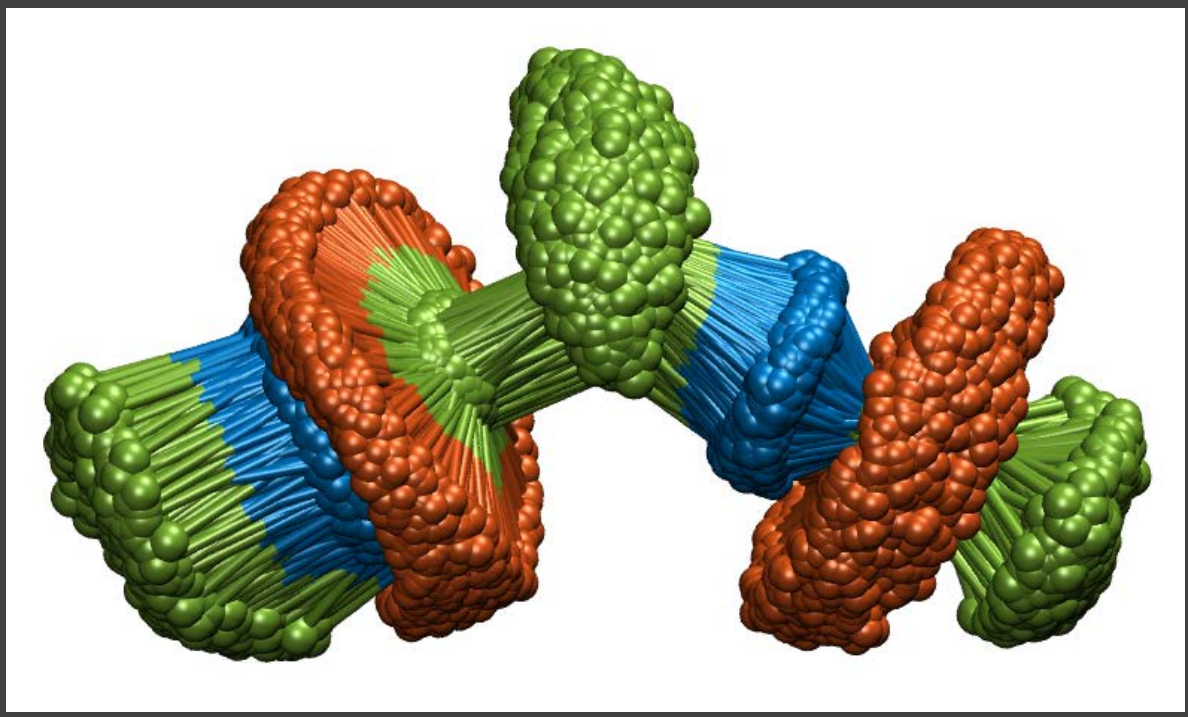
MACHINE LEARNING IN MULTISCALE MOLEDING AND SIMULATION OF MOLECULAR SYSTEMS

Behrooz Hashemian



Doctoral Thesis
Barcelona, April 2015

MACHINE LEARNING IN MULTISCALE MOLEDING AND SIMULATION OF MOLECULAR SYSTEMS

Behrooz Hashemian

Doctoral Thesis
Advisor: Marino Arroyo
Barcelona, April 2015

School of Mathematics and Statistics
Program of PhD in Applied Mathematics

Acknowledgements

I have not reached where I am now merely by my own efforts. Not only because of the interdisciplinary essence of my work, which requires a variety of knowledge and skills from different fields, but also because working together with other people has boosted my individual abilities and has gave me unique experiences. It is my pleasure to thank people that have helped me during this journey.

First and foremost, I want to thank Prof. Marino Arroyo, not only as my brilliant, knowledgeable, hardworking advisor, but also as a polite, disciplined, open-minded individual that he is. I owe him a particular debt of thanks for all his supports and guidance.

I want to express my gratitude to the faculty of LaCàN and Department of Applied Mathematics III, particularly Prof. Sonia Fernàndez, Prof. Jose Muños, Prof. Irene Arias, Prof. Josep Sarrate, Prof. Pedro Díez, and Prof. Antonio Huerta. I also am truly thankful to Prof. Antonio Rodriguez for all his invaluable supports.

I want to express my thankfulness for the indispensable assistance of administrative staff of LaCàN, Imma Rius and Susanna Algarra, department of applied mathematics III, Maria Angels Huguet and Carme Lòpez, and school of mathematics and statistics, Raquel Caparrós and Carme Capdevilla. I also offer thanks to LaCàN IT

support, David Ortin. Moreover, I would like to acknowledge staff of UPC International and Corporate Relations Bureau, especially Anna Maria Fabregas for all her assistance.

I would like to express my deep gratitude to Prof. Helmut Grubmüller for hosting me at Max-Planck Institute of Biophysical Chemistry and Prof. Michele Parrinello for hosting me at Department of Chemistry and Applied Biosciences of ETH-Zürich. I owe an irremediable debt of gratitude to Prof. Ignasi Vilajosana, who inspired me, showed me new doors of opportunities and supported me to reach them.

I am fortunate to be able to count so many talented and big-hearted colleagues whose company over the last years has been an honor for me: Dr. Daniel Millán, Dr. Adrian Rosolen, Dr. Amir Abdollahi, Dr. Mohammad Rahimi, Dr. Susanta Gosh, Dr. Luca Heltai, Dr. Christian Peco, Dr. David Modesto, Dr. Carlos Braga, Dr. Juan Vanegas, Kuan Zhang, Bin Li, Flaviu Simon, Aditya Vasudevan, Alejandro Torres, Dimitri Kaurin, Omid Javadzadeh, and Vahid Ziaei. I also want to thank Dr. Kazem Kamran, Dr. Mohammad Koohi, and Dr. Narges Dialami, for all the afternoon coffee times that they have shared with me during the last six years.

My deep gratitude goes to my family, whose unconditional supports have always been with me to follow my dreams and afforded me the opportunity to do so. Above all, I want to express my heartfelt appreciation to my compassionate, understanding and gifted friend, Dr. Parisa Farzam, who has been with me throughout the whole journey and together we look forward to a wonderful future.

Abstract

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. However, identifying a representative set of CVs for a given system is far from obvious, and most often relies on physical intuition or partial knowledge about the systems. An inappropriate choice of CVs is misleading and can lead to inefficient sampling. Thus, there is a need for systematic approaches to effectively identify CVs.

In recent years, machine learning techniques, especially nonlinear dimensionality reduction (NLDR), have shown their ability to automatically identify the most important collective behavior of molecular systems. These methods have been widely used to visualize molecular trajectories. However, in general they do not provide a differentiable mapping from high-dimensional configuration space to their low-dimensional representation, as required in enhanced sampling methods, and they cannot deal with systems with inherently nontrivial conformational manifolds.

In the fist part of this dissertation, 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 and show how it can be non-intrusively combined with off-the-shelf enhanced sampling methods, here the adaptive biasing force method. SandCV identifies the system's conformational manifold, handles out-of-manifold conformations by a closest point projection, and exactly computes the Jacobian of the resulting CVs. We also illustrate how enhanced sampling simulations with SandCV can explore regions that were poorly sampled in the original molecular ensemble.

We then demonstrate that NLDR methods face serious obstacles when the underlying CVs 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. Additionally, we characterize flexibility of alanine dipeptide molecule and show that it evolves around a flat torus in four-dimensional space.

In the final part of this thesis, we propose a novel method, atlas of collective variables, that systematically overcomes topological obstacles, ameliorates the geometrical distortions and thus allows NLDR techniques to perform optimally in molecular simulations. This method automatically partitions the configuration space and treats each partition separately. Then, it connects these partitions from the statistical mechanics standpoint.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The functionality of biomolecules such as proteins has been shown to be essentially ruled by their structure and their ability to change it (Osadchy and Kolodny, 2011). Thus characterizing the relationship between structure and function has become an active path of research in many areas of science. Despite recent developments in experimental methods such as x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy (Esteban-Martín et al., 2010), experiments provide information about conformational changes at an atomic scale to a limited extent. Computational methods such as molecular dynamics (MD) provide complementary information on the molecular mechanisms (Sotomayor and Schulten, 2007) and have become a substantial tool to guide and interpret experiments.

Molecular dynamics (Alder and Wainwright, 1957, 1959) follows the time evolution of a system, modeled at the atomic level, by solving numerically the Newton's equations of motion with certain initial and boundary conditions. For a system with N atoms, the state of each atom is specified by the set of position and momentum vectors $\{\boldsymbol{r}, \boldsymbol{p}\}$, which leads to a $6N$-dimensional phase space. However, a system does not explore all the phase space: its trajectory is given by the

system's Hamiltonian $H(\boldsymbol{r}, \boldsymbol{p})$. A Hamiltonian is said to be separable when the energetic contributions of the momentum and position can be independently added, which is the case for the most molecular systems of interest. In this case, the Hamiltonian is written as

$$H(\boldsymbol{r}, \boldsymbol{p}) = V(\boldsymbol{r}) + K(\boldsymbol{p}), \tag{1.1}$$

where V is the potential energy and K is the kinetic energy. In general, K is a quadratic form involving the particle masses. In contrast, the potential energy is not straightforward to calculate and accounts for all the particle interactions. In practice, it is often approximated by empirical potentials (or force fields) based on experiments and quantum mechanical calculations.

Statistical mechanics, on the other hand, relates the thermodynamical properties of molecules to microscopic states through probability theory. This theory thus provides a connection between MD simulations and observable properties. In biophysical systems, the number of particles (N), the volume (V) and temperature (T) can be often considered to be fixed, and thus the microscopic states are distributed according to the canonical ensemble, also called NVT ensemble (Frenkel and Smit, 2002). The canonical probability measure is written as

$$\mu_{NVT}(d\boldsymbol{r}d\boldsymbol{p}) = \frac{e^{-\beta H(\boldsymbol{r},\boldsymbol{p})}d\boldsymbol{r}d\boldsymbol{p}}{Z_\mu}, \tag{1.2}$$

where $\beta = 1/k_B T$ is the inverse temperature (k_B is Boltzmann constant and T is the temperature) and Z_μ is the normalization constant, called partition function (Lelièvre et al., 2010),

$$Z_\mu = \int_{D\times\mathbb{R}^{3N}} e^{-\beta H(\boldsymbol{r},\boldsymbol{p})}d\boldsymbol{r}d\boldsymbol{p}. \tag{1.3}$$

where D is the configuration space and $D \times \mathbb{R}^{3N}$ is the phase space. For

separable Hamiltonian, the partition function can be written as

$$Z_\mu = Z_\nu Z_\kappa, \quad Z_\nu = \int_D e^{-\beta V(\boldsymbol{r})} d\boldsymbol{r}, \quad Z_\kappa = \left(\frac{2\pi}{\beta}\right)^{3N/2} \prod_{i=1}^{N} m_i^{3/2}. \tag{1.4}$$

A core concept in thermodynamics and in modern studies of biomolecular systems is *free energy*, which is defined as the natural logarithm of the partition function,

$$A = -\frac{1}{\beta} \ln Z_\mu. \tag{1.5}$$

In many applications the interesting quantity is the free energy difference between macroscopic states of the system, which allows us to quantify the relative likelihood of different states. This *relative free energy* does not depend on the momentum contribution in the partition function and only relies on the configurational sampling,

$$A = -\frac{1}{\beta} \ln \int_D e^{-\beta V(\boldsymbol{r})} d\boldsymbol{r} + \text{constant}. \tag{1.6}$$

However, MD simulations encounter a phenomenal challenge to sufficiently sample configuration space to accurately calculate the macroscopic quantities. This challenge arises since the time step for numerical stability needs to be in the order of a few femtoseconds, and on the other hand many phenomena of interest, such as conformational transitions, occur at the millisecond and up time scales. This time-scale disparity in systems with metastability makes it very difficult to obtain accurate sampling, and consequently hinders the connection between simulations and experiments. To overcome this issue, a number of enhanced sampling methods have been proposed, such as umbrella sampling (Torrie and Valleau, 1977), non-equilibrium work methods (Jarzynski, 1997), metadynamics (Laio and Parrinello, 2002)

and adaptive biasing force (ABF) (Darve et al., 2008) (see Chipot and Pohorille (2007) for a comprehensive review). Some of these methods use an estimated relative free energy to guide the system in such a way that it avoids oversampled states and instead visits poorly sampled states. A macroscopic state is defined as the collection of all possible configurations $\boldsymbol{r}$ satisfying the constraint $\mathcal{C}(\boldsymbol{r}) = \boldsymbol{\xi}$, where $\mathcal{C}$ is a differentiable mapping from high-dimensional configuration space to a low-dimensional space. The free energy of these states reads as

$$A(\boldsymbol{\xi}) = -\frac{1}{\beta} \ln \int_D e^{-\beta V(\boldsymbol{r})} \delta(\mathcal{C}(\boldsymbol{r}) - \boldsymbol{\xi}) d\boldsymbol{r}. \tag{1.7}$$

$\mathcal{C}(\boldsymbol{r})$ is called a collective variables (CV), reaction coordinate, order parameter, or slow variable depending on the context, and $\boldsymbol{\xi}$ is the low-dimensional representation of a state. For example, $\boldsymbol{r}$ can represent the 22 atoms in alanine dipeptide, and $\boldsymbol{\xi}$ two dihedral angles as illustrated in Figure 1.1(a). Figure 1.1(b,c) graphically depicts how a CV foliates the high-dimensional space of molecular configurations, by representing with color surfaces the manifolds defined by $\mathcal{C}(\boldsymbol{r})$ =constant. Physically, all molecular configurations within such a surface are represented by the same $\boldsymbol{\xi}$, and therefore should correspond to similar states of the system. Ideally, the system should not exhibit transverse metastability, which refers to metastability within the $\mathcal{C}(\boldsymbol{r})$ =constant manifolds, and all the complexity should be along the CV.

The choice of CVs, typically guided by experience or intuition, is very important and should satisfy several requirements. First, CVs should capture as much as possible the metastability of the system, leaving a simple landscape in the remaining transversal coordinates (Sutto et al., 2012). Otherwise, enhanced sampling methods become ineffective, exhibit hysteresis, and may not converge. Second, CVs should be as few as possible to be meaningful and efficient. This last requirement is

not only convenient, but also agrees with the actual behavior of large biomolecules, in which all-atom descriptions of conformations exhibit a large degree of redundancy (Brown et al., 2008; Hegger et al., 2007).

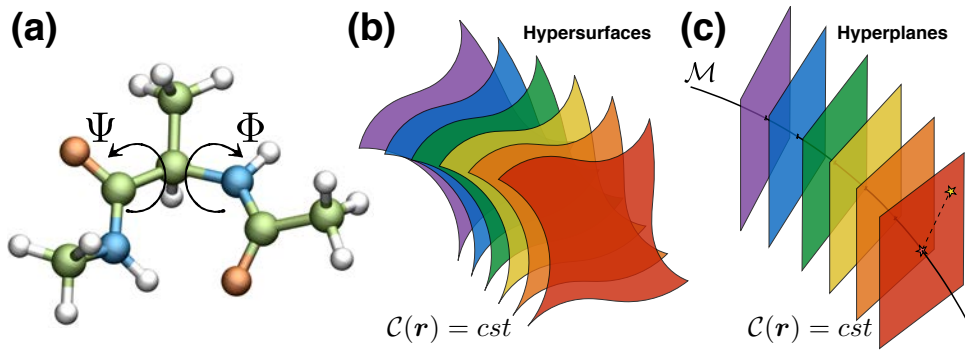


Fig. 1.1 Alanine dipeptide molecule described by collective variables given by two backbone dihedral angles (a). Level sets of collective variables, $\mathcal{C}(\boldsymbol{r})$ =cst, in general (b) are nonlinear hypersurfaces foliating the configuration space; however, in SandCV they are hyperplanes perpendicular to the intrinsic manifold (c).

Identifying appropriate CVs for complex systems is an open question and urges for a systematic approach. To this end machine learning methods are emerging as a means to identify in an unsupervised way CVs from molecular ensembles representative of molecular flexibility. Among such methods, linear dimensionality reduction techniques, such as principal component analysis (PCA) (Pearson, 1901), have become a standard tool to analyze MD trajectories and also to enhance conformational sampling as in essential dynamics (Amadei et al., 1993; Daidone and Amadei, 2012), by building linear data-driven CVs. In these methods, the molecular systems are assumed to essentially lie on a linear manifold defined by linear combinations of the principal components.

More recently, it has been argued that, because of steric constraints, rigid bonds, and molecular rotations, this linear manifold paradigm may

not be appropriate, shifting the focus towards nonlinear data-driven CVs (Das et al., 2006). According to this view, molecules would essentially evolve close to nonlinear manifolds of reduced dimensionality, sometimes called intrinsic manifold, which can be identified with nonlinear dimensionality reduction (NLDR) methods (Lee and Verleysen, 2007). This approach has been applied to postprocess molecular trajectories (Brown et al., 2008; Ceriotti et al., 2011; Ferguson et al., 2011b; Plaku et al., 2007; Stamati et al., 2010), finding nonlinearly correlations and compact nonlinear representations of high-dimensional configurations, which retain the essential variability in the original data.

A fundamental limitation of NLDR techniques, unlike PCA, is that they do not provide an explicit and differentiable mapping between atomic position and CVs. Such a mapping is required to evaluate the atomic forces resulting from a bias in the space of CVs, as done in many enhanced sampling methods. NLDR techniques merely find a low-dimensional embedding of the molecular conformations. Some recent works partially address this deficiency of NLDR methods. Ferguson et al. (2011a) adapts diffusion maps to bias simulations with umbrella sampling, and Tribello et al. (2012) introduces a field-overlap procedure to combine sketch-map with metadynamics and accelerate conformational exploration. Spiwok and Králová (2011) defines such mapping by utilizing a generalized path collective variables (Branduardi et al., 2007) to enhance sampling with metadynamics.

In Chapter 2 of this dissertation, we introduce a general and flexible method to define smooth and nonlinear data-driven collective variables (SandCV). By combining NLDR, a smooth parametrization of the intrinsic manifold, and geometric operations, we obtain a robust and efficient method that produces differentiable CVs. The method can handle highly curved intrinsic manifolds, non-uniform sampling, and even bridge over poorly sampled regions. SandCV is non-intrusive with

regards to the enhance sampling method, and can be easily integrated in standard MD codes (Phillips et al., 2005; Van Der Spoel et al., 2005) in conjunction with free energy calculation libraries (Bonomi et al., 2009). We show its effectiveness with a benchmark system, alanine dipeptide, and combine it with ABF for enhanced sampling and free energy calculation.

In Chapter 3, we provide details about the algorithms involved in SandCV along with the implementation of adaptive biasing force method, which requires only first derivatives of CVs for enhanced sampling.

In Chapter 4, 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. 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.

To overcome these obstacles to a single-chart description of molecular flexibility, in Chapter 5 we propose a framework to describe statistical mechanics of molecular systems in terms of an atlas of partially overlapping CVs. We then present a data-driven method based on nonlinear dimensionality reduction to systematically build atlas of CVs from ensembles representative of molecular flexibility. We demonstrate the effectiveness of this method with two model systems: alanine dipeptide and β-D-Glucopyranose.

Each of these chapters constitutes studies partially or fully based on the following publications:

- Behrooz Hashemian, Daniel Millán, and Marino Arroyo. Modeling and enhanced sampling of molecular systems with smooth and nonlinear data-driven collective variables. *The Journal of Chemical Physics*, 139(21):214101, 2013
  [**Chapter 2**]

- Behrooz Hashemian, Daniel Millán, and Marino Arroyo. SandCV++: a toolbox for data-driven collective variables. *In preparation*, 2015a
  [**Chapter 3**]

- Behrooz Hashemian and Marino Arroyo. Topological obstructions in the way of data-driven collective variables. *The Journal of Chemical Physics*, 142(4):044102, 2015
  [**Chapter 4**]

- Behrooz Hashemian, Daniel Millán, and Marino Arroyo. Charting molecular free-energy landscapes with an atlas of collective variables. *In preparation*, 2015b
  [**Chapter 5**]

Chapter 2

Smooth and nonlinear data-driven collective variables

2.1 Introduction

Molecular dynamics (MD) simulations provide atomic resolution of important processes involving biomolecules, which complement experimental observations (Sotomayor and Schulten, 2007) and can help understand the relation between conformational changes and function (Osadchy and Kolodny, 2011). 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 (Henzler-Wildman et al., 2007). 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 (Dror et al., 2010). 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 (Brown et al., 2008; Das et al., 2006; Ferguson et al., 2010), develop techniques complementary to nonlinear dimensionality reduction methods (Lee and Verleysen, 2007) 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, and are at the core of a myriad of enhanced sampling methods, including metadynamics (Laio and Parrinello, 2002), non-equilibrium work methods (Jarzynski, 1997), or the adaptive biasing force (ABF) method (Darve et al., 2008), which we implement here (see Chipot and Pohorille (2007) 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 (E et al., 2002; Jónsson et al., 1998; Olender and Elber, 1996; Passerone et al., 2003; Ren et al., 2005). Path

collective variables provide two CVs, along and perpendicular to a given transition path (Branduardi et al., 2007). 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) (Amadei et al., 1993; Daidone and Amadei, 2012). Besides being routinely used to post-process molecular simulations, PCA has been used to drive enhanced sampling in combination with metadynamics (Spiwok et al., 2007, 2008).

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 (Brown et al., 2008; Das et al., 2006), sometimes referred to as slow or intrinsic manifold (Ferguson et al., 2010), 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 (Noji et al., 1997). This field has been fueled by the emergence of nonlinear dimensionality reduction (NLDR) techniques in the field of statistical learning (Lee and Verleysen, 2007), which automatically identify nonlinear correlations hidden in high-dimensional data. Locally linear embedding (LLE) (Roweis and Saul, 2000) and Isomap (Tenenbaum et al., 2000) 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 (Das et al., 2006). Brown et al. (2008) 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 (Coifman et al., 2008), which account for the underlying Fokker-Planck operator. Ferguson et al. (2010) obtains nonlinear CVs of n-alkane chains through diffusion maps, and the approach is subsequently refined in Rohrdanz et al. (2011). 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. Ceriotti et al. (2011) 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 (Sammon, 1969). See Rohrdanz et al. (2013) 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 (der Maaten et al., 2009), 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. Brown et al. (2008) 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 forward and backward mappings and is not geometric in nature. Ferguson et al. (2011a) adapts diffusion maps to bias simulations with umbrella sampling, and Tribello et al. (2012) introduces a field-overlap procedure to combine sketch-map with metadynamics and accelerate conformational exploration. Spiwok and Králová (2011) generalizes path collective variables (Branduardi et al., 2007) to higher dimensions, defines smooth CVs from the output of Isomap for cyclo-octane (Brown et al., 2008), 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 (Millán and Arroyo, 2013), point-set surface parametrization (Millán et al., 2013), or stereotyped cell motility (Arroyo et al., 2012).

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 (Earl and Deem, 2005; Tribello et al., 2010), or even from experiments (Fenwick et al., 2011).

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 Section 2.2. 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 (Phillips et al., 2005; Van Der Spoel et al., 2005) in conjunction with free energy calculation libraries (Bonomi et al., 2009). In Section 2.3, 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 Section 2.4.

2.2 Methods

2.2.1 Problem statement

The methods presented here address the following problem. Given prior knowledge of a molecular system in terms of ensemble of M conformations given by the Cartesian coordinates of N atoms, $R = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_a, \ldots, \boldsymbol{r}_M\} \subset \mathbb{R}^{3N}$, the goal is to define a smooth function, referred to as collective variables, $\mathcal{C} : \mathbb{R}^{3N} \to \mathbb{R}^d$ mapping any out-of-sample molecular conformation $\boldsymbol{r} \in \mathbb{R}^{3N}$ into a low-dimensional representation $\boldsymbol{\xi} = \mathcal{C}(\boldsymbol{r}) \in \mathbb{R}^d$. 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 subsequent sections. We first identify the intrinsic manifold underlying the

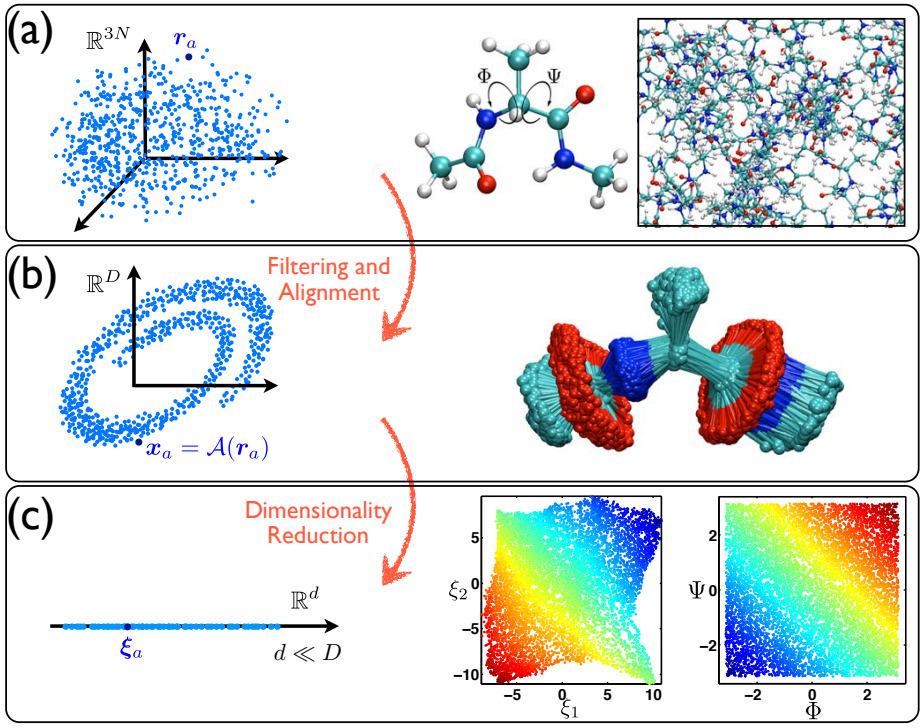


Fig. 2.1 Illustration of the main stages to identify the nonlinear intrinsic manifold, both in abstract terms (left) and also for a specific example involving alanine dipeptide (right). In the left plots, the points $\boldsymbol{r}_a$, $\boldsymbol{x}_a$, and $\boldsymbol{\xi}_a$ are different representations 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 an 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.

molecular ensemble with nonlinear dimensionality reduction methods in Section 2.2.2. These methods operate at the discrete level, and for this reason we then build a smooth representation of this manifold in in Section 2.2.3. Finally, we map any out-of-sample conformation to low-dimensions by projecting it onto the intrinsic manifold, as elaborated in Section 2.2.4. We present these steps in an abstract way, but with concrete reference to the system studied here, alanine dipeptide. This small molecule, shown in Figure 2.1(a), has been extensively studied and is a benchmark for free energy calculation methods (Branduardi et al., 2007; Darve et al., 2008; Ren et al., 2005; Rohrdanz et al., 2011; Spiwok et al., 2008). It is particularly well suited for our purposes because its exhibits metastability, good CVs are known (the dihedral angles Φ and Ψ), and these are highly nonlinear.

2.2.2 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 2.1(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 $\boldsymbol{r}_a \in \mathbb{R}^{3N}$, 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 conformations.

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 (Kroonenberg et al., 2003), 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 $\boldsymbol{r}_a \in \mathbb{R}^{3N}$ to a smooth contact map (Bonomi et al., 2008), or resorting to internal coordinates (Brown et al., 2008). 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 2.1(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 $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_a, \ldots, \boldsymbol{x}_M\} \subset \mathbb{R}^D$.

Adopting the intrinsic manifold paradigm to model the molecular system (Amadei et al., 1993; Das et al., 2006; García, 1992; Hegger et al., 2007), we seek to identify this nonlinear and low-dimensional structure underlying X with NLDR methods. We resort to Isomap (Tenenbaum et al., 2000), although the procedure presented here is not specific to this method. Isomap builds on multi-dimensional scaling (MDS), which given a matrix of pairwise distances $dist_D(\boldsymbol{x}_a, \boldsymbol{x}_b)$ in high dimension, finds an optimal low dimensional embedding of the points in X, denoted by $\Xi = \{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_a, \ldots, \boldsymbol{\xi}_M\} \subset \mathbb{R}^d$, such that the matrix of pairwise distances in low dimension, given by $dist_d(\boldsymbol{\xi}_a, \boldsymbol{\xi}_b)$, 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 $dist_{D,G}(\boldsymbol{x}_a, \boldsymbol{x}_b)$ 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 2.1(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 2.1(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 (Grassberger and Procaccia, 1983). 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 (Lee and Verleysen, 2007). 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.

2.2.3 Parametrizing the intrinsic manifold

From the output of NLDR, Ξ, we introduce now a smooth parametrization of the intrinsic manifold, as illustrated in Figure 2.2. 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 $\mathcal{M} : \Omega \subset \mathbb{R}^d \to \mathbb{R}^D$ of the form

$$\mathcal{M}(\boldsymbol{\xi}) = \sum_{i=1}^{L} p_i(\boldsymbol{\xi}) \boldsymbol{y}_i, \tag{2.1}$$

where $p_i(\boldsymbol{\xi})$ are smooth basis functions associated to a set of landmarks $\boldsymbol{\eta}_i$, see Figure 2.2. We denote by Ω a region in $\mathbb{R}^d$ delimited by the points in Ξ. We select the coefficients of the linear combination $\boldsymbol{y}_i$ by fitting $\mathcal{M}$ to the data in a least-squares sense, i.e. by minimizing

$$\sum_{a=1}^{M} \left| \boldsymbol{x}_a - \mathcal{M}(\boldsymbol{\xi}_a) \right|^2, \tag{2.2}$$

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 $p_i(\boldsymbol{\xi})$ 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

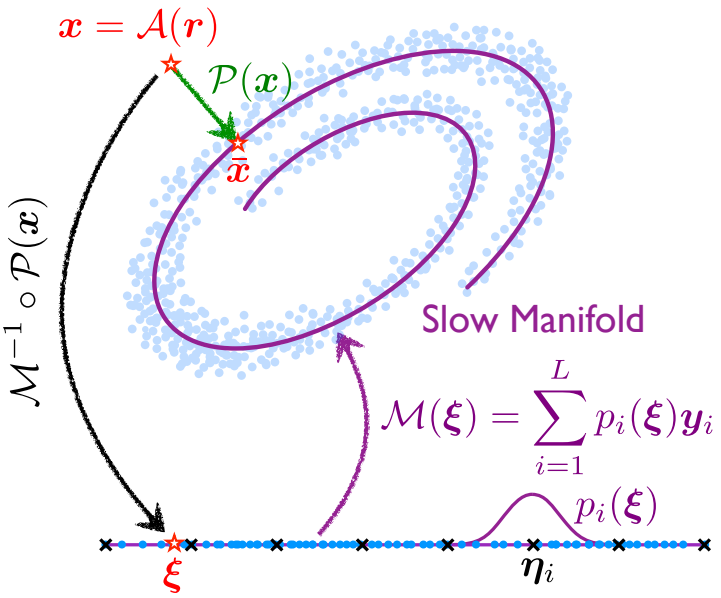


Fig. 2.2 The intrinsic manifold underlying the set of M aligned configurations (light-blue points) is parametrized from the low-dimensional embedding with a linear combination of L ($\ll M$) basis functions $p_i(\boldsymbol{\xi})$, resulting in the purple line. The coefficients $\boldsymbol{y}_i$ are obtained thought a least-squares fit. An out-of-sample point $\boldsymbol{x}$ is labeled in low-dimensional space by first obtaining its closest-point projection on the manifold, $\bar{\boldsymbol{x}}$, and then finding its pre-image through $\mathcal{M}$.

image of $\mathcal{M}$ together with the original ensemble.

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

$$p_i(\boldsymbol{\xi}) = \frac{e^{-\beta_i|\boldsymbol{\xi}-\boldsymbol{\eta}_i|^2+\boldsymbol{\lambda}\cdot(\boldsymbol{\xi}-\boldsymbol{\eta}_i)}}{\sum_j e^{-\beta_j|\boldsymbol{\xi}-\boldsymbol{\eta}_j|^2+\boldsymbol{\lambda}\cdot(\boldsymbol{\xi}-\boldsymbol{\eta}_j)}}, \tag{2.3}$$

where the parameter β_i sets the width of the basis functions locally, and $\boldsymbol{\lambda}$ 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 (Arroyo and Ortiz, 2006; Rosolen et al., 2010). 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 $\boldsymbol{\lambda} = 0$, then the basis functions result in the Shepard approximants (Wendland, 2005) used in path collective variables (Branduardi et al., 2007). The local maximum-entropy basis functions are smooth, can accurately represent point-set manifolds (Millán et al., 2013), and can deal with non-uniform sets of landmarks in any dimension d. However, many other choices are possible for parametrizing $\mathcal{M}$.

2.2.4 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 on it in general. A CV must be able to assign a label $\boldsymbol{\xi}$, representative of the state of the system, to any out-of-sample aligned configuration $\boldsymbol{x}$. The closest-point projection onto the intrinsic manifold, denoted by $\mathcal{P}(\boldsymbol{x})$, 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 (Ruuth and Merriman, 2008). 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

$$\mathcal{C}(\boldsymbol{r}) = \mathcal{M}^{-1} \circ \mathcal{P} \circ \mathcal{A}(\boldsymbol{r}), \tag{2.4}$$

as depicted graphically in Figure 2.2. Thus, by definition, the sub-manifolds of constant $\mathcal{C}(\boldsymbol{r})$ in configuration space are perpendicular to the intrinsic manifold. Although this expression is conceptually illuminating, in practice the SandCV is evaluated by minimizing

$$\left|\mathcal{M}(\boldsymbol{\xi}) - \mathcal{A}(\boldsymbol{r})\right|^2, \tag{2.5}$$

with respect to $\boldsymbol{\xi}$. Numerically, we resort to Newton's method, possibly with a few quasi-Newton iterations with line-search (Nocedal and Wright, 2006), to solve this d-dimensional optimization problem.

Equation (2.4) 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

$$\underbrace{\mathbf{D}\mathcal{C}(\boldsymbol{r})}_{d\times(3N)} = \underbrace{\mathbf{D}\mathcal{M}^{-1}(\bar{\boldsymbol{x}})}_{d\times D}\ \underbrace{\mathbf{D}\mathcal{P}(\boldsymbol{x})}_{D\times D}\ \underbrace{\mathbf{D}\mathcal{A}(\boldsymbol{r})}_{D\times(3N)}, \tag{2.6}$$

where we have highlighted the dimensions of the Jacobian matrices and for conciseness we introduce $\boldsymbol{x} = \mathcal{A}(\boldsymbol{r})$ and $\bar{\boldsymbol{x}} = \mathcal{P}(\boldsymbol{x})$. The Jacobian of alignment is method-dependent. For Procrustes superimposition we refer to Section 3.2.1. In Section 3.5.1, we derive an exact expression for the $d \times D$ matrix $\mathbf{D}\mathcal{M}^{-1}(\bar{\boldsymbol{x}})\mathbf{D}\mathcal{P}(\boldsymbol{x})$ in Eq. (2.6), given by

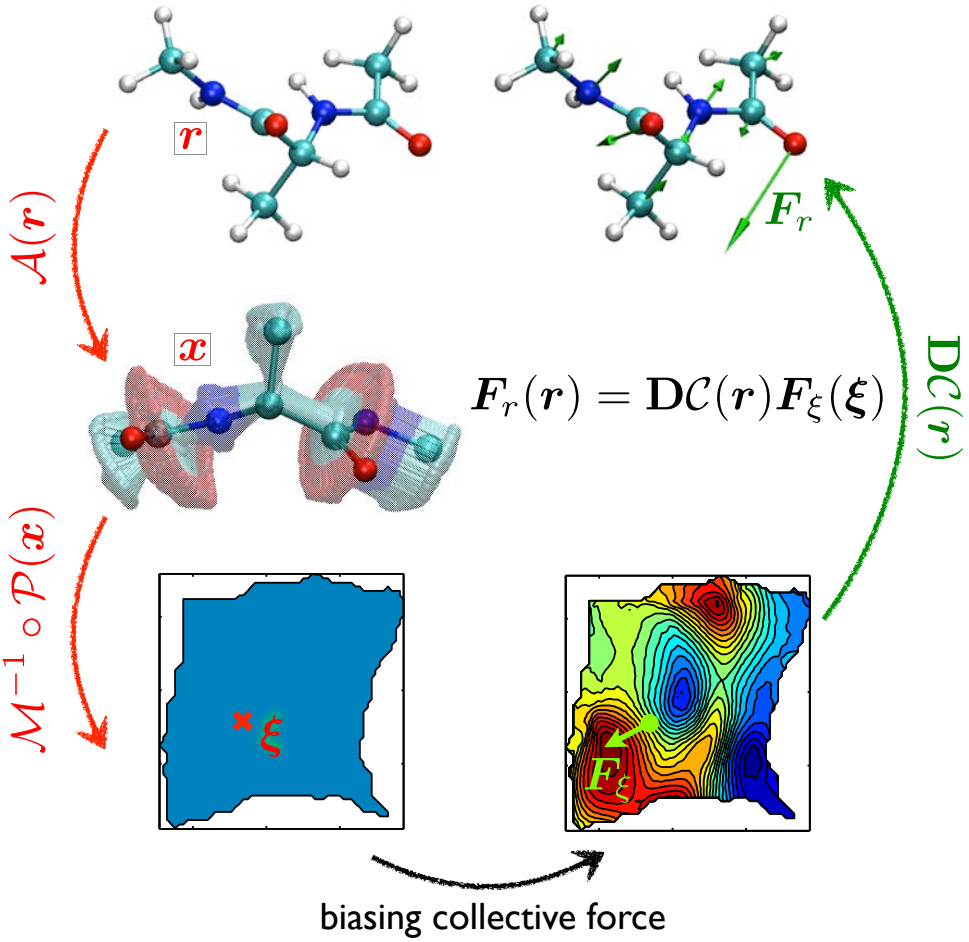


Fig. 2.3 Illustration of how bias forces in SandCV space are applied to the molecule in an enhanced sampling MD simulation. In each time-step, the all-atom configuration $\boldsymbol{r}$ is aligned to a common frame, and then mapped to the low-dimensional embedding, $\boldsymbol{\xi}$, through the closest-point projection on the intrinsic manifold. The bias force $\boldsymbol{F}_\xi$ 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.

$$\mathbf{D}\mathcal{M}^{-1}(\bar{\boldsymbol{x}})\ \mathbf{D}\mathcal{P}(\boldsymbol{x}) = \Big\{ \mathbf{D}\mathcal{M}^T(\boldsymbol{\xi})\mathbf{D}\mathcal{M}(\boldsymbol{\xi}) - \mathbf{D}^2\mathcal{M}(\boldsymbol{\xi})(\boldsymbol{x}-\bar{\boldsymbol{x}}) \Big\}^{-1} \mathbf{D}\mathcal{M}^T(\bar{\boldsymbol{x}}). \tag{2.7}$$

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

To illustrate how $\mathbf{D}\mathcal{C}(\boldsymbol{r})$ is needed in enhanced sampling methods, consider we want to bias the MD simulation with a potential defined in CV space, $U(\boldsymbol{\xi})$. 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, $U \circ \mathcal{C}(\boldsymbol{r})$. The force on the CVs is then $\boldsymbol{F}_\xi = -\partial U/\partial\boldsymbol{\xi}$. By the chain rule, we can map these forces to the atoms of the molecule, $\boldsymbol{F}_r(\boldsymbol{r}) = \mathbf{D}\mathcal{C}(\boldsymbol{r})\boldsymbol{F}_\xi(\boldsymbol{\xi})$, as illustrated in Figure 2.3. In the ABF method used here (Darve et al., 2008), the force $\boldsymbol{F}_\xi$ is an estimation of the thermodynamic force on the CVs.

2.2.5 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 Brown et al. (2008) 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 will be discusses in Chapter 4, and a general remedy to overcome such obstacles in molecular systems, called atlas of collective variables, is proposed in Chapter 5, where we employ recursive partitioning of the ensemble X with specialized algorithms (Karypis and Kumar, 1998) 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. However, in the current chapter, we have adopted an ad hoc approach to deal with the topology of the intrinsic manifold of alanine dipeptide, presented in the next section.

2.3 Results and discussions

We exercise the proposed method studying alanine dipeptide (N–acetyl–N'–methyl–L–alanylamide), also known as dialanine, 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 Φ and Ψ have been shown to be effective collective variables, although the significance of other dihedral angles has been

examined (Bolhuis et al., 2000; Ferguson et al., 2011a).

We show 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 preform 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 free energy surfaces (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 (Phillips et al., 2005) molecular dynamics code with the CHARM22 force field (MacKerell et al., 1998) and a Langevin thermostat. For the simulations in explicit water, we use the particle mesh Ewald method (Essmann et al., 1995) for long-range electrostatic forces and periodic boundary conditions. We implement SandCV in a stand-alone C++ code, which is interfaced with another C++ code that implements the vectorial version of the ABF method (Darve et al., 2008) and communicates with NAMD through a TCL interface to obtain configurations and return forces on the atoms.

2.3.1 Isomap low-dimensional embedding

We initially consider an ideal sampling of the intrinsic manifold,

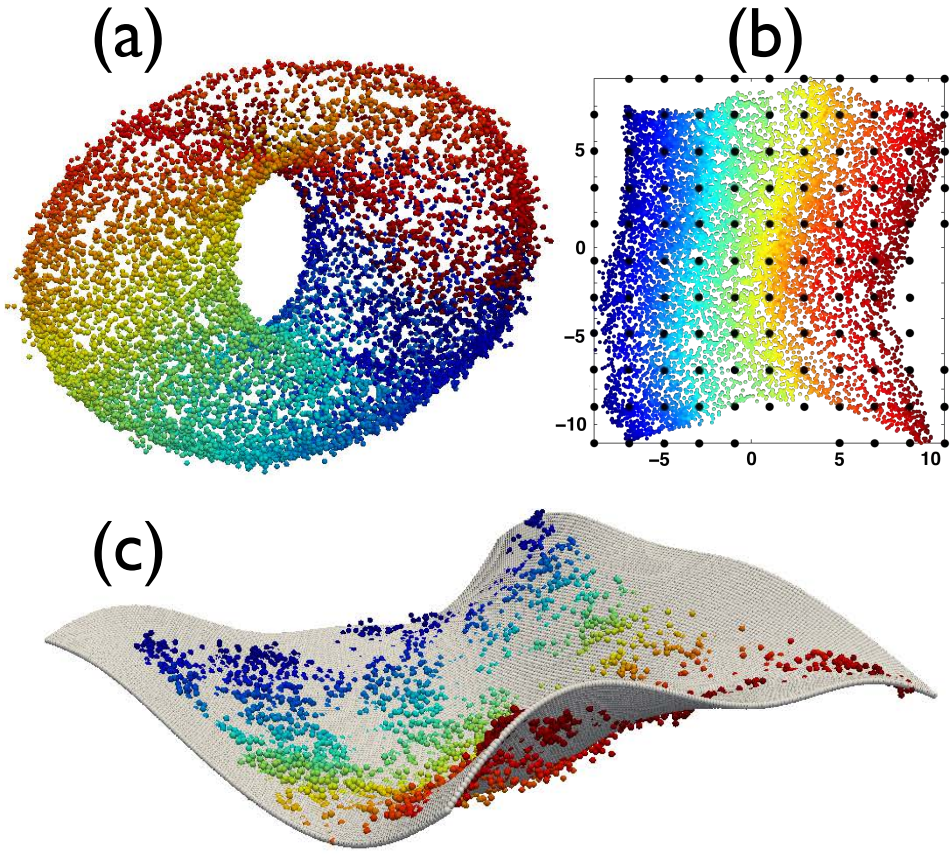


Fig. 2.4 Low-dimensional embedding of alanine dipeptide in vacuum with Isomap. (A) The three-dimensional embedding of alanine dipeptide shows it has the topology of a torus, and allows us to identify tearing curves to simplify the topology. (B) Two-dimensional embedding after tearing. The landmark points used in the parametrization are marked as black dots. (C) Smooth parametrization of the intrinsic manifold (gray surface) visualized on the three-dimensional Isomap embedding after tearing. The color map represents the dihedral angle Φ.

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 Section 2.2.5, 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 2.4(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 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 Section 2.2.5, we have suggested a general method to deal with general manifolds, which is beyond the scope of the present thesis.

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 2.4(b). Figure 2.4(c) illustrates the smooth parametrization of the intrinsic manifold as a surface in the three-dimensional torn Isomap embedding.

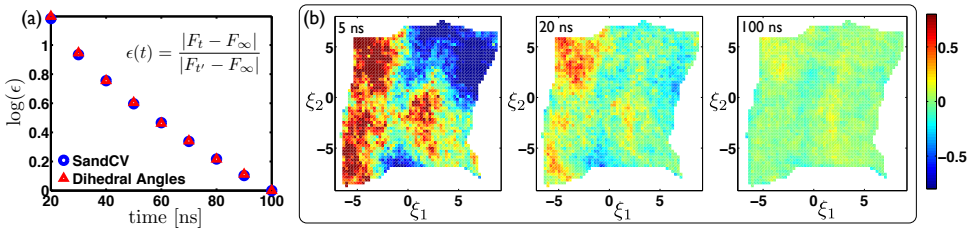


Fig. 2.5 Convergence of ABF simulations for alanine dipeptide in vacuum. (A) Convergence of the thermodynamic forces along the CVs for a simulation biased along dihedral angles (red triangles) and along SandCV (blue circles). We plot the error in the forces (ϵ) relative to the equilibrium forces F_∞ 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. This shows how the ABF algorithm biases the MD trajectory and results in nearly uniform sampling.

2.3.2 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 (Darve et al., 2008). Furthermore, equilibrium properties such as the thermodynamic force on the CVs should converge with simulation time. Figure 2.5 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 2.5(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 (Lelièvre et al., 2010). The different panels in Figure 2.5(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 (Sutto et al., 2012). 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.

2.3.3 Free energy comparisons

Table 2.1 Free Energy differences at the points marked in Figure 2.6. The units are in kcal/mol. ($1k_BT = 0.596$ kcal/mol) and the point number 1 is taken as the reference with 0.0 value.

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

Free energy surfaces (FES) are subjective in that they fundamentally depend on the CVs along which they vary, and are not insensitive to reparametrizations of CV space (E et al., 2005). Although this fact does not have consequences on physical observables such as rates of conformational changes (Frenkel, 2013), it complicates a meaningful

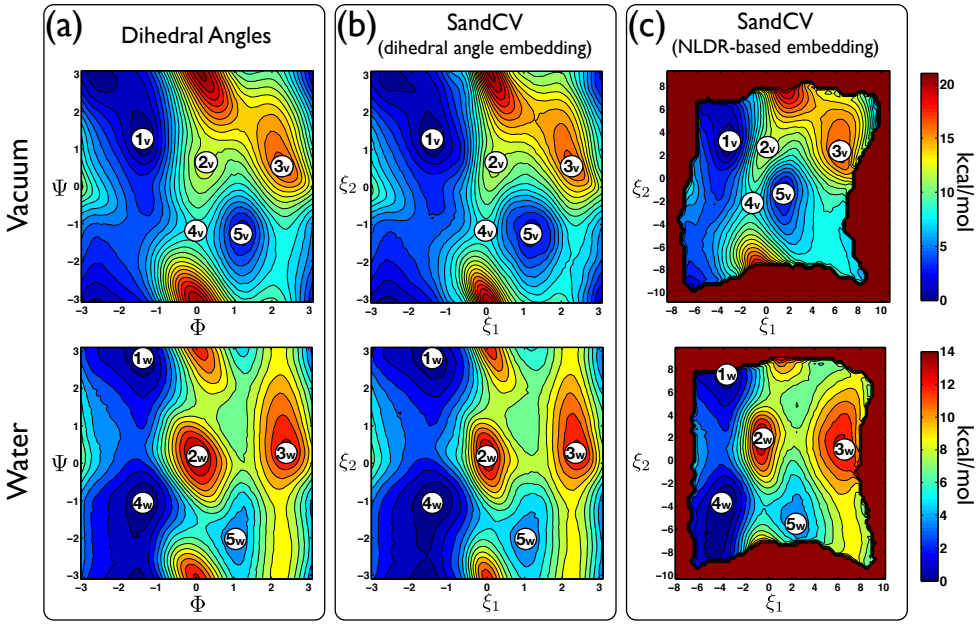


Fig. 2.6 Comparison of the free energy surface of alanine dipeptide in vacuum (v) and water (w) along three different sets of CVs. (A) Backbone dihedral angles (Φ and Ψ), (B) SandCV based on the two-dimensional embedding given by the dihedral angles, and (C) SandCV based on a two-dimensional Isomap embedding.

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 2.6(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 2.5. We parametrize the intrinsic manifold with about 11×11 landmark points, and choose $\beta = 1/h^2$ for the basis functions in Eq. (2.3), 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, Figure 2.6(a) and 2.6(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 2.1 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 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_BT. 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.

2.3.4 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 realistic 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 310K. 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 $7 \cdot 10^6$ 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 trajectory, chosen in such a way that the Euclidean distance between any pair of aligned configurations within a trajectory is larger than a cut-off. Second, the resulting $1.4 \cdot 10^6$ 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 Section 2.3.1 to tear the manifold.

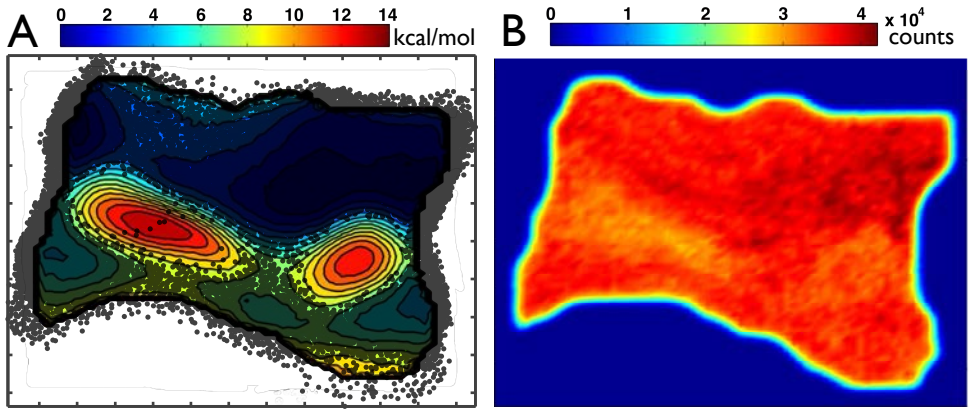


Fig. 2.7 SandCV on a training set with unexplored regions. The low-dimensional embedding of the training set exhibits large holes (A), which correspond to regions of high free energy. An ABF simulation based on SandCV can bridge over these regions and end up with nearly uniform sampling in the space of CVs, as shown by the distribution of configurations in CV space (B).

Figure 2.7(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 $\boldsymbol{\eta}_i$ and taking $\beta = 1/h^2$ in Eq. (2.3), the smooth representation of the intrinsic manifold in Eq. (2.1) 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 2.7(b). The resulting FES is represented in Figure 2.7(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 exploration, 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.

2.3.5 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 2.6(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 2.8(a) shows a reference FES in water computed with the dihedral angles, while Figure 2.8(b) shows the FES 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.

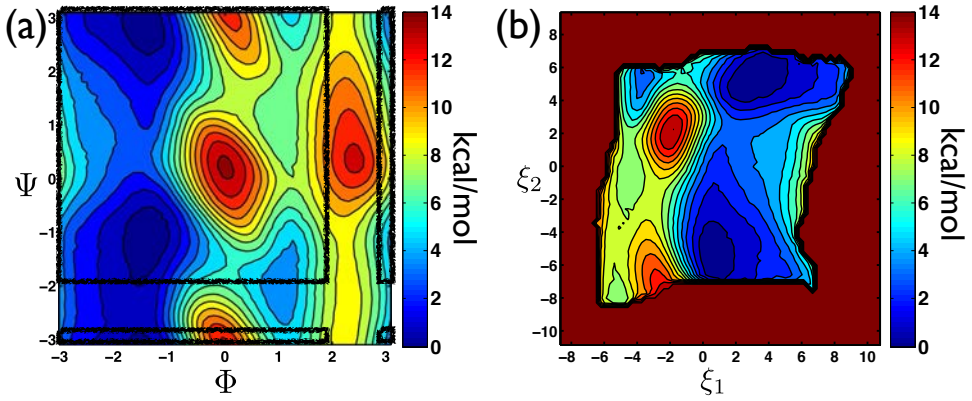


Fig. 2.8 Exploring the transferability of SandCV. The free energy of alanine dipeptide in water, shown in (A) in terms of dihedral angles, is computed along SandCV based on an ensemble in vacuum (B). For comparison, the FES in (B) should be rotated by 180°, and the support of (B) has been highlighted in (A).

2.4 Conclusion

We have introduced a general method to model molecular systems with smooth and nonlinear data-driven collective variables (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 $\mathcal{C}(\boldsymbol{r})$ 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 Φ and Ψ. 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 (Ceriotti et al., 2013). 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 with maximum entropy approximants. The last block 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. The embedding in low dimensions can also combine physical insight and statistical learning techniques. Complex systems may exhibits nontrivial topologies that cannot be treated with nonlinear dimensionality reduction, as we discuss in Chapter 4, and will require a systematic approach to describe the intrinsic manifold and the CVs through multiple charts, which is proposed in Chapter 5.

Chapter 3

SandCV++: a toolbox for data-driven collective variables

3.1 Introduction

The proposed method in the Chapter 2 for constructing smooth and data-driven collective variables (SandCV) requires a toolbox of algorithms, from machine learning to optimization to enhanced sampling. I have implemented all these methods and algorithms in a stand-alone C++ library, called SandCV++. The input of this library is a training set of ensemble of molecular conformations, which is fully or partially representative of the molecule's flexibility. Using this training set, it first computes a low-dimensional embedding, characterizing the most important behavior of the molecular system. Then, using local-maximum entropy approximants, it constructs a slow manifold in high-dimensional space, which is the best representation of original dataset in the least-square sense. This slow manifold is smoothly parametrized in the low-dimensional space, and finally produces a functional set of collective variables (CVs) that can be

use in enhanced sampling simulations. SandCV++ communicates with molecular dynamics code through a TCL interface, although it can also be interfaced with Python. It can also be patched to the MD code in the same way as some popular free energy methods such as Plumed (Tribello et al., 2014). SandCV++ comes with a built-in enhanced sampling method, which is a particular implementation of the vectorial adaptive biasing force method (Darve et al., 2008).

SandCV++ consists of four main modules (molecular alignment, dimensionality reduction, slow manifold parametrization, and closest point projection), some auxiliary modules and a vectorial adaptive biasing force. The algorithm behind each module is explained in the following sections.

3.2 Molecular alignment

In molecular simulations, the rigid-body motion of the molecule is often not constrained and thus similar configurations occupy distant positions in the Cartesian coordinates. This is problematic for methods that rely on distances between Cartesian representations of conformations to find the similarities between structures. To resolve this problem, we implement Procrustes superimposition that removes the rigid-body motion by optimally aligning the configurations to a reference one. We also discuss alignment using a smooth contact map. The alignment procedure needs to be differentiable to be useful in SandCV.

3.2.1 Procrustes superimposition

We represent any given configuration $\boldsymbol{r}_a \in \mathbb{R}^{3N}$ 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 $\mathbf{R} \in \mathbb{R}^{3\times3}$ and

translation vector $\mathbf{t} \in \mathbb{R}^{1\times 3}$ by minimizing the cost function

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

where $\mathbf{X}_{ref}$ is a reference configuration and $rep : \mathbb{R}^{1\times 3} \to \mathbb{R}^{N_s\times 3}$ 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

$$\mathbf{M} = \left(\mathbf{X}_{ref} - rep[\boldsymbol{\mu}_{ref}]\right)^T \left(\mathbf{X} - rep[\boldsymbol{\mu}]\right), \tag{3.1}$$

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

$$\mathbf{R} = \mathbf{V}\mathbf{U}^T, \quad \mathbf{t} = \boldsymbol{\mu}_{ref} - \boldsymbol{\mu}\mathbf{R} \ . \tag{3.2}$$

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

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

$$\begin{aligned} &\mathbf{V}^T \frac{\partial \mathbf{M}^T}{\partial X_{ir}} \mathbf{U} - \mathbf{U}^T \frac{\partial \mathbf{M}}{\partial X_{ir}} \mathbf{V} = \\ &\left[\mathbf{V}^T \frac{\partial \mathbf{V}}{\partial X_{ir}} - \mathbf{U}^T \frac{\partial \mathbf{U}}{\partial X_{ir}} \right] \mathbf{S} + \mathbf{S} \left[\frac{\partial \mathbf{U}^T}{\partial X_{ir}} \mathbf{U} - \frac{\partial \mathbf{V}^T}{\partial X_{ir}} \mathbf{V} \right]. \end{aligned} \tag{3.3}$$

By defining

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

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 (Papadopoulo and Lourakis, 2000), we rewrite Eq. (3.3) as

$$\mathbf{\Omega}(X_{ir}) \mathbf{S} + \mathbf{S} \mathbf{\Omega}(X_{ir}) = \mathbf{V}^T \frac{\partial \mathbf{M}^T}{\partial X_{ir}} \mathbf{U} - \mathbf{U}^T \frac{\partial \mathbf{M}}{\partial X_{ir}} \mathbf{V}. \tag{3.4}$$

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

$$\Omega_{pq}(X_{ir}) = \frac{1}{S_p + S_q} \left[V_{mp} \frac{\partial M_{nm}}{\partial X_{ir}} U_{nq} - U_{mp} \frac{\partial M_{mn}}{\partial X_{ir}} V_{nq} \right], \tag{3.5}$$

without summation over p and q, where

$$\frac{\partial M_{mn}}{\partial X_{ir}} = \left(X^{im}_{ref} - \mu^m_{ref} \right) \delta_{nr}.$$

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. (3.2), we only need to calculate the derivatives of $\mathbf{R}$, which follow from

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

3.2.2 Smooth contact map

Contact maps (Holm et al., 1993; Vendruscolo et al., 2000) have been extensively used as a simplified representations of protein conformations, capturing its important structural features. Contact maps are matrices such that m_{ij} is 1 if residues i and j are closer than a given threshold and 0 otherwise. Because they are defined from distances between atoms of the molecule, contact map representations are invariant with respect to rigid body transformations. In addition to the common binary contact map, one can also consider a smoother variation of the entries m_{ij} as a function of the distance between the corresponding residues. Shibberu et al. (2010) introduced contact maps with C^0 continuity and used them for protein structure alignment (Shibberu and Holder, 2011). However, this contact map is not smooth enough for our purposes. We have used within SandCV the smooth contact map introduced in Bonomi et al. (2008), along with C^∞ alternatives using sigmoid and log-sum-exp functions. In this approach, the derivative of the alignment map is

trivial, see Appendix B for more details. A smooth contact map is illustrated in Figure 3.1.

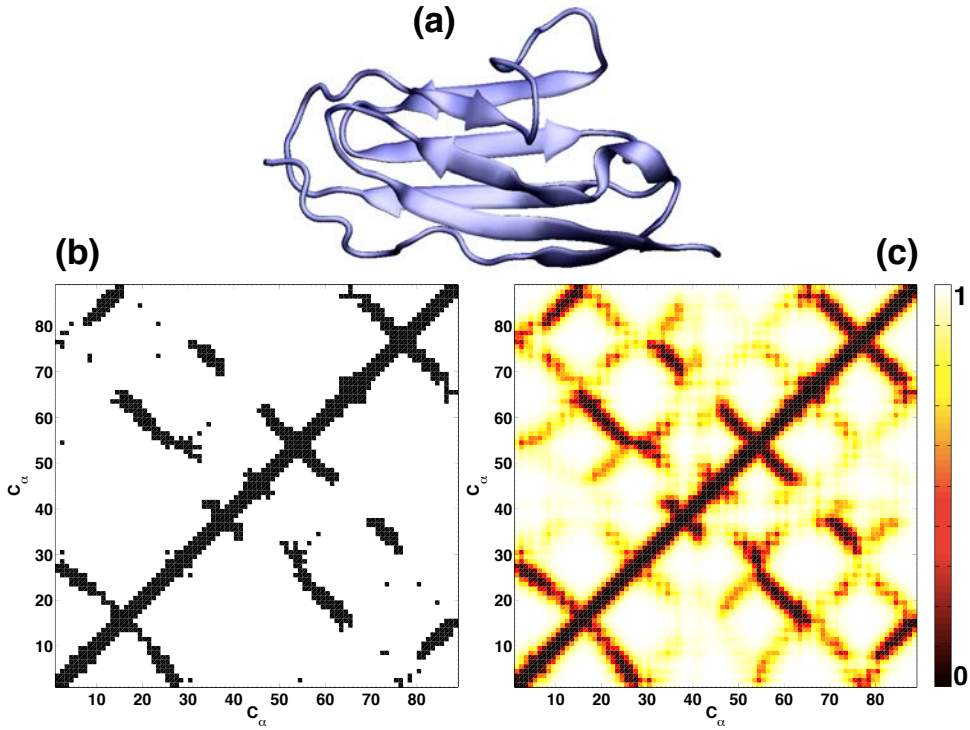


Fig. 3.1 Usual contact map (b) verse smooth contact map (c) description of Titin I21 domain (a), considering alpha carbons (C_α) and using cut-off value of 8.5Å.

3.3 Dimensionality reduction

An important part of SandCV is determine the important features in the molecular system through machine learning methods such as dimensionality reduction. This task can be done by employing any suitable dimensionality reduction technique, such as Isomap (Tenenbaum et al., 2000), LLE (Roweis and Saul, 2000), and diffusion

map (Coifman and Lafon, 2006), from any available implementation such as scikit-learn in Python (Pedregosa et al., 2011), RDRToolbox in R, and dimensionality reduction toolbox in MATLAB (der Maaten et al., 2009). However, to have a self-contained toolbox, we have implemented an efficient version of Isomap that has been proven to work well in molecular systems (Brown et al., 2008; Das et al., 2006; Hashemian et al., 2013; Spiwok and Králová, 2011). This implementation consists of following steps.

1. Constructing neighborhood graph:
   We define the weighted graph G over the high-dimensional data points by connecting points i and j if j is one of the K nearest neighbors of i, and setting the edge weight to $D(i,j)$, the distance between aligned representations of the molecule. We have a lot of freedom to select this distance based on our understanding about the system. Examples include the Cartesian distance between configurations aligned with Procrustes analysis, or distances between smooth contact maps, see Section 3.2.

2. Computing shortest paths:
   We resort to Dijkstra algorithm (Dijkstra, 1959), which is asymptotically the fastest known single-source shortest-path algorithm for arbitrary directed graphs with unbounded non-negative weights, to evaluate the approximate shortest path between each node, and thus construct the graph-based geodesic distance matrix $D_G(i,j)$.

3. Evaluating low-dimensional embedding:
   Using classical multidimensional scaling, we evaluate the eigenvectors and eigenvalues of $HD_G^2H/2$, where H is the centering matrix $H_{ij} = \delta_{ij} - 1/N$, using LAPACK (Anderson et al., 1999) routines. Considering λ_p be the p-th eigenvalue (in the decreasing

order) and v_p^i be the i-th component of the p-th eigenvector, then the p-th component of the low-dimensional vector y_i is $\sqrt{\lambda_p} v_p^i$.

3.4 Slow manifold parametrization

As described in Section 2.2.3, we parametrize the high-dimensional slow-manifold as

$$\mathcal{M}(\boldsymbol{\xi}) = \sum_{i=1}^{L} p_i(\boldsymbol{\xi}) \boldsymbol{y}_i, \tag{3.6}$$

where $p_i(\boldsymbol{\xi})$ are smooth basis functions and $\boldsymbol{y}_i$ are coefficients that are calculated by fitting $\mathcal{M}$ to the data in a least-squares sense,

$$\sum_{a=1}^{M} \left| \boldsymbol{x}_a - \mathcal{M}(\boldsymbol{\xi}_a) \right|^2, \tag{3.7}$$

which involves solving a $L \times L$ linear system of equations. In SandCV++, this equation is solved by utilizing Cholesky factorization implemented in LAPACK routines (dpotrf and dpotrs).

For smooth basis functions, we took advantage of local maximum-entropy(Arroyo and Ortiz, 2006) given by

$$p_i(\boldsymbol{\xi}) = \frac{e^{-\beta_i |\boldsymbol{\xi} - \boldsymbol{\eta}_i|^2 + \boldsymbol{\lambda} \cdot (\boldsymbol{\xi} - \boldsymbol{\eta}_i)}}{\sum_j e^{-\beta_j |\boldsymbol{\xi} - \boldsymbol{\eta}_j|^2 + \boldsymbol{\lambda} \cdot (\boldsymbol{\xi} - \boldsymbol{\eta}_j)}}, \tag{3.8}$$

where $\boldsymbol{\eta}_i$ is a set of landmarks, arbitrarily distributed in the CV space with a typical distance of h. The β_i defines the width of the basis functions and in practice for molecular systems can be set to $1/h^2$. $\boldsymbol{\lambda}$ is a Lagrange multiplier to impose exact reproducibility of affine functions, which can be obtained by minimizing the denominator using Newton-Raphson method with possibly line search. The defined slow-manifold can be tuned to filter the out-of-manifold noise while not smearing out the important features of the intrinsic manifold by

adjusting the β_i parameters and the distribution of landmarks.

3.5 Closest point projection

In Section 2.2.4, we discussed how to find the CV value for a new configuration by projecting it on the slow manifold. This results in a nonlinear optimization problem with d unknowns. To solve this nonlinear minimization, we used the Newton-Raphson method with a line search algorithm and also implemented as an alternative a trust-region algorithm from Intel Math Kernel Library (MKL).

3.5.1 Jacobian of the closest-point projection

As argued in Section 2.2.4, the Jacobian of SandCV follows from

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

where $\mathbf{D}\mathcal{A}(\boldsymbol{r})$ can be computed as indicated in Section 3.2.1. On the other hand, from the inverse function theorem $\mathbf{D}\mathcal{M}^{-1}(\bar{\boldsymbol{x}})$ is the pseudo-inverse of the matrix $\mathbf{D}\mathcal{M}$:

$$\mathbf{D}\mathcal{M}^{-1}(\bar{\boldsymbol{x}}) = \left[\mathbf{D}\mathcal{M}^T(\boldsymbol{\xi})\mathbf{D}\mathcal{M}(\boldsymbol{\xi})\right]^{-1}\mathbf{D}\mathcal{M}^T(\boldsymbol{\xi}), \tag{3.9}$$

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

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

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

where $\boldsymbol{n}(\boldsymbol{x})$ is a normal vector to the manifold at $\bar{\boldsymbol{x}}$ and $d(\boldsymbol{x})$ is the

distance to the manifold. Taking the derivative of this equation, we find

$$\mathbf{D}\mathcal{P}(\boldsymbol{x}) = \mathbf{I}_D - d(\boldsymbol{x})\mathbf{D}\boldsymbol{n}(\boldsymbol{x}) - \boldsymbol{n}(\boldsymbol{x}) \otimes \mathbf{D}d(\boldsymbol{x}). \tag{3.10}$$

Multiplying Eq. (3.10) by $\mathbf{D}\mathcal{M}^{-1}(\bar{\boldsymbol{x}})$, recalling Eq. (3.9), 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. $\boldsymbol{n}(\boldsymbol{x}) = \mathcal{N} \circ \mathcal{M}^{-1} \circ \mathcal{P}(\boldsymbol{x})$, we have

$$\mathbf{D}\mathcal{M}^{-1}\mathbf{D}\mathcal{P} = \left[\mathbf{D}\mathcal{M}^T\mathbf{D}\mathcal{M}\right]^{-1}\mathbf{D}\mathcal{M}^T\left\{\mathbf{I}_D - d(\boldsymbol{x})\mathbf{D}\mathcal{N}\mathbf{D}\mathcal{M}^{-1}\mathbf{D}\mathcal{P}\right\}.$$

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

$$\mathbf{D}\mathcal{M}^{-1}\mathbf{D}\mathcal{P} = \left\{\mathbf{I}_d + d\left[\mathbf{D}\mathcal{M}^T\mathbf{D}\mathcal{M}\right]^{-1}\mathbf{D}\mathcal{M}^T\mathbf{D}\mathcal{N}\right\}^{-1}\mathbf{D}\mathcal{M}^{-1}, \tag{3.11}$$

and by further using of Eq. (3.9) and simplifying, we find

$$\mathbf{D}\mathcal{M}^{-1}\mathbf{D}\mathcal{P} = \left\{\mathbf{D}\mathcal{M}^T\mathbf{D}\mathcal{M} + d\ \mathbf{D}\mathcal{M}^T\mathbf{D}\mathcal{N}\right\}^{-1}\mathbf{D}\mathcal{M}^T, \tag{3.12}$$

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,

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

and therefore $\mathbf{D}\mathcal{M}^T\mathbf{D}\mathcal{N} = -\mathbf{D}^2\mathcal{M}^T\mathcal{N}$. Plugging this expression into Eq. (3.12), we finally obtain

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

3.6 Adaptive biasing forces

The free energy of a system can be estimated by thermodynamic integration of the average forces along collective variables $\boldsymbol{\xi} = \mathcal{C}_1(\boldsymbol{r})$ (Darve and Pohorille, 2001),

$$\mathbf{D}A_1(\boldsymbol{\xi}) = -\left\langle \boldsymbol{F}_{\xi}^{sys}(t^k) \right\rangle, \qquad \boldsymbol{\xi} \in b_k$$

with

$$\boldsymbol{F}_{\xi}^{sys}(t_i^k) = \frac{\partial}{\partial t}\left(M_{\xi} \frac{\partial \mathcal{C}_1}{\partial t} \right)\Bigg|_{t_i^k}, \tag{3.13}$$

where b_k is the k-th bin in CV's space, t_i^k is the time when i-th sample resides in k-th bin, and $M_{\xi}^{-1} = \mathbf{D}\mathcal{C}_1 M^{-1} \mathbf{D}\mathcal{C}_1^T$ where M is the mass matrix. However, if the energy barriers along $\boldsymbol{\xi}$ are significant, this method of calculating free energy will face the inefficient sampling. To overcome this problem, the adaptive biasing force (ABF) algorithm (Darve et al., 2008) adds a properly chosen, position-dependent biasing force to the system that allows for a self-diffusion in the CV space.

This low-dimensional biasing force, $\boldsymbol{F}_{\xi}^{bias}$, is calculated at each time step by running average of the force acting along $\boldsymbol{\xi}$ in the k-th bin assuming that $\mathcal{N}(N,k)$ is the number of samples collected in bin k after N steps in a simulation,

$$\boldsymbol{F}_{\xi}^{bias}(N,k) = -\frac{R\left(\mathcal{N}(N,k)\right)}{\mathcal{N}(N,k)} \sum_{i=1}^{\mathcal{N}(N,k)} \boldsymbol{F}_{\xi}^{sys}(t_i^k), \tag{3.14}$$

where $R\left(\mathcal{N}(N,k)\right) = \min(1, \mathcal{N}(N,k)/N_0)$ is a ramp function to avoid the crude estimation of the biasing force in poorly sampled bins. Afterwards we should subtract the applied biasing force from the

modified system force to recover the original system force,

$$\boldsymbol{F}_{\xi}^{sys}(t_i^k) = \frac{\partial}{\partial t}\left(M_{\xi}\frac{\partial \xi}{\partial t}\right)\Bigg|_{t_i^k} - \boldsymbol{F}_{\xi}^{bias}(N-1,k).$$

In this method, the free energy of system can be estimated at each bin by

$$\mathbf{D}A_1(\boldsymbol{\xi}) = -\left\langle \boldsymbol{F}_{\xi}^{sys} \right\rangle = -\left\langle \boldsymbol{F}_{\xi}^{mod} - \boldsymbol{F}_{\xi}^{bias} \right\rangle. \tag{3.15}$$

It can be shown from Eq. (3.14) and Eq. (3.15) that after a brief equilibration the biasing force converges to the $-\left\langle \boldsymbol{F}_{\xi}^{sys} \right\rangle$ and thus the average of modified system force acting along $\boldsymbol{\xi}$, $\boldsymbol{F}_{\xi}^{mod}$ becomes zero. This allows for a self-diffusion motion along the CV in the modified system and

$$\mathbf{D}A_1(\boldsymbol{\xi}) \approx \boldsymbol{F}_{\xi}^{bias}. \tag{3.16}$$

Some implementation of ABF require second derivative of CVs, which is cumbersome and costly. Here, we implement a vectorial ABF that requires only first derivative of CVs and ask only for the position vectors of atoms from MD code. Algorithm 3.1 shows the pseudocode to compute the biasing force with the ABF method.

Algorithm 3.1 A second-derivative-free and vectorial implementation of adaptive biasing force (ABF) method.

Require: Retrieve the atomic mass matrix $\boldsymbol{M}$ and the atomic position vector $\boldsymbol{r}$ from the molecular dynamics code.

Require: Calculate the projection onto the collective variables space $\boldsymbol{\xi} = \mathcal{C}(\boldsymbol{r})$ and its Jacobean $\boldsymbol{J}_\xi = \mathbf{D}\mathcal{C}(\boldsymbol{r})$

function ABF::CalculateForce($\boldsymbol{r}$,$\boldsymbol{M}$,$\boldsymbol{\xi}$,$\boldsymbol{J}_\xi$)

$\boldsymbol{v} \leftarrow (\boldsymbol{r} - \boldsymbol{r}_{old})/\Delta t$ ▷ calculate the atomic velocities

$\boldsymbol{r}_{old} \leftarrow \boldsymbol{r}$ ▷ update the old atomic positions

$\boldsymbol{v}_\xi \leftarrow \boldsymbol{J}_\xi \boldsymbol{v}$ ▷ project the velocity into the CV space

$\boldsymbol{M}_\xi \leftarrow (\boldsymbol{J}_\xi \boldsymbol{M}^{-1} \boldsymbol{J}_\xi^T)^{-1}$ ▷ calculate the reduced mass matrix

$\boldsymbol{p}_\xi \leftarrow \boldsymbol{M}_\xi \boldsymbol{v}_\xi$ ▷ calculate the momenta in CV space

$\boldsymbol{F}_\xi^{total} \leftarrow (\boldsymbol{p}_\xi - \boldsymbol{p}_\xi^{old})/\Delta t$ ▷ projection of the total forces exerted on the system

$\boldsymbol{p}_\xi^{old} \leftarrow \boldsymbol{p}$ ▷ update the old momentum

$k \leftarrow$ bin corresponding to $\boldsymbol{\xi}$

if $\mathcal{N}(k) = 0$ **then**

$\boldsymbol{F}_\xi^{bias}(k) \leftarrow 0.0$

else

$\boldsymbol{F}_\xi^{bias}(k) \leftarrow -R(\mathcal{N}(k))/\mathcal{N}(k) \boldsymbol{F}_\xi^{hist}(k)$ ▷ evaluate the biasing force in CV space

end if

$\boldsymbol{F}^{bias}(k) \leftarrow \boldsymbol{F}_\xi^{bias}(k) \boldsymbol{J}_\xi$ ▷ convert the low-dimensional biasing force to the force on individual atoms

$\mathcal{N}(k) \leftarrow \mathcal{N}(k) + 1$ ▷ update number of samples in the k-th bin

$\boldsymbol{F}_\xi^{sys}(k) \leftarrow \boldsymbol{F}_\xi^{total} - \boldsymbol{F}_\xi^{oldbias}$ ▷ Correcting the system to exclude the previous biasing force

$\boldsymbol{F}_\xi^{hist}(k) \leftarrow \boldsymbol{F}_\xi^{hist}(k) + \boldsymbol{F}_\xi^{total} - \boldsymbol{F}_\xi^{oldbias}$ ▷ add the system force into the histogram

$\boldsymbol{F}_\xi^{oldbias} \leftarrow \boldsymbol{F}_\xi^{bias}(k)$ ▷ update old biasing force

return $\boldsymbol{F}^{bias}$

end function

Chapter 4

Topological obstructions in the way of data-driven collective variables

4.1 Introduction

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 (Borhani and Shaw, 2012), the task of selecting appropriate CVs becomes highly nontrivial (Pietrucci and Laio, 2009). 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) (Pearson, 1901). 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 (Amadei et al., 1993; David and Jacobs, 2014; de Groot et al., 2001; Lange and Grubmüller, 2006; Maisuradze et al., 2009), understand molecular flexibility (Teodoro et al., 2003) and enhance sampling in molecular dynamics (Michielssens et al., 2012; Spiwok et al., 2007). 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 (García, 1992; Hegger et al., 2007; Noji et al., 1997).

Advances in the field of statistical learning, notably in nonlinear dimensionality reduction (NLDR) techniques (Hinton and Salakhutdinov, 2006; Roweis and Saul, 2000; Tenenbaum et al., 2000), 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 (Brown et al., 2008; Ceriotti et al., 2011; Das et al., 2006; Ferguson et al., 2010; Stamati et al., 2010), although some systems evolve on non-manifold sets (Ceriotti et al., 2013). Different NLDR techniques have been applied to molecular systems, including Isomap (Tenenbaum et al., 2000), locally linear embedding (Roweis and Saul, 2000), autoencoder networks (Hinton and Salakhutdinov, 2006), diffusion map (Coifman and Lafon, 2006) or LSDMap (Rohrdanz et al., 2011). 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 (Ferguson et al., 2011a; Hashemian et al., 2013; Spiwok and Králová, 2011; Tribello et al., 2012).

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 (Lee and Verleysen, 2007). 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 (Duan et al., 2014; Ferguson et al., 2011a; Rohrdanz et al., 2011; Stamati et al., 2010; Xue et al., 2013). 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.

4.2 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 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 (Lee and Verleysen, 2007). 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 4.1(a). This example illustrates Whitney's embedding theorem (Whitney, 1936), 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.

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 4.1(b).

Figure 4.2(a,b) shows a torus and a sphere, which are

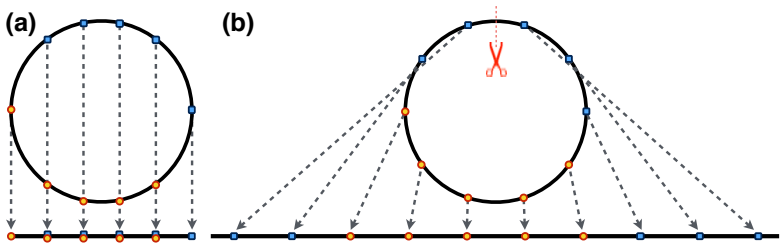


Fig. 4.1 Due to their different topology, it is impossible to embed a circle into a line (a); however, by tearing the circle at one point, this topological obstacle can be circumvented (b).

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 4.2(c).

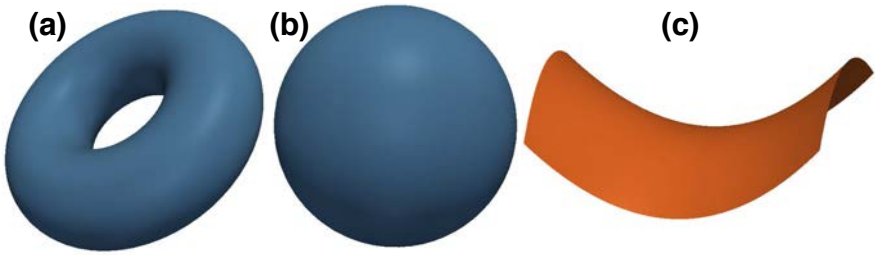


Fig. 4.2 Surfaces ($d = 2$) of different topology. Conformational changes of molecules with two significant dihedral angles evolve around a torus (a), while six-membered rings carbohydrates, like β-D-Glucopyranose (Biarnés et al., 2007), have a sphere-like intrinsic manifold (b).

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 (Jákli et al., 2012). 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 (Rohrdanz et al., 2011; Stamati et al., 2010; Xue et al., 2013). Furthermore, because of this topological obstruction, NLDR techniques suggest an excessive number of CVs relative to the intrinsic dimension (Ferguson et al., 2011a).

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 4.3(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 4.3(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 (Roweis and Saul, 2000)), and two distance-preserving methods that use different notions of distance (Diffusion map (Coifman and Lafon, 2006) and Isomap (Tenenbaum et al., 2000)). 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 (Duan et al., 2014). A straightforward way to remove the topological obstructions is to consider a trimmed ensemble of conformations, which lies within the dashed rectangle in Figure 4.3(a), at the expense of throwing away a significant number of conformations. As shown in Figure 4.3(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 4.3(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 4.3(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 chapter.

4.3 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 (Hashemian et al., 2013). After removing hydrogen atoms and alignment, we embed the molecular ensemble in three-dimensions, see Figure 4.4(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 4.4(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 4.4(c). Figure 4.4(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

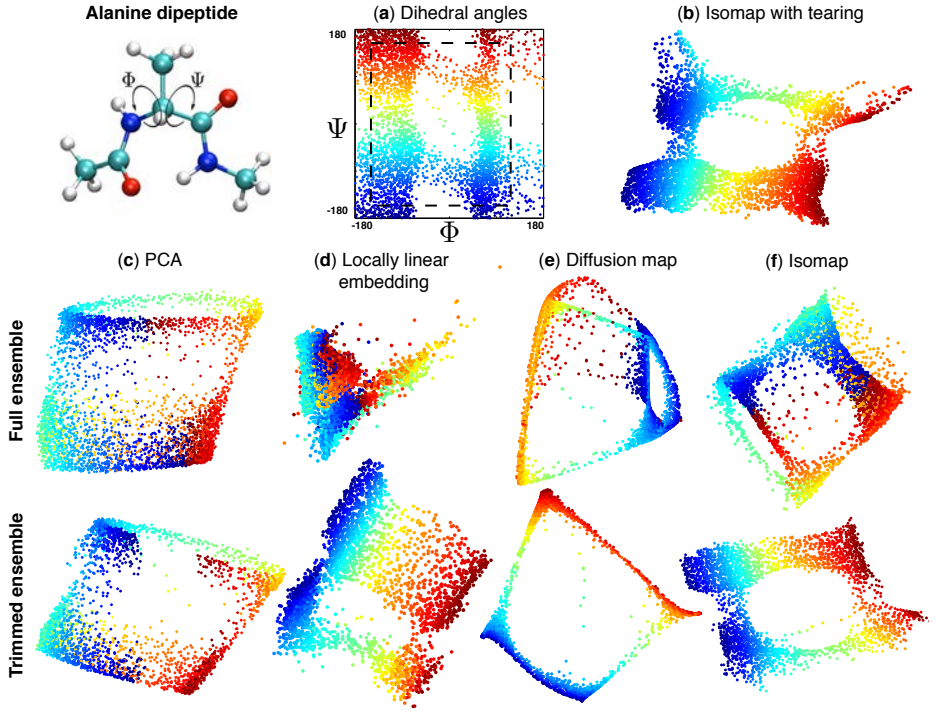


Fig. 4.3 Dimensionality reduction of an ensemble of molecular configurations of alanine dipeptide, obtained from multiple short-run simulations and visualized in dihedral space, Φ and Ψ (a). A trimmed ensemble delimited by the dashed rectangle is also considered to avoid topological obstructions. Dimensionality 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).

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.

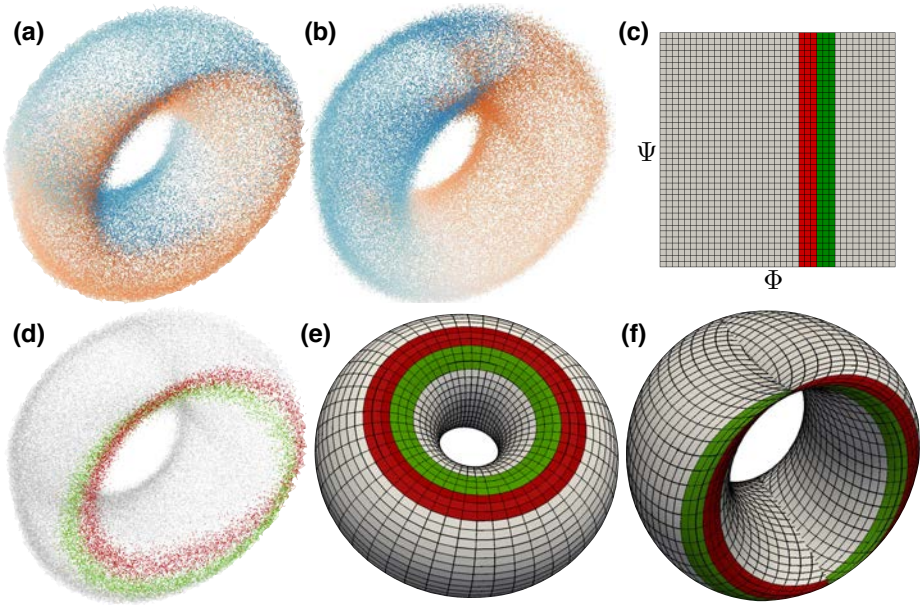


Fig. 4.4 Topology and geometry of molecular flexibility of alanine dipeptide. A well-sampled molecular ensemble is processed by Isomap to obtain a three-dimensional representation of the conformational changes, where the colormap is the value of the backbone dihedral angles, Φ (a) and Ψ (b). Two adjacent 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 dipeptide 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 (McCleary, 2012). A consequence of this fact is that any three-dimensional embedding will distort the metric, as illustrated by the grid in Figure 4.4(e). Interestingly, the three-dimensional projection of the four-dimensional isometric embedding of the flat torus shown in Figure 4.4(f) is very similar to the embedding provided by Isomap, see Figure 4.4(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.

4.4 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 4.1 and 4.3. 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 (Hashemian et al., 2013), if the systems is low-dimensional enough, or by more systematic algorithms that find essential loops and disconnect them (Lee and Verleysen, 2005). 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 (Hashemian et al., 2013), and modifications of metadynamics to avoid artifacts at boundaries in CV space have been developed (Crespo et al., 2010).

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 (Kohonen, 1982; Malsburg, 1973) or Generative Topographic Mapping (Bishop et al., 1998). 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 is discussed in detail in Chapter 5, also reduces the metric-induced distortions of the low-dimensional embeddings. Furthermore, systematic partitioning may enable the analysis of systems exhibiting non-manifold behavior (Martin et al., 2010). 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 (Hegger et al., 2007). Thus, one can expect that statistical learning methods applied to proteins may help understand these complex systems with a few collective variables (Piana and Laio, 2008). Once freed from topological obstructions and geometrical distortion, data-driven strategies to define CVs may deliver their full potential.

Chapter 5

Charting molecular free-energy landscapes with an atlas of collective variables

5.1 Introduction

Collective variables (CVs) provide a coarse-grained, low-dimensional description of molecular conformations, and thus help us rationalize complex molecular mechanisms. If CVs are able to separate the metastable states of the system, then they can be used to enhance sampling in molecular simulations and efficiently compute free energy surfaces (Chipot and Pohorille, 2007). Furthermore, since CVs often represent the slowly relaxing degrees of freedom of the system, they may be the basis of reduced models for conformational dynamics (Lu and Vanden-Eijnden, 2014; Salvalaglio et al., 2014). A poor selection of the CVs, however, mixes metastable states and results in hysteresis and lack of convergence of enhanced sampling algorithms (Rosta et al., 2009; Sutto et al., 2012).

Machine learning methods, in particular nonlinear dimensionality reduction, are emerging as a promising approach to identify systematically reliable CVs from ensembles representative of molecular flexibility (Brown et al., 2008; Das et al., 2006; Ferguson et al., 2011a; Hashemian et al., 2013; Spiwok and Králová, 2011; Spiwok et al., 2015; Tribello et al., 2012). We have recently realized, however, that such methods face fundamental obstacles when trying to represent conformational manifolds with general topologies, resulting for instance from cyclic motions around a dihedral angle. In such cases, manifold learning methods fail by collapsing distant conformations in their low-dimensional embedding (Hashemian and Arroyo, 2015). Furthermore, the curvature of the conformational manifold may severely distort, and even collapse, the low-dimensional representation of collective motions. In cartography, similar issues are dealt with by mapping the globe using a collection (atlas) of charts, each describing with appropriate detail a region of limited extent. Here, we examine whether this idea to describe geographical landscapes can be transposed to conformational landscapes. After all, it is natural to expect that a complex biomolecule may be best described by different CVs (a different chart) in different regions of its conformational space. However, free energy formalisms and enhanced sampling methods have only been developed considering a single chart description, or multiple isolated locally valid CVs (Tribello et al., 2010).

Here, we present a statistical mechanics framework that allows us to describe a molecular system using an atlas of partially overlapping CVs. We show that such a multiple-chart description is not an obstacle to seamlessly compute thermodynamical observables or to enhance sampling in molecular dynamics (MD) simulations. We further propose a simple but powerful data-driven algorithm based on the intrinsic manifold model to construct atlas of CVs and combine it with enhanced

sampling methods. We apply the proposed methods to two model systems with nontrivial conformational topology: alanine dipeptide and β-D-Glucopyranose.

5.2 Theory

Consider a N-atom system with separable Hamiltonian and potential energy $\mathcal{V}(\boldsymbol{r})$. Suppose that we are given a CV, i.e. a smooth surjective mapping from conformation space $D = \mathbb{R}^{3N}$ to a low-dimensional space, which is only defined in a region $D_1 \subset D$ and denoted by $\mathcal{C}_1 : D_1 \to \mathbb{R}^d$, where $d \ll 3N$. The free energy along this CV, up to additive constant, is

$$A_1(\boldsymbol{\xi}) = -\frac{1}{\beta} \ln \int_{D_1} e^{-\beta \mathcal{V}(\boldsymbol{r})} \delta(\mathcal{C}_1(\boldsymbol{r}) - \boldsymbol{\xi}) \, d\boldsymbol{r}, \tag{5.1}$$

where $\delta(\cdot)$ is the Dirac delta distribution (Chipot and Pohorille, 2007; Lelièvre et al., 2010), and $1/\beta = kT$ is the Boltzmann constant times temperature. Applying co-area formula (Hartmann et al., 2011), the free energy can be written as an integral over the level set $\mathcal{L}_1(\boldsymbol{\xi})$ (the points $\boldsymbol{r} \in D_1$ such that $\mathcal{C}_1(\boldsymbol{r}) = \boldsymbol{\xi}$, see Figure 5.1),

$$A_1(\boldsymbol{\xi}) = -\frac{1}{\beta} \ln \int_{\mathcal{L}_1(\boldsymbol{\xi})} e^{-\beta \mathcal{V}(\boldsymbol{r})} \mathrm{vol}(\mathbf{D}\mathcal{C}_1)^{-1} \, d\boldsymbol{\sigma}_{\boldsymbol{\xi}}, \tag{5.2}$$

where $\mathbf{D}\mathcal{C}_1$ is the Jacobian matrix of the CV, $d\boldsymbol{\sigma}_{\boldsymbol{\xi}}$ is the volume element of $\mathcal{L}_1(\boldsymbol{\xi})$ and $\mathrm{vol}(\mathbf{D}\mathcal{C}_1) = \sqrt{|\mathbf{D}\mathcal{C}_1 \mathbf{D}\mathcal{C}_1^T|}$. $|\cdot|$ stands for the determinant.

Analogously, we consider a different CV, $\boldsymbol{\eta} = \mathcal{C}_2(\boldsymbol{r})$, defined in another region of configuration space $D_2 \subset D$, which partially overlaps with D_1, i.e. $D_1 \cap D_2 \neq \emptyset$. Examining Eq. (5.2), it is clear that the free energies along these two CVs can only be related in the overlapping region if their respective level sets can be mapped to each other. For this to be the case, there must exist a bijective transition mapping

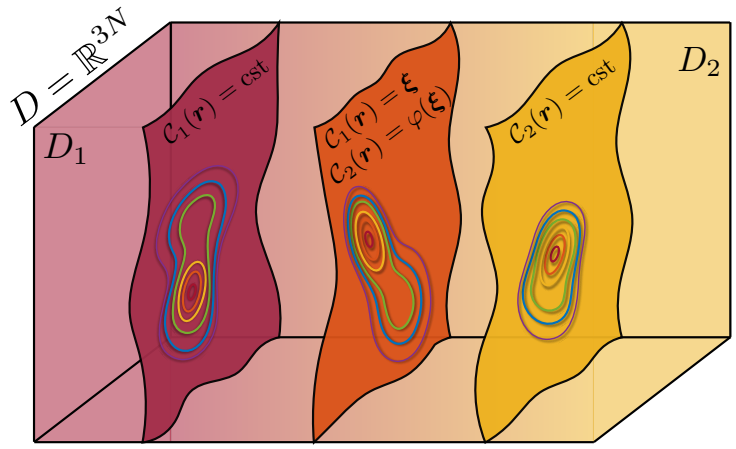


Fig. 5.1 Graphical illustration of high-dimensional configuration space D, described in the left by collective variable $\mathcal{C}_1$, in the right by $\mathcal{C}_2$, and in the central overlapping region by both. The surfaces illustrate the level sets. The level set in the middle can be described either as $\mathcal{C}_1(\boldsymbol{r}) = \boldsymbol{\xi}$ or as $\mathcal{C}_2(\boldsymbol{r}) = \varphi(\boldsymbol{\xi})$.

$\varphi : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\mathcal{C}_2(\boldsymbol{r}) = \varphi \circ \mathcal{C}_1(\boldsymbol{r}) \tag{5.3}$$

for $\boldsymbol{r} \in D_1 \cap D_2$. This mapping provides a connection between different CVs and a means to build a meaningful global statistical mechanics description of the system. Indeed, assuming that such a transition mapping $\boldsymbol{\eta} = \varphi(\boldsymbol{\xi})$ exists, using the chain rule on Eq. (5.3), and noting that $\text{vol}(\mathbf{D}\varphi\mathbf{D}\mathcal{C}_1) = |\mathbf{D}\varphi| \text{ vol}(\mathbf{D}\mathcal{C}_1)$ (Ben-Israel, 1992), we can write Eq. (5.2) for the second CV as

$$\begin{aligned} A_2(\boldsymbol{\eta}) &= -\frac{1}{\beta} \ln \int_{\mathcal{L}_2(\boldsymbol{\eta})} e^{-\beta \mathcal{V}} \left[|\mathbf{D}\varphi| \text{ vol}(\mathbf{D}\mathcal{C}_1) \right]^{-1} d\boldsymbol{\sigma}_{\boldsymbol{\eta}} \\ &= -\frac{1}{\beta} \ln \int_{\mathcal{L}_1(\boldsymbol{\xi})} e^{-\beta \mathcal{V}} \text{vol}(\mathbf{D}\mathcal{C}_1)^{-1} d\boldsymbol{\sigma}_{\boldsymbol{\xi}} + \frac{1}{\beta} \ln |\mathbf{D}\varphi(\boldsymbol{\xi})| \\ &= A_1(\boldsymbol{\xi}) + \frac{1}{\beta} \ln |\mathbf{D}\varphi(\boldsymbol{\xi})|. \end{aligned} \tag{5.4}$$

where we have used the facts that $\mathcal{L}_2(\boldsymbol{\eta}) = \mathcal{L}_1(\boldsymbol{\xi})$ and that $\mathbf{D}\varphi$ only depends on $\boldsymbol{\xi}$ and hence can be factored outside of the integral.

Equation (5.4) provides a connection between the free energies relative to each of the two CVs in the overlapping region, and highlights the fact that the free energy is not invariant with respect to reparametrizations of CV space (Frenkel, 2013; Lelièvre et al., 2010). The last term in Eq. (5.4) is reminiscent of a Fixman potential (Fixman, 1974; Hartmann et al., 2011). Recalling that both A_1 and A_2 are computed up to an additive constant, Eq. (5.4) provides a compatibility relation between these two constants.

Let us compute now thermodynamic observables, such as relative probabilities of states, in a multiple CV framework. Consider two conformations characterized by regions A and B belonging to D_1 and D_2 respectively. Consider also an auxiliary state $C \subset D_1 \cap D_2$. The relative probabilities between states A and C, and between C and B can be computed as

$$\frac{p(A)}{p(C)} = \frac{\int_{\mathcal{C}_1(A)} e^{-\beta A_1(\boldsymbol{\xi})} d\boldsymbol{\xi}}{\int_{\mathcal{C}_1(C)} e^{-\beta A_1(\boldsymbol{\xi})} d\boldsymbol{\xi}}, \quad \frac{p(C)}{p(B)} = \frac{\int_{\mathcal{C}_2(C)} e^{-\beta A_2(\boldsymbol{\eta})} d\boldsymbol{\eta}}{\int_{\mathcal{C}_2(B)} e^{-\beta A_2(\boldsymbol{\eta})} d\boldsymbol{\eta}}.$$

Consequently, the relative probability between $A \subset D_1$ and $B \subset D_2$ takes the form

$$\frac{p(A)}{p(B)} = \frac{\int_{\mathcal{C}_1(A)} e^{-\beta A_1(\boldsymbol{\xi})} d\boldsymbol{\xi}}{\int_{\mathcal{C}_2(B)} e^{-\beta A_2(\boldsymbol{\eta})} d\boldsymbol{\eta}} \underbrace{\frac{\int_{\mathcal{C}_2(C)} e^{-\beta A_2(\boldsymbol{\eta})} d\boldsymbol{\eta}}{\int_{\mathcal{C}_1(C)} e^{-\beta A_1(\boldsymbol{\xi})} d\boldsymbol{\xi}}}_{=1}. \tag{5.5}$$

Invoking Eq. (5.4) and using the change of variable formula, the second term is one and this relative probability can be computed without reference to state C, showing that the statistical mechanics of the system can be seamlessly formulated across multiple CVs. The arguments drawn here extend directly to an atlas of partially overlapping CVs,

provided that transition mappings exist.

We examine next the impact of having multiple CVs on accelerated free energy calculations. In many enhanced sampling methods, such as metadynamics (Laio and Parrinello, 2002) or the adaptive biasing force method (ABF) (Darve et al., 2008), an approximation to the thermodynamic force at $\boldsymbol{\xi} = \mathcal{C}(\boldsymbol{r})$ along the CV, $\boldsymbol{F}^{CV}(\boldsymbol{\xi}) \approx -\mathbf{D}A(\boldsymbol{\xi})$, is mapped onto a force on the atoms that biases the dynamics

$$\boldsymbol{F}^{bias}(\boldsymbol{r}) \;=\; \left[\boldsymbol{F}^{CV} \circ \mathcal{C}(\boldsymbol{r})\right]\; \mathbf{D}\mathcal{C}(\boldsymbol{r}). \tag{5.6}$$

In the ideal situation, $\boldsymbol{F}^{CV} = -\mathbf{D}A(\boldsymbol{\xi})$, the enhanced sampling trajectory undergoes free diffusion along the CVs irrespective of free energy barriers. Consider now a two-CV atlas and $\boldsymbol{r} \in D_1 \cap D_2$. As a result of Eq.(5.4), even in the ideal situation, the biasing forces corresponding to each CV are essentially different. Indeed,

$$\begin{aligned} \boldsymbol{F}_1^{bias} = -\mathbf{D}A_1\mathbf{D}\mathcal{C}_1 \;&=\; -\left[\mathbf{D}A_2\mathbf{D}\varphi - \frac{1}{\beta}\mathbf{D}\left(\ln|\mathbf{D}\varphi|\right)\right]\mathbf{D}\mathcal{C}_1 \\ &=\; \boldsymbol{F}_2^{bias} - \frac{1}{\beta}\mathbf{D}\left(\ln|\mathbf{D}\varphi|\right)\mathbf{D}\mathcal{C}_1, \end{aligned} \tag{5.7}$$

where the last term is non-zero in general. Thus, during an enhanced sampling simulation and for $\boldsymbol{r}$ in the overlapping region, the algorithm must make a choice for $\boldsymbol{F}^{bias}(\boldsymbol{r})$, which cannot simultaneously represent the thermodynamic force along $\mathcal{C}_1$ and $\mathcal{C}_2$. As a result, an adaptive enhanced sampling algorithm cannot converge in the sense of free diffusion simultaneously in all overlapping CVs. However, as we show later, this fact is irrelevant in practice and computed free energies do converge.

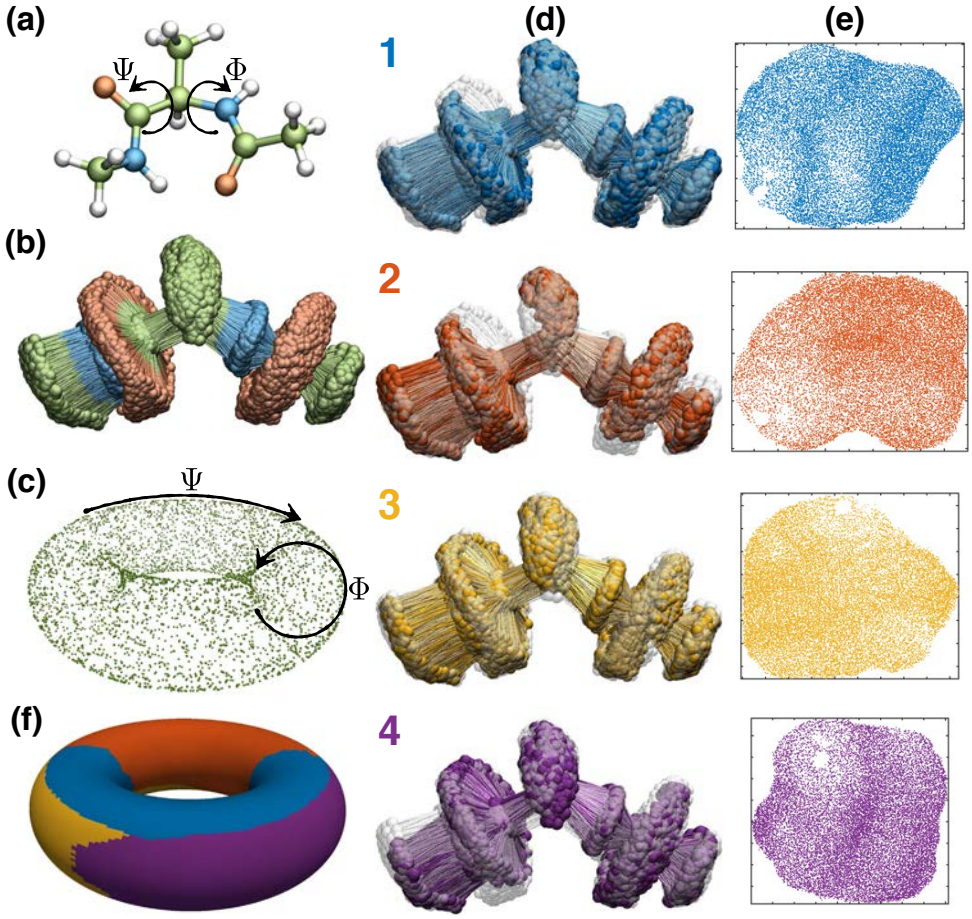


Fig. 5.2 Data-driven algorithm to create an atlas of collective variables for alanine dipeptide (a). Starting from an ensemble of molecular configurations (b), which can be represented in three dimensional space using the dihedral angles (Φ and Ψ) as a torus (c), we systematically partition the configurations into four slightly overlapping partitions (d,f). Then, by applying dimensionality reduction techniques, we embed the conformations in each partition (devoid of topological obstructions) in low-dimensions (e).

5.3 Method

Thus, we have shown that it is possible to build a sound statistical mechanics description of a molecule in terms of an atlas of CVs, provided that transition mappings exist as in Eq. (5.3). However, how to practically define such an atlas of CVs? We first address the issue of transition mappings by resorting to the intrinsic manifold model for molecular systems. In this model, it is assumed that the dynamics of the molecule take place close to a manifold (in general nonlinear) of dimension d embedded in the $D-$dimensional conformational space (Brown et al., 2008; Das et al., 2006; Ferguson et al., 2010; Hashemian et al., 2013; Spiwok et al., 2015). Although it may seem a rather strong assumption, this notion has been shown to be useful in a number systems.

Suppose that we have identified the intrinsic manifold and parametrized it as $\boldsymbol{r} = \mathcal{M}(\boldsymbol{\xi})$. In (Hashemian et al., 2013), we proposed a systematic and robust method to build a CV from $\mathcal{M}$. Starting from a conformation in high-dimensions $\boldsymbol{r}$ (in general off the manifold), we first compute its closest-point projection onto the intrinsic manifold $\mathcal{P}(\boldsymbol{r})$ and then find the pre-image through the mapping $\mathcal{M}$, i.e. $\boldsymbol{\xi} = \mathcal{C}(\boldsymbol{r}) = \mathcal{M}^{-1} \circ \mathcal{P}(\boldsymbol{r})$. In practice, the evaluation of $\mathcal{C}$ requires solving a straightforward nonlinear minimization problem with d unknowns, and $\mathbf{D}\mathcal{C}$ can be computed explicitly (Hashemian et al., 2013). Importantly, this framework provides *ab initio* compatible collective variables, along with the transition mappings. Indeed, consider for instance that the intrinsic manifold is described by two overlapping parametrizations $\mathcal{M}_1$ and $\mathcal{M}_2$. Because of the closest-point construction, the level sets $\mathcal{L}_1$ and $\mathcal{L}_2$ (see Figure 5.1) are hyper-planes perpendicular to the intrinsic manifold in its vicinity, and therefore depend on geometry but not on the specific choice of parametrization. The transition mapping is simply $\varphi = \mathcal{M}_2^{-1} \circ \mathcal{M}_1$, see Figure 5.3.

Building on this conceptual framework, we describe next a data-driven algorithm that automatically constructs an atlas of CVs for systems with general intrinsic manifold topology, without prior knowledge about the system other than an ensemble of conformations representative of molecular flexibility. We illustrate this method with alanine dipeptide, a standard benchmark with a good known CV in terms of two dihedral angles Figure 5.2(a). Our algorithm relies on nonlinear dimensionality reduction (Brown et al., 2008; Das et al., 2006; Ferguson et al., 2010; Hashemian et al., 2013; Lee and Verleysen, 2007) to identify the nonlinear intrinsic manifold from the molecular ensemble shown in Figure 5.2(b). Because of the periodic nature of dihedral angles, the intrinsic manifold underlying this system has the topology of a torus, Figure 5.2(c), posing an obstacle to a single data-driven CV identified by dimensionality reduction (Hashemian and Arroyo, 2015). To describe this manifold with multiple parametrizations, we first split the ensemble into a few disjoint pieces as visualized in Figure 5.2(f) on the torus representation of the system. This partitioning, however, is performed in high dimensions, by first building a connectivity graph of the ensemble based on the K-nearest neighbors to each conformation, and then using systematic graph partitioning algorithms (Karypis and Kumar, 1998). This partitioning is performed recursively until each piece is flat enough and devoid of topological obstructions to be tractable by nonlinear dimensionality methods (Millán et al., 2013). Figure 5.2(d) shows the conformations in each of the four partitions identified by the algorithm, slightly enlarged to provide sufficient overlap. From this point on, the smooth and nonlinear data-driven collective variables (SandCV) method (Hashemian et al., 2013) is directly applicable to each partition. In this method, high-dimensional conformations are embedded in low-dimensions (here 2) with Isomap (Tenenbaum et al., 2000), a nonlinear dimensionality reduction method, see Figure 5.2(e).

From each of these embeddings, a smooth parametrization of a portion of the intrinsic manifold, $\mathcal{M}_\alpha$, is constructed, and the corresponding CVs are defined as $\mathcal{C}_\alpha = \mathcal{M}_\alpha^{-1} \circ \mathcal{P}_\alpha$. One last ingredient is required to locate the active CV as the system navigates through the conformational landscape. For this, we build in high-dimensions a partition of unity, i.e. a collection of non-negative functions $\psi_\alpha(\boldsymbol{r})$ that are 1 in the interior of each partition D_α, smoothly decay to zero at its boundary, and $\sum_\alpha \psi_\alpha(\boldsymbol{r}) = 1$ everywhere as illustrated in Figure 5.3. We assign conformation $\boldsymbol{r}$ to a chart following the criterion $\max_\alpha \psi_\alpha(\boldsymbol{r})$, and apply $\boldsymbol{F}_\alpha^{bias}(\boldsymbol{r})$ in enhanced sampling simulations accordingly. See Supplemental Material A for a detailed description of the algorithm.

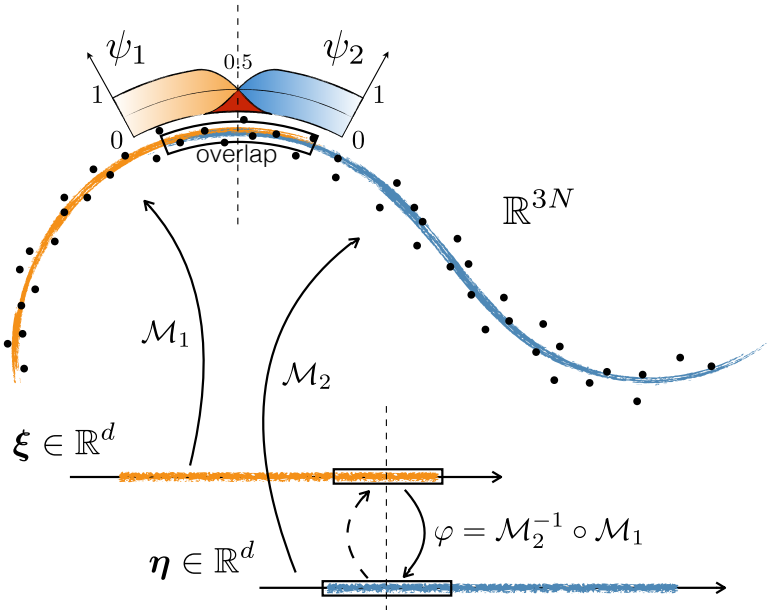


Fig. 5.3 Schematic view of atlas of collective variables. The configurations in $\mathbb{R}^D$ space are denoted by black dots. The configuration space is divided into partially overlapping regions. In each region, the intrinsic manifold is parametrized separately. In overlapping regions, a transition map is naturally defined between adjacent regions.

5.4 Results

We demonstrate next that the concept of atlas of SandCVs can be practically implemented into a standard MD code to perform enhanced sampling simulations and free-energy calculations. MD simulations were performed in version 2.8 of NAMD (Phillips et al., 2005) with Langevin thermostat at $300K$. The atlas of SandCVs and the adaptive biasing force algorithm (Darve et al., 2008) for enhanced sampling are implemented in C++ and communicate with NAMD through a TCL interface to obtain configurations and return forces on the atoms.

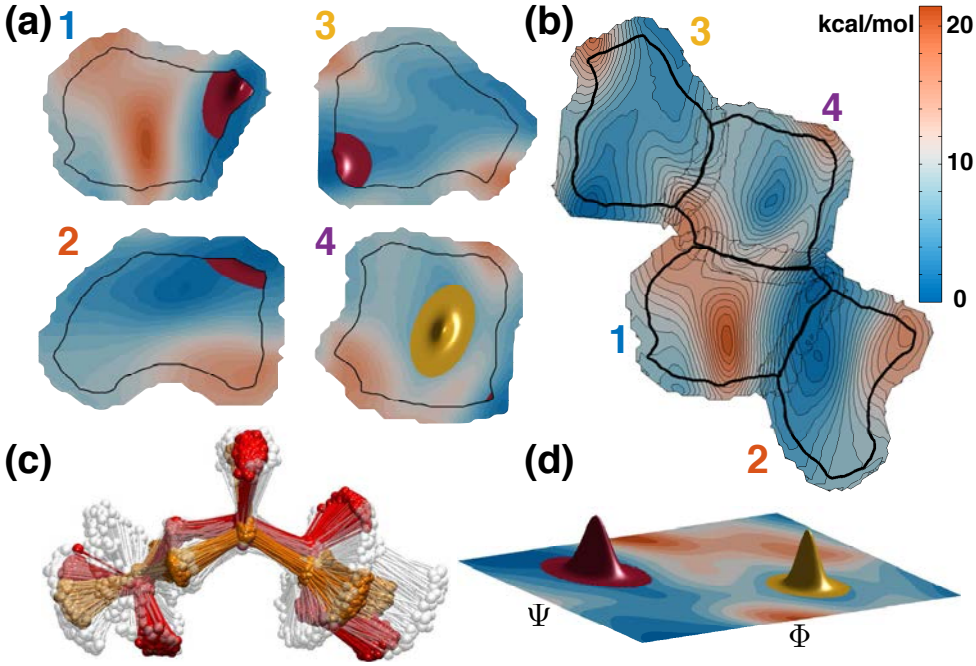


Fig. 5.4 (a) Free energy surfaces for each of the four CVs resulting from the systematic partitioning in Figure 5.2. (b) Coherent juxtaposition of these free energy profiles. (c) Sampling of two states, C_{7eq} and C_{7ax}, of the molecule. (d) Projection of these states on dihedral space.

We focus first on alanine dipeptide. Figure 5.4(a) shows the

free energy surfaces (FES) along each of the four CVs systematically identified by the algorithm, see Figure 5.2. The enhanced sampling MD trajectory seamlessly transitioned between charts and resulted in converged FES. Using Eq. (5.4) in the overlapping regions, we fix compatible additive constants to the free energy in each of the partitions. We discuss next the role of the Fixman correction, Eq. (5.4), in this example. Since the statistical error in the evaluation of the free energy is in the order of kT, Eq. (5.4) shows that a significant Fixman contribution would require that $|\mathbf{D}\varphi| > e$, which for $d = 2$ would require for instance a stretching of adjacent CV spaces by a factor of 1.6 along each coordinate. Interestingly, because of the nature of Isomap, the partitioning of the data-set, and the fact that the underlying intrinsic manifold is essentially flat (a flat torus) (Hashemian and Arroyo, 2015), the four embeddings for this molecule are nearly isometric: distances between low-dimensional points in Figure 5.2(e) are similar to distances of the corresponding conformations in high-dimensions, see Figure 5.2(d), and therefore close to the low-dimensional distances in an adjacent patch. As a result, $|\mathbf{D}\varphi| \approx 1$ and in this example the Fixman correction is negligible. Furthermore, recalling Eq. (5.7), the biasing forces of adjacent charts nearly coincide, and we can graphically "glue" the different patches to visualize a globally coherent FES, as illustrated in Figure 5.4(b).

We note however that such joint representation is not possible in general, nor necessary to compute thermodynamic observables. To show this, we consider two states A and B, computationally described by two ensembles obtained by restrained MD simulations around conformations C_{7eq} and C_{7ax} of the molecule, see Figure 5.4(c). For reference, these two states are graphically represented in dihedral space where we also compute the FES, Figure 5.4(d). When represented in the atlas of CVs, Figure 5.4(a), we note that C_{7eq} spreads over each of the four charts.

Using Eq. (5.5) sampled at 1 million configurations for each state, we find using the atlas of CVs that the probability of C_{7eq} relative to C_{7ax} is $P_{atlas} = 56.406$. The excellent agreement with the same quantity computed in dihedral space, $P_{dihed} = 56.415$, illustrates that the atlas of CVs provides a seamless statistical mechanics framework over multiple CVs.

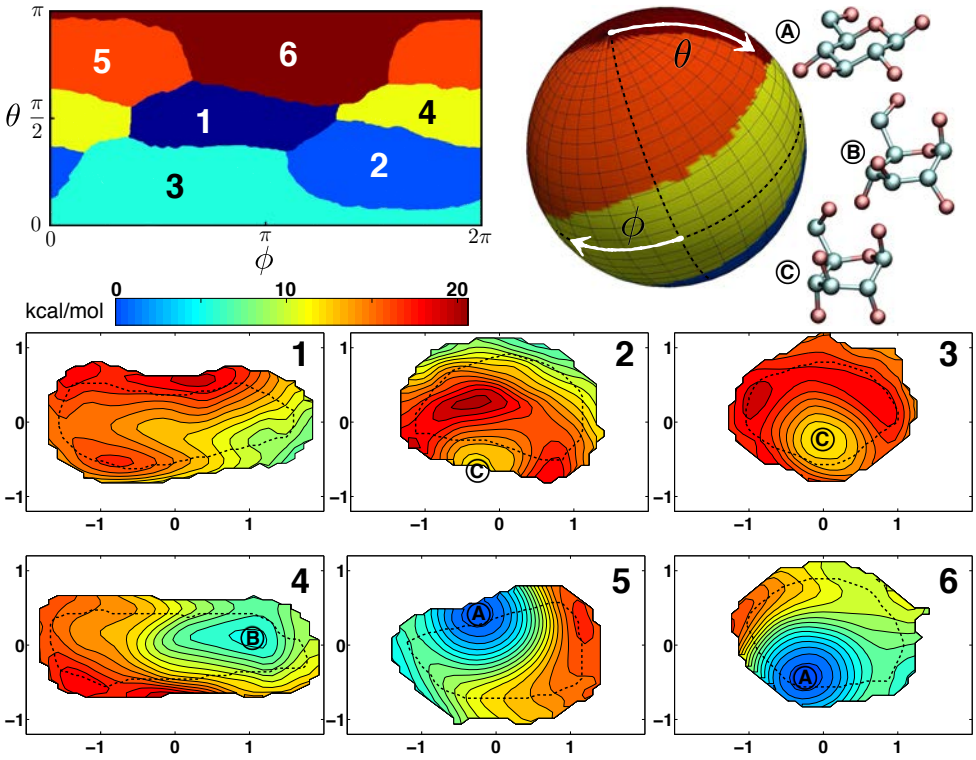


Fig. 5.5 Atlas of collective variables for β-D-Glucopyranose, a six-membered ring carbohydrates. Systematic partitioning of this molecule requires six patches to overcome topological obstructions and alleviate geometrical distortions (a), which result in six free-energy profiles (b).

We apply the proposed methodology to a different benchmark molecule, β-D-Glucopyranose (Branduardi and Faraldo-Gómez, 2013;

Spiwok et al., 2010), known to have the topology of the sphere (Biarnés et al., 2007). For this six-membered ring, the Cremer-Pople puckering coordinates (Cremer and Pople, 1975) reduce to polar coordinates (Q, θ, ϕ), and adequately describe its molecular flexibility with small fluctuations about the "radius" Q. The systematic partitioning of a well-sampled ensemble A leads here to six different charts, as shown in Figure 5.5(a) in (θ, ϕ) and (Q, θ, ϕ) spaces. Applying enhanced sampling along this atlas of CVs, we find the FES shown in Figure 5.5(b).

5.5 Summary

We have shown that molecular conformations and free energy landscapes can be described by an atlas of collective variables, in the same way that geographical landscapes are described by atlas of charts. Furthermore, this concept can be practically implemented in combination with the intrinsic manifold model of molecular flexibility by simply partitioning the intrinsic manifold. We have proposed a data-driven algorithm based on nonlinear dimensionality reduction methods to systematically build atlas of CVs for systems exhibiting conformational manifolds of nontrivial topologies, which preclude single-chart data-driven CVs. More generally, the proposed framework may significantly expand the applicability of systematic methods to identify CVs and that of enhanced-sampling methods, by providing a globally meaningful thermodynamic description and enhanced trajectory from locally appropriate CVs.

Chapter 6

Conclusion

In this dissertation, I introduced a general method to model molecular systems with smooth and nonlinear data-driven collective variables (SandCV). These CVs can be non-intrusively combined with off-the-shelf enhanced sampling molecular dynamics methods. I exemplified SandCV with alanine dipeptide, a benchmark for free energy calculations. I demonstrated the effectiveness of the method by providing numerical evidence of the convergence of enhanced sampling simulations. SandCV is implemented in a standalone C++ toolbox composed of distinct algorithmic blocks: alignment of the molecule, intrinsic manifold identification through nonlinear dimensionality reduction methods, smooth parametrization of the intrinsic manifold, and the closest-point projection of out-of-sample configurations onto the intrinsic manifold.

I also addressed a fundamental obstacle to construct data-driven CVs, due to the inability of dimensionality reduction methods to deal with complex topologies. These topological obstruction has been largely overlooked in the field and has misled interpretations.

Finally, I proposed a general remedy to overcome topological obstructions and alleviate geometrical distortion. This methods, called atlas of collective variables, provides a statistical mechanics framework

and a concrete algorithmic implementation to describe molecular systems with a collection of partially overlapping collective variables.

Future work

- The results of Chapter 2 showing that SandCV can bridge over poorly sampled regions, together with the notion of atlas of CVs, suggest that it may be possible to build increasingly detailed descriptions of complex conformational landscapes starting from locally-valid CVs. Along this line of thought, it may be adequate to resort to non-manifold descriptions of conformational space, with higher dimensional resolution in some regions and dimensionally narrow paths in others. These extensions should be motivated by specific studies on more complex proteins and their conformational flexibility.

- We are currently investigating unfolding pathways of I21 domain of Titin, a protein responsible for muscle flexibility, with data-driven approaches. In addition to identifying its collective behavior, we are utilizing SandCV to collectively unfold and refold it to its native states. Moreover, we can consider enhanced sampling simulations in the SandCV space along an unfolding pathway. By monitoring the distance to the manifold, which accounts for all other missing transversal variables, we expect to characterize free energy funnel of Titin folding.

- Describing long time-scale dynamical processes is another challenging topic in molecular simulations in which interesting dynamics take place as the system moves from one free energy basin to another through infrequent rare events in the time scales of often exceeding the milliseconds. To this end, SandCV can be employed in tandem with variety of methods from Transition

State Theory (TST) reliant approaches, like conformation flooding (Grubmüller, 1995) and Hyperdynamics (Voter, 1997), to methods that go beyond the limitation of TST, like Transition Path Sampling (Chandler, 1978) and Transition Interface Sampling (van Erp et al., 2003), to efficiently calculate the dynamics from affordable simulations.

- PLUMED is an open source library for free energy calculations in molecular systems, including many different state-of-the-art enhanced sampling methods and analytical and computational collective variables. It works together with many of the most popular molecular dynamics engines and is gaining ground in the field. Integrating SandCV in PLUMED will enable the researchers to easily use SandCV using the same familiar user interface to exploit the capabilities of this data-driven approach in a wide range of molecular systems.

Appendix A

More details on atlas of collective variables

A.1 Partitioning configuration space

The atlas of collective variables approach relies on a systematic partitioning of the slow-manifold into pieces that are tractable with dimensionality reduction techniques (open sets). Consider a smooth d-manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$ and sampled by a set of points $R = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N\} \subset \mathbb{R}^D$. Following Millán et al. (2013), the goal is to numerically represent $\mathcal{M}$ from the data in R through a collection of overlapping parametrizations, "patches", and make computations on it.

The partitioning process can be described as

1. Partition the set of scattered points (viewed as geometric markers) into L groups, on the basis of proximity by using METIS domain decomposition with a k-nearest neighbor graph (Karypis and Kumar, 1998). METIS tries to partition a graph in equals size subdomains with minimal shared boundary lengths.

   These L groups of points can be represented with index sets

$\mathcal{I}_\kappa, \kappa = 1, \ldots, L$ with $\cup_{\kappa=1}^{L} \mathcal{I}_\kappa = \{1, 2, \ldots, N\}$ and $\mathcal{I}_\kappa \cap \mathcal{I}_\iota = \emptyset$ when $\kappa \neq \iota$. As it will become clear below, there is a one-to-one correspondence between these groups of geometric markers and the local parametrizations of the manifold.

2. For each partition we create an enlarged index set $\mathcal{J}_\kappa$ by combining the indices in $\mathcal{I}_\kappa$ and the indices from their nearest neighbors defined by a cutoff distance $d_{\texttt{cut}}$, such that

$$\mathcal{J}_\kappa = \{a \in \{1, \ldots, N\} \mid |\boldsymbol{r}_a - \boldsymbol{r}_b| < d_{\texttt{cut}} \ \forall \ b \in \mathcal{I}_\kappa\}.$$

Here and elsewhere, $|\cdot|$ denotes the Euclidean norm. The enlarged index set obeys $\cup_{\kappa=1}^{L} \mathcal{J}_\kappa = \{1, 2, \ldots, N\}$, but now $\mathcal{J}_\kappa \cap \mathcal{J}_\iota \neq \emptyset$. The intersection of these index sets defines the overlapping amongst patches.

3. A dimensionality reduction technique is applied to each one of the enlarged sets $P_\kappa = \{\boldsymbol{r}_a\}_{a \in \mathcal{J}_\kappa} \subset \mathbb{R}^D$ to find their low-dimensional embedding $\Xi_\kappa = \{\boldsymbol{\xi}_a\}_{a \in \mathcal{J}_\kappa} \subset \mathbb{R}^d$. In our simulations we have used Isomap, which is a NLDR method designed to find an isometric low-dimensional representation by approximately preserving the geodesic distance on the manifold.

4. The quality of the resulting embedding is measured through a reconstruction error computed as

$$e_a = \frac{|\boldsymbol{r}_a - \sum_{b \in \mathcal{K}_a} w_{ab} \, \boldsymbol{r}_b|}{|\boldsymbol{r}_a|} \qquad \forall a = 1, \ldots, N,$$

where $\mathcal{K}_a$ is the index set with the first k-nearest neighbors of the a-th configuration in the low-dimensional space, and w_{ab} are the weights that best linearly reconstruct $\boldsymbol{\xi}_a$ from its k-nearest

neighbors (Roweis and Saul, 2000), obtained by

$$\min_{w_{ab}} \left| \boldsymbol{\xi}_a - \sum_{b \in \mathcal{K}_a} w_{ab}\, \boldsymbol{\xi}_b \right| \qquad \text{subject to} \qquad \sum_{b \in \mathcal{K}_a} w_{ab} = 1.$$

We perform the neighbor search in low-dimensional space because (1) it is more efficient, and more importantly, because (2) this allows us to detect data collapse in the nonlinear dimensionality process.

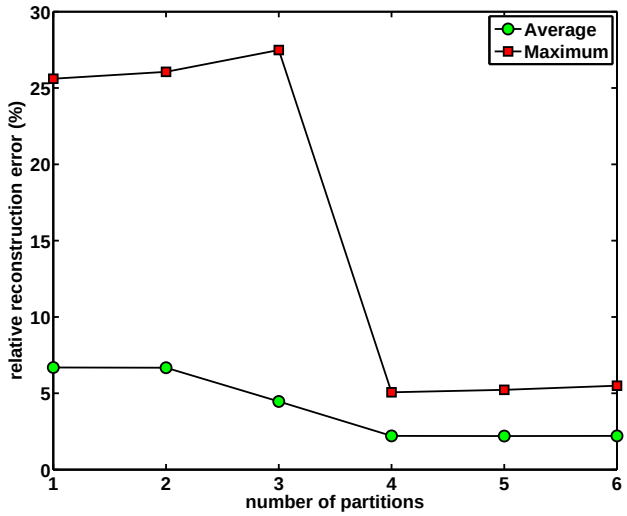


Fig. A.1 Average and maximum relative reconstruction error in the partitioning of alanine dipeptide. It is clear from this plot that the maximum norm best discriminates the number of required partitions given by the algorithm, four in this case.

5. We monitor in each partition the reconstruction error. We require that $\max\{e_a\}_{a \in \mathcal{J}_\kappa} < Tol_e$ for a partition to be acceptable, where Tol_e is a numerical tolerance typically below 0.1. If the reconstruction error exceeds the tolerance, the partition is

recursively subdivided until the reconstruction requirement is met. Figure A.1 shows the reconstruction error for different number of partitions of an ensemble of alanine dipeptide.

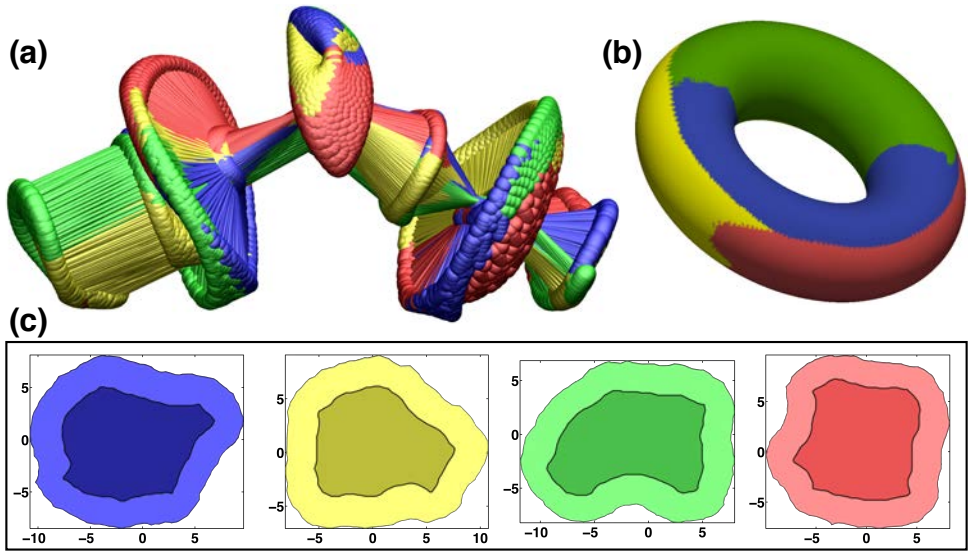


Fig. A.2 Partitioning a molecular ensemble of alanine dipeptide. The partitioning procedure recursively proceeds until all the partitions are tractable with nonlinear dimensionality reduction (NLDR) methods, which leads to four partitions in configuration space. Although the partitions have overlap, using partition of unity we can divide the configuration space into non-overlapping regions (a). These partitions also depicted over a torus of dihedral angles to highlight the topology of the intrinsic manifold (b). The low-dimensional embedding of each of the original four partitions, calculated separately with NLDR methods is shown in (c). The non-overlapping regions are outlined with darker colors.

6. After the partitioning process is finished and the low-dimensional embedding for each partition is available, we build a smooth parametrization for each patch using smooth basis functions $p_a(\boldsymbol{\xi})$

associated to nodes from a uniform grid of landmarks $\boldsymbol{\eta}_i \in \mathbb{R}^d$,

$$\begin{aligned} \mathcal{M}_\kappa : \Omega_\kappa \subset \mathbb{R}^d &\longrightarrow \mathbb{R}^D \\ \boldsymbol{\xi} &\longmapsto \sum_{a \in \mathcal{J}_\kappa} p_a(\boldsymbol{\xi})\, \boldsymbol{y}_a. \end{aligned} \tag{A.1}$$

The high-dimensional control points $\boldsymbol{y}_a \in \mathbb{R}^D$ are chosen such that the reconstruction error is minimized in a least-squares sense. Full details are presented in Chapter 2. Thus, we end up with a collection of partially overlapping parametrizations of the intrinsic manifold, Figure A.2.

We consider here the local maximum-entropy (LME) basis functions. See (Arroyo and Ortiz, 2006; Millán et al., 2011; Rosolen et al., 2010) for the LME formulation, properties, and the evaluation of the basis functions and their derivatives.

We consider a Shepard partition of unity associated with the geometric markers. Given a set of non-negative reals $\{\beta_a\}_{a=1,2,\ldots,N}$, we define the Shepard partition of unity with Gaussian weight associated to the set Q as

$$w_a(\boldsymbol{r}) = \frac{\exp(-\beta_a |\boldsymbol{r} - \boldsymbol{r}_a|^2)}{\sum_{b=1}^N \exp(-\beta_b |\boldsymbol{r} - \boldsymbol{r}_b|^2)}. \tag{A.2}$$

To obtain a coarser partition of unity representative of a partition, we aggregate the partition of unity functions as

$$\psi_\kappa(\boldsymbol{r}) = \sum_{a \in \mathcal{I}_\kappa} w_a(\boldsymbol{r}). \tag{A.3}$$

These non-negative functions form a partition of unity in $\mathbb{R}^D$. For very large β_a, these functions tend to the characteristic functions of the Voronoi cells in high-dimension associated to the group of points given by $\mathcal{I}_\kappa$. They can thus be viewed as a smooth regularization of these

characteristic functions (Millán et al., 2013). Figure A.3 illustrates the four partition of unity functions in the torus representation of alanine dipeptide. Employing these partition of unity functions, we can also divide the whole configuration space into non-overlapping regions by associating each configuration $\boldsymbol{r}$ to the patch κ which maximizes $\psi_\kappa(\boldsymbol{r})$.

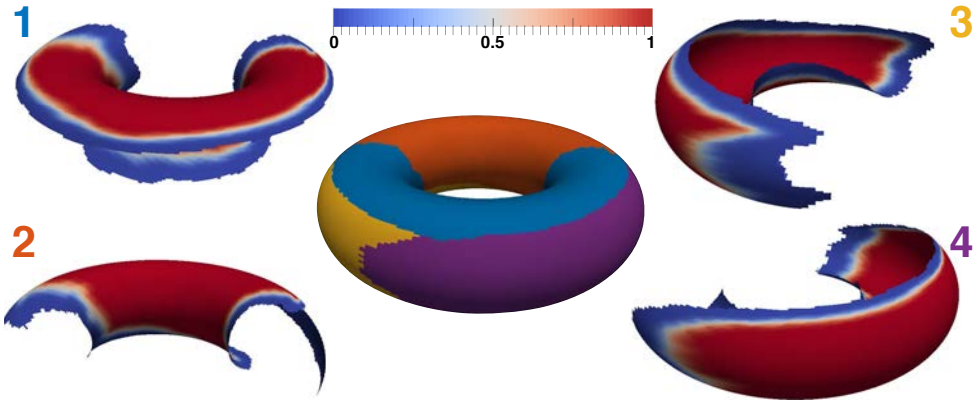


Fig. A.3 Partition of unity for alanine dipeptide. Partition of unity functions are represented on the torus defined by dihedral angles. It is worth emphasizing that these functions are evaluated in high-dimensional configuration space and the torus representation is for visualization purposes only.

A.2 Adaptive biasing force for atlas of collective variables

For enhanced sampling molecular dynamics simulation with atlas of CVs, we need to know for any given configuration $\boldsymbol{r}$ along the trajectory which charts are "active" (there could be multiple because of overlap). For these active charts, we need to evaluate the CVs and choose one of the charts to compute the biasing force to be applied on the system. We identify active charts by evaluating the partition of unity functions

and checking if they are larger than a small threshold. We select the distinguished chart as that with maximum partition of unity function at $\boldsymbol{r}$.

A.2.1 Adaptive biasing force in a single chart

The free energy of a system can be estimated by thermodynamic integration of the average forces along collective variables $\boldsymbol{\xi} = \mathcal{C}_1(\boldsymbol{r})$ (Darve and Pohorille, 2001),

$$\mathbf{D}A_1(\boldsymbol{\xi}) = -\left\langle \boldsymbol{f}_1^{sys}(t^k) \right\rangle, \qquad \boldsymbol{\xi} \in b_k,$$

with

$$\boldsymbol{f}_1^{sys}(t_i^k) = \frac{\partial}{\partial t}\left(M_\xi \frac{\partial \mathcal{C}_1}{\partial t} \right)\Bigg|_{t_i^k}, \tag{A.4}$$

where b_k is the k-th bin in CV space, t_i^k is the time when i-th sample resides in k-th bin, and $M_\xi^{-1} = \mathbf{D}\mathcal{C}_1 M^{-1} \mathbf{D}\mathcal{C}_1^T$ where M is the mass matrix. However, if the energy barriers along $\boldsymbol{\xi}$ are significant, this method of calculating free energy will face the inefficient sampling. To overcome this problem, the adaptive biasing force (ABF) algorithm (Darve et al., 2008) adds a properly chosen, position-dependent biasing force to the system that allows for a self-diffusion in the CV space.

This low-dimensional biasing force, $\boldsymbol{F}_1^{CV}$, is calculated at each time step as the running average of the force acting along $\boldsymbol{\xi}$ in the k-th bin assuming that $\mathcal{N}(N,k)$ is the number of samples collected in bin k after N steps in a simulation,

$$\boldsymbol{F}_1^{CV}(N,k) = -\frac{R\left(\mathcal{N}(N,k)\right)}{\mathcal{N}(N,k)} \sum_{i=1}^{\mathcal{N}(N,k)} \boldsymbol{f}_1^{sys}(t_i^k), \tag{A.5}$$

where $R\left(\mathcal{N}(N,k)\right) = \min(1, \mathcal{N}(N,k)/N_0)$ is a ramp function to avoid the crude estimation of the biasing force in poorly sampled bins.

Afterwards we should subtract the applied biasing force from the modified system force to recover the original system force,

$$\boldsymbol{f}_1^{sys}(t_i^k) = \frac{\partial}{\partial t}\left(M_\xi \frac{\partial \xi}{\partial t}\right)\Bigg|_{t_i^k} - \boldsymbol{F}_1^{CV}(N-1,k).$$

In this method, the free energy derivatives can be estimated in each bin by

$$\mathbf{D}A_1(\boldsymbol{\xi}) = -\left\langle \boldsymbol{f}_1^{sys} \right\rangle = -\left\langle \boldsymbol{f}_1^{mod} - \boldsymbol{F}_1^{CV} \right\rangle. \tag{A.6}$$

It can be shown from Eq. (A.5) and Eq. (A.6) that after a brief equilibration the biasing force converges to the $-\langle \boldsymbol{f}_1^{sys} \rangle$ and thus the average of the modified system force acting along $\boldsymbol{\xi}$, $\boldsymbol{f}_1^{mod}$ becomes zero. This allows for a self-diffusion motion along the CV in the modified system,

$$\mathbf{D}A_1(\boldsymbol{\xi}) \approx \boldsymbol{F}_1^{CV}. \tag{A.7}$$

A.2.2 Adaptive biasing forces in multiple charts

We exploit the fact that in the ABF method, the biasing force is in fact arbitrary, and thus the choice resulting in self-diffusion along CV is just a convenient one (allowing us to check convergence) but by no means a strict requirement for free-energy calculations. Relying on the partition of unity (PU) functions, we modify ABF as follows. If one of the PU functions is equal to one at the current configuration, then all other partition of unity functions are zero and there is a single active chart. Thus, there is no overlap and usual vectorial ABF can be used. However, if several PU functions are nonzero, we are in an overlapping region with multiple active charts. In this case, we first tag the chart with the largest PU function as the master chart and the rest of charts with nonzero PU function as slave charts. Then, we calculate the biasing

force from the master chart, say $\mathcal{C}_1$, as

$$\begin{aligned} \boldsymbol{F}_1^{CV}(\boldsymbol{\xi}) &= -\mathbf{D}A_1(\boldsymbol{\xi}), \\ \boldsymbol{F}_1^{bias}(\boldsymbol{r}) &= \boldsymbol{F}_1^{CV}(\boldsymbol{\xi})\mathbf{D}\mathcal{C}_1(\boldsymbol{r}). \end{aligned}$$

Finally, we let the slave charts know about the exerted biasing force, by projecting it into all the slave CVs. Denoting $\mathcal{C}_2$ as a slave chart, we have

$$\begin{aligned} \boldsymbol{F}_2^{CV}(\boldsymbol{\eta}) &= \boldsymbol{F}_1^{bias}(\boldsymbol{r})\mathbf{D}\mathcal{C}_2^{-1}(\boldsymbol{r}) \\ &= \boldsymbol{F}_1^{CV}(\boldsymbol{\xi})\mathbf{D}\mathcal{C}_1(\boldsymbol{r})\mathbf{D}\mathcal{C}_2^{-1}(\boldsymbol{r}). \end{aligned}$$

This projected biasing force, $\boldsymbol{F}_2^{CV}(\boldsymbol{\eta})$, is stored to estimate the free energy gradient in slave charts. In SandCV, this transition of biasing forces between charts can be readily performed by using the transition map, $\mathbf{D}\varphi(\boldsymbol{\eta})$,

$$\boldsymbol{F}_2^{CV}(\boldsymbol{\eta}) = \boldsymbol{F}_1^{CV}(\boldsymbol{\xi})\mathbf{D}\varphi^{-1}(\boldsymbol{\eta}). \tag{A.8}$$

We note that regions where a chart is a master span all the configuration space, but regions with a single active chart (where the method is plain ABF) do not. Active but slave charts do not have any influence on the running simulation and only gather statistics about the bias force given by the master to evaluate the free energy.

A.3 Molecular simulation details of six-membered ring

We obtain a well sampled trajectory of β-D-Glucopyranose by performing a 10 ns metadynamics simulation with ϕ and θ as collective variables with Guassian height of 0.1 and sigma of 0.1, starting from chair conformation ($^4C_1$), and using Glycam force field (Kirschner et al.,

2008) at 300K using Langevin thermostat. We set time step to 0.2fs and we sample at each 500 steps (50fs), obtaining 100K configurations. We then use this trajectory as the training set to build an atlas of SandCVs. Systematic partitioning of this trajectory provides us with 6 different partitions and we build for each partition a set of SandCVs. Afterwards, we run a 100ns adaptive biasing force simulation using the constructed atlas of SandCVs with time step of 0.5 fs and 32 bins in each individual CV although the simulation does not need to fill all the bins since the embedding is not square. Then by post-processing the force histogram from this simulation, we calculate the free energy surfaces of conformational changes of β-D-Glucopyranose.

Appendix B

Smooth contact maps

Here, we introduce the definition and derivatives of two smooth contact map using sigmoid and log-sum-exp functions.

Sigmoid function

The sigmoid function takes the form

$$f(u) = \frac{1}{1 + e^{-u}} \tag{B.1}$$

By changing the parameters as following, we can define a differentiable version of the widely-used contact map

$$[C(\kappa, \epsilon)]_{ij} = \frac{1}{2} \left\{ 1 - \tanh \left[\frac{1}{\epsilon} \left(\|r_i - r_j\| - \kappa \right) \right] \right\} \tag{B.2}$$

This Sigmoid function measures the proximity of residues i and j. The parameter κ is, roughly, the maximum distance at which two residues can influence each other, and ϵ estimates the certainty of the cut-off distance (κ).

If $\epsilon = \kappa$, we can make $[C(\kappa, \epsilon)]_{ij}$ arbitrarily small for $i \neq j$ as κ decreases to zero. Hence, we can assume $C(\kappa, \epsilon)$ is diagonally dominant,

which implies $C(\kappa, \epsilon)$ is positive definite.

$$\begin{cases} \lim_{\kappa \to 0} \left[C(\kappa, \epsilon)\right]_{ij} = 0 & for \quad i \neq j \\ \lim_{\kappa \to 0} \left[C(\kappa, \epsilon)\right]_{ij} = \frac{1}{2} \left\{1 + \tanh(1)\right\} & for \quad i = j \end{cases}$$

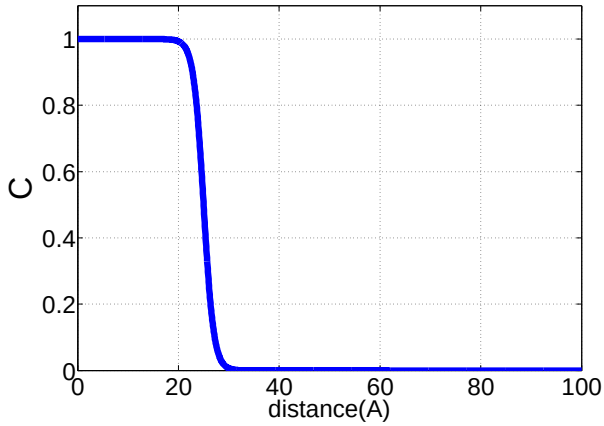


Fig. B.1 Smooth contact map with Sigmoid function

Derivatives: We define u_{ij} as follow

$$\begin{aligned} u_{ij} &= \frac{1}{\epsilon} \left(\| r_i - r_j \| - \kappa \right) \\ &= \frac{1}{\epsilon} \left(\left(\sum_{\beta=1}^{3} (r_{i\beta} - r_{j\beta})^2 \right)^{\frac{1}{2}} - \kappa \right) \end{aligned} \tag{B.3}$$

and thus

$$\begin{aligned} \left[C(\kappa, \epsilon)\right]_{ij} &= \frac{1}{2} \left(1 - \tanh\left(u_{ij}\right)\right) \\ &= \left(1 + e^{2u_{ij}}\right)^{-1}. \end{aligned} \tag{B.4}$$

Because u is only function of r, by applying chain rule we have

$$\frac{\partial C_{ij}}{\partial r_{k\beta}} = \frac{\partial C_{ij}}{\partial u_{pq}} \frac{\partial u_{pq}}{\partial r_{k\beta}}. \tag{B.5}$$

The derivative of contact map with respect to u is then

$$\begin{aligned} \frac{\partial C_{ij}}{\partial u_{ij}} &= \frac{\partial}{\partial u_{ij}} \left(\frac{1}{2} \left(1 - \tanh\left(u_{ij} \right) \right) \right) \\ &= -\frac{1}{2} \left(1 - \tanh^2\left(u_{ij} \right) \right) \\ &= -\frac{1}{2} \left(1 - \tanh\left(u_{ij} \right) \right) \left(1 + \tanh\left(u_{ij} \right) \right) \\ &= -2 \; C_{ij} \; \left(1 - C_{ij} \right) \end{aligned} \tag{B.6}$$

and derivative of u with respect to r is

$$\frac{\partial u_{ij}}{\partial r_{k\beta}} = \begin{cases} \dfrac{r_{k\beta} - r_{j\beta}}{\epsilon \| r_k - r_j \|} & if \quad k = i, k \neq j \\ \dfrac{r_{k\beta} - r_{i\beta}}{\epsilon \| r_i - r_k \|} & if \quad k = j, k \neq i \\ 0 & \text{otherwise} \end{cases} \tag{B.7}$$

Plugging Eq. (B.6) and Eq. (B.7) into Eq. (B.5) we have

$$\frac{\partial C_{ij}}{\partial r_{k\beta}} = \begin{cases} 2C_{kj}(C_{kj} - 1) \dfrac{(r_{k\beta} - r_{j\beta})}{\epsilon \| r_k - r_j \|} & if \quad k = i, k \neq j \\ 2C_{ik}(C_{ik} - 1) \dfrac{(r_{k\beta} - r_{i\beta})}{\epsilon \| r_i - r_k \|} & if \quad k = j, k \neq i \\ 0 & \text{otherwise} \end{cases} \tag{B.8}$$

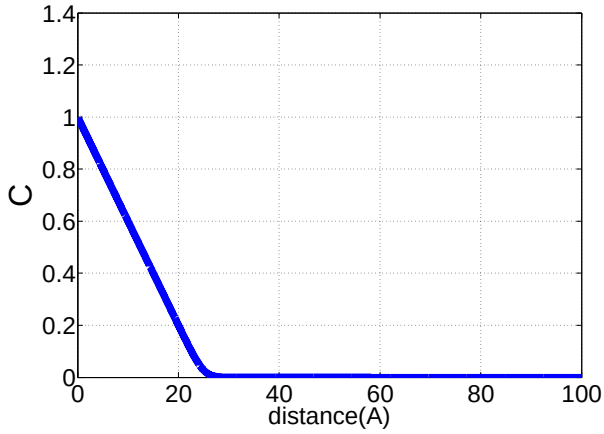


Fig. B.2 Smooth contact map with log-sum-exp function

Log-sum-exp Function

Considering general definition of log-sum-exp function

$$f(u) = \log\left(\sum_{a=1}^{N} e^{f_a(u)}\right) \tag{B.9}$$

we define two functions as follows,

$$\begin{aligned} f_1 &= 0, \\ f_2 &= -\left(\|r_i - r_j\| - \kappa\right). \end{aligned}$$

After plugging these function into Eq. (B.9) and normalizing them, we have

$$[C(\kappa,\epsilon)]_{ij} = \frac{\epsilon}{\kappa}\left\{\log\left(1 + e^{-\frac{1}{\epsilon}(\|r_i - r_j\| - \kappa)}\right)\right\} \tag{B.10}$$

where the parameter κ is again a rough cut-off distance of two residues influence, and the value of ϵ tunes the certainty of this cut-off distance.

Derivatives: Using the same definition of u_{ij} as Eq. (B.3), we have

$$[C(\kappa,\epsilon)]_{ij} = \frac{\epsilon}{\kappa}\left\{\log\left(1+e^{-u_{ij}}\right)\right\}, \tag{B.11}$$

and its derivatives with respect to u are

$$\begin{aligned}\frac{\partial C_{ij}}{\partial u_{ij}} &= \frac{\partial}{\partial u_{ij}}\left(\frac{\epsilon}{\kappa}\left\{\log\left(1+e^{-u_{ij}}\right)\right\}\right)\\ &= -\frac{\epsilon\, e^{-u_{ij}}}{\kappa\left(1+e^{-u_{ij}}\right)} = \frac{-\epsilon}{\kappa\left(1+e^{u_{ij}}\right)}\\ &= \frac{\epsilon}{\kappa}\left(\frac{1}{1+e^{-u_{ij}}}-1\right) = \frac{\epsilon}{\kappa}\left(e^{-\frac{\kappa}{\epsilon}C_{ij}}-1\right).\end{aligned} \tag{B.12}$$

Finally by using Eq. (B.12) and recalling Eq. (B.7) and Eq. (B.5), the derivative of log-sum-exp contact map with respect to r takes the following form,

$$\frac{\partial C_{ij}}{\partial r_{k\beta}} = \begin{cases} \left(e^{-\frac{\kappa}{\epsilon}C_{kj}}-1\right)\dfrac{(r_{k\beta}-r_{j\beta})}{\kappa\,\|r_k-r_j\|} & if \quad k=i, k\neq j\\ \left(e^{-\frac{\kappa}{\epsilon}C_{ik}}-1\right)\dfrac{(r_{k\beta}-r_{i\beta})}{\kappa\,\|r_i-r_k\|} & if \quad k=j, k\neq i\\ 0 & \text{otherwise}\end{cases} \tag{B.13}$$

References

B. J. Alder and T. E. Wainwright. Phase Transition for a Hard Sphere System. *The Journal of Chemical Physics*, 27(1957):1208, 1957.

B. J. Alder and T. E. Wainwright. Studies in Molecular Dynamics. I. General Method. *The Journal of Chemical Physics*, 31(1959):459, 1959.

Andrea Amadei, A. B. Linssen, and Herman J. C. Berendsen. Essential dynamics of proteins. *Proteins*, 17(4):412–25, 1993.

E Anderson, Z Bai, C Bischof, S Blackford, J Demmel, J Dongarra, J Du Croz, A Greenbaum, S Hammarling, A McKenney, and D Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999. ISBN 0-89871-447-8 (paperback).

Marino Arroyo and Michael Ortiz. Local maximum-entropy approximation schemes: a seamless bridge between finite elements and meshfree methods. *International Journal for Numerical Methods in Engineering*, 65(13):2167–2202, 2006.

Marino Arroyo, Luca Heltai, Daniel Millán, and Antonio DeSimone. Reverse engineering the euglenoid movement. *Proceedings of the*

National Academy of Sciences of the United States of America, 109 (44):17874–9, 2012.

Adi Ben-Israel. A volume associated with m x n matrices. *Linear Algebra and its Applications*, 167:87–111, 1992.

Xevi Biarnés, Albert Ardèvol, Antoni Planas, Carme Rovira, Alessandro Laio, and Michele Parrinello. The conformational free energy landscape of beta-D-glucopyranose. Implications for substrate preactivation in beta-glucoside hydrolases. *Journal of the American Chemical Society*, 129(35):10686–93, 2007.

Christopher M Bishop, Markus Svensén, and Christopher K I Williams. GTM: The Generative Topographic Mapping. *Neural Computation*, 10(1):215–234, 1998.

Peter G. Bolhuis, Christoph Dellago, and David Chandler. Reaction coordinates of biomolecular isomerization. *Proceedings of the National Academy of Sciences of the United States of America*, 97(11):5877–82, 2000.

Massimiliano Bonomi, Davide Branduardi, Francesco L. Gervasio, and Michele Parrinello. The unfolded ensemble and folding mechanism of the C-terminal GB1 beta-hairpin. *Journal of the American Chemical Society*, 130(42):13938–44, 2008.

Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provasi, Paolo Raiteri, Davide Donadio, Fabrizio Marinelli, Fabio Pietrucci, Ricardo a. Broglia, and Michele Parrinello. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications*, 180(10): 1961–1972, 2009.

David W. Borhani and David E. Shaw. The future of molecular dynamics simulations in drug discovery. *Journal of Computer-Aided Molecular Design*, 26:15–26, 2012.

Davide Branduardi and José D. Faraldo-Gómez. String Method for Calculation of Minimum Free-Energy Paths in Cartesian Space in Freely Tumbling Systems. *Journal of Chemical Theory and Computation*, 9(9):4140–4154, 2013.

Davide Branduardi, Francesco Luigi Gervasio, and Michele Parrinello. From A to B in free energy space. *The Journal of Chemical Physics*, 126(5):054103, 2007.

W Michael Brown, Shawn Martin, Sara N Pollock, Evangelos a Coutsias, and Jean-Paul Watson. Algorithmic dimensionality reduction for molecular structure analysis. *The Journal of Chemical Physics*, 129 (6):064118, 2008.

Michele Ceriotti, Gareth A. Tribello, and Michele Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences of the United States of America*, 108(32):13023–8, 2011.

Michele Ceriotti, Gareth a. Tribello, and Michele Parrinello. Demonstrating the Transferability and the Descriptive Power of Sketch-Map. *Journal of Chemical Theory and Computation*, 9(3): 1521–1532, 2013.

David Chandler. Statistical mechanics of isomerization dynamics in liquids and the transition state approximation. *The Journal of Chemical Physics*, 68(6):2959, 1978.

Christophe Chipot and Andrew Pohorille, editors. *Free Energy*

Calculations, volume 86. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-38447-2.

R R Coifman, I G Kevrekidis, S Lafon, M Maggioni, and B Nadler. Diffusion Maps, Reduction Coordinates, and Low Dimensional Representation of Stochastic Systems. *Multiscale Modeling & Simulation*, 7(2):842–864, 2008.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

D Cremer and J A Pople. A General definition of ring puckering coordinates. *Journal of the American Chemical Society*, 97(6): 1354–1358, 1975.

Yanier Crespo, Fabrizio Marinelli, Fabio Pietrucci, and Alessandro Laio. Metadynamics convergence law in a multidimensional system. *Physical Review E*, 81(5):055701, 2010.

Isabella Daidone and Andrea Amadei. Essential dynamics: foundation and applications. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(5):762–770, 2012.

Eric Darve and Andrew Pohorille. Calculating free energies using average force. *The Journal of Chemical Physics*, 115(20):9169, 2001.

Eric Darve, David Rodríguez-Gómez, and Andrew Pohorille. Adaptive biasing force method for scalar and vector free energy calculations. *The Journal of Chemical Physics*, 128(14):144120, 2008.

Payel Das, Mark Moll, Hernán Stamati, Lydia E Kavraki, and Cecilia Clementi. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proceedings of the National Academy of Sciences of the United States of America*, 103 (26):9885–90, 2006.

Charles C David and Donald J Jacobs. Principal component analysis: a method for determining the essential dynamics of proteins. In *Methods in molecular biology*, volume 1084, pages 193–226. 2014. ISBN 9781627036573.

B L de Groot, X Daura, a E Mark, and H Grubmüller. Essential dynamics of reversible peptide folding: memory-free conformational dynamics governed by internal hydrogen bonds. *Journal of molecular biology*, 309(1):299–313, 2001.

Laurens Van der Maaten, Eric Postma, and Jaap van den Henrik. Dimensionality reduction: A comparative review. *Review Literature And Arts Of The Americas*, pages 1–35, 2009.

E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

Ron O Dror, Morten ØJensen, David W Borhani, and David E Shaw. Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations. *The Journal of General Physiology*, 135(6):555–62, 2010.

Mojie Duan, Minghai Li, Li Han, and Shuanghong Huo. Euclidean sections of protein conformation space and their implications in dimensionality reduction. *Proteins*, 82(10):2585–96, 2014.

Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. String method for the study of rare events. *Physical Review B*, 66(5):052301, 2002.

Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. Finite temperature string method for the study of rare events. *The Journal of Physical Chemistry B*, 109(14):6688–93, 2005.

David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910, 2005.

Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, 103(19):8577, 1995.

Santi Esteban-Martín, Robert Bryn Fenwick, and Xavier Salvatella. Refinement of ensembles describing unstructured proteins using NMR residual dipolar couplings. *Journal of the American Chemical Society*, 132(13):4626–32, 2010.

R Bryn Fenwick, Santi Esteban-Martín, Barbara Richter, Donghan Lee, Korvin F a Walter, Dragomir Milovanovic, Stefan Becker, Nils A Lakomek, Christian Griesinger, and Xavier Salvatella. Weak long-range correlated motions in a surface patch of ubiquitin involved in molecular recognition. *Journal of the American Chemical Society*, 133(27):10336–9, 2011.

Andrew L Ferguson, Athanassios Z Panagiotopoulos, Pablo G Debenedetti, and Ioannis G Kevrekidis. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 107(31):13597–602, 2010.

Andrew L Ferguson, Athanassios Z Panagiotopoulos, Pablo G Debenedetti, and Ioannis G Kevrekidis. Integrating diffusion maps with umbrella sampling: application to alanine dipeptide. *The Journal of Chemical Physics*, 134(13):135103, 2011a.

Andrew L Ferguson, Athanassios Z Panagiotopoulos, Ioannis G Kevrekidis, and Pablo G Debenedetti. Nonlinear dimensionality

reduction in molecular simulation: The diffusion map approach. *Chemical Physics Letters*, 509(1-3):1–11, 2011b.

Marshall Fixman. Classical Statistical Mechanics of Constraints: A Theorem and Application to Polymers. *Proceedings of the National Academy of Sciences of the United States of America*, 71 (8):3050–3053, 1974.

D. Frenkel. Simulations: The dark side. *The European Physical Journal Plus*, 128(1):10, 2013.

Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, second edition, 2002. ISBN 9780122673511.

Angel García. Large-amplitude nonlinear motions in proteins. *Physical Review Letters*, 68(17):2696–2699, 1992.

P Grassberger and I Procaccia. Measuring the strangeness of strange attractors. *Physica D*, 9(1-2):189–208, 1983.

Helmut Grubmüller. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Physical Review E*, 52(3):2893–2906, 1995.

C. Hartmann, J. C. Latorre, and G. Ciccotti. On two possible definitions of the free energy for collective variables. *The European Physical Journal Special Topics*, 200(1):73–89, 2011.

Behrooz Hashemian and Marino Arroyo. Topological obstructions in the way of data-driven collective variables. *The Journal of Chemical Physics*, 142(4):044102, 2015.

Behrooz Hashemian, Daniel Millán, and Marino Arroyo. Modeling and enhanced sampling of molecular systems with smooth and nonlinear

data-driven collective variables. *The Journal of Chemical Physics*, 139(21):214101, 2013.

Behrooz Hashemian, Daniel Millán, and Marino Arroyo. SandCV++: a toolbox for data-driven collective variables. *In preparation*, 2015a.

Behrooz Hashemian, Daniel Millán, and Marino Arroyo. Charting molecular free-energy landscapes with an atlas of collective variables. *In preparation*, 2015b.

Rainer Hegger, Alexandros Altis, Phuong Nguyen, and Gerhard Stock. How Complex Is the Dynamics of Peptide Folding? *Physical Review Letters*, 98(2):028102, 2007.

Katherine A Henzler-Wildman, Ming Lei, Vu Thai, S Jordan Kerns, Martin Karplus, and Dorothee Kern. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171): 913–6, 2007.

G E Hinton and R R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science (New York, N.Y.)*, 313(5786): 504–7, 2006.

L. Holm, C. Sander, and Others. Protein structure comparison by alignment of distance matrices. *Journal of molecular biology*, 233: 123–123, 1993.

Imre Jákli, Svend J. Knak Jensen, Imre G. Csizmadia, and András Perczel. Variation of conformational properties at a glance. True graphical visualization of the Ramachandran surface topology as a periodic potential energy surface. *Chemical Physics Letters*, 547: 82–88, 2012.

C. Jarzynski. Nonequilibrium Equality for Free Energy Differences. *Physical Review Letters*, 78(14):2690–2693, 1997.

H Jónsson, G Mills, and K W Jacobsen. *Nudged elastic band method for finding minimum energy paths of transition*, chapter 16, pages 385–404. World Scientific, 1998.

George Karypis and Vipin Kumar. A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392, 1998.

Karl N Kirschner, Austin B Yongye, Sarah M Tschampel, Jorge González-Outeiriño, Charlisa R Daniels, B Lachele Foley, and Robert J Woods. GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *Journal of computational chemistry*, 29(4): 622–55, 2008.

Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

P M Kroonenberg, W J Dunn, and J J F Commandeur. Consensus molecular alignment based on generalized procrustes analysis. *Journal of Chemical Information and Computer Sciences*, 43(6):2025–32, 2003.

Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12562–6, 2002.

Oliver F Lange and Helmut Grubmüller. Can principal components yield a dimension reduced description of protein dynamics on long time scales? *The Journal of Physical Chemistry B*, 110(45):22842–52, 2006.

John Aldo Lee and Michel Verleysen. Nonlinear dimensionality reduction of data manifolds with essential loops. *Neurocomputing*, 67:29–53, 2005.

John Aldo Lee and Michel Verleysen. *Nonlinear Dimensionality Reduction*. Springer New York, New York, NY, 2007. ISBN 978-0-387-39350-6.

Tony Lelièvre, Mathias Rousset, and Gabriel Stoltz. *Free Energy Computations*, volume F. Imperial College Press, London, 2010. ISBN 978-1-908978-75-2.

Jianfeng Lu and Eric Vanden-Eijnden. Exact dynamical coarse-graining without time-scale separation. *The Journal of Chemical Physics*, 141 (4), 2014.

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. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998.

Gia G Maisuradze, Adam Liwo, and Harold a Scheraga. Principal component analysis for protein folding dynamics. *Journal of molecular biology*, 385(1):312–29, 2009.

Chr. Malsburg. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14(2):85–100, 1973.

Shawn Martin, Aidan Thompson, Evangelos a Coutsias, and Jean-Paul Watson. Topology of cyclo-octane energy landscape. *The Journal of Chemical Physics*, 132(23):234115, 2010.

J McCleary. *Geometry from a Differentiable Viewpoint*. Cambridge University Press, second edition, 2012. ISBN 9781139788472.

Servaas Michielssens, Titus S van Erp, Carsten Kutzner, Arnout Ceulemans, and Bert L de Groot. Molecular dynamics in principal component space. *The Journal of Physical Chemistry B*, 116(29): 8350–4, 2012.

Daniel Millán and Marino Arroyo. Nonlinear manifold learning for model reduction in finite elastodynamics. *Computer Methods in Applied Mechanics and Engineering*, 261-262:118–131, 2013.

Daniel Millán, Adrian Rosolen, and Marino Arroyo. Thin shell analysis from scattered points with maximum-entropy approximants. *International Journal for Numerical Methods in Engineering*, 85(6): 723–751, 2011.

Daniel Millán, Adrian Rosolen, and Marino Arroyo. Nonlinear manifold learning for meshfree finite deformation thin-shell analysis. *International Journal for Numerical Methods in Engineering*, 93(7): 685–713, 2013.

Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer Science, New York, second edition, 2006.

H Noji, R Yasuda, M Yoshida, and K Kinosita. Direct observation of the rotation of F1-ATPase. *Nature*, 386(6622):299–302, 1997.

Roberto Olender and Ron Elber. Calculation of classical trajectories with a very large time step: Formalism and numerical examples. *The Journal of Chemical Physics*, 105(20):9299, 1996.

Margarita Osadchy and Rachel Kolodny. Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America*, 108(30):12301–6, 2011.

T Papadopoulo and M I A Lourakis. Estimating the Jacobian of the Singular Value Decomposition: Theory and Applications. In *Computer Vision - ECCV 2000*, volume 1842, pages 554–570. Springer, 2000. ISBN 978-3-540-45054-2.

Daniele Passerone, Matteo Ceccarelli, and Michele Parrinello. A concerted variational strategy for investigating rare events. *The Journal of Chemical Physics*, 118(5):2025, 2003.

Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, 1901.

F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, Matthieu Brucher, M Perrot, and E Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, 26(16):1781–802, 2005.

Stefano Piana and Alessandro Laio. Advillin Folding Takes Place on a Hypersurface of Small Dimensionality. *Physical Review Letters*, 101 (20):208101, 2008.

Fabio Pietrucci and Alessandro Laio. A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures: Application to SH3 and GB1. *Journal of Chemical Theory and Computation*, 5 (9):2197–2201, 2009.

Erion Plaku, Hernan Stamati, Cecilia Clementi, and Lydia E Kavraki. Fast and reliable analysis of molecular motion using proximity relations and dimensionality reduction. *Proteins: Structure, Function, and Bioinformatics*, 907:897–907, 2007.

Weiqing Ren, Eric Vanden-Eijnden, Paul Maragakis, and Weinan E. Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide. *The Journal of Chemical Physics*, 123(13):134109, 2005.

Mary a Rohrdanz, Wenwei Zheng, Mauro Maggioni, and Cecilia Clementi. Determination of reaction coordinates via locally scaled diffusion map. *The Journal of Chemical Physics*, 134(12):124116, 2011.

Mary a Rohrdanz, Wenwei Zheng, and Cecilia Clementi. Discovering mountain passes via torchlight: methods for the definition of reaction coordinates and pathways in complex macromolecular reactions. *Annual review of physical chemistry*, 64(December 2012):295–316, 2013.

Adrian Rosolen, Daniel Millán, and Marino Arroyo. On the optimum support size in meshfree methods: A variational adaptivity approach with maximum-entropy approximants. *International Journal for Numerical Methods in Engineering*, 82(7):868–895, 2010.

Edina Rosta, H. Lee Woodcock, Bernard R. Brooks, and Gerhard Hummer. Artificial reaction coordinate "tunneling" in free-energy calculations: The catalytic reaction of RNase H. *Journal of Computational Chemistry*, 30(11):1634–1641, 2009.

S T Roweis and L K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–6, 2000.

Steven J. Ruuth and Barry Merriman. A simple embedding method for solving partial differential equations on surfaces. *Journal of Computational Physics*, 227(3):1943–1961, 2008.

Matteo Salvalaglio, Pratyush Tiwary, and Michele Parrinello. Assessing the Reliability of the Dynamics Reconstructed from Metadynamics. *Journal of Chemical Theory and Computation*, 10(4):1420–1425, 2014.

J.W. Sammon. A Nonlinear Mapping for Data Structure Analysis. *IEEE Transactions on Computers*, C-18(5):401–409, 1969.

Yosi Shibberu and Allen Holder. A spectral approach to protein structure alignment. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 8(4):867–75, 2011.

Yosi Shibberu, Allen Holder, and Kyla Lutz. Fast Protein Structure Alignment. *Structure*, (i):152–165, 2010.

Marcos Sotomayor and Klaus Schulten. Single-molecule experiments in vitro and in silico. *Science*, 316(5828):1144–8, 2007.

Vojtech Spiwok, Petra Lipovová, and Blanka Králová. Metadynamics in essential coordinates: free energy simulation of conformational changes. *The Journal of Physical Chemistry B*, 111(12):3073–6, 2007.

Vojtech Spiwok, Blanka Králová, and Igor Tvaroska. Continuous metadynamics in essential coordinates as a tool for free energy modelling of conformational changes. *Journal of molecular modeling*, 14(11):995–1002, 2008.

Vojtech Spiwok, Blanka Králová, and Igor Tvaroska. Modelling of beta-D-glucopyranose ring distortion in different force fields: a metadynamics study. *Carbohydrate research*, 345(4):530–7, 2010.

Vojtěch Spiwok and Blanka Králová. Metadynamics in the conformational space nonlinearly dimensionally reduced by Isomap. *The Journal of Chemical Physics*, 135(22):224504, 2011.

Vojtěch Spiwok, Pavel Oborský, Jana Pazúriková, Aleš Křenek, and Blanka Králová. Nonlinear vs. linear biasing in Trp-cage folding simulations. *The Journal of Chemical Physics*, 142(11):115101, 2015.

Hernán Stamati, Cecilia Clementi, and Lydia E Kavraki. Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. *Proteins*, 78(2):223–35, 2010.

Ludovico Sutto, Simone Marsili, and Francesco Luigi Gervasio. New advances in metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(5):771–779, 2012.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science (New York, N.Y.)*, 290(5500):2319–23, 2000.

Miguel L Teodoro, George N Phillips, and Lydia E Kavraki. Understanding protein flexibility through dimensionality reduction. *Journal of computational biology : a journal of computational molecular cell biology*, 10(3-4):617–34, 2003.

G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

Gareth a Tribello, Michele Ceriotti, and Michele Parrinello. A self-learning algorithm for biased molecular dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 107(41):17509–14, 2010.

Gareth A Tribello, Michele Ceriotti, and Michele Parrinello. Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 109(14):5196–201, 2012.

Gareth a. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, 2014.

David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. C. Berendsen. GROMACS: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–18, 2005.

Titus S. van Erp, Daniele Moroni, and Peter G. Bolhuis. A novel path sampling method for the calculation of rate constants. *The Journal of Chemical Physics*, 118(17):7762, 2003.

Michele Vendruscolo, Rafael Najmanovich, and Eytan Domany. Can a Pairwise Contact Potential Stabilize Native Protein. *Proteins*, 148 (August 1999):134–148, 2000.

Arthur Voter. Hyperdynamics: Accelerated Molecular Dynamics of Infrequent Events. *Physical Review Letters*, 78(20):3908–3911, 1997.

Holger Wendland. *Scattered data approximation*. Cambridge University Press, Cambridge, 2005. ISBN 9780521843355.

Hassler Whitney. Differentiable manifolds. *Annals of Mathematics*, 37 (3):645–680, 1936.

Yuzhen Xue, Peter J. Ludovice, Martha a. Grover, Lilia V. Nedialkova, Carmeline J. Dsilva, and Ioannis G. Kevrekidis. State reduction in molecular simulations. *Computers & Chemical Engineering*, 51: 102–110, 2013.