCA was identified as a state j in dPCA+). Represented as a “matching wheel,” Fig. 7(b) reveals the mutual correspondence of dPCA and dPCA+ macrostates. The maybe most striking result is the obvious splitting of the first two states in dPCA into four distinct states (1, 8 and 2, 6) in dPCA+. The Ramacolor plots of the dPCA+ states show that residues Ser-2 and Asp-3 make the difference, which separate native and stable intermediate structures with compact and elongated N-terminals. Studying the probability distributions of (ϕ, ψ) and their sine and cosine transforms of Ser-2, Fig. S9 of the [supplementary material](#) shows that the discrepancy is again caused by artifacts due to the sine and cosine projections. On the other hand, states 4, 7 and 9 of dPCA+, encoding entropic and intermediate structures with partly folded and unfolded sections, show a more or less direct correspondence to dPCA states III, IV, and VIII. Overall, we find that dPCA+ is able to encode higher structural resolution in less principal components.

We finally wish to demonstrate that the above featured methodology also leads to a physically appealing representation of the free energy landscape of HP-35. To this end, Fig. 8 compares the free energy landscape along the first two principal components of dPCA+ to a Markov state model constructed from the 12 macrostates. As discussed in detail

in Refs. 39 and 47, the folding dynamics of HP-35 can be well described by three main basins of the free energy which comprise native, intermediate, and unfolded metastable conformational states, respectively. Quite remarkably, we find that the location and the connectivity of these basins and the underlying metastable states in the Markov state model is directly reflected in the dPCA+ free energy landscape. While, of course, the positions of the states in the Markov state model are somewhat arbitrary, we note that the model—especially with respect to the folded to unfolded transition—was solemnly constructed based on transition probabilities and structural similarity of the macrostates. This confirms the well-established (but hardly achieved) promise that an energy landscape with well-chosen reaction coordinates directly reveals the main states and barriers as well as reaction pathways of the considered process.

V. CONCLUDING REMARKS

To develop a suitable dimensionality reduction method for high-dimensional circular data, we have reconsidered PCA in terms of geometrical concepts and focused on the geometrical description of a flat torus comprising the circular data. This led to the discussion of possible distance measures for data points distributed on a torus as well as of the problem of projecting data onto the principal subspaces. By analyzing various methods suggested so far,^{16–22} we have found that it is mainly the projection problem (and not the definition of distance and mean) that undermines previous approaches to a PCA on the torus.

Considering the—in parts rather involved—underlying mathematical formulation, we ended up with a surprisingly simple solution of the problem. That is, we have shown that the residual projection error can be minimized by transforming the data such that the maximal gap of the sampling is shifted to the periodic boundary. At the same time, the transformation was found to minimize the (periodicity-induced) error of the estimation of the covariance matrix. Because the transformed data can be treated as linear, we subsequently may compute the covariance matrix and its eigendecomposition in a standard manner. By avoiding nonlinear transformations or deformations of the original topology (as in Refs. 16–22), the new approach termed dPCA+ avoids artificial extra dimensions or distortions of the underlying probability distribution and is therefore appealing in its conceptual simplicity and computational efficiency.

The main underlying assumption of a significant gap in the data distribution has been tested by applying the method to the description of backbone dihedral angles of proteins. Adopting Aib₉ and HP-35 as well-established model systems of complex conformational dynamics, we have found that dPCA+ represents a significant improvement of the previously used dPCA (hence the name). That is, the new approach offers a direct interpretation of covariances and principal components in terms of the angular variables. It furthermore allows for a robust and well-defined construction of metastable states, which is essential for the successful construction of Markov state models. For the two systems considered, the first few principal components of dPCA+ can be considered as a

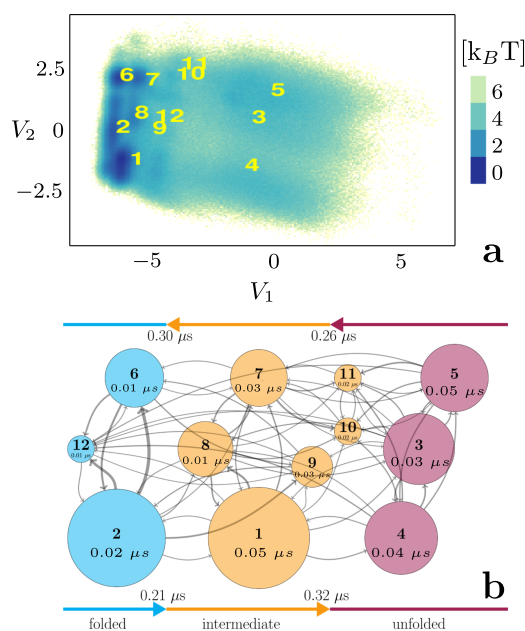


FIG. 8. (a) Two-dimensional representation of the free energy landscape of HP-35 along the first two principal components of dPCA+. Numbers indicate the location of the metastable conformational states of the system. (b) Markov state model built from these states, showing states of the folded, intermediate, and unfolded basin in blue, yellow, and purple, respectively. States are annotated by their lifetime, their size indicates their population, the thickness of the arrows indicates the number of transitions, and the colored bars show the cumulative transition times between the basins.

low-dimensional reaction coordinate that accurately portrays the folding free energy landscape and reveals main states, barriers, and reaction pathways.

While in this work we have focused on aspects of the PCA, we wish to stress that the main idea of the new approach is much more general and can also be applied to other dimensionality reduction methods for circular data. In particular, the shifting according to the maximal gap should also cure an essential part of the projection problem in other approaches, including nonlinear methods,⁷ and various versions of independent component analysis.^{8,9}

SUPPLEMENTARY MATERIAL

See [supplementary material](#) for illustration of a double-peak distribution on a unit circle (Fig. S1), details of the PCAs on Aib₉ including cumulative fluctuations (Fig. S2), eigenvector contents (Fig. S3), autocorrelation functions (Fig. S4), Ramachandran plots (Fig. S5), as well as details of the PCAs on HP-35 including one-dimensional (Fig. S6) and two-dimensional (Fig. S7) free energy landscapes, autocorrelation functions (Fig. S8) and probability densities of residue 2 (Fig. S9).

ACKNOWLEDGMENTS

We thank Sebastian Buchenberg and Matthias Ernst for numerous instructive and helpful discussions and D. E. Shaw Research for sharing their trajectories of HP-35.

APPENDIX: SOFTWARE AND DATA AVAILABILITY

The dPCA+ method was implemented in the open source software *FastPCA*, freely available at <https://github.com/lettis/FastPCA>. *FastPCA* has also been embedded in the *prodyna* R-library, a toolkit for dimensionality reduction, clustering, and visualization of protein dynamics data. *prodyna* is available at <https://github.com/lettis/prodyna>.

- ¹A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis* (John Wiley & Sons, New York, 2001).
- ²I. T. Jolliffe, *Principal Component Analysis* (Springer, New York, 2002).
- ³P. Benner, V. Mehrmann, and D. C. Sorensen, *Dimension Reduction of Large-Scale Systems* (Springer, New York, 2005).
- ⁴I. Borg and P. J. Groenen, *Modern Multidimensional Scaling: Theory and Applications* (Springer, New York, 2005).
- ⁵J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction* (Springer, New York, 2007).
- ⁶G. R. Bowman, V. S. Pande, and F. Noe, *An Introduction to Markov State Models* (Springer, Heidelberg, 2013).
- ⁷M. A. Rohrdanz, W. Zheng, and C. Clementi, "Discovering mountain passes via torchlight: Methods for the definition of reaction coordinates and pathways in complex macromolecular reactions," *Annu. Rev. Phys. Chem.* **64**, 295 (2013).
- ⁸O. F. Lange and H. Grubmüller, "Generalized correlation for biomolecular dynamics," *Proteins: Struct., Funct., Bioinf.* **62**, 1053 (2006).
- ⁹G. Perez-Hernandez, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noe, "Identification of slow molecular order parameters for Markov model construction," *J. Chem. Phys.* **139**, 015102 (2013).
- ¹⁰A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, "Essential dynamics of proteins," *Proteins: Struct., Funct., Genet.* **17**, 412 (1993).
- ¹¹B. L. de Groot, X. Daura, A. E. Mark, and H. Grubmüller, "Essential dynamics of reversible peptide folding: Memory-free conformational dynamics governed by internal hydrogen bonds," *J. Mol. Biol.* **309**, 299 (2001).

- ¹²G. Kurz, I. Gilitschenski, and U. D. Hanebeck, "Recursive nonlinear filtering for angular data based on circular distributions," in *American Control Conference (ACC), 2013* (IEEE, 2013), pp. 5439–5445.
- ¹³S. C. Lovell, I. W. Davis, W. B. Arendall, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson, "Structure validation by α geometry: ϕ , ψ and $c\beta$ deviation," *Proteins: Struct., Funct., Bioinf.* **50**, 437 (2003).
- ¹⁴K. V. Mardia and P. E. Jupp, *Directional Statistics* (John Wiley & Sons, 2009).
- ¹⁵D. M. D. van Aalten, B. L. de Groot, J. B. C. Finday, H. J. C. Berendsen, and A. Amadei, "A comparison of techniques for calculating protein essential dynamics," *J. Comput. Chem.* **18**, 169 (1997).
- ¹⁶Y. Mu, P. H. Nguyen, and G. Stock, "Energy landscape of a small peptide revealed by dihedral angle principal component analysis," *Proteins: Struct., Funct., Bioinf.* **58**, 45 (2005).
- ¹⁷A. Altis, P. H. Nguyen, R. Hegger, and G. Stock, "Dihedral angle principal component analysis of molecular dynamics simulations," *J. Chem. Phys.* **126**, 244111 (2007).
- ¹⁸K. Sargsyan, J. Wright, and C. Lim, "GeoPCA: A new tool for multivariate analysis of dihedral angles based on principal component geodesics," *Nucl. Acids Res.* **40**, e25 (2012).
- ¹⁹K. Sargsyan, J. Wright, and C. Lim, "Corrigendum to GeoPCA: A new tool for multivariate analysis of dihedral angles based on principal component geodesics," *Nucl. Acids Res.* **43**, 10571 (2015).
- ²⁰S. Huckemann and H. Ziezold, "Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces," *Adv. Appl. Prob.* **38**, 299 (2006).
- ²¹A. Nodehi, M. Gholizadeh, and A. Heydari, "Dihedral angles principal geodesic analysis using nonlinear statistics," *J. Appl. Stat.* **42**, 1962 (2015).
- ²²B. Eltzner, S. Huckemann, and K. V. Mardia, "Torus principal component analysis with an application to RNA structures," e-print [arXiv:1511.04993](https://arxiv.org/abs/1511.04993) (2015).
- ²³A. Altis, M. Otten, P. H. Nguyen, R. Hegger, and G. Stock, "Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis," *J. Chem. Phys.* **128**, 245102 (2008).
- ²⁴L. Riccardi, P. H. Nguyen, and G. Stock, "Free energy landscape of an RNA hairpin constructed via dihedral angle principal component analysis," *J. Phys. Chem. B* **113**, 16660 (2009).
- ²⁵G. G. Maisuradze, A. Liwo, and H. A. Scheraga, "Principal component analysis for protein folding dynamics," *J. Mol. Biol.* **385**, 312 (2009).
- ²⁶A. Jain, R. Hegger, and G. Stock, "Hidden complexity of protein energy landscape revealed by principal component analysis by parts," *J. Phys. Chem. Lett.* **1**, 2769 (2010).
- ²⁷D. A. Potoyan and G. A. Papoian, "Energy landscape analyses of disordered histone tails reveal special organization of their conformational dynamics," *J. Am. Chem. Soc.* **133**, 7405 (2011).
- ²⁸J. C. Miner, A. A. Chen, and A. E. García, "Free-energy landscape of a hyperstable RNA tetraloop," *Proc. Natl. Acad. Sci. U. S. A.* **113**, 6665 (2016).
- ²⁹G. M. Hocky, J. L. Baker, M. J. Bradley, A. V. Sinititskiy, E. M. De La Cruz, and G. A. Voth, "Cations stiffen actin filaments by adhering a key structural element to adjacent subunits," *J. Phys. Chem. B* **120**, 4558 (2016).
- ³⁰C. R. Watts, A. J. Gregory, C. P. Frisbie, and S. Lovas, "Structural properties of amyloid β (1–40) dimer explored by replica exchange molecular dynamics simulations," *Proteins* **85**, 1024 (2017).
- ³¹S. Buchenberg, N. Schaudinnus, and G. Stock, "Hierarchical biomolecular dynamics: Picosecond hydrogen bonding regulates microsecond conformational transitions," *J. Chem. Theory Comput.* **11**, 1330 (2015).
- ³²S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "Protein folding kinetics and thermodynamics from atomistic simulation," *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17845 (2012).
- ³³K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philos. Mag.* **2**, 559 (1901).
- ³⁴The projection of the data points onto this main principal component axis destroys any properties of "neighborhood," i.e., two points which are very close to each other on the torus (i.e., the data space) may be arbitrarily far apart from each other when projected (according to closest distance) onto this axis. Furthermore, on this principal component the data points will in general be distributed over an infinite length.
- ³⁵K. Hinsen, "Comment on 'Energy landscape of a small peptide revealed by dihedral angle principal component analysis,'" *Proteins: Struct., Funct., Bioinf.* **64**, 795 (2006).

- ³⁶Y. Mu, P. H. Nguyen, and G. Stock, "Reply to the comment on 'Energy landscape of a small peptide revealed by dihedral angle principal component analysis,'" *Proteins: Struct., Funct., Bioinf.* **64**, 798 (2006).
- ³⁷Due to a lack of rigor, in particular with respect to notation, it is not really clear from the article how exactly GeoPCA is performed. Some formulae of Ref. 18 refer to scalar products of D -dimensional vectors of data points with $(D + 1)$ -dimensional vectors in the embedding space, which is clearly not what the authors had in mind. The existing computer program for this analysis⁵³ seems to be based on the usual representation of a sphere by generalizations of Euler angles. However, the restriction to principal dimensions being great circles (geodesics) will yield satisfactory results only in very special cases of data structures.
- ³⁸A. Jain and G. Stock, "Hierarchical folding free energy landscape of HP35 revealed by most probable path clustering," *J. Phys. Chem. B* **118**, 7750 (2014).
- ³⁹F. Sittel and G. Stock, "Robust density-based clustering to identify metastable conformational states of proteins," *J. Chem. Theory Comput.* **12**, 2426 (2016).
- ⁴⁰D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "Gromacs; fast, flexible and free," *J. Comput. Chem.* **26**, 1701 (2005).
- ⁴¹W. F. van Gunsteren, S. R. Billeter, A. A. Eising, P. H. Hünenberger, P. Krüger, A. E. Mark, W. R. P. Scott, and I. G. Tironi, *Biomolecular Simulation: The GROMOS96 Manual and User Guide* (Vdf Hochschulverlag AG an der ETH Zürich, Zürich, 1996).
- ⁴²I. G. Tironi and W. F. van Gunsteren, "A molecular dynamics simulation study of chloroform," *Mol. Phys.* **83**, 381 (1994).
- ⁴³V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters," *Proteins: Struct., Funct., Bioinf.* **65**, 712 (2006).
- ⁴⁴R. B. Best and G. Hummer, "Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides," *J. Phys. Chem. B* **113**, 9004 (2009).
- ⁴⁵K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99sb protein force field," *Proteins: Struct., Funct., Bioinf.* **78**, 1950 (2010).
- ⁴⁶W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.* **79**, 926 (1983).
- ⁴⁷A. Jain and G. Stock, "Identifying metastable states of folding proteins," *J. Chem. Theory Comput.* **8**, 3810 (2012).
- ⁴⁸H. Frauenfelder, S. Sligar, and P. Wolynes, "The energy landscapes and motions of proteins," *Science* **254**, 1598 (1991).
- ⁴⁹J. N. Onuchic, Z. L. Schulten, and P. G. Wolynes, "Theory of protein folding: The energy landscape perspective," *Annu. Rev. Phys. Chem.* **48**, 545 (1997).
- ⁵⁰K. A. Dill and H. S. Chan, "From Levinthal to pathways to funnels: The 'new view' of protein folding kinetics," *Nat. Struct. Biol.* **4**, 10 (1997).
- ⁵¹J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noe, "Markov models of molecular kinetics: Generation and validation," *J. Chem. Phys.* **134**, 174105 (2011).
- ⁵²D. Shukla, C. X. Hernández, J. K. Weber, and V. S. Pande, "Markov state models provide insights into dynamic modulation of protein function," *Acc. Chem. Res.* **48**, 414 (2015).
- ⁵³K. Sargsyan, Y. H. Hua, and C. Lim, "Clustangles: An open library for clustering angular data," *J. Chem. Inf. Mod.* **55**, 1517 (2015).