Chemical Physics Letters 509 (2011) 1–11

Contents lists available at ScienceDirect

Chemical Physics Letters

journal homepage: www.elsevier.com/locate/cplett

FRONTIERS ARTICLE

Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach

Andrew L. Ferguson [a,1], Athanassios Z. Panagiotopoulos [a], Ioannis G. Kevrekidis [a,b], Pablo G. Debenedetti [a,*]

[a] Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544, USA
[b] Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA

ARTICLE INFO

Article history:
Available online 23 April 2011

ABSTRACT

Molecular simulation is an important and ubiquitous tool in the study of microscopic phenomena in fields as diverse as materials science, protein folding and drug design. While the atomic-level resolution provides unparalleled detail, it can be non-trivial to extract the important motions underlying simulations of complex systems containing many degrees of freedom. The *diffusion map* is a nonlinear dimensionality reduction technique with the capacity to systematically extract the essential dynamical modes of high-dimensional simulation trajectories, furnishing a kinetically meaningful low-dimensional framework with which to develop insight and understanding of the underlying dynamics and thermodynamics. We survey the potential of this approach in the field of molecular simulation, consider its challenges, and discuss its underlying concepts and means of application. We provide examples drawn from our own work on the hydrophobic collapse mechanism of *n*-alkane chains, folding pathways of an antimicrobial peptide, and the dynamics of a driven interface.

1. Introduction

The development of efficient, scalable and highly parallelized simulation algorithms together with ever increasing computational power, bolstered in recent years by the advent of the multicore era [1], has given rise to molecular simulations spanning time and length scales unthinkable merely a decade ago [2,3]. This has permitted the exploration of previously inaccessible phenomena, such as millisecond conformational rearrangements of proteins [2], or the observation of DNA translocation through a transmembrane pore [4]. Atomistically detailed classical molecular dynamics simulations furnish the positions and velocities of, and forces upon, every atom in the system over the course of nano to millisecond time periods, providing a resolution unattainable by experimental approaches, and access to length and time scales far beyond those accessible to quantum mechanical treatments.

Attendant to the exploration of large molecular systems over long time scales, is a vast increase in both the length and dimensionality of the associated simulation trajectories, further exacerbating the perennial issue of how to systematically extract the important dynamical motions and distinct conformational states underlying voluminous simulation data sets [5]. The existence of low-dimensional effective descriptions is supported, for example, by studies of proteins demonstrating the important dynamics to be confined within a handful of collective motions [6–10], permitting the conformational space explored by a 22-residue β-hairpin [11] and a 10-residue polyalanine chain [12] to be parametrized by as few as three effective degrees of freedom.

Despite the availability of unprecedented computational power, many phenomena remain beyond the length and time scales attainable with atomistically detailed simulation techniques, for instance the folding/unfolding transitions of large proteins or the assembly of multimeric enzymes into their quaternary structure. One means to reduce the computational cost of simulating large molecules and collective assembly processes is to coarse grain the molecules into lower resolution abstractions in a manner preserving their salient features, while greatly reducing the number of degrees of freedom [13]. An alternative approach exploits the Mori–Zwanzig projection operator approach to formulate a low-dimensional generalized Langevin equation in a small number of variables describing the slow dynamical modes of the system [14–16]. The 'right' variables in which to construct the low-dimensional parametrization, however, may be intuited for only the simplest systems, with dimensionality reduction of short, atomistically detailed simulations providing a means to systematically determine these variables and properly parametrize the associated Langevin equation [16–18].

The validity of a low-dimensional representation of a molecular system is founded on the assumption that the system dynamics may be well-modeled by a *diffusion process*, whereby its dynamical

* Corresponding author. Address: Department of Chemical and Biological Engineering, Princeton University, A-419 Engineering Quad, Princeton, NJ 08544, USA. Fax: +1 609 258 0211.
E-mail address: pdebene@princeton.edu (P.G. Debenedetti).
[1] Present address: Department of Chemical Engineering, MIT, Cambridge, MA 02139, USA.

0009-2614/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cplett.2011.04.066

evolution is governed by small number of slow modes to which the remaining fast degrees of freedom couple as effective stochastic noise [14,18,19]. The order parameters associated with these slow modes are 'good' in the sense that the dynamical evolution according to the stochastic differential equations written in these variables is, on sufficiently long timescales, *Markovian* – the evolution of the system depends only on the instantaneous state and not on its previous history – and *closed* – all information required to propagate the system in time is contained within this set of variables [15].

Geometrically, the existence of a low-dimensional description assumes that the molecular system exhibits a low-dimensional manifold structure [20], in the sense that the full-dimensional phase space accessible to the process is only sparsely populated. The values – or more generally, the statistics – of the fast degrees of freedom are slaved to the evolution of the slow variables, possibly due to the presence of steep free energy gradients, thereby effectively restraining the dynamical evolution of the system to a low-dimensional hypersurface which we term the *intrinsic manifold* [15,19].

The existence of such low-dimensional descriptions has been demonstrated for many complex biophysical systems, and may be considered to arise from cooperative couplings between molecular degrees of freedom leading to a separation of time scales between fundamental and slaved dynamical modes [6–8,10–12,19,21,22]. The development of tools with which to systematically extract these fundamental underlying modes is of critical importance in gaining insight and understanding of the process, parametrization of dynamically meaningful free energy surfaces, and robust classification of the metastable and stable conformational states.

The goals of this letter are to place the diffusion map in the context of the spectrum of dimensionality reduction methods, to discuss recent examples of its useful application to the interpretation of molecular simulation data, and to outline what we consider to be promising future applications of this technique in the broad area of molecular simulation.

The organization of this letter is as follows. In Section 2 we briefly survey popular contemporary dimensionality reduction methodologies. Section 3 describes the diffusion map approach in some detail, considering its application, interpretation, advantages, limitations and computational aspects of its deployment. In Section 4 we briefly survey three case studies from our own research in which diffusion maps have proved a powerful and effective tool in the analysis of molecular simulation data. Finally, in Section 5 we present our view of the current and future role of the diffusion map in the field of molecular simulation, mention some particular applications for which we believe it holds great promise, and outline current challenges and directions for the technique.

2. Dimensionality reduction techniques

In the application of dimensionality reduction methodologies to molecular simulation trajectories, one typically seeks to construct a low-dimensional description of the ensemble of configurations explored by the system over the course of the simulation. Unexplored regions of phase space may be inaccessible due to physical constraints such as excluded volume overlaps or the existence of high free energy barriers. Methods to achieve adequate sampling of the thermally accessible phase space for systems exhibiting high free energy barriers is an important area of continuing research, but one which is distinct from the goals of dimensionality reduction.

Since the intrinsic manifold is typically defined as a hypersurface in *configurational space*, its reconstruction does not consider particle velocities. Accordingly, while the dimensionality reduction techniques discussed below are often applied to molecular dynamics (MD) simulation trajectories, they are equally applicable to Monte-Carlo (MC) simulations where particle velocities are not defined. Conceptually, one may consider the simulation algorithm simply as a means to sample the thermally accessible regions of configurational space, and dimensionality reduction as a means to synthesize a low-dimensional parametrization of this space [19].

Approaches to dimensionality reduction may be broadly classified as linear or nonlinear. In the case of the former, the reduced dimensional representations generated are restricted to linear combinations of the input variables. Geometrically, this is equivalent to assuming that the purported low-dimensional manifold structure of the data in the original high-dimensional space may be well-approximated by a hyperplane. Conversely, nonlinear techniques admit low-dimensional descriptions formed by arbitrary nonlinear functions of the input variables, rendering such techniques more appropriate for systems whose dynamics lie on complex, possibly highly curved and convoluted, low-dimensional intrinsic manifolds.

Principal component analysis (PCA) [23] – also known as the Karhunen–Loève transform (KLT) – is the prototypical linear dimensionality reduction technique, which has found widespread applications in fields ranging from stock portfolio optimization [24] to analysis of evolutionary modules in three-dimensional protein structure [25]. The technique was introduced in the analysis of molecular simulation trajectories by Karplus and coworkers (under the name 'quasi-harmonic analysis') [26] and García [6], as a means to obtaining a set of orthogonal vectors spanning the 'essential subspace' [8] capturing the largest amplitude dynamical motions contained within the trajectory. A particularly attractive feature of PCA is that the linear transformation mapping the input atomic coordinates into the low-dimensional essential subspace is explicit, permitting unambiguous physical interpretation of the collective dynamical motions associated with the order parameters defining the low-dimensional essential subspace.

In some instances, linear techniques may adequately capture complex dynamics by embedding inherently nonlinear intrinsic manifolds in higher dimensional hyperplanes [11,21,27], but in other cases such approaches have been shown to fail more severely, providing inadequate preservation of local structures [28] and separation of stable conformational states [10,11]. In that they do not assume the *a priori* validity of a hyperplane approximation, nonlinear techniques are expected to be more robust, parsimonious and globally valid than linear approaches employing the same number of variables [10,11,29].

Variants of PCA have been developed to lift the linear restriction, including kernel PCA [30] and what was described as 'nonlinear PCA' [31,32]. The former assumes the availability of an appropriate nonlinear transformation with which to 'pre-treat' the data, whereas the latter employs neural networks to uncover appropriate nonlinear transformations. Full correlation analysis (FCA) [27] is a relatively new technique which seeks to capture nonlinear and multivariable correlations among input variables by minimizing the Shannon mutual information.

The majority of recent advances in nonlinear dimensionality reduction have, however, focused on the development of *manifold learning* techniques, which infer the global geometry of the intrinsic manifold by integrating local information about its structure into a coherent global description [33]. The physical interpretation of the low-dimensional representations furnished by these techniques is more challenging than for linear approaches, since the explicit (nonlinear) functions relating the input and output variables are typically unavailable. This is an important issue to which we will return below.

Several manifold learning algorithms have emerged in recent years, examples of which include local tangent space alignment (LTSA) [34], local linear embedding (LLE) and its variants [35,36], Isomap and its variants [10,37,38], semidefinite embedding (SDE)/maximum variance unfolding (MVU) [39], Laplacian eigenmaps [40], Hessian eigenmaps [41], and diffusion maps [28,42]. These methodologies share many common features [20,42], but under some assumptions to be discussed below, it has been shown that the low-dimensional order parameters furnished by diffusion maps may be interpreted as descriptors of the underlying dynamical motions contained within the data set [33]. This feature is particularly attractive in the analysis of molecular simulation trajectories, permitting the synthesis of dynamically meaningful low-dimensional representations, and facilitating the inference of transition mechanisms between conformational states.

3. The diffusion map

3.1. Introduction

Diffusion maps were introduced by Coifman and coworkers in 2005 as a tool for the multiscale analysis of high-dimensional data sets [33]. The technique has been rapidly adopted by many researchers, finding diverse applications in many fields, such as harmonic analysis, graph partitioning, and tomographic image reconstruction [19,22,28,43–47].

In recent years, we have applied and developed the diffusion map technique to synthesize low-dimensional *embeddings* of high-dimensional molecular simulation trajectories [19,21,22,29]. As we shall see, the embeddings are constructed by the diffusion map in such a manner that configurational microstates which are kinetically close – in the sense that they are connected by a large number of short pathways – are placed nearby one another, whereas states which are only connected by a relatively small number of long pathways – perhaps due to the presence of a dynamical bottleneck – are placed far apart [33]. Furthermore, under some modest assumptions, the order parameters spanning the low-dimensional embedding characterize the important dynamical motions underlying the temporal evolution of the molecular system [33].

To make these ideas concrete, and illuminate the means by which diffusion maps embody these features, we consider the application of the approach to a hypothetical simulation trajectory of a molecular system comprising P atoms. The instantaneous (classical mechanical) configurational microstate of the system may be specified by a $3P$-dimensional state vector recording, for example, the Cartesian coordinates of the constituent atoms. Correspondingly, a simulation trajectory comprising N frames or snapshots may be represented as an N-by-$3P$ observation matrix tracking the instantaneous state of the system as it moves through the $3P$-dimensional configurational space. In the diffusion map approach, the ordering of the snapshots (rows) and observation/input variables (columns) is immaterial, and snapshots need not be collected at uniform time intervals.

The particular simulation algorithm employed may simply be regarded as a means to populate the accessible configurational space [19], where the precise region explored, and therefore the low-dimensional parametrization extracted, depends on the thermodynamic ensemble in which the simulations are conducted, and the sampling efficiency of the algorithm. In the following outline of the technique, which broadly follows that presented in Refs. [19,21], we shall consider the hypothetical system to consist of explicitly modeled atoms, but the technique is also extensible to higher-level models consisting of united atoms, coarse-grained interaction sites or multi-molecular aggregates. Correspondingly, the approach is also applicable to simulation trajectories generated by algorithms beyond conventional MD and MC, such as coarse-grained Brownian dynamics, dissipative particle dynamics, kinetic Monte-Carlo [22] (c.f. Section 4.3) or even systems of coupled deterministic or stochastic differential equations [45].

3.2. Calculation of similarity distances

The initial step in the application of the diffusion map is the calculation of similarity distances $d(i,j)$ between all pairs of $3P$-dimensional snapshots $i,j = 1 \ldots N$ in the observation matrix. For the low-dimensional order parameters synthesized by the diffusion map to be good descriptors of the important dynamical motions of the high-dimensional molecular system, these similarity distances should be a good measure of the ease with the system may evolve from the microstate defined by one snapshot, to that corresponding to another.

Ideally one would like to employ a dynamic measure of interstate transition rates, but for more than a small number of snapshots this would require a prohibitively expensive calculation by a computationally intensive technique such as transition interface sampling [48]. Instead, one typically employs a structural metric capturing the short time diffusive motions of the system from one configurational microstate to another. It is from this structural proxy for the short-time dynamic proximity of the constituent simulation snapshots that the diffusion map synthesizes a global description of the important underlying modes contained in the data.

For simulations of a single solute, either in isolation, or in implicit or explicit solvent, the translationally and rotationally minimized root mean squared deviation (rmsd) between the solute atom coordinates is a natural choice for $d(i,j)$ that is expected to capture the thermal fluctuations driving the short time diffusive molecular motions [19,29]. Depending on the particular system, alternative metrics based on, for example, the earth mover's or Hamming distances may be appropriate [22]. Prior experimental or computational knowledge about the conformational dynamics may warrant the use of more elaborate metrics. For example, a study of enzyme conformational dynamics or substrate docking may employ heavier weightings on the subset of solute atoms defining the active site or binding pocket.

The reorganization of solvent molecules around a solute, or more generally of one molecule relative to another, renders the definition of meaningful distance measure explicitly accounting for multi-molecular atomic coordinates, a non-trivial matter. Nevertheless, metrics based on coarse-grained abstractions of the system to a grid have shown promise in the application of the diffusion map to collective phenomena [22,29]. We have also demonstrated that solvent effects may sufficiently strongly influence the ensemble of solute conformations sampled by the simulation to be implicitly captured by a purely solute-centric distance metric [19].

3.3. Soft thresholding

Having defined $d(i,j)$ for all snapshot pairs, the pairwise distances are then soft-thresholded by a Gaussian kernel of bandwidth ϵ, and stored in a matrix $\mathbf{A}$ with elements,

$$A_{ij} = exp\left(-\frac{d(i,j)^2}{2\epsilon}\right) \quad i,j = 1 \ldots N. \tag{1}$$

The soft-thresholding operation has the effect of retaining only short pairwise distances on the order of $\sqrt{\epsilon}$, providing a description of the local connectivity of each point with its neighbors on the surface of the intrinsic manifold. Large distances, which are not

expected to meaningfully characterize the manifold structure, are discarded.

Interpreting the fractal dimension of the data points in the $3P$-dimensional configurational space as a measure of effective dimensionality [49], Coifman et al. proposed twice the slope of the linear regime in a plot of $\log\left(\sum_{ij}A_{ij}\right)$ against $\log(\epsilon)$ as an estimate of the dimensionality of the intrinsic manifold, with appropriate values of ϵ delimited by the extent of this linear region [43].

Since the diffusion map proceeds by integrating local structural information into a unified global reconstruction of the intrinsic manifold, the trajectory must be *well-connected*, in the sense that each snapshot be accessible from any other by a series of 'small hops' of $O(\sqrt{\epsilon})$ in the distance metric [21]. In the case of disconnected data – as may arise, for example, in replica exchange molecular dynamics simulations (REMD) [21], or due to poor sampling of high free energy barrier regions [29] – the diffusion map may not generate a single useful global approximation to the underlying manifold, but rather synthesize distinct embeddings of each disconnected region. By filling in the gaps in the sampled conformational ensemble, perhaps through the use of biased sampling techniques [29], or by discarding the disconnected regions in the data [21], a meaningful low-dimensional embedding may be recovered.

3.4. The diffusion map embedding

The rows of the **A** matrix are then normalized to yield the **M** matrix, a right-stochastic Markov transition matrix with elements,

$$M_{ij} = \frac{A_{ij}}{\sum_{j=1}^{N} A_{ij}} \quad i,j = 1 \ldots N. \tag{2}$$

By virtue of its Markovian nature, the top eigenvalue/eigenvector pair of **M** is trivial, with $\psi_1 = 1$ and $\vec{\Psi}_1 = \vec{1}$, where $\vec{1}$ is the all-ones vector. The *diffusion map embedding* is defined as the mapping of the ith snapshot into the ith components of each of the top k non-trivial eigenvectors of the **M** matrix,

$$snapshot_i \mapsto \left(\vec{\Psi}_2(i), \vec{\Psi}_3(i), \ldots \vec{\Psi}_{k+1}(i)\right). \tag{3}$$

The low-dimensional embedding defined by the mapping is a reconstruction of the intrinsic manifold underlying the molecular system, data-mined directly from the simulation trajectory. We present a simple illustration of an application of the approach to the canonical 'Swiss roll' data set in Figure 1. (We observe that a more generalized time dependent version of the diffusion map embedding may be defined, as discussed in Ref. [15].)

3.5. Properties of the embedding

If the dynamical system is well-approximated as a diffusion process, and the pairwise similarity metric is a good measure of the short time microscopic diffusive motions, then the diffusion map embedding possesses two attractive attributes. Firstly, Euclidean distances in the diffusion map embedding incorporating all $(N-1)$ eigenvectors correspond to *diffusion distances* in the full-dimensional configurational space, where the latter measures the ease with which the system can dynamically transition between two microstates [33,50]. States connected by a large number of short paths possess small diffusion distances, whereas those linked by only a few long routes have large values of this measure. For systems in which the top k eigenvectors are significant (see the following section), Euclidean distances in embeddings in the top $k < (N-1)$ eigenvectors provide good approximations of the diffusion distance [50].

Secondly, in the limit of $N \to \infty$ and $\epsilon \to 0$ the eigenvectors $\{\vec{\Psi}_i\}_{i=2}^{N}$ converge to the eigenfunctions of an effective Fokker–Planck operator [51] (modulo a factor of 2 in the associated potential), describing a diffusion process over the low-dimensional free energy surface explored by the simulation trajectory [19,33,43]. In other words, the top eigenvectors of **M** are discrete approximations to the top eigenfunctions of the spectral solution of the diffusion process, and therefore describe the slowest diffusive modes and dictate the long-time dynamics of the system [33,50].

We note that correspondence to other continuous-space operators may be obtained by performing alternative normalizations of the **A** matrix in Eq. (2) [33]. The normalization corresponding to the Laplace–Beltrami operator explicitly compensates for non-uniform distribution of data points over the intrinsic manifold, effectively separating the geometry of the manifold from the topography of its free-energy surface [43].

3.6. Estimation of the intrinsic dimensionality

The effective dimensionality of the system, and therefore the specification of an appropriate number of eigenvectors to incorporate into the diffusion map embedding (Eq. (3)), may be inferred by three means: the presence of a spectral gap in the eigenvalue spectrum of the **M** matrix [44], the aforementioned fractal measure of the intrinsic manifold dimensionality [43], or the *plateau dimension* of the diffusion map embedding [49]. Regarding the diffusion map embedding as a mapping of the simulation snapshots onto the surface of the intrinsic manifold, the dimensionality of this (possibly highly convoluted) surface may be estimated by computing the (fractal) dimensionality of the mapping defined by Eq. (3). We perform this evaluation by computing the correlation dimension [49] of mappings incorporating successively more eigenvectors, and determine the dimensionality at which this measure levels out, remaining unchanged with the inclusion of additional eigenvectors. Following Grassberger and Procaccia [49], we define this measure as the plateau dimension, and take it as an estimation of the effective dimensionality of the intrinsic manifold [19].

In some instances, the observation of a spectral gap may be obscured by (the statistics of) the components of certain higher order eigenvectors being slaved to, and therefore uniquely specified by, the components of certain other lower order eigenvectors. Despite the mutual orthogonality of the eigenvectors of **M**, such functional dependencies may arise from distinct eigenvectors characterizing the same low-dimensional dynamical mode. We have previously drawn an analogy with multivariate Fourier series, in which $cos(x)$ and $cos(2x)$ correspond to the same direction in space, but are nonetheless orthogonal Fourier components [19]. The existence of such a functional dependency is manifest in collapse of the data onto an effectively one-dimensional curve in scatter plots of the components of one eigenvector against another (c.f. caption to Figure 2). In the absence of a clear gap in the eigenvalue spectrum, the detection of such dependencies may facilitate the identification of an *effective spectral gap* [44].

Care must be taken in interpreting eigenvalue spectra possessing a series of gaps, where global conformational modes may not be entirely contained within the eigenvectors prior to the first gap [44]. Spectra with no gaps suggests that dimensionality reduction will fail, since the system is not well-approximated by a low-dimensional diffusion process, and that no low-dimensional manifold underlies the distribution of data points in high-dimensional phase space [44].

3.7. Free energy surfaces and transition mechanisms

The free energy surface (FES) supported by the intrinsic manifold may be computed by collecting a multidimensional histogram approximation to the probability distribution of points in the top k

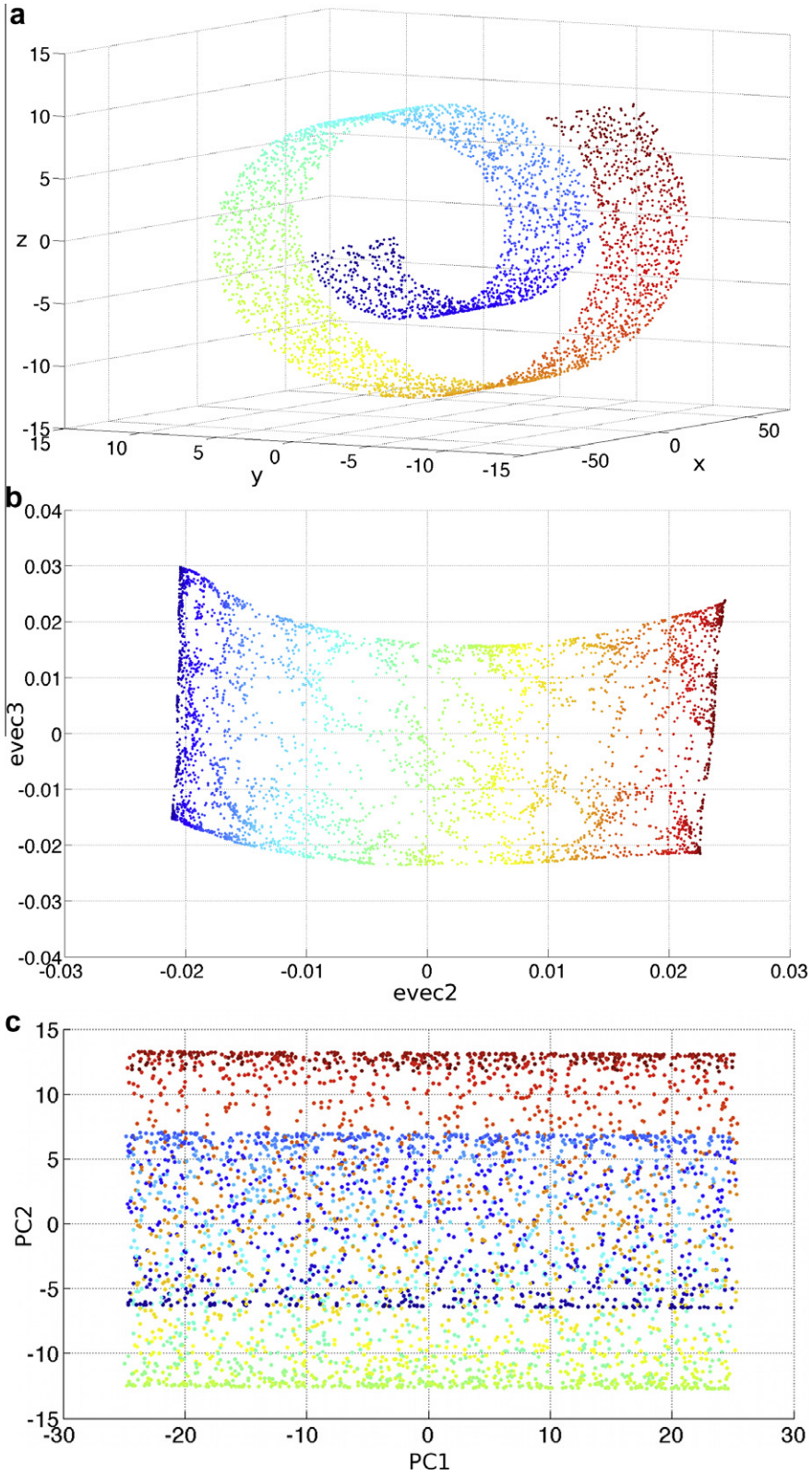

Figure 1. Application of diffusion maps and PCA to the 'Swiss roll' data set in which data reside on a two-dimensional surface in three-dimensional space. (a) In the context of molecular simulation, the (x, y, z) coordinate triplets may be considered to describe the Cartesian coordinates of a single point particle, and the two-dimensional manifold to represent a surface to which it is effectively restrained by a nonlinear coupling between its degrees of freedom. The intrinsic dimensionality of the system is therefore one less than that of the ambient space in which it dynamically evolves. Points are colored according to their geodesic distances along the spiral. (b) Application of the diffusion map approach and embedding of the data into the top two non-trivial eigenvectors synthesizes a meaningful reconstruction of the underlying intrinsic manifold of the system. Diffusion maps provide a means to extract the underlying (nonlinear) two-dimensional intrinsic manifold from an set of observations of the particle position. (c) The two-dimensional embedding of the data into the top two principal components determined by application of PCA to the data. This embedding does not adequately parametrize the underlying manifold, instead synthesizing an effective projection of the Swiss roll in (a) into the x,y-plane.

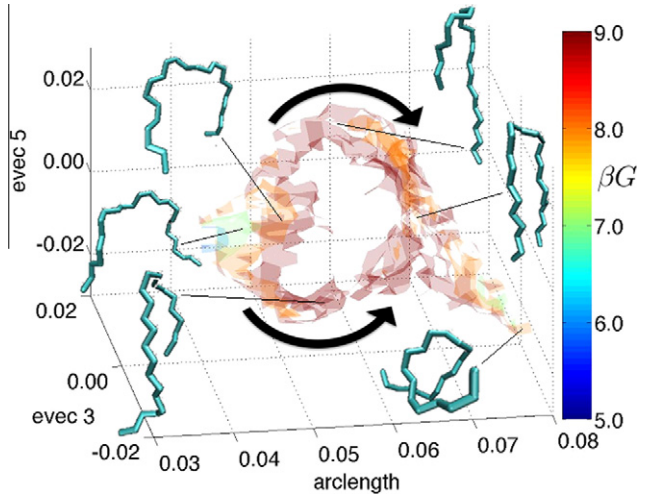

Figure 2. Free energy surface constructed over the intrinsic manifold of a solvated $C_{24}H_{50}$ *n*-alkane chain. *G* is the Gibbs free energy and $\beta = k_B T$, where k_B is Boltzmann's constant and *T* is temperature. In this instance we observed redundancy between *evec2* and *evec4*, revealed by data collapse onto a one-dimensional curve in a scatter plot of the components of one eigenvector against the other. Correspondingly, the [*evec2*, *evec4*] embedding coordinates were instead defined by the distance of the point along this one-dimensional curve, termed the *arclength*. Arrows indicate the low-free energy collapse pathways from extended to collapsed chain conformations that initiate by the development of an asymmetric kink towards the head or tail of the molecule. The dynamic interpretability of diffusion map embeddings permitted inference of the mechanism of collapse by superposition of representative molecular conformations from the simulation trajectory onto the low-free energy pathways. Figure adapted from Ref. [19].

eigenvectors, $\widehat{P}\left(\{\vec{\Psi}\}_{i=2}^{k+1}\right)$, and inverting this distribution to obtain the FES,

$$\beta F = -\ln \widehat{P}\left(\{\vec{\Psi}\}_{i=2}^{k+1}\right) + C, \tag{4}$$

where $\beta = 1/k_B T$, k_B is Boltzmann's constant, *T* is the temperature, *F* is the free energy potential corresponding to the ensemble in which the simulations were performed, and *C* is an arbitrary additive constant. An example of a three-dimensional FES constructed over the intrinsic manifold extracted from simulations of a $C_{24}H_{50}$ *n*-alkane chain is presented in Figure 2.

Since the leading eigenvectors correspond to the slow underlying dynamical motions, the low free energy pathways linking local free energy minima on the surface of the intrinsic manifold may be used to infer conformational transition mechanisms. Du et al. [52] define the *reaction coordinate* as the precise low-free energy path along which the system evolves from one conformational state to another, with only small oscillations in the degrees of freedom orthogonal to the reaction path. In contrast, the *transition coordinate* is defined as the motion between (meta) stable conformational states that exhibits the largest relaxation time, and may involve large fluctuations in degrees of freedom orthogonal to the path. Since the top eigenvectors of the diffusion map correspond to the slowest dynamical motions of the system, the low free energy paths in diffusion map embeddings may be associated with transition coordinates, providing a good parametrization of the evolution of the system between its (meta) stable states. The conformational changes associated with motions of the system along transition pathways may be used to infer transition mechanisms, with the fine details of the transition identifiable by superposition of representative system configurations onto the free energy surface (c.f. Figures 2 and 3) [19].

The dynamic interpretability of diffusion map embeddings permits mechanistic information to be extracted by data mining equilibrium simulation trajectories. The validity of this approach requires that the simulation trajectories adequately sample the stable and metastable states of interest, and the pathways by which they are connected. For systems exhibiting high free energy barriers, unbiased simulation trajectories may only surmount free energy barriers on the order of $\sim k_B T$, and are not expected to adequately sample barrier regions between states. To ameliorate this difficulty, we have recently developed a variant of the diffusion map approach appropriate for biased simulation data which we term the *umbrella adapted diffusion map approach* [29]. By reformulating the diffusion map eigenvalue problem, we have shown that the diffusion map order parameters of the *unbiased* system may be obtained by appropriate reweighting and analysis of *biased* simulation data, facilitating the efficient synthesis of low-dimensional embeddings for systems possessing rugged free energy landscapes.

Given the important distinction between *reaction* and *transition* coordinates, we emphasize that while the diffusion map may be used to efficiently infer transition mechanisms, approaches such as transition path sampling [53], forward flux sampling [54] or geometry optimization [55] must be employed to extract precise reaction coordinates, rates and committor probabilities. We note, however, the potential synergy of nonlinear dimensionality reduction and path sampling, where the former may be used to robustly identify initial and final conformational states, order parameters through which they may be defined and transition pathways along which path sampling should be conducted [19,29].

3.8. Eigenvector interpretation

In contrast with linear dimensionality reduction techniques such as PCA, where the linear transformation from the ambient to low-dimensional space is explicitly available, the primary (in our view) weakness of the diffusion map approach is the unavailability of the explicit nonlinear mapping between the input observables – the atomic coordinates – and the diffusion map order parameters – the components of the eigenvectors of the **M** matrix. This deficiency is characteristic of nonlinear approaches, where the physical interpretation of the eigenvectors may only be inferred by correlation with candidate physical variables, and the superposition of representative system configurations onto the low-dimensional embedding (c.f. Figures 2 and 3) [19,21].

More systematic means to ascertain this physical correspondence may employ techniques for high-throughput screening of putative physical variables such as those suggested by Ma and Dinner [56] and Peters et al. [57,58]. The E-Isomap adaptation of the Isomap technique by Li et al. [37,59] expresses the coordinates of each ambient space data point as a sum of localized, nonlinear basis functions, and then computes an explicit expression for the low-dimensional mapping by solving a linear regression problem between the ambient and low-dimensional space [59]. Nevertheless, the explicit expression for the mapping strongly depends on the choice of basis functions, and remains difficult to interpret physically.

In the 'equation free' approach pioneered by Kevrekidis and coworkers, complex dynamical systems comprising many degrees of freedom are cast as low-dimensional effective equations (here, generalized Langevin equations) in a small number of variables governing the long-time evolution of the system [60]. Parameters in the resulting Langevin equations are then determined by performing appropriately initialized short bursts of detailed simulation in the full-dimensional space [16,17,61]. The ability of the diffusion map to systematically identify the important modes underlying a dynamical system suggested a means to systematically identify 'good' variables in which to construct the low-dimensional descriptions, and was incorporated into the methodology in the 'variable-free/equation-free' approach [45]. The absence of an explicit physical interpretation of the diffusion map eigenvectors

has typically resulted in the formulation of the low-dimensional Langevin equations in proxy physical variables [45], but interpolative approaches have been developed to permit the development of descriptions directly in the diffusion map variables themselves (i.e. the eigenvectors of the **M** matrix) [62].

Closely related to the absence of an explicit mapping is the issue of *out of sample extension*, or situating a new data point within an existing low-dimensional N-point embedding without explicitly computing the diffusion mapping of the augmented $(N+1)$-point system. The Nyström extension provides a means to incorporate new points into the intrinsic manifold defined by the top eigenvectors of the N-point system, but performs poorly for points located further than a distance of $\sqrt{\epsilon}$ from the manifold, and is therefore not practicable for arbitrarily located points [22,63].

3.9. Computational issues

For large **M** matrices, the eigenvectors corresponding to the few leading eigenvalues may be efficiently computed by power iteration, using, for example, the Implicitly Restarted Arnoldi Method [64,65] implemented as serial and parallel FORTRAN routines in the ARPACK [65] and PARPACK libraries [66], and underlying the MATLAB 'eigs' function [67].

Subsequent to the computation of the pairwise distances between snapshots, the computation of the diffusion map embedding depends only on the number of trajectory snapshots N, possessing no dependence on the size and dimensionality of the molecular system. Matrix storage scales as N^2, and the algorithmic complexity of the eigenvector computation between N and N^2 depending on the matrix structure [65,68]. Due to the high efficiency of the ARPACK routines, the maximum number of snapshots in a trajectory is typically set by RAM limitations rather than execution time. For example, a modestly sized 30 000 snapshot trajectory requires over 3 GB of RAM to be held in working memory during the eigenvector computation, whereas 500 000 snapshots require almost 1 TB.

Shared memory systems are ideally suited to large matrix applications, although the use of the parallelized ARPACK routines through an MPI interface permits the use of distributed memory architectures. For very large matrices, the time required to read the large matrix from file into memory can dominate that of the eigenvector computation, although this deficiency may be effectively addressed by the use of libraries and hardware supporting parallel I/O.

As an example of the timings involved in a typical application of the diffusion map approach, we consider a 36 786 snapshot molecular dynamics trajectory of the 22-atom alanine dipeptide in explicit water [29]. Specifying the pairwise distances between snapshots as the rotationally and translationally minimized rmsd between peptide conformations, the construction of the 36 786-by-36 786 **M** matrix required 27 h on a single core 2.66 GHz Intel Xeon processor. The eigenvector calculation was conducted over three nodes each containing a 2.77 GHz Intel Core 2 Quad processor, requiring $\sim$100 min to serially load the matrix into memory, and $\sim$20 min to compute the top 25 eigenvectors using the PARPACK libraries.

4. State of the field

A primary research focus in recent years has been the application and adaptation of the diffusion map approach in the analysis of molecular simulation data. In this section we briefly survey some results which have emerged from our efforts in this area, in an attempt to illustrate the broad applicability and power of diffusion maps in developing deeper insight and understanding of molecular phenomena. For more details of these studies, we refer the readers to the original publications.

4.1. Hydrophobic collapse of n-alkane chains

N-alkanes are prototypical hydrophobic polymers, the study of which has implications for the understanding of the role of hydrophobicity in the dynamics and thermodynamics of proteins and peptides [69]. We constructed dynamically meaningful embeddings of molecular simulation trajectories of *n*-alkane chains in explicit water to determine the mechanism of hydrophobic collapse [19]. Previous work by Chandler and coworkers explored the collapse of idealized hydrophobic polymers [70,71], but to our knowledge this is the first prediction of the collapse pathway where realistic models were used for both the hydrocarbons and water. This work employed the rotationally and translationally minimized rmsd between the atomic coordinates of the chains as a similarity measure between snapshot pairs. Despite ostensibly discarding all solvent degrees of freedom, the influence of the solvent was shown to be sufficiently strongly 'encoded' in the chain configurations sampled that its effects were manifested in the resulting diffusion map embedding. Full details of this study are provided in Ref. [19].

We determined the intrinsic dimensionality of a $C_{24}H_{50}$ chain in water to be approximately three, and Figure 2 presents the associated effective free energy surface constructed over its three-dimensional intrinsic manifold. By correlating the diffusion map order parameters with candidate physical variables, we identified a correspondence of the three diffusion map order parameters to the degree of chain collapse, the location of a kink in the chain, and the handedness of the helicity of the chain. Superposing representative chain conformations onto the free energy surfaces revealed the (dynamically meaningful) low-free energy pathway for chain collapse to proceed by the development of a kink towards the head or tail of the chain, which slides toward the center of the chain to form a tight symmetric hairpin, followed by subsequent collapse into a globular helical conformation. The symmetric collapse pathway corresponding to motions directly through the center of the 'doughnut' shown in Figure 2 is disfavored by the high free energy cost associated with the collective expulsion of multiple confined solvent molecules [72].

4.2. Spontaneous lasso formation in an antimicrobial peptide

Microcin J25 (MccJ25) is a 21-residue antimicrobial peptide, the native state of which is an intriguing 'lassoed' β-hairpin, in which the N-terminus wraps around the C-terminus in a counter-clockwise manner, covalently sealing it inside an 8-residue ring [73] (Figure 3a). The structural rigidity of this fold imparts great resistance to thermal and chemical denaturation, making this a motif of interest in protein engineering and design [74,75]. However, the maturation mechanism of MccJ25, and the precise functions of its attendant maturation enzymes, remain poorly understood [21].

To provide an inferential understanding of MccJ25 biosynthesis, we applied diffusion maps to long replica exchange molecular dynamics simulations of the 21-residue linear peptide in isolation to determine the extent to which, and pathways by which, the native lasso structure was spontaneously approached in the absence of the maturation machinery [21]. Full details of this work are provided in Ref. [21].

Using the rotationally and translationally minimized rmsd between peptide atomic coordinates as a pairwise similarity measure, we extracted the three-dimensional intrinsic manifold presented in Figure 3b. The embedding exhibits a triply-branched structure corresponding to three distinct folding pathways from the global free energy minimum located in the vicinity of structure *g*. The lower route corresponds to global hydrophobic collapse of the peptide chain (structure *e*), the upper route to the formation

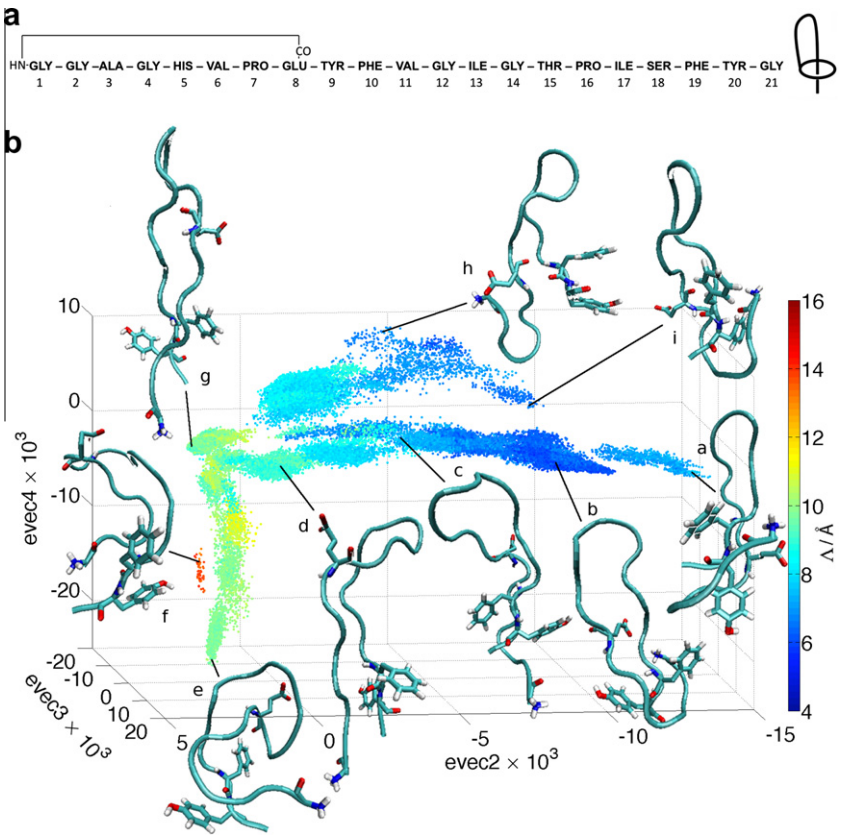

Figure 3. Application of diffusion maps to microcin J25 (MccJ25). (a) Primary sequence of MccJ25 showing the location of the isopeptide bond between Gly_1 and Glu_8 residues forming the 8-residue ring. The cartoon to the right of the sequence illustrates the three-dimensional native fold with the C-terminal strand threaded through the N-terminal ring. (b) The three-dimensional diffusion map embedding resulting from a 95 ns REMD simulation of the 21-residue antimicrobial 'lasso' peptide microcin J25 (MccJ25) in explicit water. The central pathway in the triply-branched intrinsic manifold corresponds to the spontaneous evolution of the global free energy minimum structure, *G*, to a non-native lasso conformation, *b*, in which the N-terminal Gly_1 threads between the Phe_{19} and Tyr_{20} encircling the C-terminus in a clockwise manner. In contrast, the native state possesses a counter-clockwise topology. The data points in the embedding are colored according to a parameter Λ which is a measure of the identity of the residue in the turn position of the β-hairpin. Λ values of $\sim$4 indicate the native Ile_{13} residue in the turn position, while higher values are associated with ratcheting of the β-hairpin away from its native conformation. For visual clarity only the Gly_1, Glu_8, Phe_{19} and Tyr_{20} residues are explicitly represented upon the peptide backbone, and solvent molecules have been omitted. See text for a discussion of the two other branches of the intrinsic manifold. Figure adapted from Ref. [21].

of an improperly wrapped lasso conformation (structure *i*) and the central path to the spontaneous adoption of a non-native lasso conformation with the N-terminus encircling the C-terminus in the opposite direction to that in the native state (structure *b*). Despite initializing the simulations in the vicinity of the native lasso conformation, no pathways to a native lasso topology were observed, suggesting a possible role for the maturation machinery in enforcing the correct topology of the lasso motif.

The data points in Figure 3b are colored according to a parameter Λ, which is a proxy measure for the identity of the residue in the β-turn position. Small ($\sim$4) values of Λ correspond to the native Ile_{13} in this position, while larger values correspond to a shift in the location of the β-turn. (See Ref. [21] for details.) Only by observing a good correlation between *evec2* and Λ, and inspection of the representative peptide structure projected onto the embedding were we able to assign physical meaning to this eigenvector. Similarly, *evec4* was observed to show good correlation with the Ψ dihedral angle of the Glu_8 residue. Our inability to correlate *evec3* with any candidate physical variable provides a pointed illustration of the principal (in our view) current limitation of the diffusion map approach.

4.3. Description of the dynamics of a driven interface

To study the dynamics of, for example, the motion of polycrystalline grain boundaries, we [22] developed a simplified two-dimensional lattice Ising model describing the motion of a domain wall separating bulk regions of up and down spins, and driven by the influence of an external magnetic field. Mobile impurities are modeled as interstitial particles that are attracted to the domain wall. A typical system snapshot is presented in Figure 4a. The system was evolved using a kinetic Monte-Carlo algorithm, and system snapshots saved for analysis through the diffusion map approach. Full details of the model are presented in Refs. [22,76].

The definition of a similarity metric characterizing the short time diffusive motions of the system was less apparent in this work than for the molecular systems described above. We developed a metric capturing the shape and local impurity concentration on the wall by 'smearing out' the concentration of each impurity over its neighboring lattice sites and computing the minimal difference in the domain wall impurity profiles in each snapshot pair subject to periodic realignment. We observed a spectral gap after the second non-trivial eigenvalue, suggesting an intrinsic dimensionality of two and informing our construction of the two-dimensional diffusion map embeddings in Figure 4b and c. For full details of the application of the diffusion map approach, we refer the reader to Ref. [22].

In Figure 4b the data points are colored according to the domain wall roughness, while in Figure 4c we have colored them by the number of impurities located on the domain wall. Visual inspection suggests that the antisymmetric combination of *evec2* and

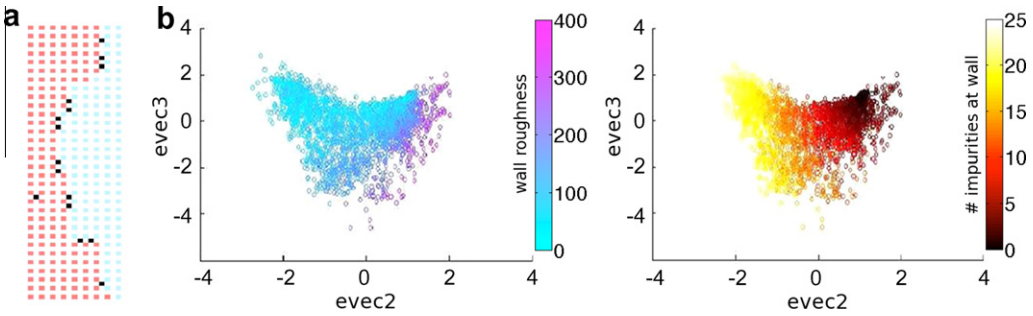

Figure 4. Analysis of the dynamical evolution of a driven domain wall between bulk regions of up and down spins in a two-dimensional lattice Ising model with mobile impurities. An external magnetic field drives the wall to the right. (a) A typical system snapshot where red squares denote lattice sites containing up spins, blue contain down spins and black sites represent interstitial mobile impurities which are attracted to the domain wall. The lattice is vertically periodic, and horizontally infinite. (b and c) Two-dimensional diffusion map embeddings of kinetic Monte-Carlo simulations of the dynamical evolution of the driven interface colored according to (b) domain wall roughness and (c) number of impurities located at the wall. For the precise definition of wall roughness, see Ref. [22]. Visual inspection suggests the antisymmetric, resp. symmetric, combination of *evec2* and *evec3* to be well-correlated with wall roughness, resp. wall impurity concentration. Together with a non-singular Jacobian of the transformation between the leading eigenvectors and features of the physical domain, this strongly suggests the dynamics of the domain wall to be governed primarily by the roughness and its local impurity concentration. Figure adapted from Ref. [22]. Copyright (2009) by the American Physical Society.

evec3 is well-correlated with wall roughness, while their symmetric combination shows good correlation with the impurity concentration at the wall. The Jacobian of the transformation between the two eigenvectors and physical observables does not become singular over the manifold, further intimating the existence of a bijection between the leading eigenvectors and features of the physical domain. These results suggest the dynamics of the domain wall to be well-approximated by an effective two-dimensional description formulated in the degree of wall roughness and the impurity concentration at the interface.

5. Conclusions and outlook

The diffusion map is a powerful nonlinear dimensionality reduction technique, which is rapidly finding broad applications in the systematic synthesis of dynamically meaningful low-dimensional representations of molecular simulation data. The approach provides an efficient and inexpensive means to gain insight into the important dynamical motions underlying molecular phenomena. In contrast to linear dimensionality reduction techniques, the diffusion map is not restricted to the construction of low-dimensional embeddings from exclusively linear combinations of the observed input variables. This attractive feature permits the development of parametrizations expected to be globally valid over the configurational space sampled by the simulation, and renders the technique more appropriate for complex molecular systems expected to possess highly nonlinear intrinsic manifolds. Furthermore, nonlinear embeddings are expected to be lower-dimensional and therefore more parsimonious than embeddings synthesized by linear methodologies [11].

Whereas linear dimensionality reduction techniques require the specification of a state vector characterizing the system configuration at each simulation snapshot, the diffusion map requires only scalar pairwise distances between snapshot pairs. This is potentially useful where a natural basis with which to describe the system is unavailable, but where differences between system configurations are well-defined. For example, consider a protein-threading problem in which Monte-Carlo simulations applying residue point mutations are employed to find low-energy primary sequences for a specified three-dimensional structure [77]. Dimensionality reduction may be of use to develop a coarse-grained description of the primary sequence in terms of blocks of closely associated residues [25]. In this instance, no natural ordering of the 20 amino acids exists, but Hamming distances defining the fraction of conserved residues between sequence pairs would provide a natural measure of sequence similarity.

Under the dual assumptions that the molecular system is well-described by a small number of slowly evolving dynamical modes to which the remaining degrees of freedom are effectively slaved, and that the similarity measure between snapshot pairs is a good measure of the short time diffusive motions, the order parameters furnished by the diffusion map (i.e. eigenvectors of the **M** matrix) are good descriptors of the slow dynamical motions of the molecular system. The dynamical interpretability this imparts is a particularly attractive feature of the technique that permits mechanistic information to be inferred from the low-dimensional diffusion map embeddings. However, whereas path-based methods seek to extract the *reaction coordinate* linking two conformational states, the diffusion map seeks to identify one or more *transition coordinates*; the former describes the precise pathway followed by the system, and the latter only the slowest evolving dynamical modes of the system [52].

The (umbrella adapted) diffusion map approach represents an efficient means to systematically identify good descriptors of the important dynamical modes underlying (biased) molecular simulations that provide good sampling of the thermally accessible phase space. Nevertheless, only path-based methodologies have the means to extract the fine details of a conformational transition, validate reaction coordinates by the computation of committor probabilities, and track the precise course of the reaction tube through phase space [78]. In this regard, diffusion maps may be a useful precursor to path sampling approaches, facilitating the robust identification of the reactant and product basins and the determination of transition paths along which sampling should proceed [19,29,79].

In our view, the principal limitation of the diffusion map approach is the absence of an explicit mapping between the observed/input variables and the low-dimensional diffusion map order parameters. Approximate mappings between the ambient space and low-dimensional embeddings may be developed by regressing the low-dimensional embeddings upon the original high-dimensional coordinates [59], or training of artificial neural networks [80], but the interpretation of such expressions typically remains opaque. Currently, the only means to assign physical meaning to diffusion map order parameters is to correlate them with combinations of candidate physical variables, a process which may be efficiently accelerated by high-throughput analysis methods such as those developed by Peters et al. [57,81] and Ma and Dinner [56]. The development of more systematic means to assign

physical meaning to diffusion map order parameters is a current research focus that we, and others, are pursuing.

We envisage a strong future for diffusion maps in the analysis of molecular simulation data, providing an inexpensive and efficient means to develop low-dimensional descriptions where linear methodologies, such as principal component analysis, are inadequate or inappropriate. One particularly interesting direction is the use of the diffusion map to systematically identify a small number of variables in which to construct low-dimensional dynamical descriptions of the system, which, once correctly parametrized, may permit access to time and length scales orders of magnitude larger than those attainable by conventional molecular simulation [22,45]. Another potentially intriguing application we believe warrants further investigation is the synergy between the diffusion map and transition path sampling. Finally, we see tremendous potential for diffusion maps in the analysis of collective phenomena such as self-assembly, where their inherently multibody nature arguably renders the heuristic or intuitive determination of appropriate order parameters less transparent than single-molecule processes. We believe the systematic identification of the important variables underlying such phenomena to be of great importance in the understanding and control of important biological processes such as viral capsid assembly [82] and cellular pore formation [83], and in facilitating the rational design of organic and inorganic building blocks with which to robustly self-assemble materials with designed properties [84].

Jargon Box

Configurational space
the typically very high-dimensional space recording the instantaneous value of each degree of freedom contributing to a system's potential energy

Configurational microstate
the state of a microscopic system defined by the instantaneous value of each degree of freedom that contributes to a system's potential energy; associated with a unique point in *configurational space*

Simulation trajectory
a particular realization of the dynamical evolution of the system specified as a succession of *configurational microstates*

Diffusion distance
a measure of the ease with which the dynamical system may evolve from one *configurational microstate* to another; microstates connected by a large number of short pathways are linked by small diffusion distances, whereas those connected by few, long pathways are separated by large diffusion distances

Diffusion process
a set of coupled stochastic differential equations describing the dynamical evolution of a number of, in this context, slow variables under the action of a gradient field and coupled to Gaussian white noise, in this context, as an implicit representation of the interaction with the remaining fast degrees of freedom in the system

Intrinsic manifold
a low-dimensional hypersurface in the full-dimensional configurational space to which the evolution of a dynamical system is effectively restrained

Low-dimensional embedding
a reduced dimensional description of the region of full-dimensional *configurational space* sampled by a *simulation trajectory*

(Meta)stable conformational state
a collection of nearby *configurational microstates* in the full-dimensional *configurational space* which reside in a local free energy basin

Reaction coordinate
an order parameter tracing the minimum free-energy pathway along which the system evolves from one *(meta) stable conformational state* to another

Transition coordinate
an order parameter describing the slowest component of the dynamical evolution of a system between two *(meta) stable conformational states*; since the reaction coordinate is generally composed of multiple dynamical modes, the reaction and transition coordinates are typically not coincident

Acknowledgements

A.Z.P. acknowledges financial support from the Department of Energy, Office of Basic Energy Sciences (Grant No. DE-SC-0002128) and the Princeton Center for Complex Materials (Grant No. DMR-0819860). The work of I.G.K. was partially supported by the Department of Energy (Grant No. DE-SC0002097). P.G.D. gratefully acknowledges support from the National Science Foundation (Collaborative Research in Chemistry Grant No. CHE-0908265).

References

[1] J.E. Stone, D.J. Hardy, I.S. Ufimtsev, K. Schulten, J. Mol. Graphics Modell. 29 (2010) 116.
[2] D.E. Shaw et al., Science 330 (2010) 341.
[3] P.L. Freddolino, F. Liu, M. Gruebele, K. Schulten, Biophys. J. 94 (2008) L75.
[4] J. Mathé, A. Aksimentiev, D.R. Nelson, K. Schulten, A. Meller, Proc. Natl. Acad. Sci. USA 102 (2005) 12377.
[5] W.M. Brown, S. Martin, S.N. Pollock, E.A. Coutsias, J.P. Watson, J. Chem. Phys. 129 (2008) 064118.
[6] A.E. García, Phys. Rev. Lett. 68 (1992) 2696.
[7] P.I. Zhuravlev, C.K. Materese, G.A. Papoian, J. Phys. Chem. B 113 (2009) 8800.
[8] A. Amadei, A.B.M. Linssen, H.J.C. Berendsen, Proteins Struct. Funct. Genet. 17 (1993) 412.
[9] O.F. Lange, H. Grubmüller, J. Phys. Chem. B 110 (2006) 22842.
[10] P. Das, M. Moll, H. Stamati, L.E. Kavraki, C. Clementi, Proc. Natl. Acad. Sci. USA 103 (2006) 9885.
[11] H. Stamati, C. Clementi, L.E. Kavraki, Proteins Struct. Funct. Genet. 78 (2009) 223.
[12] R. Hegger, A. Altis, P.H. Nguyen, G. Stock, Phys. Rev. Lett. 98 (2007) 028102.
[13] G. Voth, Coarse-Graining of Condensed Phase and Biomolecular Systems, CRC, 2008.
[14] R. Zwanzig, Nonequilibrium Statistical Mechanics, Oxford University Press, New York, USA, 2001.
[15] R.R. Coifman, I.G. Kevrekidis, S. Lafon, M. Maggioni, B. Nadler, Multiscale Model. Simul. 7 (2008) 842.
[16] S. Yang, J.N. Onuchic, A.E. García, H. Levine, J. Mol. Biol. 372 (2007) 756.
[17] D.I. Kopelevich, A.Z. Panagiotopoulos, I.G. Kevrekidis, J. Chem. Phys. 122 (2005) 044908.
[18] G. Hummer, I.G. Kevrekidis, J. Chem. Phys. 118 (2003) 10762.
[19] A.L. Ferguson, A.Z. Panagiotopoulos, P.G. Debenedetti, I.G. Kevrekidis, Proc. Natl. Acad. Sci. USA 107 (2010) 13597.
[20] L. Cayton, Algorithms for manifold learning, Technical Report CS2008-0923, Department of Computer Science, University of California at San Diego, San Diego, CA, 2005. Available from: <http://people.kyb.tuebingen.mpg.de/lcayton/resexam.pdf> (accessed November 3, 2010).
[21] A.L. Ferguson, S. Zhang, I. Dikiy, A.Z. Panagiotopoulos, P.G. Debenedetti, A.J. Link, Biophys. J. 99 (2010) 3056.
[22] B.E. Sonday, M. Haataja, I.G. Kevrekidis, Phys. Rev. E 80 (2009) 031102.
[23] I.T. Jolliffe, Principal Component Analysis, second edn., Springer, New York, 2002.

[24] Y. Fujiwara, W. Souma, H. Murasato, H. Yoon, in: H. Takayasu (Ed.), Practical Fruits of Econophysics, Springer, Tokyo, 2006, p. 226.
[25] N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, Cell 138 (2009) 774.
[26] T. Ichiye, M. Karplus, Proteins Struct. Funct. Genet. 11 (1991) 205.
[27] O.F. Lange, H. Grubmüller, Proteins Struct. Funct. Bioinform. 70 (2008) 1294.
[28] R.R. Coifman, S. Lafon, Appl. Comput. Harmon. Anal. 21 (2006) 5.
[29] A.L. Ferguson, A.Z. Panagiotopoulos, P.G. Debenedetti, I.G. Kevrekidis, J. Chem. Phys. 134 (2011) 135103.
[30] B. Schölkopf, A. Smola, K. Müller, in: W. Gerstner, A. Germond, M. Hasler, J. Nicoud (Eds.), International Conference on Artificial Neural Networks—ICANN. Lecture Notes in Computer Science, vol. 1327, Springer, Berlin, 1997, p. 583.
[31] M.A. Kramer, AIChE J. 37 (1991) 233.
[32] P.H. Nguyen, Proteins Struct. Funct. Bioinform. 65 (2006) 898.
[33] R.R. Coifman et al., Proc. Natl. Acad. Sci. USA 102 (2005) 7426.
[34] Z. Zhang, H. Zha, J. Shanghai Univ. (English Edn.) 8 (2004) 406.
[35] Z. Zhang, J. Wang, Adv. Neural Inform. Process. Syst. 19 (2007) 1593.
[36] S.T. Roweis, L.K. Saul, Science 290 (2000) 2323.
[37] J.B. Tenenbaum, V. de Silva, J.C. Langford, Science 290 (2000) 2319.
[38] V. De Silva, J. Tenenbaum, Adv. Neural Inform. Process. Syst. (2003) 721.
[39] K. Weinberger, L. Saul, Int. J. Comput. Vision 70 (2006) 77.
[40] M. Belkin, P. Niyogi, Adv. Neural Inform. Process. Syst. 1 (2002) 585.
[41] D. Donoho, C. Grimes, Proc. Natl. Acad. Sci. USA 100 (2003) 5591.
[42] M. Belkin, P. Niyogi, Neural Comput. 15 (2003) 1373.
[43] R.R. Coifman, Y. Shkolnisky, F.J. Sigworth, A. Singer, IEEE Trans. Image Process. 17 (2008) 1891.
[44] R.R. Coifman, I.G. Kevrekidis, S. Lafon, M. Maggioni, B. Nadler, Multiscale Model. Simul. 7 (2008) 842.
[45] R. Erban et al., J. Chem. Phys. 126 (2007) 155103.
[46] S. Lafon, A.B. Lee, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1393.
[47] A. Singer, R. Erban, I.G. Kevrekidis, R.R. Coifman, Proc. Natl. Acad. Sci. USA 106 (2009) 16090.
[48] D. Moroni, T.S. Van Erp, P.G. Bolhuis, Phys. A Stat. Mech. Appl. 340 (2004) 395.
[49] P. Grassberger, I. Procaccia, Phys. D Nonlinear Phenom. 9 (1983) 189.
[50] B. Nadler, S. Lafon, R.R. Coifman, I.G. Kevrekidis, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), Advances in Neural Information Processing Systems 18, Neural Information Processing Systems (NIPS), MIT Press, Cambridge, MA, 2006, p. 955.
[51] H. Risken, The Fokker–Planck Equation: Methods of Solution and Applications, second edn., Springer-Verlag, Berlin, Heidelberg, 1989.
[52] R. Du, V.S. Pande, A.Y. Grosberg, T. Tanaka, E.S. Shakhnovich, J. Chem. Phys. 108 (1998) 334.
[53] P.G. Bolhuis, D. Chandler, C. Dellago, P.L. Geissler, Ann. Rev. Phys. Chem. 53 (2002) 291.
[54] R.J. Allen, D. Frenkel, P.R. ten Wolde, J. Chem. Phys. 124 (2006) 194111.
[55] H.B. Schlegel, J. Comput. Chem. 24 (2003) 1514.
[56] A. Ma, A.R. Dinner, J. Phys. Chem. B 109 (2005) 6769.
[57] B. Peters, B.L. Trout, J. Chem. Phys. 125 (2006) 054108.
[58] B. Peters, G.T. Beckham, B.L. Trout, J. Chem. Phys. 127 (2007) 034109.
[59] C.-G. Li, J. Guo, G. Chen, X.-F. Nie, Z. Yang, in: Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, pp. 3201.
[60] I.G. Kevrekidis, C.W. Gear, J.M. Hyman, P.G. Kevrekidis, O. Runborg, C. Theodoropoulos, Commun. Math. Sci. 1 (2003) 715.
[61] R. Erban, I.G. Kevrekidis, D. Adalsteinsson, T.C. Elston, J. Chem. Phys. 124 (2006) 084106.
[62] C.R. Laing, T.A. Frewen, I.G. Kevrekidis, Nonlinearity 20 (2007) 2127.
[63] Y. Bengio, J.F. Paiement, P. Vincent, in: S. Thrun, L. Saul, B. Schölkopf (Eds.), Advances in Neural Information Processing Systems 16, MIT Press, Cambridge, MA, 2004, p. 177.
[64] W.E. Arnoldi, Quart. Appl. Math. 9 (1951) 17.
[65] R.B. Lehoucq, D.C. Sorensen, C. Yang, ARPACK Users' Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods, Soc. Ind. Appl. Math. (SIAM), Philadelphia, 1998.
[66] K.J. Maschhoff, D.C. Sorensen, in: J. Wasniewski, J. Dongarra, K. Madsen, D. Olesen (Eds.), Third International Workshop on Applied Parallel Computing, Industrial Computation and Optimization, vol. 1184 of Lecture Notes in Computer Science, Springer, New York, 1996, pp. 478.
[67] The MathWorks Inc., MATLAB 7: Mathematics, Natick, MA, 2010.
[68] M. Sadkane, Numer. Math. 64 (1993) 181.
[69] A.L. Ferguson, P.G. Debenedetti, A.Z. Panagiotopoulos, J. Phys. Chem. B 113 (2009) 6405.
[70] P.R. ten Wolde, D. Chandler, Proc. Natl. Acad. Sci. USA 99 (2002) 6539.
[71] T.F. Miller, E. Vanden-Eijnden, D. Chandler, Proc. Natl. Acad. Sci. USA 104 (2007) 14559.
[72] K. Lum, D. Chandler, J.D. Weeks, J. Phys. Chem. B 103 (1999) 4570.
[73] K. Wilson et al., J. Am. Chem. Soc. 125 (2003) 12475.
[74] R.A. Salomón, R.N. Farías, J. Bacteriol. 174 (1992) 7428.
[75] S. Rebuffat, A. Blond, D. Destoumieux-Garzón, C. Goulard, J. Peduzzi, Curr. Protein Pept. Sci. 5 (2004) 383.
[76] M. Haataja, D. Srolovitz, I. Kevrekidis, Phys. Rev. Lett. 92 (2004) 160603.
[77] M.S. Shell, P.G. Debenedetti, A.Z. Panagiotopoulos, Proteins Struct. Funct. Bioinform. 62 (2006) 232.
[78] P.G. Bolhuis, C. Dellago, D. Chandler, Proc. Natl. Acad. Sci. USA 97 (2000) 5877.
[79] J. Juraszek, P.G. Bolhuis, Proc. Natl. Acad. Sci. USA 103 (2006) 15859.
[80] C. Bishop, Rev. Sci. Instrum. 65 (1994) 1803.
[81] B. Peters, G.T. Beckham, B.L. Trout, J. Chem. Phys. 127 (2007) 034109.
[82] M.F. Hagan, D. Chandler, Biophys. J. 91 (2006) 42.
[83] G. Illya, M. Deserno, Biophys. J. 95 (2008) 4163.
[84] S.C. Glotzer, Science 306 (2004) 419.

Andrew L. Ferguson received an M.Eng. in Chemical Engineering from Imperial College London in 2005, and a Ph.D. in Chemical and Biological Engineering from Princeton University in 2010. He is currently a post-doctoral associate and Ragon Fellow in the Laboratory for Computational Immunology in the Department of Chemical Engineering at MIT. His research interests lie in the development of linear and nonlinear dimensionality reduction techniques and their applications to molecular simulation and bioinformatics data sets.

Athanassios Z. Panagiotopoulos received an undergraduate degree from the National Technical University of Athens in 1982, and a Ph.D. from MIT in 1986, both in Chemical Engineering. He was a postdoctoral fellow in Physical Chemistry at the University of Oxford, a faculty member at Cornell and the University of Maryland, and is currently the Susan Dod Brown Professor of Chemical and Biological Engineering at Princeton University. He is the recipient of the AIChE Colburn Award and the Prausnitz Award, and was elected to the U.S. National Academy of Engineering in 2004. He is the author of 'Essential Thermodynamics' (2011).

Ioannis G. Kevrekidis received a Dipl. Eng. in Chemical Engineering from the National Technical University of Athens in 1982 and an MA in Mathematics and a Ph.D. in Chemical Engineering from the University of Minnesota in 1986. A Director's Postdoctoral Fellow at Los Alamos National Laboratory for a year, he is currently the Pomeroy and Betty Perry Smith Professor of Engineering at Princeton University, and a member of the Program in Applied and Computational Mathematics. He has received the AIChE Colburn and Wilhelm Awards, a senior Humboldt prize and the SIAM/DS Crawford prize; in 2010 he became a SIAM Fellow.

Pablo G. Debenedetti received his BS degree from the University of Buenos Aires, Argentina, in 1978, and his Ph.D. from MIT in 1985. He joined Princeton University in 1985, and is currently the Class of 1950 Professor of Engineering and Applied Science, Professor of Chemical and Biological Engineering, and Vice Dean of the School of Engineering and Applied Science. The author of 'Metastable Liquids' (1996), he is the recipient of the AIChE Professional Progress and Walker awards, and the ACS Hildebrand Award. He is a member of the National Academy of Engineering and the American Academy of Arts and Sciences.