

# Modeling and enhanced sampling of molecular systems with smooth and nonlinear data-driven collective variables

Behrooz Hashemian, Daniel Millán, and Marino Arroyo<sup>a)</sup>

*LaCàN, Universitat Politècnica de Catalunya – BarcelonaTech, Campus Nord, 08034 Barcelona, Spain*

(Received 16 July 2013; accepted 24 October 2013; published online 2 December 2013)

Collective variables (CVs) are low-dimensional representations of the state of a complex system, which help us rationalize molecular conformations and sample free energy landscapes with molecular dynamics simulations. Given their importance, there is need for systematic methods that effectively identify CVs for complex systems. In recent years, nonlinear manifold learning has shown its ability to automatically characterize molecular collective behavior. Unfortunately, these methods fail to provide a differentiable function mapping high-dimensional configurations to their low-dimensional representation, as required in enhanced sampling methods. We introduce a methodology that, starting from an ensemble representative of molecular flexibility, builds smooth and nonlinear data-driven collective variables (SandCV) from the output of nonlinear manifold learning algorithms. We demonstrate the method with a standard benchmark molecule, alanine dipeptide, and show how it can be non-intrusively combined with off-the-shelf enhanced sampling methods, here the adaptive biasing force method. We illustrate how enhanced sampling simulations with SandCV can explore regions that were poorly sampled in the original molecular ensemble. We further explore the transferability of SandCV from a simpler system, alanine dipeptide in vacuum, to a more complex system, alanine dipeptide in explicit water. © 2013 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4830403>]

## I. INTRODUCTION

Molecular dynamics (MD) simulations provide atomic resolution of important processes involving biomolecules, which complement experimental observations<sup>1</sup> and can help understand the relation between conformational changes and function.<sup>2</sup> MD can in principle establish a link between atomic motions and thermodynamic observables. Yet, in practice this goal is not easily realized. Leaving aside the accuracy of current force fields, the predictive ability of MD is mainly limited by sampling. Indeed, while femtosecond time steps are required for accurate and stable time integration, important phenomena such as molecular conformational changes involve a hierarchy of time scales spanning milliseconds and up.<sup>3</sup> This huge disparity, caused generically by metastability, makes the accurate sampling of the equilibrium distribution, and hence the evaluation of thermodynamics observables, extremely challenging even in highly specialized supercomputing platforms.<sup>4</sup> An additional issue in molecular simulations of complex systems is processing and extracting meaningful information out of large amounts of data contained in the numerical trajectories. To deal with these difficulties, we adopt a nonlinear intrinsic manifold model for molecular systems,<sup>5–7</sup> develop techniques complementary to nonlinear dimensionality reduction methods<sup>8</sup> to define smooth collective variables based on molecular ensembles, and enhance sampling with these variables.

Collective variables (CVs), also called reaction coordinates, order parameters, or slow variables depending on the

context, are low-dimensional representations of the state of a molecular system. CVs often capture the concerted nature of molecular conformational changes. They organize our understanding of the system, e.g., through a low-dimensional free energy surface (FES), and are at the core of a myriad of enhanced sampling methods, including metadynamics,<sup>9</sup> non-equilibrium work methods,<sup>10</sup> or the adaptive biasing force (ABF) method,<sup>11</sup> which we implement here (see Ref. 12 for a comprehensive review).

For simple systems, experience or intuition can guide the selection of CVs, which can take the form of distances between molecular groups or dihedral angles. However, for most systems of interest, this choice is far from obvious, which has motivated many attempts to systematize the selection of CVs. When a specific transition between two metastable states is considered, a number of methods have been proposed to identify transition paths.<sup>13–17</sup> Path collective variables provide two CVs, along and perpendicular to a given transition path.<sup>18</sup> To examine broader or higher-dimensional regions of conformational space, CVs based on linear combinations of modes have been proposed. These include CVs based on the normal modes of the linearized potential energy, or on statistical learning methods applied to a training set of molecular conformations, as in essential dynamics relying on principal component analysis (PCA).<sup>19,20</sup> Besides being routinely used to post-process molecular simulations, PCA has been used to drive enhanced sampling in combination with metadynamics.<sup>21,22</sup>

In recent years, it has been noted that molecular motion often occurs to a good approximation on nonlinear low-dimensional manifolds of dimension  $d \ll 3N$  where  $N$  is the number of particles,<sup>5,6</sup> sometimes referred to as slow

<sup>a)</sup> Author to whom correspondence should be addressed. Electronic mail: [marino.arroyo@upc.edu](mailto:marino.arroyo@upc.edu)

or intrinsic manifold.<sup>7</sup> Although in general it is far from obvious that one should expect such a manifold to be an inherent feature of complex molecular systems, modeling these in terms of nonlinear manifolds has shown to be fruitful in many instances. Such nonlinearity may arise from steric interactions amongst different protein domains, or upon relative rotations of subunits about molecular hinges.<sup>23</sup> This field has been fueled by the emergence of nonlinear dimensionality reduction (NLDR) techniques in the field of statistical learning,<sup>8</sup> which automatically identify nonlinear correlations hidden in high-dimensional data. Locally linear embedding (LLE)<sup>24</sup> and Isomap<sup>25</sup> are amongst the oldest and most successful methods, which have been applied in a wide variety of problems in science and engineering. In essence, these methods represent a set of high-dimensional data points in low-dimensions by trying to preserve some notion of similarity between the high-dimensional points.

Isomap has been shown to distill functionally meaningful nonlinear coordinates, and has been used to post-process an equilibrated trajectory of a coarse-grained protein.<sup>5</sup> Reference 6 presents a comprehensive comparison of NLDR methods applied to cyclo-octane conformations, and shows that low-dimensional embeddings may be non-manifold objects. Rather than geometric similarity, other authors focus on diffusion distances,<sup>26</sup> which account for the underlying Fokker-Planck operator. Reference 7 obtains nonlinear CVs of *n*-alkane chains through diffusion maps, and the approach is subsequently refined in Ref. 27. Diffusion maps provide a deep understanding of the physics, but their accurate estimation requires a good sampling of the equilibrium distribution, which may limit their applicability. Reference 28 recognizes that often MD trajectories densely sample basins around conformers connected by sparsely sampled paths, and proposes a new iterative NLDR method adapted to such ensembles called sketch-map, in the spirit of earlier variants of multidimensional scaling.<sup>29</sup> See Ref. 30 for a recent review.

A fundamental limitation of NLDR techniques in the present context is that they merely provide a low-dimensional representation of the molecular conformations present in the training molecular ensemble. Unlike PCA, most NLDR methods employed for studying molecular conformations are discrete in nature,<sup>31</sup> and do not provide a differentiable mapping between arbitrary atomic positions and CVs, required in enhanced sampling methods to evaluate the atomic forces resulting from a bias in the space of CVs. In order to provide a method that can be generally applied to discrete reduced-dimensionality embeddings, and make NLDR techniques easily applicable to modeling molecular systems with nonlinear manifolds, the goal of the present work is to develop techniques complementary to NLDR that take their output and automatically generate differentiable CVs. Some recent works point in the same direction, but the topic is far from being settled. Reference 6 construct mappings between low and high dimensions following the ideas of LLE, but it is not clear that such mappings are differentiable or can be evaluated at conformations outside the convex hull of the training molecular ensemble. This reference also implements neural networks autoencoder, which provides for

Reference 32 adapts diffusion maps to bias simulations with umbrella sampling, and Ref. 33 introduces a field-overlap procedure to combine sketch-map with metadynamics and accelerate conformational exploration. Reference 34 generalizes path collective variables<sup>18</sup> to higher dimensions, defines smooth CVs from the output of Isomap for cyclo-octane,<sup>6</sup> and reports on promising but not converged enhanced sampling simulations. Our work is similar in scope to this reference, by taking the output of Isomap to define smooth CVs and perform enhanced sampling. In different contexts, we have previously proposed techniques to smoothly represent intrinsic manifolds identified by NLDR, including the reduced modeling of mechanical systems,<sup>35</sup> point-set surface parametrization,<sup>36</sup> or stereotyped cell motility.<sup>37</sup>

Here, we introduce a general and flexible method to define smooth and nonlinear data-driven collective variables (SandCV). The input of this method is a molecular ensemble representative of the system's geometric variability, which does not need to be thermodynamically meaningful. Such an ensemble can be obtained from a variety of conformation exploration methods in MD,<sup>38,39</sup> or even from experiments.<sup>40</sup> By combining existing NLDR methods, a smooth parametrization of the intrinsic manifold, and geometric operations, we obtain a robust and general method that produces differentiable CVs, presented in Sec. II. SandCV is non-intrusive with regards to the enhanced sampling method, imposes a negligible computational overhead, and can be easily integrated in standard MD codes<sup>41,42</sup> in conjunction with free energy calculation libraries.<sup>43</sup> In Sec. III, we show its effectiveness with a benchmark system, alanine dipeptide, and combine it with ABF for enhanced sampling and free energy calculation. The conclusions are collected in Sec. IV.

## II. METHODS

### A. Problem statement

The methods presented here address the following problem. Given prior knowledge of a molecular system in terms of an ensemble of  $M$  conformations given by the Cartesian coordinates of  $N$  atoms,  $R = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_a, \dots, \mathbf{r}_M\} \subset \mathbb{R}^{3N}$ , the goal is to define a smooth function, referred to as collective variables,  $\mathcal{C} : \mathbb{R}^{3N} \rightarrow \mathbb{R}^d$  mapping any out-of-sample molecular conformation  $\mathbf{r} \in \mathbb{R}^{3N}$  into a low-dimensional representation  $\boldsymbol{\xi} = \mathcal{C}(\mathbf{r}) \in \mathbb{R}^d$ . These collective variables should quantitatively represent the state of the system. They should also be amenable to enhanced sampling MD techniques, i.e., explicit expressions of its derivatives should be available, and their evaluation should be robust and computationally efficient.

Our strategy for defining the CVs is data-driven, and hinges on the intrinsic manifold model for molecular systems and on statistical learning methods. We proceed in several steps detailed in Secs. II B–II D. We first identify the intrinsic manifold underlying the molecular ensemble with nonlinear dimensionality reduction methods in Sec. II B. These methods operate at the discrete level, and for this reason we then build a smooth representation of this manifold in Sec. II C. Finally, we map any out-of-sample conformation to low-dimensions by projecting it onto the intrinsic manifold, as elaborated in Sec. II D. We present these steps in a

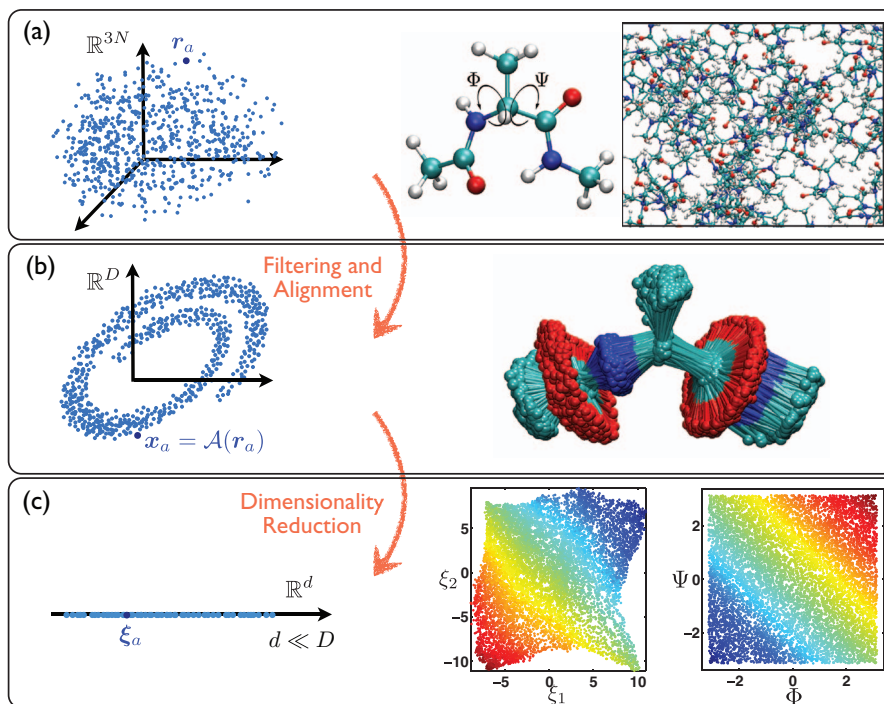


FIG. 1. Illustration of the main stages to identify the nonlinear intrinsic manifold, both in abstract terms (left) and also for a specific example involving alanine dipeptide (right). In the left plots, the points  $r_a$ ,  $x_a$ , and  $\xi_a$  are different representations of a given configuration. A molecular ensemble representing the geometric variability of the molecule (a) is first represented in a way that eliminates irrelevant atoms, translations, and rotations, which obscure the comparison of conformations, resulting in (b). We refer to this step as filtering and alignment, although it may involve resorting to internal coordinates. In (b), the features of the underlying intrinsic manifold may be already identified. The aligned configurations are then embedded in low dimensions by a NLDR method, here Isomap, preserving as much as possible the geodesic distance between high-dimensional configurations, thus revealing the structure of the intrinsic manifold (c). For reference, we compare the embedding coordinates  $\{\xi_1, \xi_2\}$  to the dihedral angles commonly used as CVs for this molecule. The rainbow coloring is the sum of the two dihedral angles in both embeddings.

concrete reference to the system studied here, alanine dipeptide. This small molecule, shown in Figure 1(a), has been extensively studied and is a benchmark for free energy calculation methods.<sup>11,17,18,22,27</sup> It is particularly well suited for our purposes because it exhibits metastability, good CVs are known (the dihedral angles  $\Phi$  and  $\Psi$ ), and these are highly nonlinear.

## B. Identifying the intrinsic manifold

Dimensionality reduction techniques try to identify the correlations hidden in a high-dimensional data set, the *training set*, in order to represent the data with less redundancy in low dimensions. Figure 1(a) provides an instance of molecular ensemble for alanine dipeptide, together with a schematic representation of the configurations in  $\mathbb{R}^{3N}$ . The low-dimensional representation provides a better understanding of the system, and can be more easily visualized. Most dimensionality reduction methods try to preserve the similarity between the points in high dimension. Before applying these techniques to molecular conformations described by the Cartesian coordinates of the atomic positions  $r_a \in \mathbb{R}^{3N}$ , one should note that such vectors cannot be directly compared to assess conformational similarity since a translation or rotation of the atomic positions leaves the conformation unchanged. Furthermore, some light atoms such as hydrogens present a very large variability and do not help in representing

Alignment is a standard procedure to remove rigid body transformations and correctly assess shape similarity between molecular configurations. Some alignment methods optimally superimpose each configuration in the ensemble to a reference configuration. Here, we use Procrustes superimposition without scaling and reflection,<sup>44</sup> applied on a filtered conformation consisting only of the backbone atoms of the molecule. Different subsets of atoms or groups of them may be more appropriate for other systems. Other alignment procedures are possible, such as transforming the Cartesian coordinates  $r_a \in \mathbb{R}^{3N}$  to a smooth contact map,<sup>45</sup> or resorting to internal coordinates.<sup>6</sup> Since MD codes typically apply forces in Cartesian coordinates, alignment maps to be used in conjunction with enhanced sampling MD techniques need to be differentiated with respect to the Cartesian coordinates, as elaborated later. Figure 1(b) illustrates the filtering and alignment procedure, which we symbolically denote as an operator  $\mathcal{A} : \mathbb{R}^{3N} \rightarrow \mathbb{R}^D$ . For Procrustes analysis on the  $N_B$  backbone atoms,  $D = 3N_B$ . The figure also suggests that alignment may reveal the intrinsic manifold of the molecule. After alignment, the molecular ensemble  $R$  is transformed to the set of points  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_a, \dots, \mathbf{x}_M\} \subset \mathbb{R}^D$ .

Adopting the intrinsic manifold paradigm to model the molecular system,<sup>5,19,46,47</sup> we seek to identify this nonlinear and low-dimensional structure underlying  $X$  with NLDR methods. We resort to Isomap,<sup>25</sup> although the procedure presented here is not specific to this method. Isomap builds on

pairwise distances  $dist_D(\mathbf{x}_a, \mathbf{x}_b)$  in high dimension, finds an optimal low dimensional embedding of the points in  $X$ , denoted by  $\Xi = \{\xi_1, \xi_2, \dots, \xi_a, \dots, \xi_M\} \subset \mathbb{R}^d$ , such that the matrix of pairwise distances in low dimension, given by  $dist_d(\xi_a, \xi_b)$ , is as close as possible to the high-dimensional counterpart. Algorithmically, finding this embedding involves linear algebra operations on the distance matrix. The key idea behind Isomap is replacing the Euclidean distance by the geodesic distance, that is, the length of the shortest path within the manifold connecting two points, when computing pairwise distances in high dimensions. To make this feasible, Isomap approximates the geodesic distance on a graph. The first step in Isomap is to build a weighted graph  $G$  whose vertices are the points  $X$ , and whose edges are the connections between  $K$  nearest neighbors, weighted by the length of these connections. In a second step, the geodesic distance between any pair of points  $dist_{D,G}(\mathbf{x}_a, \mathbf{x}_b)$  is approximated by the length of the shortest path connecting them in the graph. With this distance matrix capturing the low-dimensional geometry of the manifold, the embedding is obtained through the classical MDS procedure. Figure 1(c) illustrates how Isomap provides a discrete mapping between the input point set  $X \subset \mathbb{R}^D$  and the output point set  $\Xi \subset \mathbb{R}^d$ . Interestingly, the colormap in Figure 1(c) highlights the similarity between the Isomap embedding for alanine dipeptide and the embedding based on dihedral angles, although there is a nonlinear transformation between them. Note also that the Isomap embedding uses more collective information since it involves 10 atoms, instead of the 5 involved in the two dihedrals.

The estimation of the intrinsic dimensionality  $d$ , an input in NLDR algorithms, is not obvious for most systems, and is scale dependent in general.<sup>48</sup> The low-dimension  $d$  is selected by the user on the basis of previous knowledge about the system, of intrinsic dimension detection methods, or of computational convenience.<sup>8</sup> Note that although the configurations lying on a  $d$  dimensional nonlinear manifold can also be represented as a linear superposition of modes, the number of linear dimensions (dimension of the affine hull of  $X$ ) is necessarily larger than  $d$ .

### C. Parametrizing the intrinsic manifold

From the output of NLDR,  $\Xi$ , we introduce now a smooth parametrization of the intrinsic manifold, as illustrated in Figure 2. In this figure, the aligned configurations are represented in high dimensions by light blue points, which essentially lie on a nonlinear manifold, and their embedding in low-dimensions is represented by darker blue points lying on a segment. To represent mathematically and numerically this manifold (purple line), we define a parametrization  $\mathcal{M} : \Omega \subset \mathbb{R}^d \rightarrow \mathbb{R}^D$  of the form

$$\mathcal{M}(\xi) = \sum_{i=1}^L p_i(\xi) \mathbf{y}_i, \quad (1)$$

where  $p_i(\xi)$  are smooth basis functions associated to a set of landmarks  $\eta_i$  represented by black crosses, see Figure 2. We denote by  $\Omega$  a region in  $\mathbb{R}^d$  delimited by the points in  $\Xi$ . We solve for the coefficients  $\mathbf{y}_i$  by fitting

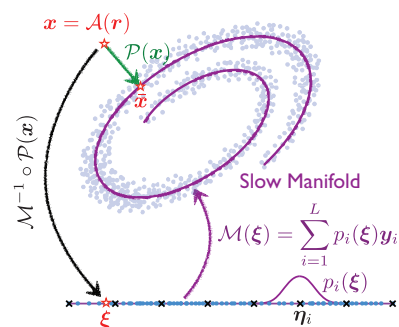


FIG. 2. The intrinsic manifold underlying the set of  $M$  aligned configurations (light-blue points) is parametrized from the low-dimensional embedding with a linear combination of  $L$  ( $\ll M$ ) basis functions  $p_i(\xi)$ , resulting in the purple line. The coefficients  $\mathbf{y}_i$  are obtained through a least-squares fit. An out-of-sample point  $\mathbf{x}$  is labeled in low-dimensional space by first obtaining its closest-point projection on the manifold,  $\bar{\mathbf{x}}$ , and then finding its pre-image through  $\mathcal{M}$ .

$\mathcal{M}$  to the data in a least-squares sense, i.e., by minimizing

$$\sum_{a=1}^M |\mathbf{x}_a - \mathcal{M}(\xi_a)|^2, \quad (2)$$

which involves solving a  $L \times L$  linear system of equations. Here and elsewhere,  $|\cdot|$  denotes the Euclidean norm.

The support of the basis functions  $p_i(\xi)$  should be wide enough to filter the out-of-manifold variability, but not too wide to blunt the features of the intrinsic manifold. The support of these basis functions should also observe sampling, to avoid ill-conditioning of the least-squares fit associated to narrow functions in poorly sampled regions. While systematic procedures are desirable, this support is chosen here heuristically, and then verified by visually inspecting the image of  $\mathcal{M}$  together with the original ensemble.

Here, we use local maximum-entropy basis functions given by

$$p_i(\xi) = \frac{e^{-\beta_i |\xi - \eta_i|^2 + \lambda \cdot (\xi - \eta_i)}}{\sum_j e^{-\beta_j |\xi - \eta_j|^2 + \lambda \cdot (\xi - \eta_j)}}, \quad (3)$$

where the parameter  $\beta_i$  sets the width of the basis functions locally, and  $\lambda$  is a Lagrange multiplier that enforces that the basis functions reproduce exactly affine functions, which can be found by minimizing the denominator in the equation above.<sup>49,50</sup> Denoting by  $h$  the typical spacing between landmark points, if  $\beta_i h^2$  is very large, then the basis functions become narrow and faceted, converging to the barycentric coordinates of the underlying Delaunay simplicial complex. If instead  $\beta_i h^2$  is small, then the basis functions become very wide and smooth. If  $\lambda = 0$ , then the basis functions result in the Shepard approximants<sup>51</sup> used in path collective variables.<sup>18</sup> The local maximum-entropy basis functions are smooth, can accurately represent point-set manifolds,<sup>36</sup> and can deal with non-uniform sets of landmarks in any dimension  $d$ . However, many other choices are possible for parametrizing  $\mathcal{M}$ .

### D. SandCV: Putting it all together

Although the dynamics of the molecule closely follows the intrinsic manifold, represented numerically by the set  $\mathcal{M}(\Omega)$ , configurations are not constrained to lie on the manifold.



CV must be able to assign a label  $\xi$ , representative of the state of the system, to any out-of-sample aligned configuration  $\mathbf{x}$ . The closest-point projection onto the intrinsic manifold, denoted by  $\mathcal{P}(\mathbf{x})$ , is a very natural geometric concept that accomplishes this. The closest-point projection onto a smooth manifold is itself smooth in a neighborhood of the manifold and away from the boundaries.<sup>52</sup> In practical terms, the fluctuations around the intrinsic manifold should remain small compared to its local curvature for  $\mathcal{P}$  to remain differentiable. Thus, we define the SandCV as the composition of three maps

$$\mathcal{C}(\mathbf{r}) = \mathcal{M}^{-1} \circ \mathcal{P} \circ \mathcal{A}(\mathbf{r}), \quad (4)$$

as depicted graphically in Figure 2. Thus, by definition, the sub-manifolds of constant  $\mathcal{C}(\mathbf{r})$  in configuration space are perpendicular to the intrinsic manifold. Although this expression is conceptually illuminating, in practice the SandCV is evaluated by minimizing

$$|\mathcal{M}(\xi) - \mathcal{A}(\mathbf{r})|^2, \quad (5)$$

with respect to  $\xi$ . Numerically, we resort to Newton's method, possibly with a few quasi-Newton iterations with line-search,<sup>53</sup> to solve this  $d$ -dimensional optimization problem.

Equation (4) is also useful to derive the Jacobian of the SandCV. Applying the chain rule, and denoting by  $\mathbf{D}$  the matrix of partial derivatives of a mapping, the Jacobian of the proposed CV can be computed as the product of three matrices

$$\underbrace{\mathbf{DC}(\mathbf{r})}_{d \times (3N)} = \underbrace{\mathbf{DM}^{-1}(\bar{\mathbf{x}})}_{d \times D} \underbrace{\mathbf{DP}(\mathbf{x})}_{D \times D} \underbrace{\mathbf{DA}(\mathbf{r})}_{D \times (3N)}, \quad (6)$$

where we have highlighted the dimensions of the Jacobian matrices and for conciseness we introduce  $\mathbf{x} = \mathcal{A}(\mathbf{r})$  and  $\bar{\mathbf{x}} = \mathcal{P}(\mathbf{x})$ . The Jacobian of alignment is method-dependent. For Procrustes superimposition, we refer to Appendix A. In Appendix B, we derive an exact expression for the  $d \times D$  matrix  $\mathbf{DM}^{-1}(\bar{\mathbf{x}})\mathbf{DP}(\mathbf{x})$  in Eq. (6), given by

$$\mathbf{DM}^{-1}(\bar{\mathbf{x}})\mathbf{DP}(\mathbf{x}) = \{\mathbf{DM}^T(\xi)\mathbf{DM}(\xi) - \mathbf{D}^2\mathcal{M}(\xi)(\mathbf{x} - \bar{\mathbf{x}})\}^{-1}\mathbf{DM}^T(\bar{\mathbf{x}}). \quad (7)$$

It is clear that if either the intrinsic manifold is flat ( $\mathbf{D}^2\mathcal{M} = 0$ ) or  $\mathbf{x}$  is on the manifold, the derivative of the closest point projection is the identity and this expression simplifies to the pseudo-inverse of  $\mathbf{DM}$ , as indicated by the inverse function theorem. In practice, we find that correctly accounting for  $\mathbf{DP}$  is essential to accurately compute  $\mathbf{DC}(\mathbf{r})$ , and that the computational overhead of this procedure in every time-step of the MD simulation is negligible.

To illustrate how  $\mathbf{DC}(\mathbf{r})$  is needed in enhanced sampling methods, consider we want to bias the MD simulation with a potential defined in CV space,  $U(\xi)$ . This is the case in umbrella sampling or metadynamics. The bias can be seen as a potential in terms of all-atom configurations by composing it with the CVs,  $U \circ \mathcal{C}(\mathbf{r})$ . The force on the CVs is then  $\mathbf{F}_\xi = -\partial U / \partial \xi$ . By the chain rule, we can map these forces to the atoms of the molecule,  $\mathbf{F}_r(\mathbf{r}) = \mathbf{DC}(\mathbf{r})\mathbf{F}_\xi(\xi)$ , as illustrated in Figure 3. In the ABF method used here,<sup>11</sup> the force  $\mathbf{F}_\xi$  is an estimation of the thermodynamic force on the CVs,

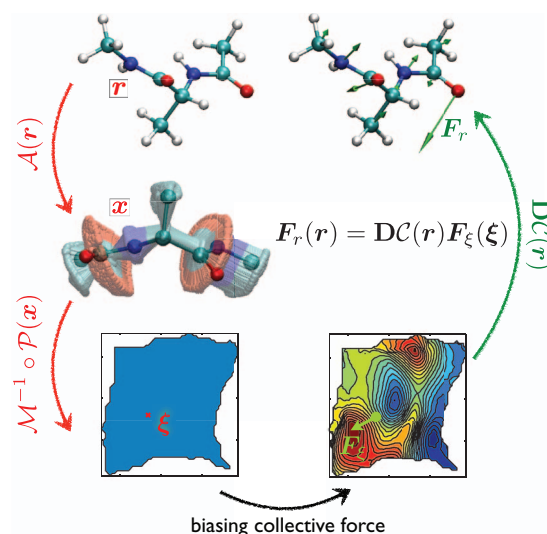


FIG. 3. Illustration of how bias forces in SandCV space are applied to the molecule in an enhanced sampling MD simulation. In each time-step, the all-atom configuration  $\mathbf{r}$  is aligned to a common frame, and then mapped to the low-dimensional embedding,  $\xi$ , through the closest-point projection on the intrinsic manifold. The bias force  $\mathbf{F}_\xi$  is then evaluated in low-dimension, e.g., as an approximation of the derivative of the free energy, and mapped to the atoms with the Jacobian of SandCV.

## E. Dealing with general manifolds

Isomap and other NLDR algorithms can only succeed in identifying  $d$ -manifolds of simple topology, which admit an embedding in  $\mathbb{R}^d$ . Some molecules evolve on manifolds with complex topology, due, for instance, to the cyclic nature of rotations about bonds. In fact, alanine dipeptide is an instance of such a system, and as shown later, its intrinsic manifold has the topology of a two-dimensional torus. A consequence pointed out in Ref. 6 is that low-dimensional embeddings may become non-manifold even if the system evolves on a well-defined manifold embedded in high-dimensions. In encountering topological obstructions, NLDR methods collapse configurations that are distinct in high-dimensions. Even for manifolds of simple topology but with significant intrinsic curvature, NLDR methods such as Isomap may provide highly distorted embeddings of poor quality. All these difficulties are potentially serious, and arise when one attempts to describe globally (with a single chart) the intrinsic manifold.

These difficulties were noted in Ref. 36, where a general computational method to address them was proposed, and applied to point-set surface processing. By recursively partitioning the ensemble  $X$  with specialized algorithms,<sup>54</sup> it is easy to ensure that each partition has simple topology and admits an embedding in its intrinsic dimension without excessive distortion. By applying NLDR and a smooth parametrization to each of these partitions, the manifold can be described by an atlas of charts, which can then be glued using a partition of unity. This general methodology is in principle applicable to molecular systems as well. However, in the current exploratory work, we have adopted an *ad hoc* approach to deal with the topology of the intrinsic manifold of alanine dipeptide, presented in Sec. III.

### III. RESULTS AND DISCUSSIONS

We exercise the proposed method studying alanine dipeptide (N-acetyl-N'-methyl-L-alanylamine), also known as di-alanine, both in vacuum and in explicit water. As mentioned earlier, this small peptide has become a testbed for free energy calculations. The backbone dihedral angles  $\Phi$  and  $\Psi$  have been shown to be effective collective variables, although the significance of other dihedral angles has been examined.<sup>32,55</sup>

We first describe the implementation of Isomap to alanine dipeptide, which requires addressing the topology of its intrinsic manifold. We then build the SandCV from the resulting low-dimensional embedding, and perform enhanced sampling simulations using the ABF method. We show the effectiveness of SandCV as a smooth CV by showing the convergence of the enhanced sampling method. These simulations provide FES, which are then compared with those computed along the dihedral angles. To show the possibilities of SandCV in more realistic situations with non-ideal sampling of the intrinsic manifold, we apply the methodology starting from a training set obtained by a crude exploration method, which does not visit significant regions of configuration space. We show that SandCV, combined with the ABF method, can bridge over these gaps and explore these regions. Finally, we examine the transferability of SandCV obtained under simple simulation conditions (vacuum) to more complex conditions (explicit water).

All simulations were performed with version 2.8 of the NAMD<sup>41</sup> molecular dynamics code with the CHARM22 force field<sup>56</sup> and a Langevin thermostat. For the simulations in explicit water, we use the particle mesh Ewald method<sup>57</sup>

for long-range electrostatic forces and periodic boundary conditions. We implement SandCV in a stand-alone C++ code that implements the vectorial version of the ABF method<sup>11</sup> and communicates with NAMD through a TCL interface to obtain configurations and return forces on the atoms.

#### A. Isomap low-dimensional embedding

We initially consider an ideal sampling of the intrinsic manifold, obtained by running two 100 ns simulations of alanine dipeptide in water and vacuum at 300 K, and sampling configurations every 10 ps, resulting in 10 000 configurations. In these simulations, sampling was enhanced with the ABF method along the two dihedral angles, resulting in a nearly uniform sampling in dihedral space.

As discussed in Sec. II D, the nontrivial topology of the manifold underlying some datasets is a generic obstacle for dimensionality reduction. For alanine dipeptide, due to the periodicity of the dihedral angles, the intrinsic manifold has the topology of the two-dimensional torus. Consequently, dimensionality reduction techniques will collapse distant parts of the manifold, thereby failing to identify it properly, unless  $d \geq 3$ . Figure 4(a) shows the three-dimensional Isomap embedding for the ideally sampled ensemble of alanine dipeptide in vacuum. The results in explicit water are similar. Such representation is not dimensionally optimal, as the intrinsic dimension is 2, and does not fill a region in the low-dimensional embedding. Yet, it is very useful because it allows us to visually identify tearing curves on the manifold. We use this information to eliminate edges in the Isomap graph

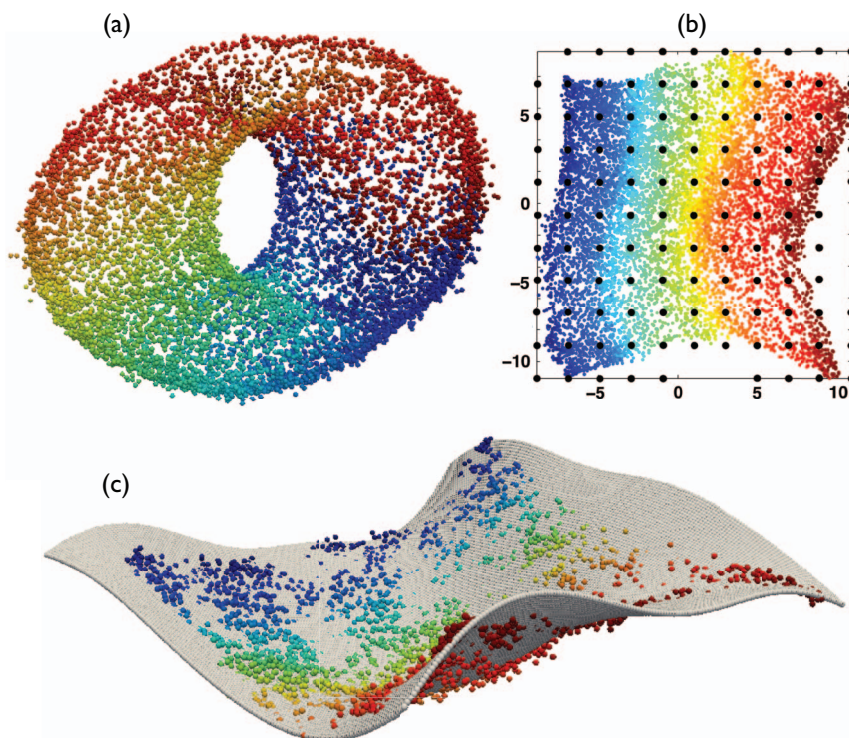


FIG. 4. Low-dimensional embedding of alanine dipeptide in vacuum with Isomap. (a) The three-dimensional embedding of alanine dipeptide shows it has the topology of a torus, and allows us to identify tearing curves to simplify the topology. (b) Two-dimensional embedding after tearing. The landmark points used in the parametrization are marked as black dots. (c) Smooth parametrization of the intrinsic manifold (gray surface) visualized on the three-dimensional Isomap

$G$  connecting vertices separated by the tearing curves. This *ad hoc* method is effective in the present system, but may be insufficient in others. In Sec. II E, we have suggested a general method to deal with general manifolds, which is beyond the scope of the present paper.

The procedure we adopt here results in a two-dimensional embedding that respects the local geometric structure of the intrinsic manifold, yet introduces artificial boundaries, see Figure 4(b). Figure 4(c) illustrates the smooth parametrization of the intrinsic manifold as a surface in the three-dimensional teared Isomap embedding.

## B. Enhanced sampling with SandCV

We illustrate next how SandCV is successfully coupled with an enhanced sampling algorithm, here ABF, to compute free energies. The success of enhanced sampling strategies can be established by the uniformity of sampling along the CVs as the simulation proceeds.<sup>11</sup> Furthermore, equilibrium properties such as the thermodynamic force on the CVs should converge with simulation time. Figure 5 provides numerical evidence of the convergence of two ABF simulations: one based on dihedral angles, and another one based on SandCV. We present the results for alanine dipeptide in vacuum, but those in water are similar. Figure 5(a) shows the convergence of the thermodynamic force as a function of simulation time. It can be seen that ABF simulations based on either dihedrals or SandCV exhibit similar convergence, and the semi-logarithmic scale highlights the exponential convergence, as theoretically predicted.<sup>58</sup> The different panels in Figure 5(b) show how the histograms of conformations in SandCV space converge to a uniform distribution, as expected. These results show that SandCV, based on statistical learning, captures the metastability of the system, since it is known that enhanced sampling methods become ineffective if the remaining transversal coordinates exhibit metastability.<sup>59</sup> This fact is not surprising, since we have already noted that the Isomap embedding closely mimics an embedding based on dihedral angles. This simulation also shows that all the on-line operations behind SandCV executed in every time-step of the simulation (parametrization of the intrinsic manifold, closest-point projection, and Jacobian of the CV) can be robustly implemented in a standard MD code.

## C. Free energy comparisons

Free energy surfaces are subjective in that they fundamentally depend on the CVs along which they vary, and are not insensitive to reparametrizations of CV space.<sup>60</sup> Although this fact does not have consequences on physical observables such as rates of conformational changes,<sup>61</sup> it complicates a meaningful comparison between FES along different CVs. However, since we have found that SandCVs based on NLDR closely correlate with dihedral angles, we attempt this comparison next.

We consider three types of CVs: (1) the usual dihedral angles, (2) the SandCV based on the two-dimensional embedding given by the dihedral angles, and (3) the SandCV based on the two-dimensional Isomap embedding, as described above. In (2), we can easily retain periodicity of the CV. In (3), we tear the manifold to simplify its topology and place a corral potential around the boundary of the embedding to confine the trajectory within the region of interest. The corral potential is not biased by the enhanced sampling method. It should be high enough so that trajectories do not escape the region of interest, and narrow enough not to shrink excessively this region. The iso-contours near the boundary of the free-energy landscapes in Figure 6(c) give an idea of the width of this corral potential in our simulations. By analyzing the system in vacuum and in water, we end up with six different sets of CVs. The corresponding FES are computed with 100 ns ABF simulations, which we have shown to converge in Figure 5. We parametrize the intrinsic manifold with about  $11 \times 11$  landmark points, and choose  $\beta = 1/h^2$  for the basis functions in Eq. (3), where  $h$  is the typical spacing between landmark points.

Dihedral CVs and the SandCV based on the dihedral embedding are not necessarily in direct correspondence. For instance, the former only involves 5 atoms of the molecule, while the latter involves the 10 backbone atoms of the alignment. Yet, Figures 6(a) and 6(b) show that the resulting FES are very similar, both in vacuum and in water. The FES obtained with the SandCV based on Isomap exhibits the same features, but nonlinearly mapped from dihedral space.

Table I provides a quantitative comparison of free energy discrepancies between the main features of the FES, including free energy basins, hills, and saddle points. All FES

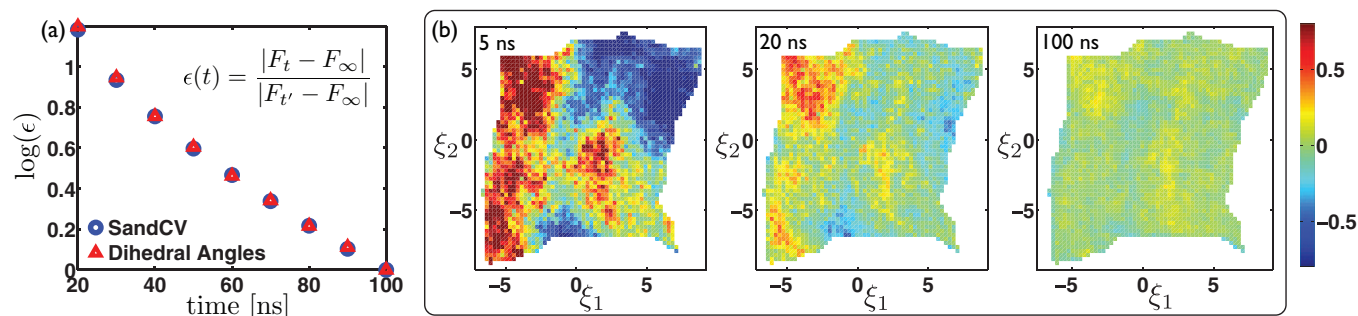


FIG. 5. Convergence of ABF simulations for alanine dipeptide in vacuum. (a) Convergence of the thermodynamic forces along the CVs for a simulation biased along dihedral angles (red triangles) and along SandCV (blue circles). We plot the error in the forces ( $\epsilon$ ) relative to the equilibrium forces  $F_\infty$  obtained with a reference long simulation (here 200 ns long), and normalized with the error at  $t' = 100$  ns. (b) Snapshots of the normalized deviation from uniform sampling of the histograms in the SandCV simulation, that is,  $(c_{ij} - \bar{c})/\bar{c}$  where  $c_{ij}$  is the number of samples in bin  $(i, j)$ ,  $\bar{c} = \sum_{i,j} c_{ij}/nBins$ , and  $nBins$  is the number of bins in the MD trajectory and  $nBins$  is the number of bins in the uniform sampling.



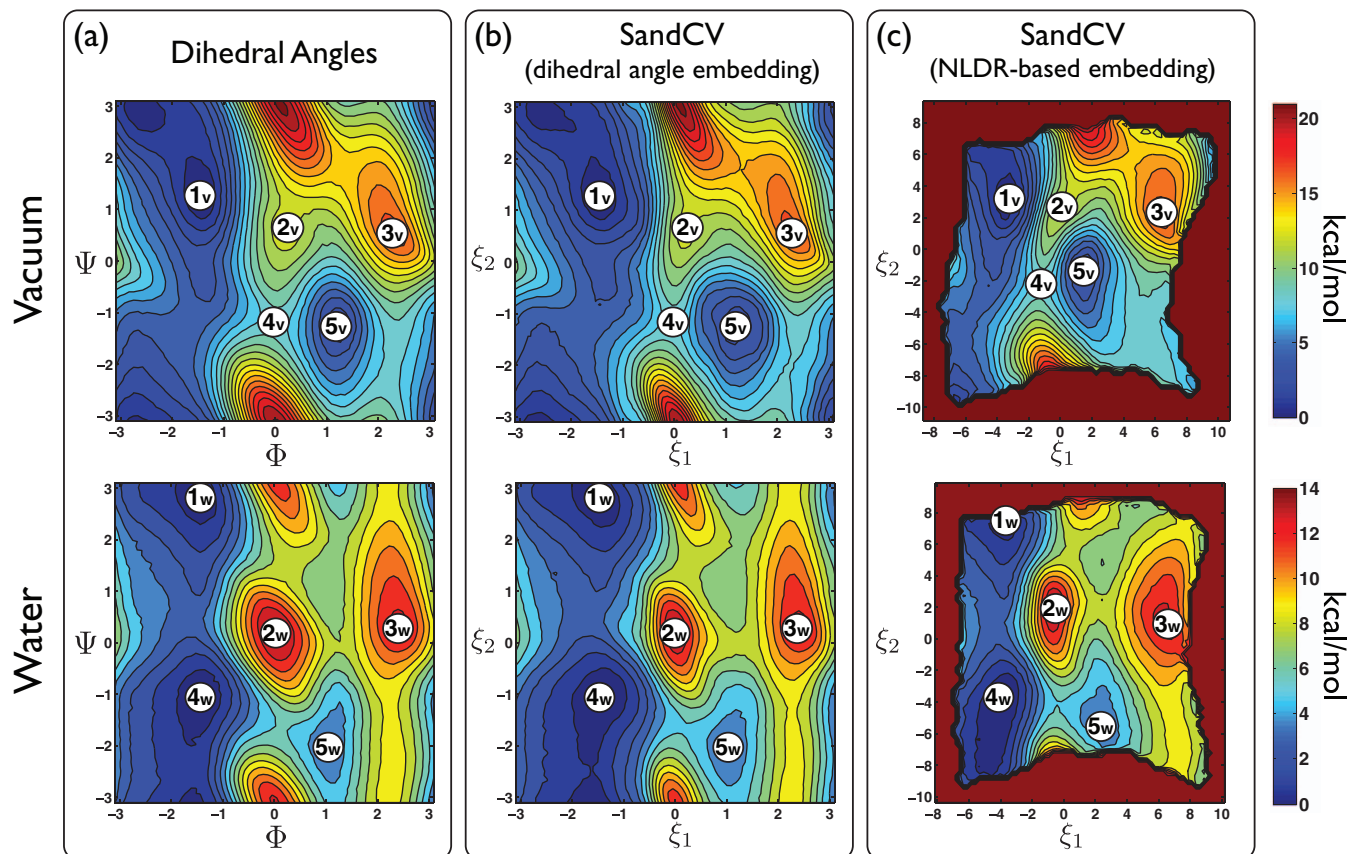


FIG. 6. Comparison of the free energy surface of alanine dipeptide in vacuum (v) and water (w) along three different sets of CVs. (a) Backbone dihedral angles ( $\Phi$  and  $\Psi$ ), (b) SandCV based on the two-dimensional embedding given by the dihedral angles, and (c) SandCV based on a two-dimensional Isomap embedding.

were shifted so that the free energy of Point 1 vanishes. The agreement of the free energy differences is remarkable, with a maximum deviation smaller than 0.3 kcal/mol, that is about  $0.5 k_B T$ . These results further emphasize the close similarity between dihedral angles and the data-driven CVs based on NLDR, which lead to nearly identical free energy differences between the main features of the FES.

#### D. SandCV on a realistic ensemble with poorly sampled regions

In practice, MD trajectories do not sample well regions of high free energy, even with configuration space exploration techniques. This is a fundamental hurdle in statistical learning approaches to identify CVs. We consider next a realis-

tic application of SandCV in combination with Isomap, in which a training set of configurations resulting from a simple exploration methodology does not sample large regions in dihedral space. We first run a set of short simulations of alanine dipeptide in water with different starting points randomly selected from a high-temperature simulation and quenched to 310 K. From 1400 starting configurations, we run short 100 ps simulations sampled every 20 fs. We end up with 1400 trajectories with 5000 configurations, and a total of  $7 \times 10^6$  configurations. Since these are too many points for a standard Isomap implementation and since that many points do not bring additional value to the geometrical description of the manifold, we decimate the data in two steps on the basis of geometric similarity. First, we select 1000 quasi-uniformly distributed configurations out of each

TABLE I. Free energy differences at the points marked in Figure 6. The units are in kcal/mol. ( $1k_B T = 0.596$  kcal/mol).

| Embedding |                                       | 1    | 2      | 3      | 4      | 5     |
|-----------|---------------------------------------|------|--------|--------|--------|-------|
| Vacuum    | Dihedral angles                       | 0.00 | 12.644 | 17.532 | 8.364  | 2.340 |
|           | SandCV with dihedral angles embedding | 0.00 | 12.438 | 17.547 | 8.351  | 2.216 |
|           | SandCV with NLDR-based embedding      | 0.00 | 12.420 | 17.534 | 8.438  | 2.309 |
| Water     | Dihedral angles                       | 0.00 | 13.908 | 12.997 | -0.084 | 3.913 |
|           | SandCV with dihedral angles embedding | 0.00 | 13.648 | 12.998 | -0.151 | 3.854 |
|           | SandCV with NLDR-based embedding      | 0.00 | 13.718 | 13.064 | -0.305 | 3.971 |



trajectory, chosen in such a way that the Euclidean distance between any pair of aligned configurations within a trajectory is larger than a cutoff. Second, the resulting  $1.4 \times 10^6$  configurations are joined and decimated with another cut-off criterion, ending up with 9163 configurations. This number of high-dimensional points is easily manageable by Isomap, which is memory intensive for large training sets, and contains all the relevant information present in the original data. We resort to the procedure described in Sec. III A to tear the manifold.

Figure 7(a) shows the two-dimensional embedding of these configurations as translucent points. The large unexplored regions (“holes”) are apparent. By placing 93 uniformly spaced landmark points  $\eta_i$  and taking  $\beta = 1/h^2$  in Eq. (3), the smooth representation of the intrinsic manifold in Eq. (1) bridges over the holes, and therefore the SandCV bridges over the corresponding regions in configuration space. By performing an ABF simulation based on this SandCV, we populate the holes and end up with a nearly uniform sampling, illustrative of the convergence of the free energy calculation, see Figure 7(b). The resulting FES is represented in Figure 7(a), and highlights the correspondence between the holes and the regions of high free energy. This experiment suggests that SandCV can be used for configuration space ex-

ploration, by bridging free energy basins separated by high barriers. Since it is likely that the description of the intrinsic manifold is poor over the holes, it is possible to proceed in two steps, by first filling these as demonstrated here, and then recomputing the SandCV with the enhanced training set. This procedure assumes that the energy barriers responsible for the holes are much smaller than those giving rise to the nonlinear manifold character of the conformation ensemble, e.g., rigid bonds, bending angles, or steric constraints.

### E. Transferability of SandCV

The previous example illustrates that producing an adequate training set can be challenging, particularly for large molecules in explicit water. A natural remedy would be to build the SandCV from a simpler model, such as a coarse-grained protein model, and then use it to enhance sampling in the full model. We explore here this idea by building a SandCV with a training set of alanine dipeptide in vacuum, as in Figure 6(c) top, and then biasing with it a simulation of the molecule in explicit water. The latter system is much more difficult to simulate due to the larger number of particles and the long-range electrostatic forces.

Figure 8(a) shows a reference FES in water computed with the dihedral angles, while Figure 8(b) shows the FES

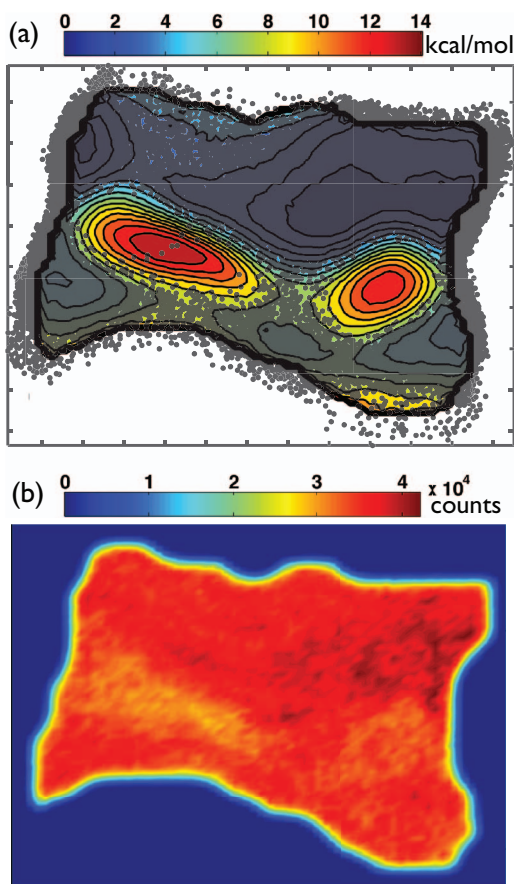


FIG. 7. SandCV on a training set with unexplored regions. The low-dimensional embedding of the training set exhibits large holes (a), which correspond to regions of high free energy. An ABF simulation based on SandCV can bridge over these regions and end up with nearly uniform sampling in the

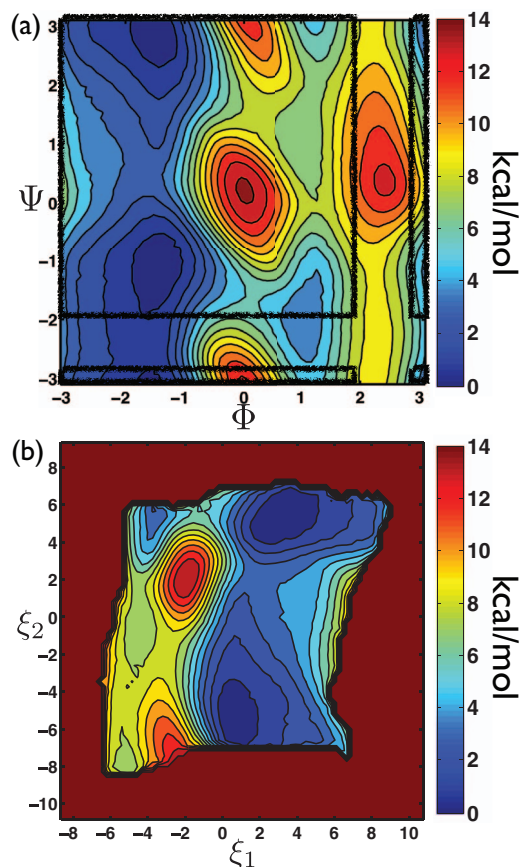


FIG. 8. Exploring the transferability of SandCV. The free energy of alanine dipeptide in water, shown in (a) in terms of dihedral angles, is computed along SandCV based on an ensemble in vacuum (b). For comparison, the FES in (b) should be rotated by 180°, and the support of (b) has been

of the molecule in water but along the vacuum SandCV. Despite the CV is defined for a different and simpler system, we find that the ABF simulation converges as in previous examples. Furthermore, the similarity of the landscape is remarkable, suggesting SandCV are transferable for this system. Broadly speaking, this suggests that even if the underlying intrinsic manifold of a simplified system is noticeably different from that of a complex system, a simulation such as that presented here can produce a good training set of configurations of the complex system, which can then be the basis of a better SandCV.

#### IV. CONCLUSION

We have introduced a general method to model molecular systems with SandCV. These CVs can be non-intrusively combined with standard enhanced sampling molecular dynamics methods. The input of the method is an ensemble representative of the flexibility of the molecule, which does not need to be thermodynamically meaningful. The geometric structure hidden in this ensemble is revealed by existing nonlinear dimensionality reduction methods (here Isomap), and then further processed to define collective variables  $\mathcal{C}(\mathbf{r})$  and its derivatives.

We exercise SandCV with alanine dipeptide both in vacuum and in explicit water. This system is a benchmark for free energy calculations and has well-known and highly nonlinear collective variables, two of its backbone dihedral angles  $\Phi$  and  $\Psi$ . We demonstrate the effectiveness of the method by providing numerical evidence of the convergence of enhanced sampling simulations based on SandCV. These simulations also show that the method can be integrated in standard MD codes and combined with an off-the-shelf enhanced sampling method. We then compare the free energy surfaces obtained with the ABF method in combination with dihedral angles, and several flavors of SandCV. This comparison shows that a systematic machine learning method such as SandCV provides a description of the system that closely mimics one based on the conventional dihedral angles.

In practice, data-driven collective variables are limited by the difficulty of producing training sets of configurations that sample the intrinsic manifold with sufficient density.<sup>62</sup> We explore this issue in two ways. First, we consider a realistic ensemble after a simple configuration exploration step, which fails to visit large regions. We show that SandCV can bridge over these regions, and then populate them in a subsequent enhanced sampling simulation. Second, we show that the initial training set can be obtained with a simplified system, for instance, alanine dipeptide in vacuum, and then the resulting SandCV transferred to a complex system, much more expensive to simulate, here alanine dipeptide in water.

SandCV provides a flexible framework composed of distinct conceptual and algorithmic blocks. The first block is the alignment of the molecule, which here is performed with Procrustes superimposition. The second block is the identification of the intrinsic manifold through nonlinear dimensionality reduction methods. Here, we use Isomap. The third block is the smooth parametrization of the intrinsic manifold, performed via the use of any arbitrary 2D or 3D

is the closest-point projection of out-of-sample configurations onto the intrinsic manifold, to label arbitrary configurations by the low-dimensional embedding coordinates. The first three of these ingredients can be replaced by alternative algorithms, adapted to specific systems or computer codes, without affecting the general methodology. For instance, we have explored alignment of proteins through smooth contact maps, and dimensionality reduction with iterative methods that minimize a nonlinear cost function.<sup>35</sup> The embedding in low dimensions can also combine physical insight and statistical learning techniques. Complex systems may require describing the intrinsic manifold and the CVs through multiple charts, as discussed in Sec. II E. Currently, we are implementing SandCV within a popular free energy calculation plugin,<sup>43</sup> which will allow us to combine this method with many free energy calculation strategies and make it available to the scientific community.

#### 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. M.A. acknowledges the support received through the prize ‘‘ICREA Academia’’ for excellence in research, funded by the Generalitat de Catalunya.

#### APPENDIX A: PROCRUSTES SUPERIMPOSITION

A differentiable alignment is an essential part of SandCV. Here, we summarize Procrustes superimposition and obtain its derivatives.

##### 1. Alignment as an optimization problem

We represent any given configuration  $\mathbf{r}_a \in \mathbb{R}^{3N}$  as a  $N_s \times 3$  matrix  $\mathbf{X}$ , where a subset of  $N_s$  atoms out of the  $N$  atoms of the full molecule have been selected. We find the optimal rotation matrix  $\mathbf{R} \in \mathbb{R}^{3 \times 3}$  and translation vector  $\mathbf{t} \in \mathbb{R}^{1 \times 3}$  by minimizing the cost function

$$\text{cost}(\mathbf{R}, \mathbf{t}) = |(\mathbf{X}\mathbf{R} + \text{rep}[\mathbf{t}]) - \mathbf{X}_{ref}|^2,$$

where  $\mathbf{X}_{ref}$  is a reference configuration and  $\text{rep}: \mathbb{R}^{1 \times 3} \rightarrow \mathbb{R}^{N_s \times 3}$  is a function that produces a matrix with  $N_s$  copies of its argument. This optimization problem can be solved resorting to the singular value decomposition (SVD). First, we define the matrix

$$\mathbf{M} = (\mathbf{X}_{ref} - \text{rep}[\boldsymbol{\mu}_{ref}])^T (\mathbf{X} - \text{rep}[\boldsymbol{\mu}]), \quad (\text{A1})$$

where  $\boldsymbol{\mu} \in \mathbb{R}^{1 \times 3}$  and  $\boldsymbol{\mu}_{ref} \in \mathbb{R}^{1 \times 3}$  are, respectively, the vectors of the average atom position of the given configuration and the reference one. Invoking the SVD,  $\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}^T$  where  $\mathbf{U}$  and  $\mathbf{V}$  are orthonormal matrices whose columns are eigenvectors of  $\mathbf{M}\mathbf{M}^T$  and  $\mathbf{M}^T\mathbf{M}$ , respectively, and  $\mathbf{S}$  is a diagonal matrix of singular values. The optimal rotation and translation matrices are then

$$\mathbf{R} = \mathbf{V}\mathbf{U}^T, \quad \mathbf{t} = \boldsymbol{\mu}_{ref} - \mathbf{R}\boldsymbol{\mu}$$

Since the SVD is not unique, it has to be chosen appropriately so that  $\mathbf{R}$  is a proper rotation, i.e.,  $\det \mathbf{R} = 1$ .

## 2. Derivatives of the singular value decomposition

We first obtain the derivatives of  $\mathbf{M}$  with respect to  $X_{ir}$ , the  $r$ th component of  $i$ th atom, in terms of the SVD

$$\frac{\partial \mathbf{M}}{\partial X_{ir}} = \frac{\partial \mathbf{U}}{\partial X_{ir}} \mathbf{S} \mathbf{V}^T + \mathbf{U} \frac{\partial \mathbf{S}}{\partial X_{ir}} \mathbf{V}^T + \mathbf{U} \mathbf{S} \frac{\partial \mathbf{V}^T}{\partial X_{ir}}.$$

By pre- and post-multiplying this equation by  $\mathbf{U}^T$  and  $\mathbf{V}$ , respectively, and recalling that they are orthonormal, we have

$$\mathbf{U}^T \frac{\partial \mathbf{M}}{\partial X_{ir}} \mathbf{V} = \mathbf{U}^T \frac{\partial \mathbf{U}}{\partial X_{ir}} \mathbf{S} + \frac{\partial \mathbf{S}}{\partial X_{ir}} + \mathbf{S} \frac{\partial \mathbf{V}^T}{\partial X_{ir}} \mathbf{V}.$$

We then subtract this equation from its transpose, and take into account that  $\mathbf{S}$  is symmetric to find

$$\begin{aligned} \mathbf{V}^T \frac{\partial \mathbf{M}^T}{\partial X_{ir}} \mathbf{U} - \mathbf{U}^T \frac{\partial \mathbf{M}}{\partial X_{ir}} \mathbf{V} \\ = \left[ \mathbf{V}^T \frac{\partial \mathbf{V}}{\partial X_{ir}} - \mathbf{U}^T \frac{\partial \mathbf{U}}{\partial X_{ir}} \right] \mathbf{S} + \mathbf{S} \left[ \frac{\partial \mathbf{U}^T}{\partial X_{ir}} \mathbf{U} - \frac{\partial \mathbf{V}^T}{\partial X_{ir}} \mathbf{V} \right]. \end{aligned} \quad (\text{A3})$$

By defining

$$\Omega(X_{ir}) = \mathbf{V}^T \frac{\partial \mathbf{V}}{\partial X_{ir}} + \frac{\partial \mathbf{U}^T}{\partial X_{ir}} \mathbf{U},$$

and considering that  $\mathbf{V}^T \frac{\partial \mathbf{V}}{\partial X_{ir}}$  and  $\frac{\partial \mathbf{U}^T}{\partial X_{ir}} \mathbf{U}$  are skew-symmetric,<sup>63</sup> we rewrite Eq. (A3) as

$$\Omega(X_{ir}) \mathbf{S} + \mathbf{S} \Omega(X_{ir}) = \mathbf{V}^T \frac{\partial \mathbf{M}^T}{\partial X_{ir}} \mathbf{U} - \mathbf{U}^T \frac{\partial \mathbf{M}}{\partial X_{ir}} \mathbf{V}. \quad (\text{A4})$$

This gives us a set of equations that can be solved if the singular values are non-degenerate. The solution can be written in the indicial notation, for  $p \neq q$ ,

$$\Omega_{pq}(X_{ir}) = \frac{1}{S_p + S_q} \left[ V_{mp} \frac{\partial M_{nm}}{\partial X_{ir}} U_{nq} - U_{mp} \frac{\partial M_{mn}}{\partial X_{ir}} V_{nq} \right], \quad (\text{A5})$$

without summation over  $p$  and  $q$ , where

$$\frac{\partial M_{mn}}{\partial X_{ir}} = (X_{im} - \mu_m) \delta_{nr}.$$

## 3. Derivatives of the Procrustes superimposition

The derivatives of the Procrustes superimposition follow from the derivatives of the rotation matrix and translation vector with respect to  $\mathbf{X}$ . Since  $\mathbf{t}$  depends only on  $\mathbf{R}$ , see Eq. (A2),

we only need to compute the derivatives of  $\mathbf{R}$ , which follow

from

$$\begin{aligned} \frac{\partial \mathbf{R}}{\partial X_{ir}} &= \frac{\partial \mathbf{V}}{\partial X_{ir}} \mathbf{U}^T + \mathbf{V} \frac{\partial \mathbf{U}^T}{\partial X_{ir}} \\ &= \mathbf{V} \left( \mathbf{V}^T \frac{\partial \mathbf{V}}{\partial X_{ir}} \right) \mathbf{U}^T + \mathbf{V} \left( \frac{\partial \mathbf{U}^T}{\partial X_{ir}} \mathbf{U} \right) \mathbf{U}^T \\ &= \mathbf{V} \left( \mathbf{V}^T \frac{\partial \mathbf{V}}{\partial X_{ir}} + \frac{\partial \mathbf{U}^T}{\partial X_{ir}} \mathbf{U} \right) \mathbf{U}^T \\ &= \mathbf{V} \Omega(X_{ir}) \mathbf{U}^T. \end{aligned}$$

## APPENDIX B: JACOBIAN OF THE CLOSEST-POINT PROJECTION

As argued in the text, the Jacobian of SandCV follows from

$$\mathbf{D}\mathcal{C}(\mathbf{r}) = \mathbf{D}\mathcal{M}^{-1}(\bar{\mathbf{x}}) \mathbf{D}\mathcal{P}(\mathbf{x}) \mathbf{D}\mathcal{A}(\mathbf{r}),$$

where  $\mathbf{D}\mathcal{A}(\mathbf{r})$  can be computed as indicated in Appendix A. On the other hand, from the inverse function theorem  $\mathbf{D}\mathcal{M}^{-1}(\bar{\mathbf{x}})$  is the pseudo-inverse of the matrix  $\mathbf{D}\mathcal{M}$

$$\mathbf{D}\mathcal{M}^{-1}(\bar{\mathbf{x}}) = [\mathbf{D}\mathcal{M}^T(\xi) \mathbf{D}\mathcal{M}(\xi)]^{-1} \mathbf{D}\mathcal{M}^T(\xi), \quad (\text{B1})$$

where  $\xi = \mathcal{M}^{-1}(\bar{\mathbf{x}})$ . Thus, we only need the Jacobian of the closest point projection  $\mathbf{D}\mathcal{P}(\mathbf{x})$ . Since this matrix always appears in the formulation of SandCV multiplied by  $\mathbf{D}\mathcal{M}^{-1}(\bar{\mathbf{x}})$ , we directly compute the product  $\mathbf{D}\mathcal{M}^{-1}(\bar{\mathbf{x}}) \mathbf{D}\mathcal{P}(\mathbf{x})$ .

In the vicinity of a smooth manifold, the closest-point projection can be written as

$$\bar{\mathbf{x}} = \mathcal{P}(\mathbf{x}) = \mathbf{x} - d(\mathbf{x}) \mathbf{n}(\mathbf{x}),$$

where  $\mathbf{n}(\mathbf{x})$  is a normal vector to the manifold at  $\bar{\mathbf{x}}$  and  $d(\mathbf{x})$  is the distance to the manifold. Taking the derivative of this equation, we find

$$\mathbf{D}\mathcal{P}(\mathbf{x}) = \mathbf{I}_D - d(\mathbf{x}) \mathbf{D}\mathbf{n}(\mathbf{x}) - \mathbf{n}(\mathbf{x}) \otimes \mathbf{D}d(\mathbf{x}). \quad (\text{B2})$$

Multiplying Eq. (B2) by  $\mathbf{D}\mathcal{M}^{-1}(\bar{\mathbf{x}})$ , recalling Eq. (B1), noting that the rows of  $\mathbf{D}\mathcal{M}^T$  are tangent to the manifold, and introducing a normal unit vector field  $\mathcal{N}$  as a function of the embedding coordinates, i.e.,  $\mathbf{n}(\mathbf{x}) = \mathcal{N} \circ \mathcal{M}^{-1} \circ \mathcal{P}(\mathbf{x})$ , we have

$$\begin{aligned} \mathbf{D}\mathcal{M}^{-1} \mathbf{D}\mathcal{P} &= [\mathbf{D}\mathcal{M}^T \mathbf{D}\mathcal{M}]^{-1} \mathbf{D}\mathcal{M}^T \\ &\quad \times \{ \mathbf{I}_D - d(\mathbf{x}) \mathbf{D}\mathcal{N} \mathbf{D}\mathcal{M}^{-1} \mathbf{D}\mathcal{P} \}. \end{aligned}$$

Noting that  $\mathbf{D}\mathcal{M}^{-1} \mathbf{D}\mathcal{P}$  appears in both sides of this equation and solving for it, we obtain

$$\begin{aligned} \mathbf{D}\mathcal{M}^{-1} \mathbf{D}\mathcal{P} &= \{ \mathbf{I}_d + d[\mathbf{D}\mathcal{M}^T \mathbf{D}\mathcal{M}]^{-1} \\ &\quad \times \mathbf{D}\mathcal{M}^T \mathbf{D}\mathcal{N} \}^{-1} \mathbf{D}\mathcal{M}^{-1}, \end{aligned} \quad (\text{B3})$$

and by further using of Eq. (B1) and simplifying, we find

$$\mathbf{D}\mathcal{M}^{-1} \mathbf{D}\mathcal{P} = \{ \mathbf{D}\mathcal{M}^T \mathbf{D}\mathcal{M} + d \mathbf{D}\mathcal{M}^T \mathbf{D}\mathcal{N} \}^{-1} \mathbf{D}\mathcal{M}^T, \quad (\text{B4})$$

where we still need to compute the Jacobian of the unit normal vector  $\mathbf{D}\mathcal{N}$ . Noting that the rows of  $\mathbf{D}\mathcal{M}^T$  are perpendicular to  $\mathcal{N}$ , we take the derivative of their inner product,

$$\mathbf{D}(\mathbf{D}\mathcal{M}^T \mathcal{N}) = \mathbf{D}^2 \mathcal{M}^T \mathcal{N} + \mathbf{D}\mathcal{M}^T \mathbf{D}\mathcal{N} = 0,$$



and therefore  $\mathbf{D}\mathcal{M}^T\mathbf{D}\mathcal{N} = -\mathbf{D}^2\mathcal{M}^T\mathcal{N}$ . Plugging this expression into Eq. (B4), we finally obtain

$$\mathbf{D}\mathcal{M}^{-1}(\bar{\mathbf{x}})\mathbf{D}\mathcal{P}(x) = \{\mathbf{D}\mathcal{M}^T(\xi)\mathbf{D}\mathcal{M}(\xi) - \mathbf{D}^2\mathcal{M}(\xi)(x - \bar{\mathbf{x}})\}^{-1}\mathbf{D}\mathcal{M}^T(\bar{\mathbf{x}}).$$

- <sup>1</sup>M. Sotomayor and K. Schulten, *Science (N.Y.)* **316**, 1144 (2007).
- <sup>2</sup>M. Osadchy and R. Kolodny, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 12301 (2011).
- <sup>3</sup>K. A. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus, and D. Kern, *Nature (London)* **450**, 913 (2007).
- <sup>4</sup>R. O. Dror, M. O. Jensen, D. W. Borhani, and D. E. Shaw, *J. Gen. Physiol.* **135**, 555 (2010).
- <sup>5</sup>P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 9885 (2006).
- <sup>6</sup>W. M. Brown, S. Martin, S. N. Pollock, E. a. Coutsias, and J.-P. Watson, *J. Chem. Phys.* **129**, 064118 (2008).
- <sup>7</sup>A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 13597 (2010).
- <sup>8</sup>J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*, Information Science and Statistics, edited by M. I. Jordan, R. Nowak, and B. Schölkopf (Springer, New York, NY, 2007).
- <sup>9</sup>A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12562 (2002).
- <sup>10</sup>C. Jarzynski, *Phys. Rev. Lett.* **78**, 2690 (1997).
- <sup>11</sup>E. Darve, D. Rodríguez-Gómez, and A. Pohorille, *J. Chem. Phys.* **128**, 144120 (2008).
- <sup>12</sup>C. Chipot and A. Pohorille, *Free Energy Calculations*, Springer Series in Chemical Physics Vol. 86, edited by C. Chipot and A. Pohorille (Springer, Berlin, 2007).
- <sup>13</sup>R. Olender and R. Elber, *J. Chem. Phys.* **105**, 9299 (1996).
- <sup>14</sup>H. Jónsson, G. Mills, and K. W. Jacobsen, “Nudged elastic band method for finding minimum energy paths of transition,” in *Classical and Quantum Dynamics in Condensed Phase Simulations*, edited by B. J. Berne, G. Ciccotti, and D. F. Coker (World Scientific, 1998), Chap. 16, pp. 385–404.
- <sup>15</sup>W. E, W. Ren, and E. Vanden-Eijnden, *Phys. Rev. B* **66**, 052301 (2002).
- <sup>16</sup>D. Passerone, M. Ceccarelli, and M. Parrinello, *J. Chem. Phys.* **118**, 2025 (2003).
- <sup>17</sup>W. Ren, E. Vanden-Eijnden, P. Maragakis, and W. E, *J. Chem. Phys.* **123**, 134109 (2005).
- <sup>18</sup>D. Branduardi, F. L. Gervasio, and M. Parrinello, *J. Chem. Phys.* **126**, 054103 (2007).
- <sup>19</sup>A. Amadei, A. B. Linssen, and H. J. Berendsen, *Proteins* **17**, 412 (1993).
- <sup>20</sup>I. Daidone and A. Amadei, *WIREs: Comput. Mol. Sci.* **2**, 762 (2012).
- <sup>21</sup>V. Spiwok, P. Lipovová, and B. Králová, *J. Phys. Chem. B* **111**, 3073 (2007).
- <sup>22</sup>V. Spiwok, B. Králová, and I. Tvaroska, *J. Mol. Model.* **14**, 995 (2008).
- <sup>23</sup>H. Noji, R. Yasuda, M. Yoshida, and K. Kinosita, *Nature (London)* **386**, 299 (1997).
- <sup>24</sup>S. T. Roweis and L. K. Saul, *Science (N.Y.)* **290**, 2323 (2000).
- <sup>25</sup>J. B. Tenenbaum, V. de Silva, and J. C. Langford, *Science* **290**, 2319 (2000).
- <sup>26</sup>R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, *Multiscale Model. Simul.* **7**, 842 (2008).
- <sup>27</sup>M. a. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, *J. Chem. Phys.* **134**, 124116 (2011).
- <sup>28</sup>M. Ceriotti, G. A. Tribello, and M. Parrinello, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 13023 (2011).
- <sup>29</sup>J. Sammon, *IEEE Trans. Comput.* **C-18**, 401 (1969).
- <sup>30</sup>M. a. Rohrdanz, W. Zheng, and C. Clementi, *Annu. Rev. Phys. Chem.* **64**, 295 (2013).
- <sup>31</sup>L. V. der Maaten, E. Postma, and J. van den Henrik, “Dimensionality reduction: A comparative review,” Technical Report No. TiCC TR 2009-005 (Tilburg Centre for Creative Computing, 2009).
- <sup>32</sup>A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, *J. Chem. Phys.* **134**, 135103 (2011).
- <sup>33</sup>G. A. Tribello, M. Ceriotti, and M. Parrinello, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 5196 (2012).
- <sup>34</sup>V. Spiwok and B. Králová, *J. Chem. Phys.* **135**, 224504 (2011).
- <sup>35</sup>D. Millán and M. Arroyo, *Comput. Methods Appl. Mech. Eng.* **261–262**, 118 (2013).
- <sup>36</sup>D. Millán, A. Rosolen, and M. Arroyo, *Int. J. Numer. Methods Eng.* **93**, 685 (2013).
- <sup>37</sup>M. Arroyo, L. Heltai, D. Millán, and A. DeSimone, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17874 (2012).
- <sup>38</sup>D. J. Earl and M. W. Deem, *Phys. Chem. Chem. Phys.* **7**, 3910 (2005).
- <sup>39</sup>G. a. Tribello, M. Ceriotti, and M. Parrinello, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 17509 (2010).
- <sup>40</sup>R. B. Fenwick, S. Esteban-Martín, B. Richter, D. Lee, K. F. a. Walter, D. Milovanovic, S. Becker, N. Lakomek, C. Griesinger, and X. Salvatella, *J. Am. Chem. Soc.* **133**, 10336 (2011).
- <sup>41</sup>J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, *J. Comput. Chem.* **26**, 1781 (2005).
- <sup>42</sup>D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, *J. Comput. Chem.* **26**, 1701 (2005).
- <sup>43</sup>M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, and R. a. Broglia, *Comput. Phys. Commun.* **180**, 1961 (2009).
- <sup>44</sup>P. M. Kroonenberg, W. J. Dunn, and J. J. F. Commandeur, *J. Chem. Inf. Comput. Sci.* **43**, 2025 (2003).
- <sup>45</sup>M. Bonomi, D. Branduardi, F. L. Gervasio, and M. Parrinello, *J. Am. Chem. Soc.* **130**, 13938 (2008).
- <sup>46</sup>A. García, *Phys. Rev. Lett.* **68**, 2696 (1992).
- <sup>47</sup>R. Hegger, A. Altis, P. Nguyen, and G. Stock, *Phys. Rev. Lett.* **98**, 028102 (2007).
- <sup>48</sup>P. Grassberger and I. Procaccia, *Physica D* **9**, 189 (1983).
- <sup>49</sup>M. Arroyo and M. Ortiz, *Int. J. Numer. Methods Eng.* **65**, 2167 (2006).
- <sup>50</sup>A. Rosolen, D. Millán, and M. Arroyo, *Int. J. Numer. Methods Eng.* **82**, 868 (2010).
- <sup>51</sup>H. Wendland, *Scattered Data Approximation* (Cambridge University Press, Cambridge, 2005).
- <sup>52</sup>S. J. Ruuth and B. Merriman, *J. Comput. Phys.* **227**, 1943 (2008).
- <sup>53</sup>J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed. (Springer Science+Business Media, LLC, New York, NY, 2006), p. 664.
- <sup>54</sup>G. Karypis and V. Kumar, *SIAM J. Sci. Comput.* **20**, 359 (1998).
- <sup>55</sup>P. G. Bolhuis, C. Dellago, and D. Chandler, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5877 (2000).
- <sup>56</sup>A. D. MacKerell, D. Bashford, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, *J. Phys. Chem. B* **102**, 3586 (1998).
- <sup>57</sup>U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, *J. Chem. Phys.* **103**, 8577 (1995).
- <sup>58</sup>T. Lelièvre, M. Rousset, and G. Stoltz, *Free Energy Computations: A Mathematical Perspective* (Imperial College Press, 2010).
- <sup>59</sup>L. Sutto, S. Marsili, and F. L. Gervasio, *WIREs: Comput. Mol. Sci.* **2**, 771 (2012).
- <sup>60</sup>W. E, W. Ren, and E. Vanden-Eijnden, *J. Phys. Chem. B* **109**, 6688 (2005).
- <sup>61</sup>D. Frenkel, *Eur. Phys. J. Plus* **128**, 10 (2013).
- <sup>62</sup>M. Ceriotti, G. a. Tribello, and M. Parrinello, *J. Chem. Theory Comput.* **9**, 1521 (2013).
- <sup>63</sup>T. Papadopoulos and M. I. A. Lourakis, in *Computer Vision - ECCV 2000*, Lecture Notes in Computer Science Vol. 1842 (Springer Berlin/Heidelberg, 2000), p. 554.