Topological obstructions in the way of data-driven collective variables

Behrooz Hashemian and Marino Arroyo[a)]
LaCàN, Universitat Politecnica de Catalunya–BarcelonaTech, Barcelona, Spain

Nonlinear dimensionality reduction (NLDR) techniques are increasingly used to visualize molecular trajectories and to create data-driven collective variables for enhanced sampling simulations. The success of these methods relies on their ability to identify the essential degrees of freedom characterizing conformational changes. Here, we show that NLDR methods face serious obstacles when the underlying collective variables present periodicities, e.g. arising from proper dihedral angles. As a result, NLDR methods collapse very distant configurations, thus leading to misinterpretations and inefficiencies in enhanced sampling. Here, we identify this largely overlooked problem and discuss possible approaches to overcome it. We also characterize the geometry and topology of conformational changes of alanine dipeptide, a benchmark system for testing new methods to identify collective variables.

PACS numbers: 87.10.Tf, 87.15.A-, 87.15.hp

Thanks to enhanced sampling techniques, it is possible to connect molecular conformations separated by high energy barriers, and accurately compute free energies in systems exhibiting metastability. The success of these techniques relies on a good set of collective variables (CVs), capturing the metastability of the system with a few degrees of freedom. CVs are commonly chosen out of experience or physical intuition. As increasingly complex systems become accessible computationally,[1] the task of selecting appropriate CVs becomes highly nontrivial.[2] This situation has motivated in recent years intense research aimed at systematic and data-driven approaches to select CVs, often relying on statistical learning methods. In particular, dimensionality reduction techniques automatically identify a reduced set of coordinates capturing the essential behavior of a complex system, starting from a pre-existing ensemble of molecular configurations, called training set.

The most widespread dimensionality reduction method is principal component analysis (PCA).[3] PCA is a linear method, which selects mutually orthogonal directions such that, by projecting the data onto a few of them, the variance of the projected data is maximized. PCA has been widely applied to characterize the essential dynamics,[4–8] understand molecular flexibility[9] and enhance sampling in molecular dynamics.[10,11] PCA and in general linear dimensionality reduction methods are very popular because of their simplicity. However, they fail to identify nonlinear correlations in the data, which are often present in molecular systems, e.g. as a result of bond rotations or steric interactions.[12–14]

Advances in the field of statistical learning, notably in nonlinear dimensionality reduction (NLDR) techniques,[15–17] were quickly embraced by the molecular simulation community to visualize trajectories, realizing that conformations often evolve close to a nonlinear manifold often called intrinsic manifold,[18–22] although some systems evolve on non-manifold sets.[23] Different NLDR techniques have been applied to molecular systems, including Isomap,[15] locally linear embedding,[16] autoencoder networks,[17] diffusion map[24] or LSDMap.[25] Building on these techniques, a number of methods have been developed to systematically define differentiable and nonlinear CVs, to be used in enhanced sampling simulations.[26–29]

Given an ensemble of molecular conformations, it is straightforward to obtain a low-dimensional representation through linear or nonlinear dimensionality reduction techniques. However, such an embedding will only be useful if the low-dimensional representation captures the essential features of the original dataset. If the low-dimensional representation collapses conformations that are distant in high-dimensions, these algorithms may induce misinterpretations or non-convergence in enhanced sampling simulations. Similar problems arise if the low-dimensional representation is not low-dimensional enough, i.e. matching the intrinsic dimension.[30] In this case, the conformations sparsely populate the reduced space.

Here, we point out a major obstacle when applying dimensionality reduction techniques to molecular simulations: topological obstructions of the intrinsic manifold. This issue has not been acknowledged in the literature, but is affecting the performance of NLDR methods in a number of recent studies.[21,25,26,31,32] We conceptually identify this problem, and illustrate its impact using a training set for alanine dipeptide, a benchmark in the field. We also take a close look at the geometry of the intrinsic manifold of this molecule. This understanding may contribute to orient the future research on systematic data-driven CVs. Finally, we suggest possible directions to overcome topological obstructions in defining adequate data-driven CVs.

DIMENSIONALITY REDUCTION AND TOPOLOGICAL OBSTRUCTIONS

A manifold of dimension d is an object that locally looks like Euclidean space $\mathbb{R}^d$. Two manifolds are said to

[a)]Electronic mail: marino.arroyo@upc.edu

have the same topology if one can be transformed into the other with a continuous deformation such as bending and stretching, but not tearing or gluing. The properties that are preserved under such deformations are called topological properties, and include connectedness, continuity and boundary.

Dimensionality reduction techniques try to find a reduced space representation in such a way that topological properties of objects in high-dimensional space preserved.[30] However, depending on the topology of the high-dimensional manifold, it may not be possible to embed it in $\mathbb{R}^d$. For instance, consider a circle ($d = 1$), which can be trivially described by a single parameter, the polar angle. Dimensionality reduction techniques will try to represent the circle as an open set in the real line, thus collapsing distant points and destroying the underlying structure, see Figure 1(a). This example illustrates Whitney's embedding theorem,[33] which states that the embedding (without self-intersection) of a d-dimensional manifold may require up to $2d+1$ dimensions. This theory guarantees that any one-dimensional manifold can be embedded in $\mathbb{R}^3$, but the minimal dimension where the manifold can be embedded will depend on the topology of the manifold. A circle requires two dimensions, while a knot requires three dimensions.

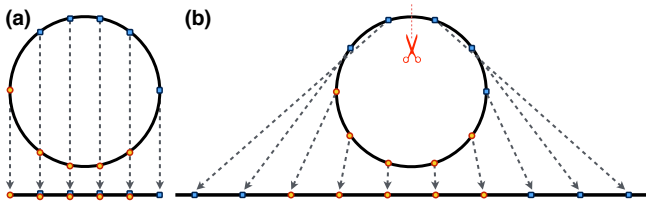

FIG. 1. Due to their different topology, it is impossible to embed a circle into a line (a); however, by tearing the circle at one point, this topological obstacle can be circumvented (b).

Thus, topology is an obstacle to embed a manifold into a space of its intrinsic dimension. However, if we change the topology of the circle by cutting it at one point, then the resulting curved segment can be easily unbent and embedded into the real line, as illustrated in Figure 1(b).

Figure 2(a,b) shows a torus and a sphere, which are two-dimensional nonlinear manifolds that cannot be embedded in less than three dimensions. As a result, NLDR methods in general will destroy their structure if they attempt to represent these surfaces in two-dimensions. In fact, NLDR methods can only embed $d-$dimensional manifolds in $\mathbb{R}^d$ if they have the topology of open sets in $\mathbb{R}^d$, thus necessarily with boundary, such as that shown in Figure 2(c).

Similar topological obstructions are encountered when examining molecular systems with dimensionality reduction methods. A notable example is alanine dipeptide. This small molecule is known to be well-described by two dihedral angles. As a result of their periodicity, the underlying intrinsic manifold has the topology of a torus, which has been exploited to visualize its free-energy landscape.[35] The consequences of this fact have not been fully acknowledged. As a result, low dimensional embeddings appear highly distorted, present loops, and partially collapse information.[21,25,31] Furthermore, because of this topological obstruction, NLDR techniques suggest an excessive number of CVs relative to the intrinsic dimension.[26]

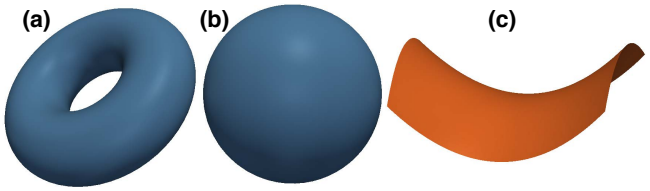

FIG. 2. Surfaces ($d = 2$) of different topology. Conformational changes of molecules with two significant dihedral angles evolve around a torus (a), while six-membered rings carbohydrates, like β-D-Glucopyranose,[34] have a sphere-like intrinsic manifold (b).

To illustrate this fact, we analyze a configurational ensemble of alanine dipeptide obtained from multiple short-run simulations, and shown in dihedral space in Figure 3(a). The color represents one of the dihedral angles. Because, the intrinsic dimension is 2, we try to embed the full ensemble in two dimensions using different dimensionality reduction methods, see Figure 3(c-f), top. As expected, PCA and a variety of NLDR methods fail to embed the ensemble without collapsing distant conformations. We have chosen in this comparison a non-metric NLDR method (Locally linear embedding[16]), and two distance-preserving methods that use different notions of distance (Diffusion map[24] and Isomap[15]). Thus, in the presence of topological obstructions, the ability of NLDR methods in general to unfold nonlinear manifolds is not exploited, and there is no clear advantage relative to PCA.[32] A straightforward way to remove the topological obstructions is to consider a trimmed ensemble of conformations, which lies within the dashed rectangle in Figure 3(a), at the expense of throwing away a significant number of conformations. As shown in Figure 3(d-f), bottom, all nonlinear methods correctly embed the data in 2D, without data collapse (color mixing). The different metric criteria underlying Diffusion map and Isomap are evident in this figure. In contrast, PCA fails to recover the 2D manifold structure, even for the trimmed ensemble, Figure 3(c) bottom. As in the example of the circle, it is possible to elegantly tear the manifold by disconnecting the connectivity structure underlying NLDR algorithms, rather than by shrinking the conformational ensemble, see Figure 3(b) for Isomap. Thus, by appropriately removing topological obstructions, the benefits of NLDR as compared to PCA become available. We further discuss systematic methods to overcome topological obstructions later in the paper.

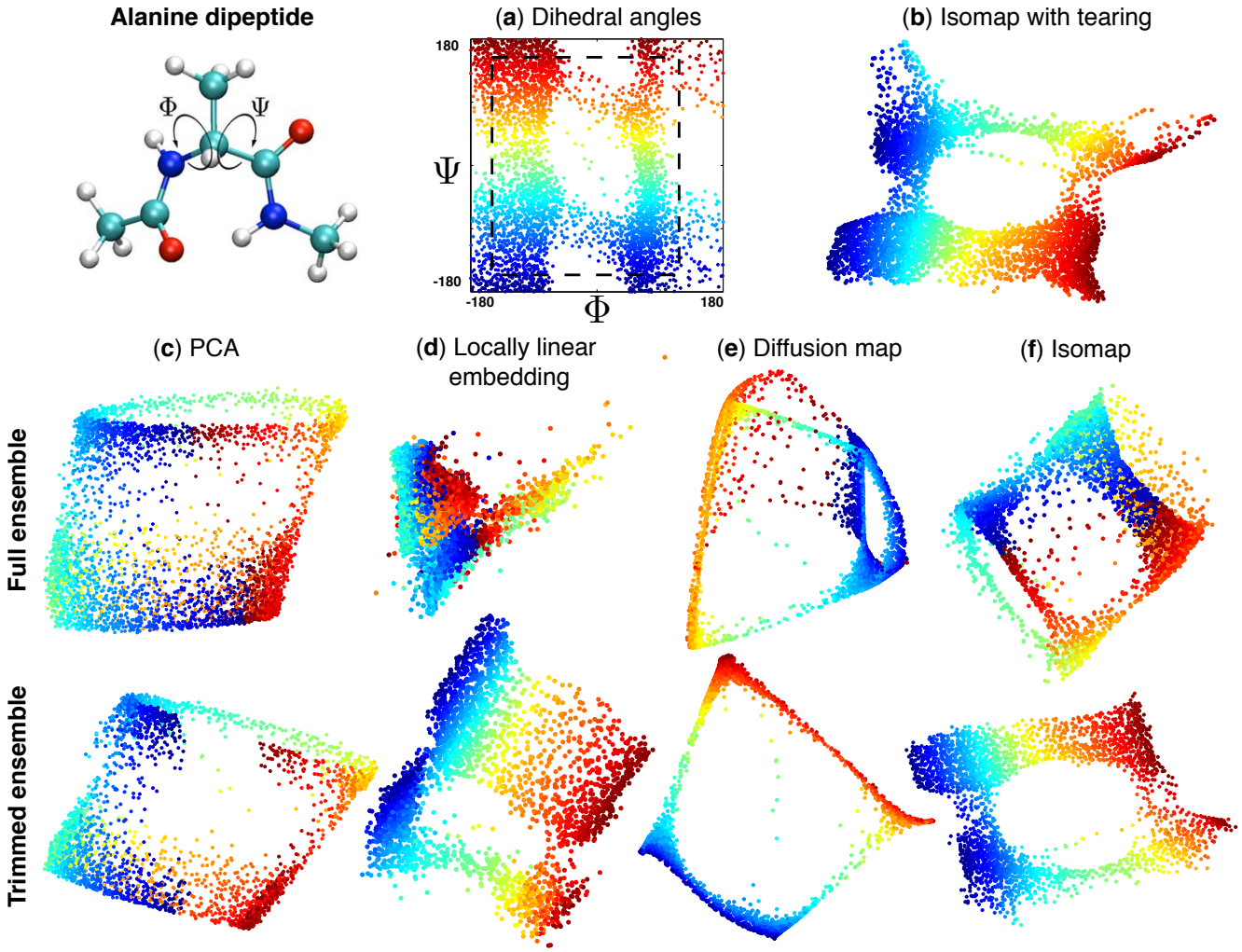

FIG. 3. Dimensionality reduction of an ensemble of molecular configurations of alanine dipeptide, obtained from multiple short-run simulations and visulized in dihedral space, Φ and Ψ (a). A trimmed ensemble delimited by the dashed rectangle is also considered to avoid topological obstructions. Dimensionality reduction methods such as PCA (c), locally linear embedding (with $k = 10$ nearest-neighbors) (d), diffusion map (with $\epsilon^2 = 0.5$ as the bandwidth of the kernel and $k = 10$ nearest-neighbors) (e), and Isomap (with $k = 10$ nearest-neighbors) (f), failed to provide a two-dimensional embedding of the full ensemble without self-intersection (mixing colors representing the backbone dihedral Ψ). In contrast, the NLDR methods successfully embedded the trimmed ensemble, (d-f) bottom. Manual tearing of the full ensemble modifying the connectivity graph of Isomap also lead to a successful embedding in two-dimensions (b).

A CLOSE LOOK AT ALANINE DIPEPTIDE CONFORMATIONAL FLEXIBILITY

We closely examine next the geometry of the intrinsic manifold underlying the conformational changes of alanine dipeptide. Because our goal here is to examine closely metric information about the intrinsic manifold, we focus now on Isomap, which tries to preserve isometry in the embeddings. We start from a well-sampled trajectory resulting from enhanced sampling.[29] After removing hydrogen atoms and alignment, we embed the molecular ensemble in three-dimensions, see Figure 4(a). This embedding strikingly resembles a torus. PCA produces very similar three-dimensional embeddings. By coloring the points representing conformations with the backbone dihedrals Φ and Ψ, the correlation between this embedding and dihedral space becomes clear, see Figure 4(a,b). However, a closer inspection reveals self-intersection of the embedded surface, with the associated collapse of conformations. To examine this, we consider two adjacent strip regions in dihedral space, and color-code them in green and red, see Figure 4(c). Figure 4(d) clearly shows that these strips cross each other in two regions, confirming the self-intersection of reduced representation.

This finding is surprising because there should not be a topological obstruction when embedding a torus in three dimensions, suggesting that the issue is not topological but rather geometrical. Indeed, dimensionality reduction

methods such as PCA or Isomap try to preserve high-dimensional distances in the low-dimensional embedding. Because manifolds cannot be isometrically embedded in general, the resulting embeddings can be distorted. If this geometric distortion is large, it could lead to (topologically avoidable) collapse of information. We further scrutinize this idea next.

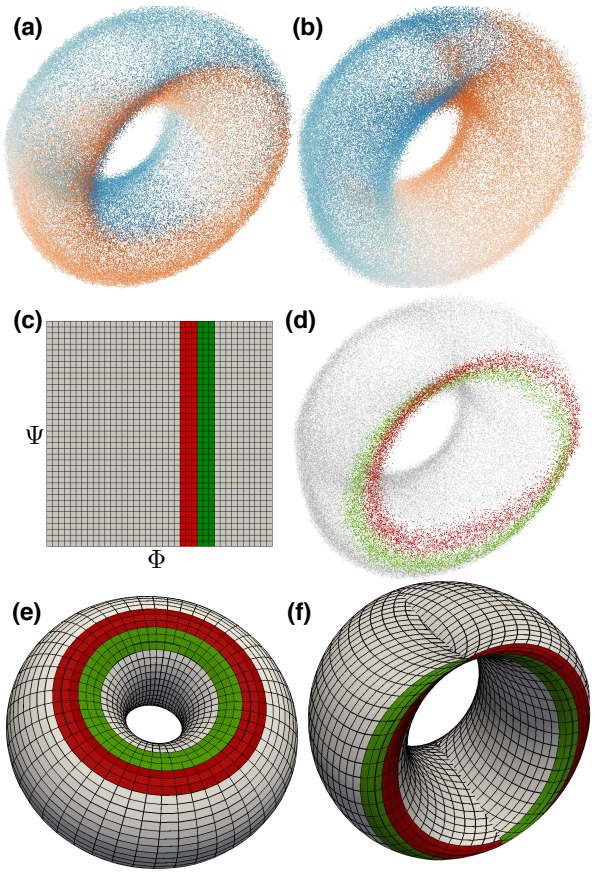

FIG. 4. Topology and geometry of molecular fexiblity of alanine dipeptide. A well-sampled molecular ensemble is processed by Isomap to obtain a three-dimensional representation of the conformational changes, where the colormap is the value of the backbone dihedral angles, Φ (a) and Ψ (b). Two adjacent strips of Φ values (c) show the self intersection of this 3D embedding (d), while in the 3D representation of a torus, there is not any self intersection(e). A three-dimensional projection of a flat torus (f) suggests the conformational changes of alanine dipetide is geometrically similar to a flat torus.

If dihedral space was an accurate representation of the molecule's flexibility, not only in terms of topology, but also in terms of geometry, then the intrinsic manifold would be a flat torus. A flat torus is a topological torus with the metric induced by the Euclidean distance in dihedral space extended by periodicity. It is known that the flat torus can only be embedded isometrically (preserving distances) in four dimensions or more.[36] A consequence of this fact is that any three-dimensional embedding will distort the metric, as illustrated by the grid in Figure 4(e). Interestingly, the three-dimensional projection of the four-dimensional isometric embedding of the flat torus shown in Figure 4(f) is very similar to the embedding provided by Isomap, see Figure 4(d). As shown by the grid, this self-intersecting representation of the surface induces a much smaller distortion of the metric. Taken together, these observations strongly suggest that self intersections in low-dimensional embeddings can not only be the result of topological obstructions, but also the result of geometrical requirements implicit in NLDR methods.

SUMMARY AND DISCUSSION

We have shown that topological obstructions of the intrinsic manifold underlying molecular flexibility can be a serious obstacle in the systematic determination of collective variables using data-driven statistical learning approaches. Focusing on the benchmark molecule alanine dipeptide, we have shown that these obstructions make it impossible to find global low-dimensional representations with minimal dimension (2 for this system) devoid of data collapse. If the embedding dimension is increased to avoid data collapse, then the reduced description becomes dimensionally inefficient and sparsely populated. We have further shown that the intrinsic manifold of alanine dipeptide metrically resembles a flat torus. When dimensionality reduction methods that try to preserve distances, such as Isomap, embed this manifold in 3D, we also observe data collapse, which this time does not have a topological origin.

The most straightforward remedy to topological obstructions is to change the topology of the intrinsic manifold by tearing, as we have illustrated in Figures 1 and 3. Tearing can be easily implemented in NLDR methods that rely on connectivity graphs by disconnecting appropriate vertices in this graph. Such an approach may be guided by data visualization,[29] if the systems is low-dimensional enough, or by more systematic algorithms that find essential loops and disconnect them.[37] This latter method does not work if the intrinsic manifold has the topology of the sphere. It should be noted that tearing the manifold may introduce artificial boundaries in CV space, which need to be dealt with computationally. Corral potentials may be used to prevent trajectories from hitting this boundary,[29] and modifications of metadynamics to avoid artifacts at boundaries in CV space have been developed.[38]

A different possibility is using NLDR methods that can be fed with a predefined topology for the low-dimensional representation, such as Self Organizing Maps[39,40] or Generative Topographic Mapping.[41] However, for complex systems, the topology may not be known a priori. Finally, a more general approach is to split systematically the high-dimensional manifold into different patches with the topology of an open set in $\mathbb{R}^d$, and then apply dimensionality reduction on each patch separately. This method, which we are currently working on, also reduces

the metric-induced distortions of the low-dimensional embeddings. Furthermore, systematic partitioning may enable the analysis of systems exhibiting non-manifold behavior.[42] In fact, this Reference shows how, by partitioning the intrinsic manifold, one can use the systematic tools of algebraic topology to characterize the structure of a molecule's conformational space. A prerequisite of algebraic topology analysis, though, is a low-dimensional embedding devoid data collapse.

An important question concerns the applicability of data-driven CVs to complex molecules such as proteins. Interestingly, it has been suggested that increasing the size of peptides makes the effective dimensionality of the molecule smaller.[14] Thus, one can expect that statistical learning methods applied to proteins may help understand these complex systems with a few collective variables.[43] Once freed from topological obstructions and geometrical distortion, data-driven strategies to define CVs may deliver their full potential.

ACKNOWLEDGMENTS

We acknowledge the support of the European Research Council under the European Community 7th Framework Programme (FP7/2007-2013)/ERC Grant agreement No 240487.

[1] D. W. Borhani and D. E. Shaw, Journal of Computer-Aided Molecular Design **26**, 15 (2012).
[2] F. Pietrucci and A. Laio, Journal of Chemical Theory and Computation **5**, 2197 (2009).
[3] K. Pearson, Philosophical Magazine Series 6 **2**, 559 (1901).
[4] A. Amadei, A. B. Linssen, and H. J. Berendsen, Proteins **17**, 412 (1993).
[5] B. L. de Groot, X. Daura, a. E. Mark, and H. Grubmüller, Journal of molecular biology **309**, 299 (2001).
[6] O. F. Lange and H. Grubmüller, The Journal of Physical Chemistry B **110**, 22842 (2006).
[7] G. G. Maisuradze, A. Liwo, and H. a. Scheraga, Journal of molecular biology **385**, 312 (2009).
[8] C. C. David and D. J. Jacobs, "Principal component analysis: a method for determining the essential dynamics of proteins." in *Methods in molecular biology (Clifton, N.J.)*, Vol. 1084 (Springer, 2014) pp. 193–226.
[9] M. L. Teodoro, G. N. Phillips, and L. E. Kavraki, Journal of computational biology : a journal of computational molecular cell biology **10**, 617 (2003).
[10] V. Spiwok, P. Lipovová, and B. Králová, The Journal of Physical Chemistry B **111**, 3073 (2007).
[11] S. Michielssens, T. S. van Erp, C. Kutzner, A. Ceulemans, and B. L. de Groot, The Journal of Physical Chemistry B **116**, 8350 (2012).
[12] A. García, Physical Review Letters **68**, 2696 (1992).
[13] H. Noji, R. Yasuda, M. Yoshida, and K. Kinosita, Nature **386**, 299 (1997).
[14] R. Hegger, A. Altis, P. Nguyen, and G. Stock, Physical Review Letters **98**, 028102 (2007).
[15] J. B. Tenenbaum, V. de Silva, and J. C. Langford, Science (New York, N.Y.) **290**, 2319 (2000).
[16] S. T. Roweis and L. K. Saul, Science **290**, 2323 (2000).
[17] G. E. Hinton and R. R. Salakhutdinov, Science (New York, N.Y.) **313**, 504 (2006).
[18] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, Proceedings of the National Academy of Sciences of the United States of America **103**, 9885 (2006).
[19] W. M. Brown, S. Martin, S. N. Pollock, E. a. Coutsias, and J.-P. Watson, The Journal of Chemical Physics **129**, 064118 (2008).
[20] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, Proceedings of the National Academy of Sciences of the United States of America **107**, 13597 (2010).
[21] H. Stamati, C. Clementi, and L. E. Kavraki, Proteins **78**, 223 (2010).
[22] M. Ceriotti, G. A. Tribello, and M. Parrinello, Proceedings of the National Academy of Sciences of the United States of America **108**, 13023 (2011).
[23] M. Ceriotti, G. a. Tribello, and M. Parrinello, Journal of Chemical Theory and Computation **9**, 1521 (2013).
[24] R. R. Coifman and S. Lafon, Applied and Computational Harmonic Analysis **21**, 5 (2006).
[25] M. a. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, The Journal of Chemical Physics **134**, 124116 (2011).
[26] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, The Journal of Chemical Physics **134**, 135103 (2011).
[27] V. Spiwok and B. Králová, The Journal of Chemical Physics **135**, 224504 (2011).
[28] G. A. Tribello, M. Ceriotti, and M. Parrinello, Proceedings of the National Academy of Sciences of the United States of America **109**, 5196 (2012).
[29] B. Hashemian, D. Millán, and M. Arroyo, The Journal of Chemical Physics **139**, 214101 (2013).
[30] J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*, edited by J. A. Lee and M. Verleysen, Information Science and Statistics (Springer New York, New York, NY, 2007).
[31] Y. Xue, P. J. Ludovice, M. a. Grover, L. V. Nedialkova, C. J. Dsilva, and I. G. Kevrekidis, Computers & Chemical Engineering **51**, 102 (2013).
[32] M. Duan, M. Li, L. Han, and S. Huo, Proteins **82**, 2585 (2014).
[33] H. Whitney, Annals of Mathematics **37**, 645 (1936).
[34] X. Biarnés, A. Ardèvol, A. Planas, C. Rovira, A. Laio, and M. Parrinello, Journal of the American Chemical Society **129**, 10686 (2007).
[35] I. Jákli, S. J. Knak Jensen, I. G. Csizmadia, and A. Perczel, Chemical Physics Letters **547**, 82 (2012).
[36] J. McCleary, *Geometry from a Differentiable Viewpoint*, 2nd ed., Geometry from a Differentiable Viewpoint (Cambridge University Press, 2012).
[37] J. A. Lee and M. Verleysen, Neurocomputing **67**, 29 (2005).
[38] Y. Crespo, F. Marinelli, F. Pietrucci, and A. Laio, Physical Review E **81**, 055701 (2010).
[39] C. Malsburg, Kybernetik **14**, 85 (1973).
[40] T. Kohonen, Biological Cybernetics **43**, 59 (1982).
[41] C. M. Bishop, M. Svensén, and C. K. I. Williams, Neural Computation **10**, 215 (1998).
[42] S. Martin, A. Thompson, E. a. Coutsias, and J.-P. Watson, The Journal of Chemical Physics **132**, 234115 (2010).
[43] S. Piana and A. Laio, Physical Review Letters **101**, 208101 (2008).