Nonlinear reconstruction of single-molecule free-energy surfaces from univariate time series

Jiang Wang[1] and Andrew L. Ferguson[2,3,*]

[1]*Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*

[2]*Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*

[3]*Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*

(Received 10 May 2015; published 21 March 2016)

The stable conformations and dynamical fluctuations of polymers and macromolecules are governed by the underlying single-molecule free energy surface. By integrating ideas from dynamical systems theory with nonlinear manifold learning, we have recovered single-molecule free energy surfaces from univariate time series in a single coarse-grained system observable. Using Takens' Delay Embedding Theorem, we expand the univariate time series into a high dimensional space in which the dynamics are equivalent to those of the molecular motions in real space. We then apply the diffusion map nonlinear manifold learning algorithm to extract a low-dimensional representation of the free energy surface that is diffeomorphic to that computed from a complete knowledge of all system degrees of freedom. We validate our approach in molecular dynamics simulations of a $C_{24}H_{50}$ n-alkane chain to demonstrate that the two-dimensional free energy surface extracted from the atomistic simulation trajectory is – subject to spatial and temporal symmetries – geometrically and topologically equivalent to that recovered from a knowledge of only the head-to-tail distance of the chain. Our approach lays the foundations to extract empirical single-molecule free energy surfaces directly from experimental measurements.

DOI: 10.1103/PhysRevE.93.032412

I. INTRODUCTION

Free-energy surfaces present a powerful tool to describe the stable states and dynamical pathways of molecules, which have been profitably employed to describe the microscopic behavior of polymers, peptides, and proteins [1–4]. The configuration of a molecule comprising N atoms can be described by a $3N$-dimensional vector of Cartesian coordinates. Interactions between the atoms in the molecule, mediated, for example, by covalent bonds, electrostatic interactions, and dispersion forces, introduce cooperative couplings between the atomic degrees of freedom that render the effective dimensionality of the molecule, m, far lower than the $3N$-dimensional atomic coordinate space [3]. In a temporal sense, a molecular system admits a low-dimensional description if—on sufficiently long time scales—its dynamical evolution is governed by a small number of collective modes to which the remaining degrees of freedom are effectively slaved [5–7]. In a geometric sense, the trajectory of the system through the $3N$-dimensional phase space is effectively restrained to an *intrinsic manifold* of much lower dimensionality [3,6]. The existence and validity of such low-dimensional descriptions has been demonstrated for many macro- and biomolecules [3,4,8–14]. For example, the effective dimensionality of the 22-atom alanine dipeptide [15] and a coarse-grained model of the 57-residue src homology 3 domain [4] have been shown to be approximately two, and that of a $C_{24}H_{50}$ n-alkane chain to be approximately three [3].

Projection of molecular configurations into this reduced dimensional space requires a mapping, $g : \mathbb{R}^{3N} \rightarrow \mathbb{R}^m$, specifying m collective variables, $\vec{\psi} = [\psi_1, \psi_2, \ldots, \psi_m]$, formed from the $3N$ degrees of freedom. For all but the simplest molecules, this mapping is expected to be highly nonlinear and unavailable from analytical theory. In recent years, a number of nonlinear machine learning approaches have been employed to systematically infer such mappings by discovering low-dimensional manifolds within high-dimensional molecular simulation trajectories [3,4,6,8–10,12].

Under the ergodic hypothesis, the distribution of molecular configurations in sufficiently long simulation trajectories is expected to follow the Boltzmann distribution. The single-molecule free-energy surface (smFES) as a hypersurface in $\mathbb{R}^{m+1}$, $F(\vec{\psi})$, can be estimated from the observed probability distribution of snapshots projected onto the manifold, $\hat{P}(\vec{\psi})$, using the statistical mechanical relationship, $F(\vec{\psi}) = -k_B T \ln \hat{P}(\vec{\psi}) + C$, where k_B is Boltzmann's constant, T is the temperature, and C is an arbitrary constant [e.g., Fig. 1(a)]. The smFES is of great value in revealing metastable and stable configurational states, dynamical pathways connecting these states, and quantitatively linking molecular chemistry to thermodynamic and dynamical behavior [3,4,6,14,16,17]. For example, free-energy surfaces were recently employed to understand the mechanism by which the anticancer drug daunomycin intercalates into B-DNA [18], and to identify a new structural intermediate in the activation pathway of c-src tyrosine kinase as a potential target for novel anticancer therapeutics [19]. With the advent of highly parallel simulation packages, powerful computer hardware, and robust dimensionality reduction algorithms, the recovery of single-molecule free-energy surfaces from molecular simulations is now routine [4,6,19], but simulations are restricted to microsecond time scales and rely upon classical force fields that are approximations to the true underlying quantum mechanical interactions. The free-energy surfaces recovered from computational studies are therefore approximations to the true smFES. It would represent a significant advance in single-molecule physics if single-molecule free-energy surfaces could instead be directly recovered from experimental data.

*Corresponding author: alf@illinois.edu

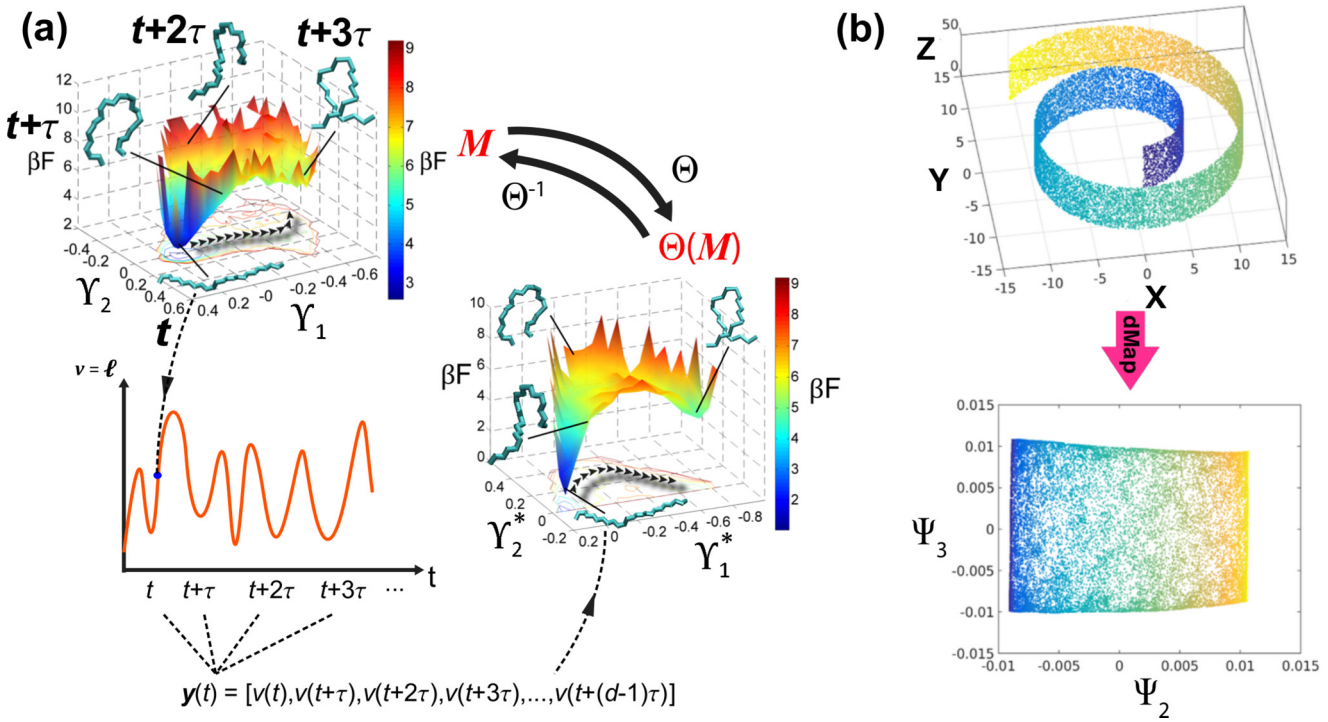

FIG. 1. Schematic overview of the single-molecule free-energy surface (smFES) reconstruction methodology. (a) Top left: The dynamical evolution of the molecular system proceeds over a low-dimensional manifold M supporting the smFES. The dynamics of the n-tetracosane polymer chain in water considered in this work are contained in a two-dimensional manifold parametrized by the collective variables $[\Upsilon_1, \Upsilon_2]$ that are nonlinear combinations of the molecular degrees of freedom. The smFES maps out the Gibbs free energy of the chain F dedimensionalized by the reciprocal temperature $\beta = 1/k_B T$ as a function of these order parameters. Computing M requires access to the atomic coordinates of the molecule that are typically only available from molecular simulations. Bottom left: Measurements of an experimentally accessible observable $v(t)$ furnish a scalar time series providing a coarse-grained characterization of the single-molecule dynamics. In this work, we consider the head-to-tail distance, ℓ, as a quantity measurable by FRET [21]. Assembling d successive measurements separated by a delay time τ produces an n-dimensional delay vector $\vec{y}(t) = [v(t), v(t+\tau), v(t+2\tau), v(t+3\tau), \ldots, v(t+(d-1)\tau)]$. By computing delay vectors over the entire time series, the scalar time series is projected into an n-dimensional delay space. Bottom right: Under quite general conditions on τ, d, and the observable $v(t)$, Takens' theorem [29–34] asserts that the manifold $\Theta(M)$ containing the delay vectors $\vec{y}(t)$ is a diffeomorphism to the manifold M containing the real space molecular dynamics, and the variables $[\Upsilon_1^*, \Upsilon_2^*]$ parametrizing $\Theta(M)$ are related by a smooth and invertible transformation Θ to those parametrizing M. Using this approach, topologically and geometrically identical reconstructions of single-molecule free-energy surfaces can be determined directly from experimental measurements. (b) The original and reconstructed manifolds M and $\Theta(M)$ exist as low-dimensional surfaces in high-dimensional space. In this work, M is a two-dimensional surface in the 72-dimensional space of Cartesian coordinates of the 24 united atoms of the polymer, and $\Theta(M)$ is a two-dimensional surface in the $(d = 20)$-dimensional delay space. We discover and extract the low-dimensional surfaces using a manifold learning technique known as diffusion maps [3,6,39,40,81,82]. Colloquially, this approach may be considered a nonlinear analog of principal components analysis that discovers low-dimensional curved hyperplanes preserving the most variance in the data. As an illustrative example [6], we show the application of diffusion maps to the "Swiss roll" data set comprising a cloud of points in $[X, Y, Z]$ defining a two-dimensional surface in three-dimensional space (top). The diffusion map discovers the latent two-dimensional manifold, and extracts it into the two collective variables $[\Psi_2, \Psi_3]$ quantifying, respectively, the location of the points along and perpendicular to the main axis of the spiral (bottom).

State-of-the-art single-molecule experimental techniques can furnish measurements of a small number of coarse-grained observables. For example, single-molecule particle tracking can furnish approximate backbone contours of linear macromolecules such as λ-DNA [20] from which molecular descriptors such as the radius of gyration or head-to-tail distance of the molecule can be extracted. Single-molecule Förster resonance energy transfer (smFRET) can supply one to three intramolecular distances between fluorescent dye molecules covalently grafted to the molecule of interest [21,22]. Given a time series in one (or more) system observables, hidden Markov models can estimate the most probable number of discrete metastable states and their interconversion rates [21,23]. Similarly, the computational mechanics approach of Crutchfield and coworkers [24,25], and its recent sophistication by Li *et al.* [26], can infer a state space network and transition probabilities from univariate measurements. An approach recently proposed by Haas *et al.* dispenses with the need to discretize the data into metastable states by inferring the parameters of a one-dimensional Langevin equation to project the smFES onto the measured observable [27]. Rather than inferring the metastable states of the molecule, or projecting the smFES onto

the measurement variable, the present work seeks to answer the following question: Is it possible to infer from a univariate time series of a single molecular observable a representation of the single-molecule free-energy surface that is geometrically and topologically equivalent to that which would have been recovered from a complete knowledge of all molecular degrees of freedom? In other words, is it possible to extract from a time series of a single molecular measurement a representation of the m collective variables, $\vec{\psi} = [\psi_1, \psi_2, \ldots, \psi_m]$, and the free-energy landscape over these variables, $F(\vec{\psi})$, that would have been computed by analyzing the $3N$-dimensional trajectory of the Cartesian coordinates of all atoms in the molecule?

By integrating ideas from dynamical systems theory, nonlinear manifold learning, and statistical mechanics, we demonstrate in molecular dynamics simulations of a polymer chain that—up to a smooth transformation and spatiotemporal symmetries—the answer to this question is in the affirmative. Attractor reconstruction seeks to infer the geometry and topology of the intrinsic manifold, M, of a dynamical system from a few system observables without knowledge of the underlying governing equations [28]. Takens' delay embedding theorem [29–34] provides a prescription to reconstruct a topologically and geometrically equivalent realization of the intrinsic manifold, $\Theta(M)$, from a scalar time series in a generic system observable by projecting the time series into a high-dimensional space in which the dynamical evolution is C^1-equivalent (i.e., related by a continuously differentiable function) to that in the original space. Villani *et al.* conducted molecular dynamics simulations of the 4-residue tropoelastin peptide in water, and employed delay embeddings of the peptide end-to-end distance to estimate the effective dimensionality of the dynamics and compute Lyupanov exponents [35,36]. Giannakis and Majda employed delay embeddings and Laplacian eigenmaps to infer the periodic, low-frequency, and intermittent spatiotemporal modes underpinning the dynamics of the upper-ocean temperature in a computational climate model [37]. Berry *et al.* integrated delay embeddings with diffusion maps to decompose high-dimensional dynamical processes into dynamical modes active at different time scales and recover the slow modes governing the long-time dynamics of coupled ordinary differential equations, 2D reaction-diffusion simulations, and videos of liquid crystal growth [38]. In this work, we use Takens' theorem to expand a univariate time series of the molecular head-to-tail distance into a high-dimensional space in which the dynamical evolution is C^1-equivalent to that of the molecule in real space, then employ diffusion maps [6,39,40] to recover a topologically equivalent reconstruction of the smFES. Takens' theorem asserts that this reconstructed smFES is—up to the removal of spatiotemporal symmetries—a diffeomorphism (i.e., related by a smooth and invertible mapping) to that which would be recovered by direct application of diffusion maps to the $3N$-dimensional simulation trajectory. We empirically verify this assertion by showing that the Jacobian determinant of the coordinate transformation between the two surfaces remains single-signed. By demonstrating in molecular simulations that we can recover a geometrically and topologically equivalent representation of the true smFES from a single experimentally accessible molecular observable, this work lays the theoretical and algorithmic foundations to infer single-molecule free-energy surfaces directly from experimental data. We present a schematic overview of our methodology in Fig. 1. We now proceed to discuss each component of the method in detail and validate our approach in an application to molecular simulations of a polymer chain.

II. RESULTS AND DISCUSSION

A. smFES from molecular dynamics simulations

We have previously applied diffusion maps to recover the smFES of n-tetracosane $C_{24}H_{50}$ in water [3]. This chemically simple homopolymer exhibits a rich conformational behavior, and serves as a prototypical model for the study of the hydrophobic effect in protein folding [41–43]. We selected this system as well-understood but nontrivial system in which to demonstrate and validate our methodology. As detailed in

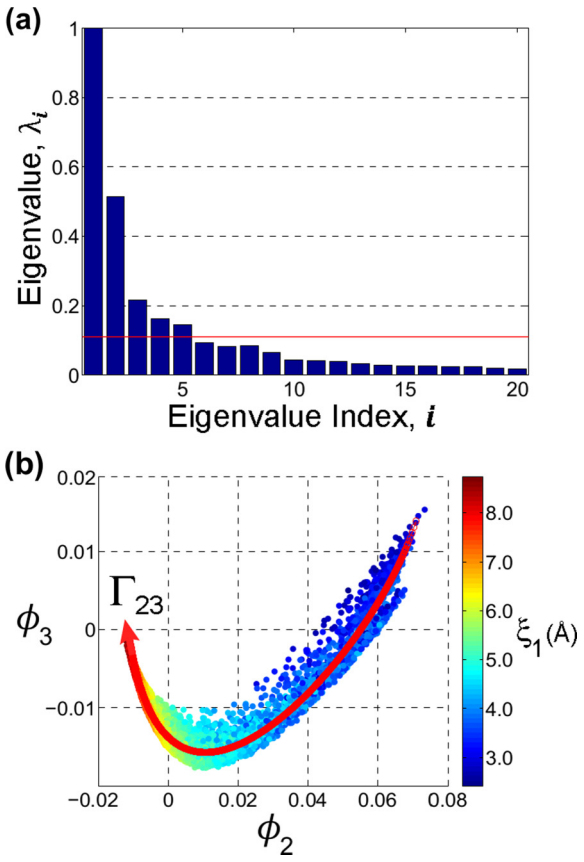

FIG. 2. Application of diffusion maps to the atomistic simulation trajectory employing a Gaussian kernel of bandwidth $\epsilon = 0.04$ specified using the approach in Ref. [83]. (a) Neglecting the trivial leading eigenvalue, $\lambda_1 = 1$, the spectral gap between λ_5 and λ_6 indicated by the horizontal line informs an embedding dimensionality of four into $[\vec{\phi}_2, \vec{\phi}_3, \vec{\phi}_4, \vec{\phi}_5]$. (b) Projection of the data into $[\vec{\phi}_2, \vec{\phi}_3]$ results in an effectively one-dimensional manifold, revealing a functional dependence between these two collective variables. We eliminate this redundancy using hierarchical nonlinear principal components analysis (h-NLPCA) to extract the effectively one-dimensional manifold that we term $\vec{\Gamma}_{23}$. Points are colored by the first principal moment of the gyration tensor ξ_1 [48].

the Methods Summary, we conducted molecular dynamics simulations of n-tetracosane in water, and applied the diffusion map nonlinear dimensionality reduction algorithm to the 72-dimensional simulation trajectory recording the Cartesian coordinates of the 24 united atoms. In brief, we computed pairwise distances between all 10 001 configurations in the molecular simulation trajectory as the root mean squared distance (RMSD) between the united atom coordinates of rotationally and translationally aligned chain configurations. By calculating a spectral decomposition of a random walk over this configurational ensemble, the diffusion map recovers the slowest modes of a diffusion process over the data identifiable as the important collective modes driving the dynamical evolution of the system [6,39,40]. The diffusion map identifies a four-dimensional manifold within the 72-dimensional space occupied by n-tetracosane defined by an embedding into the top four collective modes $[\vec{\phi}_2,\vec{\phi}_3,\vec{\phi}_4,\vec{\phi}_5]$ [Fig. 2(a)]. Following previous work, we consider the influence of the solvent degrees of freedom implicitly through their impact on the configurational ensemble sampled by the chain [3]. Consistent with previous findings, the projection of the data into $\vec{\phi}_2$ and $\vec{\phi}_3$ define an effectively one-dimensional manifold, indicating that these eigenvectors are functionally dependent collective variables describing the same dynamical mode of the system [3]. We have previously drawn the analogy with multivariate Fourier series in which $\sin(x)$ and $\sin(2x)$ are components oriented in the same spatial direction that are nonetheless orthogonal [3]. We eliminate this redundancy by applying hierarchical nonlinear principal components analysis (h-NLPCA) (Methods Summary) to the $[\vec{\phi}_2,\vec{\phi}_3]$ subspace to extract the one-dimensional manifold that we term $\vec{\Gamma}_{23}$ [Fig. 2(b)] [44,45]. An elegant alternative means to systematically detect and eliminate such so-called "repeated eigendirections" using locally linear approximations was recently proposed by Dsilva *et al.* [46]. The combined dimensionality reduction offered by sequential application of diffusion maps and h-NLPCA permits us to construct the three-dimensional embedding of the molecular dynamics trajectory into $[\vec{\Gamma}_{23},\vec{\phi}_4,\vec{\phi}_5]$ illustrated in Fig. 3. This projection defines the *intrinsic manifold*, M, of the n-tetracosane system. Temporally, the three collective variables spanning this manifold are the slow dynamical modes of the system to which the remaining degrees of freedom are effectively slaved [3,5]. Geometrically, this manifold is the three-dimensional hypersurface in phase space to which the dynamical evolution of the molecular system is effectively restrained. Representative molecular snapshots

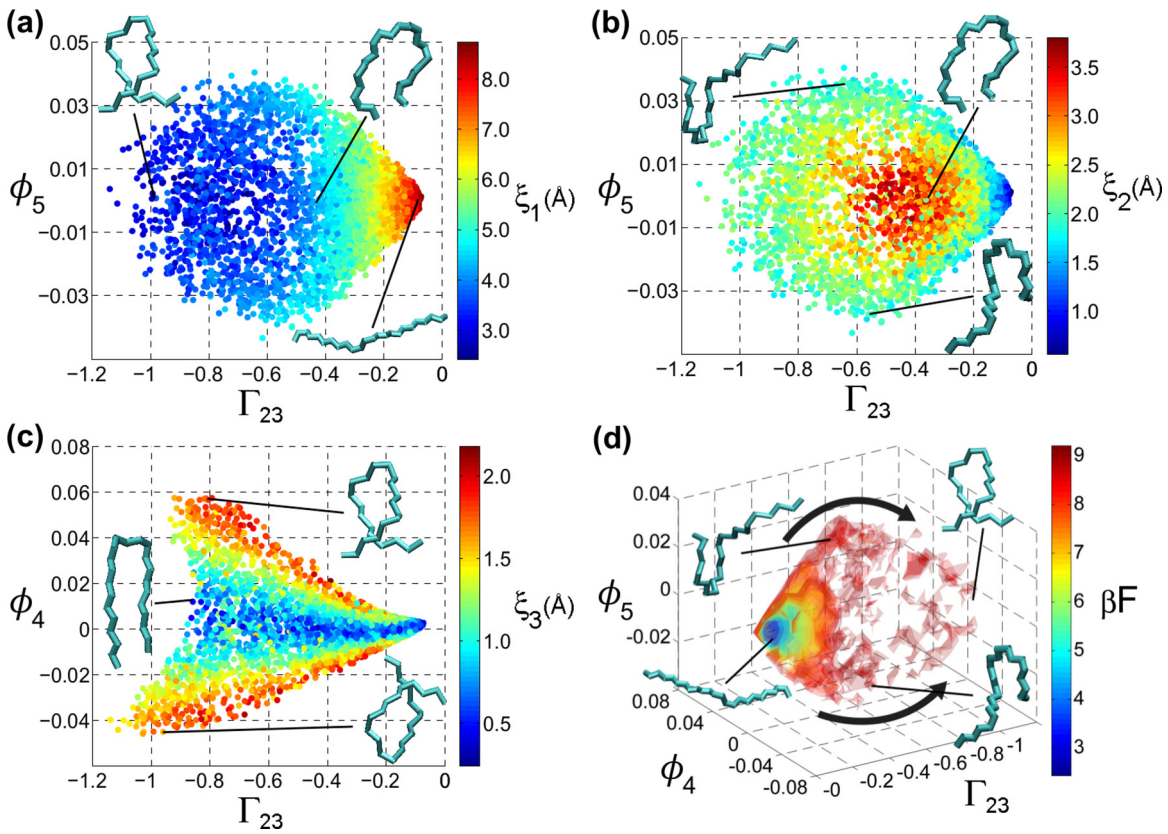

FIG. 3. Embedding of the atomistic simulation trajectory into the top three collective modes $[\vec{\Gamma}_{23},\vec{\phi}_4,\vec{\phi}_5]$ identified by diffusion maps. Projection of the 10 001 simulation snapshots into (a) the Γ_{23}-ϕ_5 projection colored by the first principal moment of the gyration tensor ξ_1, (b) the Γ_{23}-ϕ_5 projection colored by ξ_2, and (c) the Γ_{23}-ϕ_4 projection colored by ξ_3. (d) The smFES $F(\vec{\Gamma}_{23},\vec{\phi}_4,\vec{\phi}_5)$ with isosurfaces plotted at $\beta F = 3, 4, 5, 6, 7, 8, 9$, where F is the Gibbs free energy and $\beta = 1/k_B T$. The "kink-and-slide" collapse pathways are indicated by arrows, wherein a kink forms at the head or tail of the chain, the kink migrates towards the center of the chain expelling water molecules from between the arms to form a symmetric hairpin with a dry interior, then the chain condenses into a hydrophobically collapsed right- or left-handed helical coil.

are projected over this and all subsequent embeddings using VMD [47].

A known deficiency of the diffusion map, and nonlinear dimensionality reduction techniques in general, is that the low-dimensional collective variables are unknown nonlinear functions of the system degrees of freedom [6]. The gyration tensor of the n-tetracosane chain presents a useful interpretive "bridge" variable with which to correlate and develop physical insight into chain motions in $[\vec{\Gamma}_{23},\vec{\phi}_4,\vec{\phi}_5]$ [3,48]. In Figs. 3(a)–3(c) we color the projected molecular configurations according to the principal moments of the chain gyration tensor, $\{\xi_1,\xi_2,\xi_3\}$, interpretable as the length of the chain along its longest, next longest, and shortest axes [48]. The motion of the chain over this intrinsic manifold resolves the hydrophobic collapse mechanism that our previous analysis revealed to proceed by a "kink-and-slide" mechanism [3], wherein extended configurations in the global free energy minimum collapse via shifting of a loose asymmetric bend near the head or tail towards the middle of the chain to form a tight symmetric hairpin that subsequently folds into a right- or left-handed helix. Unfolding proceeds by the reverse pathway. The free-energy profile over the intrinsic manifold, M, $F(\vec{\Gamma}_{23},\vec{\phi}_4,\vec{\phi}_5) = -k_B T \ln\hat{P}(\vec{\Gamma}_{23},\vec{\phi}_4,\vec{\phi}_5)$, defines the smFES of the n-tetracosane chain in water as a surface in $\mathbb{R}^4$ presented in Fig. 3(d).

B. Spatially symmetrized smFES from molecular dynamics simulations

It is the goal of this study to employ Takens' theorem to recover a diffeomorphism of the smFES of the n-tetracosane chain from a knowledge of only the head-to-tail distance, ℓ, between the terminal united atoms of the chain. We selected ℓ as an experimentally accessible observable that can, in principle, be measured using a technique such as smFRET [21,22]. In practice—particularly for a short alkane chain—the attachment of extrinsic FRET dye molecules may perturb the molecular motions of the molecule [22], and it can be challenging to (i) attach the dyes, (ii) obtain long time series before photobleaching, (iii) achieve sub-ms time resolution, (iv) resolve adequate signal-to-noise ratios, and (v) measure distances outside 2–8 nm [21]. It is the aim of the present study to lay the theoretical foundations for the recovery of smFES in the idealized case of perfect measurements. We defer to our future work a confrontation of the important practical concerns associated with the use of real smFRET data.

As we discuss below, the technique we use to recover the smFES requires that the measured observable be *generic* in the sense that it is a function of all system degrees of freedom, and does not contain any symmetries not present in the system being observed [29,49–51]. As a function of all chain degrees of freedom (i.e., the 72 Cartesian coordinates of the 24 united atoms, up to trivial rotations and translations), ℓ satisfies the first criterion, but it does possess two symmetries absent in the molecule. First, ℓ cannot distinguish the head-to-tail sense of the molecule, such that it is invariant to head-to-tail inversions. This means, for example, that it cannot distinguish whether a kink in an asymmetrically kinked molecule occurs at the head or the tail. Second, ℓ is invariant to mirror symmetries of the chain, such that it cannot distinguish between chiral enantiomers of the same molecular configuration, and so cannot differentiate between right- and left-handed helices. The role of symmetries in dynamical systems and their observables in phase space reconstruction has been studied in detail [50–53]. The prototypical example of this symmetry is the z variable in the Lorenz equations, which cannot distinguish the symmetric wings of the Lorenz attractor [50,53,54] (cf. Appendix B1). Reconstruction of the manifold using z alone necessarily collapses together the two wings, but is an otherwise good reconstruction variable capable of producing accurate global reconstructions of the phase space [53].

In the present case, reconstruction of the intrinsic manifold from ℓ can only be performed up to head-to-tail and mirror symmetries of the chain. It is not possible, therefore, to recover a diffeomorphism of the smFES extracted from the full-dimensional molecular simulation from measurements of ℓ alone. Our objective instead should be recovery of a representation of the smFES in which these two symmetries are eliminated. We remove the symmetries by reapplying

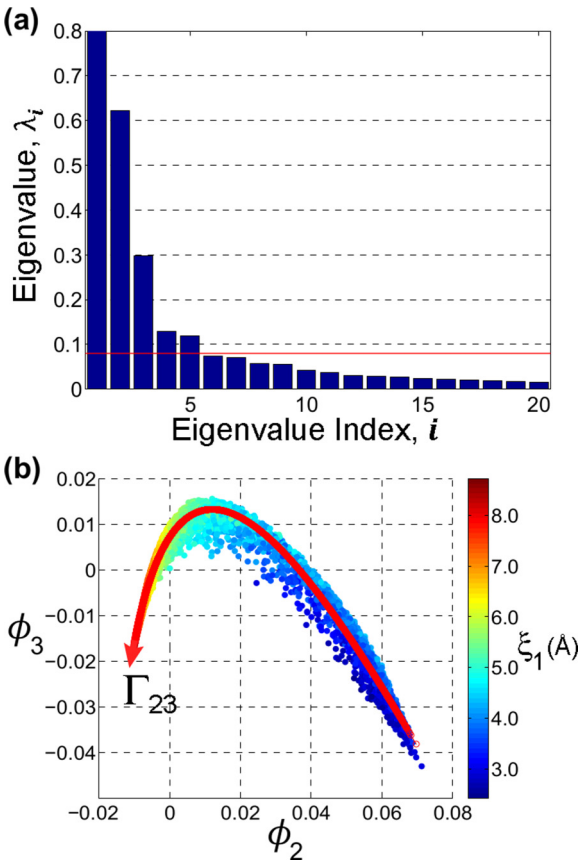

FIG. 4. Application of diffusion maps to the spatially symmetrized atomistic simulation trajectory employing a Gaussian kernel of bandwidth $\epsilon = 0.03$ specified using the approach in Ref. [83]. (a) The spectral gap between λ_5 and λ_6 indicated by the horizontal line informs an embedding dimensionality of four into $[\vec{\phi}_2,\vec{\phi}_3,\vec{\phi}_4,\vec{\phi}_5]$. (b) Projection of the data into $[\vec{\phi}_2,\vec{\phi}_3]$ results in an effectively one-dimensional manifold, informing a functional dependence between these two collective variables. We eliminate this redundancy using h-NLPCA to extract the effectively one-dimensional manifold that we term $\vec{\Gamma}_{23}$. Points are colored by the first principal moment of the gyration tensor ξ_1.

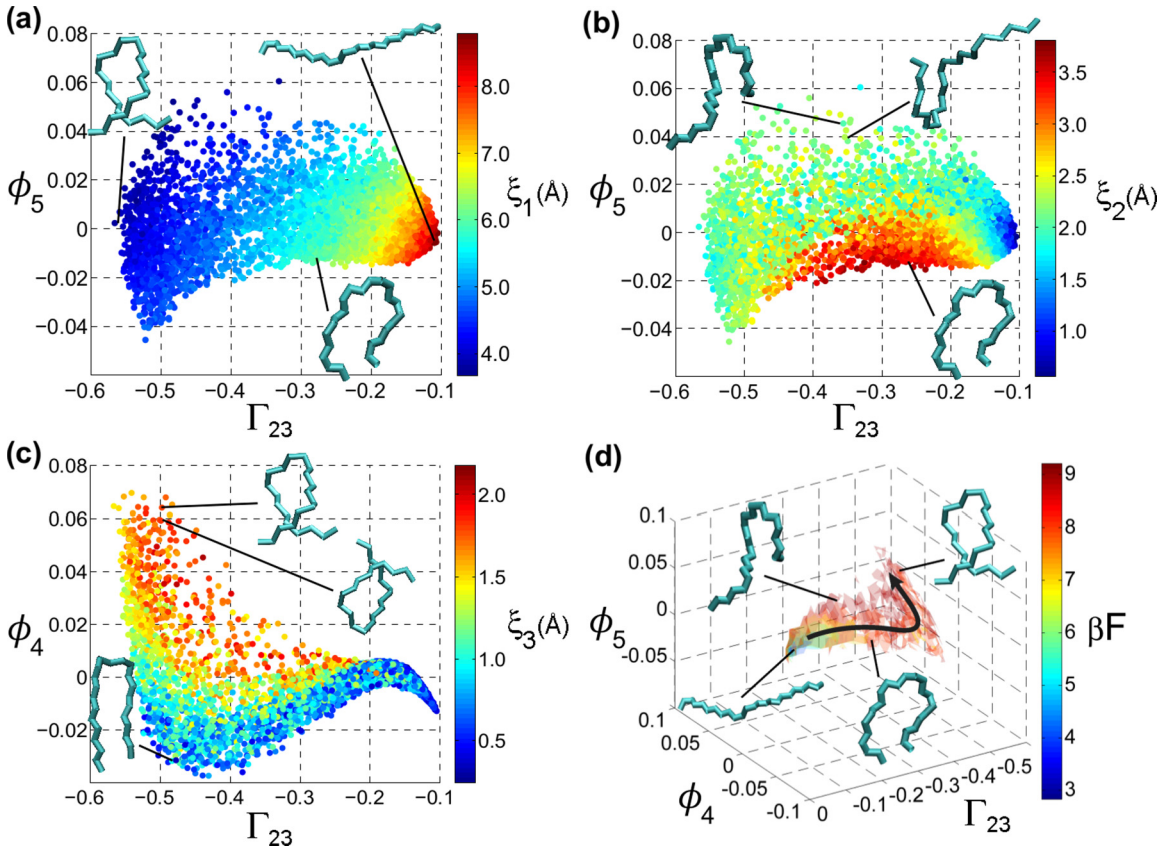

FIG. 5. Embedding of the spatially symmetrized atomistic simulation trajectory in which the head-to-tail and mirror symmetries of the molecule were removed, into the top three collective modes $[\vec{\Gamma}_{23},\vec{\phi}_4,\vec{\phi}_5]$ identified by diffusion maps. Projection of the 10 001 simulation snapshots into (a) the Γ_{23}-ϕ_5 projection colored by ξ_1, (b) the Γ_{23}-ϕ_5 projection colored by ξ_2, and (c) the Γ_{23}-ϕ_4 projection colored by ξ_3. Elimination of the head-to-tail symmetry collapses together asymmetrically kinked chain configurations, and elimination of mirror symmetry collapses together right- and left-handed helices. (d) The smFES $F(\vec{\Gamma}_{23},\vec{\phi}_4,\vec{\phi}_5)$ exists as a two-dimensional surface within the three-dimensional space spanned by $[\vec{\Gamma}_{23},\vec{\phi}_4,\vec{\phi}_5]$. Free-energy isosurfaces are plotted at $\beta F = 3, 4, 5, 6, 7, 8, 9$, and the "kink-and-slide" collapse pathway is indicated by an arrow.

diffusion maps to the molecular simulation trajectory in which we define distances between pairs of chain configurations as the rotationally and translationally aligned RMSD between the united atom coordinates minimized over head-to-tail inversion and mirror reflection. As above, the diffusion map identifies a four-dimensional intrinsic manifold in which $[\vec{\phi}_2,\vec{\phi}_3]$ are functionally dependent (Fig. 4), allowing us to apply h-NLPCA to construct the $[\vec{\Gamma}_{23},\vec{\phi}_4,\vec{\phi}_5]$ intrinsic manifold in $\mathbb{R}^3$ in Figs. 5(a)–5(c) and associated smFES in $\mathbb{R}^4$ in Fig. 5(d). The symmetrized intrinsic manifold is topologically equivalent to a "folding" in half of the original attractor in both ϕ_4 and ϕ_5, corresponding to elimination of the mirror and head-to-tail symmetries, respectively. Removing these two spatial symmetries makes ℓ an appropriate generic observable for its reconstruction since it is both a function of all chain degrees of freedom and does not contain any symmetries not present in the spatially symmetrized molecular system.

The spatially symmetrized three-dimensional intrinsic manifold exists as an effectively two-dimensional surface in the three-dimensional space spanned by $[\vec{\Gamma}_{23},\vec{\phi}_4,\vec{\phi}_5]$, providing an opportunity for further dimensionality reduction beyond that furnished by the diffusion map. Application of h-NLPCA to the embedded data identifies a new basis set of three nonlinear principal components, $[\vec{\Upsilon}_1,\vec{\Upsilon}_2,\vec{\Upsilon}_3]$, formed from nonlinear combinations of $[\vec{\Gamma}_{23},\vec{\phi}_4,\vec{\phi}_5]$ and arranged in order of decreasing variance. That 99.95% of the variance in the data lie within the top two nonlinear principal components confirms that the manifold is effectively two-dimensional, and can be projected into $[\vec{\Upsilon}_1,\vec{\Upsilon}_2] \in \mathbb{R}^2$ with essentially no loss of information (Fig. 6). We present this two-dimensional intrinsic manifold and three-dimensional smFES in Fig. 7. The "kink-and-slide" pathway for chain folding and unfolding over the intrinsic manifold remains apparent, but where elimination of the head-to-tail and mirror symmetries collapse together the head-and-tail kinked conformations, and right- and left-handed helices, respectively. It is the spatially symmetrized three-dimensional smFES in Fig. 7(d) that we seek to reconstruct from the scalar time series in ℓ.

We note that the full, unsymmetrized manifold may, in principle, be recovered by supplementing ℓ with other simultaneous measurements capable of lifting the degeneracy in head-to-tail inversion (e.g., an asymmetric intramolecular smFRET distance) and mirror symmetry (e.g., circular dichroism). It is the goal of the present work to recover the smFES from a scalar

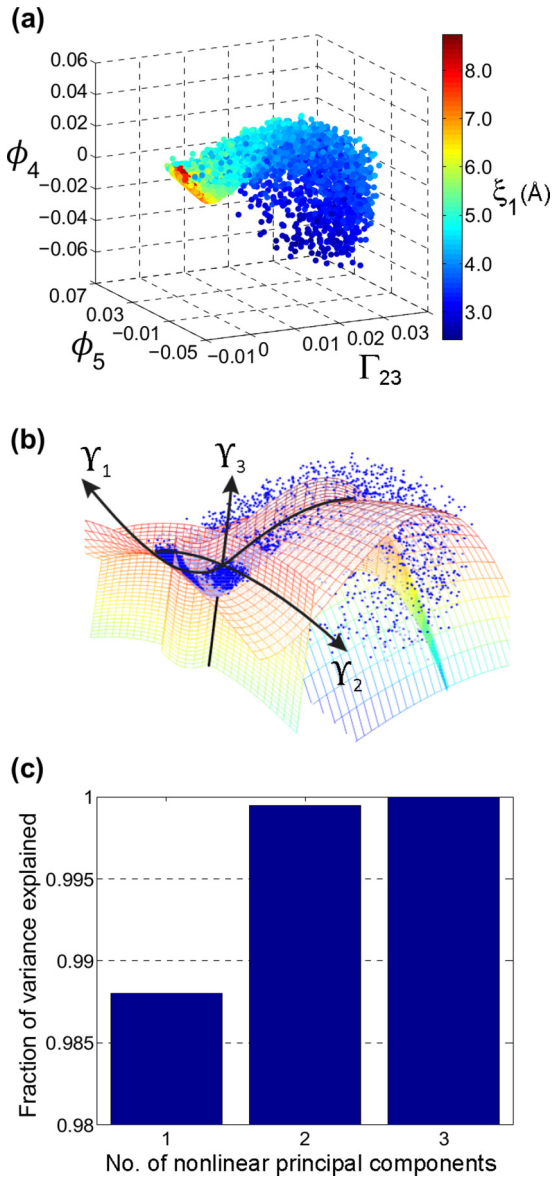

FIG. 6. Additional dimensionality reduction of the spatially symmetrized diffusion map embedding of the atomistic simulation trajectory (Fig. 5). (a) The three-dimensional diffusion map embedding of the spatially symmetrized atomistic simulation trajectory exists as an approximately two-dimensional surface in $[\vec{\Gamma}_{23},\vec{\phi}_4,\vec{\phi}_5]$. (b) Application of h-NLPCA to the embedding generates a new basis set of three nonlinear principal components, $[\vec{\Upsilon}_1,\vec{\Upsilon}_2,\vec{\Upsilon}_3]$, formed from nonlinear combinations of $[\vec{\Gamma}_{23},\vec{\phi}_4,\vec{\phi}_5]$ and arranged in order of decreasing variance. This figure was generated using the "Nonlinear PCA toolbox for Matlab" developed by Scholz [45,85]. (c) Plotting the cumulative fraction of variance explained upon incorporating additional nonlinear principal components shows that more than 99.95% of the variance in the embedding resides in the first two principal components, confirming that the manifold is effectively two-dimensional and can be projected into $[\vec{\Upsilon}_1,\vec{\Upsilon}_2] \in \mathbb{R}^2$ with essentially no loss of information.

time series, but a natural extension would be the construction of multivariate Takens' delay embeddings from multichannel measurements [54].

C. smFES from delay embeddings

Takens' delay embedding theorem is a well-established result in dynamical systems theory dating to the early 1980s [29–34], but its implications can be unintuitive. The theorem seems to state that the multidimensional free-energy landscape upon which a system evolves can be recovered from the history of a time series in a single system observable. Projecting the high-dimensional system dynamics onto a single measurement would seem to surrender any possibility of recovering the multidimensional surface. So how can Takens' theorem be rationalized? Dispensing for the moment with mathematical formality for the sake of clarity, two factors must be borne in mind. First, Takens' theorem permits recovery only of a *topologically equivalent* representation of the original landscape [32], meaning that the reconstructed landscape is related to the original landscape through a smooth transformation that may bend, stretch, or squash the manifold, but not rip it apart or stick it together in new ways [55]. The reconstructed manifold is therefore guaranteed to preserve all of the topological properties of the original, including its edges, its continuity, and its connectivity [55]. The reconstruction is not, however, guaranteed to preserve the *topography* of the manifold, since the smooth transformation may change the probability distribution over the surface and therefore perturb the heights and depths of the free-energy peaks and valleys. For our purposes, this means that the reconstructed landscape is guaranteed to faithfully identify the states of the system and the connectivity of the structural transition pathways between them, but the smooth transformation may perturb the terrain of the free-energy landscape from that over the original manifold. We are unaware of any theoretical results placing bounds on the degree to which the transformation may perturb the landscape. In this work, we quantify the topographical perturbation numerically to demonstrate that it is relatively mild for this particular system, and describe in the Conclusions our ongoing work to place analytical and/or theoretical limits on the extent of the perturbation. Second, a univariate time series provides not just a single measurement of the system state, but the entire history of that observable. Provided that the measurement is a function of all of the system degrees of freedom (i.e., it is generic) then the evolution of the system over its multidimensional free-energy surface is encoded into this univariate time trace. Keeping a sufficiently long history of system univariate observations enables Takens' delay embeddings to unambiguously pinpoint the location of the system on its multidimensional free-energy surface. It is perhaps useful to make an analogy with Markov chains: the future evolution of a mth order chain can be predicted from knowledge of the last m states visited by the system [56]. A second useful analogy is one with ordinary differential equations: the existence and uniqueness theorem states that—subject to some constraints on continuity and smoothness—an Nth order ordinary differential equation, $y^{(N)} = F(x, y', y'', \ldots, y^{(N-1)})$, possesses a unique solution, $y(x)$, for a particular specification of its initial condition, $x = x_0$, and the first $(N-1)$ derivatives at that point $y'(x_0) = \sigma_1$, $y''(x_0) = \sigma_2$, …, $y^{(N-1)}(x_0) = \sigma_{(N-1)}$ [31,57]. By keeping a sufficiently long record of the past history of $y(x)$ these derivatives may be computed by finite differences, permitting calculation of the unique solution.

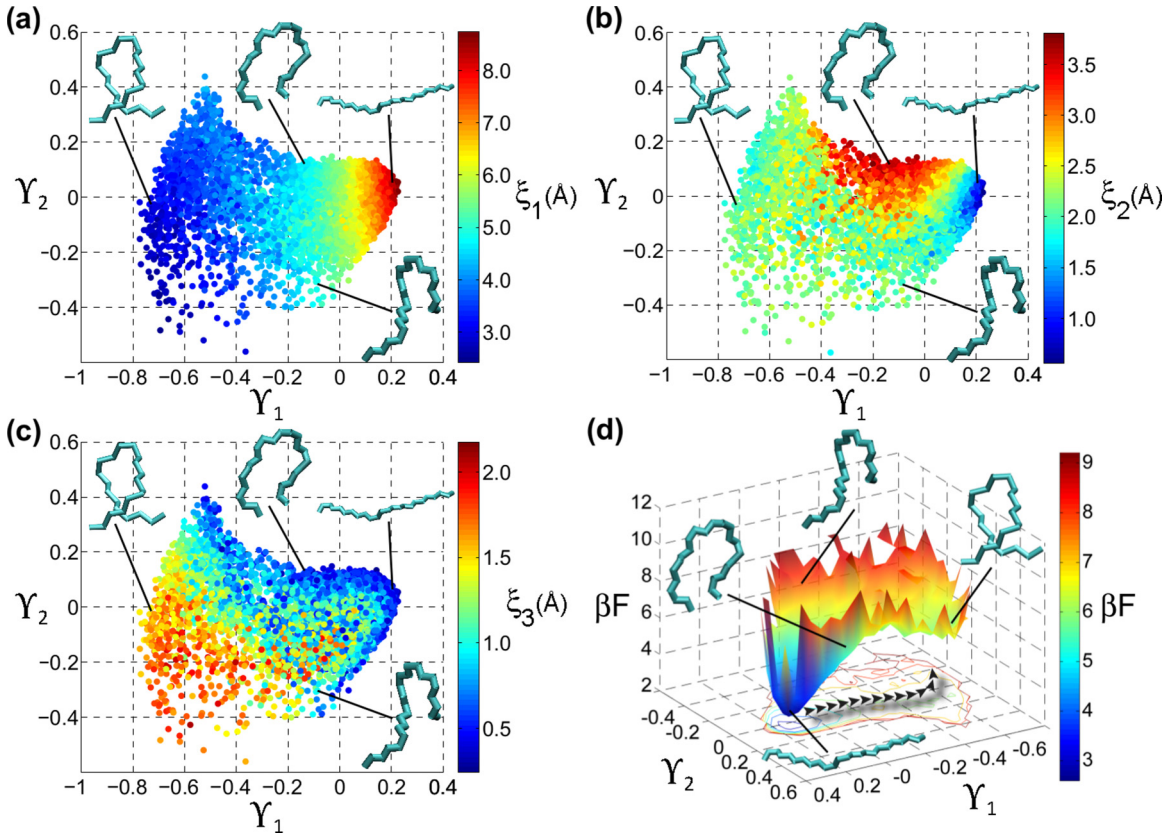

FIG. 7. Embedding of the spatially symmetrized atomistic simulation trajectory into the top two nonlinear principal components $[\vec{\Upsilon}_1, \vec{\Upsilon}_2]$ identified by sequential application of diffusion maps and h-NLPCA. Projection of the 10 001 simulation snapshots colored by (a) ξ_1, (b) ξ_2, and (c) ξ_3. (d) The smFES $F(\vec{\Upsilon}_1, \vec{\Upsilon}_2)$ over which the "kink-and-slide" collapse pathway is indicated by chevrons.

Appreciating the possibly alien nature of these ideas, we present in Appendix B two simple examples of the application of Takens' theorem to recover from univariate time series topologically equivalent representations of the multidimensional landscapes of (i) the Lorenz attractor and (ii) two-dimensional Brownian motion in a three-well potential. Below, we use this approach to recover a topologically equivalent representation of the smFES of the n-tetracosane chain presented in Fig. 7(d).

Mathematically, Takens' delay embedding theorem [29–34] provides a means to reconstruct a topologically and geometrically equivalent realization of the intrinsic manifold, $\Theta(M)$, from a scalar time series in a generic observable—not containing any symmetries that are not present in the system—by projecting the time series into a high-dimensional space in which the dynamical evolution is C^1-equivalent (i.e., related by a continuously differentiable function) to that in the original space. Θ is an invertible function mapping M to $\Theta(M)$ such that both Θ and Θ^{-1} are smooth, such that $\Theta(M)$ is geometrically and topologically equivalent to M [32,58] (Methods Summary). Given our scalar time series $\{\ell(t_i)\}_{i=1}^K$ measured at equally spaced 10 ps intervals over the course of the 100 ns molecular simulation (Fig. 8), Takens' theorem prescribes that we construct the mapping, Θ, through a *delay embedding*,

$$\vec{y}(t_i) = \Theta(\ell(t_i)) = [\ell(t_i), \ell(t_i + \tau), \ldots, \ell(t_i + (d-1)\tau)], \quad (1)$$

where τ is the delay time between successive system observations and d is the delay embedding dimensionality. By considering sufficiently many delayed observations into this projection, Takens' theorem makes the remarkable assertion that the dynamical evolution of the delay embedding becomes equivalent to that of the dynamical evolution of the molecule in its Cartesian coordinate space, with one related to the other by a smooth and invertible transformation [59]. Formally, our application of Takens' theorem is to observations of a subspace, the dynamics of the polymer chain, subject to external

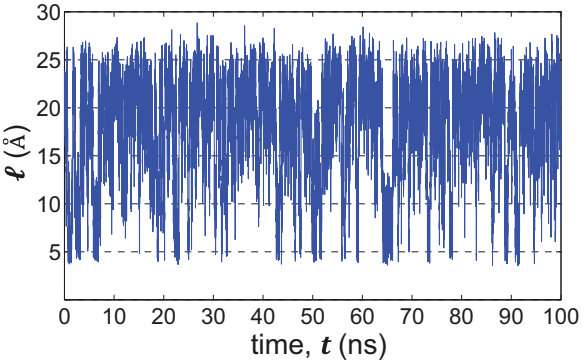

FIG. 8. The scalar time series $\{\ell(t_i)\}_{i=1}^K$ measuring the head-to-tail distance of the n-tetracosane at $K = 10\,001$ points at 10 ps intervals over the course of the 100 ns molecular simulation trajectory.

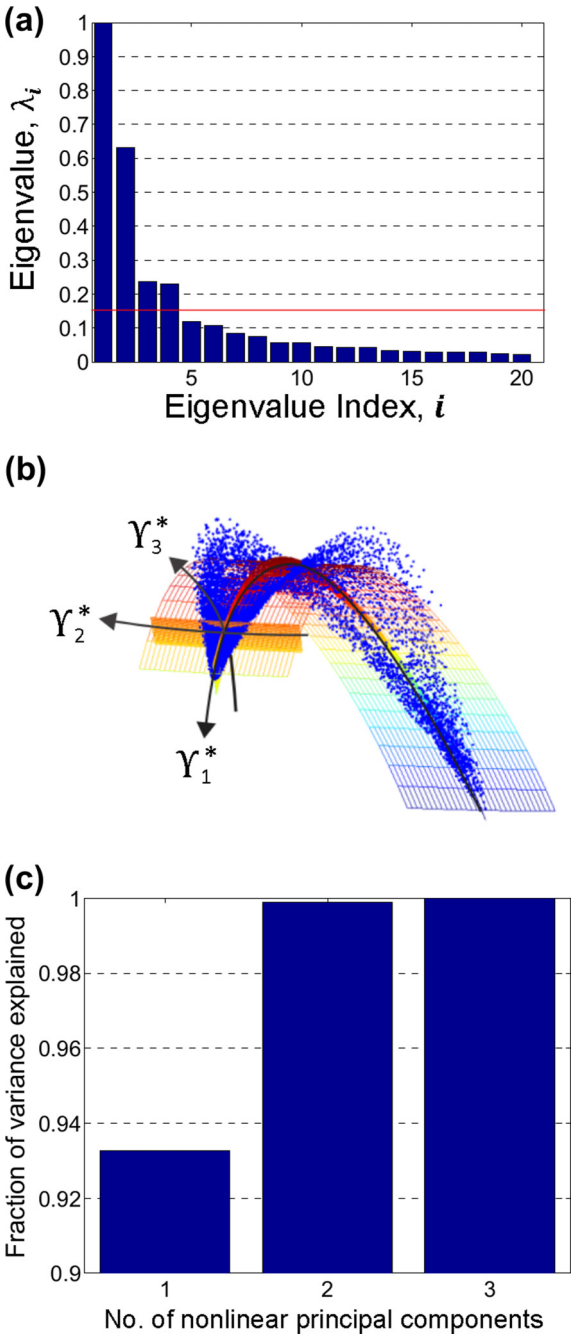

FIG. 9. Application of diffusion maps and h-NLPCA to the 20-dimensional delay embedding constructed from the scalar time series of the chain head-to-tail distance. (a) Diffusion maps were applied using a Gaussian kernel of bandwidth $\epsilon = 4.00$ specified using the approach in Ref. [83]. The gap in the eigenvalue spectrum between λ_4 and λ_5 indicated by the horizontal line informs an embedding dimensionality of three into $[\vec{\phi}_2^*,\vec{\phi}_3^*,\vec{\phi}_4^*]$. (b) Application of h-NLPCA to the embedding generates a new basis set of three nonlinear principal components, $[\vec{\Upsilon}_1^*,\vec{\Upsilon}_2^*,\vec{\Upsilon}_3^*]$, formed from nonlinear combinations of $[\vec{\phi}_2^*,\vec{\phi}_3^*,\vec{\phi}_4^*]$ and arranged in order of decreasing variance. (c) Plotting the cumulative fraction of variance explained upon incorporating additional nonlinear principal components shows that more than 99.79% of the variance in the embedding resides in the first two principal components, confirming that the manifold is effectively two-dimensional and can be projected into $[\vec{\Upsilon}_1^*,\vec{\Upsilon}_2^*] \in \mathbb{R}^2$ with essentially no loss of information.

forcing by solvent motion and the coupled thermostat and barostat (Methods Summary). Takens' theorem was originally formulated for autonomous dynamical systems independent of time and external influences [33,34], and so our application appeals to recent generalizations of Takens' theorem by Stark *et al.*, who proved it to hold, under very general conditions, for both deterministically and stochastically forced systems [33,34]. The projected time series $\{\vec{y}(t_i)\}_{i=1}^{K'}$ defines the reconstructed intrinsic manifold $\Theta(M) \in \mathbb{R}^d$. Takens' theorem assures recovery of $\Theta(M)$ for $d \geqslant (2k+1)$, where k is the dimensionality of the original system, but $k < d < (2k+1)$ can be sufficient [30,60]. The theorem places no restrictions on the value of τ. In practice, empirical tools exist to choose appropriate values of τ and d for finite data and a system of unknown dimensionality. We employ the mutual information approach of Fraser and Swinney to choose $\tau = 20$ ps [61], and the approach of Cao [62] based on the false nearest neighbors method of Kennel *et al.* [63] to select $d = 20$ (Methods Summary).

Due to the incorporation of 20 measurements of ℓ spaced at 20 ps intervals into each delay embedding vector $\vec{y}(t)$, the $K = 10\,001$ observations of ℓ produce only $K' = 9963$ points in the delay embedding. Accordingly, we assign the properties of each delay-embedded point from the mean over the points constituting the delay vector. For example, in Figs. 10, 11, and 13 we color each delay embedded point according to the principal moments of the gyration tensor averaged over the 20 molecular configurations, $\{\Xi_1,\Xi_2,\Xi_3\}$, where we use upper case to denote a multisnapshot average.

Above, we applied diffusion maps and h-NLPCA to extract the intrinsic manifold, M, from the $3N$-dimensional Cartesian coordinate space of the atomistic simulation trajectory. We employ an analogous approach to extract the reconstructed intrinsic manifold, $\Theta(M)$, from the 20-dimensional delay embedding, $\{\vec{y}(t_i)\}_{i=1}^{K'}$. Computing pairwise distances between delay embedding vectors using the Euclidean norm, the diffusion map infers a three-dimensional projection of the delay embedded data into the leading collective modes $[\vec{\phi}_2^*,\vec{\phi}_3^*,\vec{\phi}_4^*]$ [Fig. 9(a)], where we decorate the collective modes inferred from the delay embedding with an asterisk to distinguish them from those derived from the atomistic simulation. These collective order parameters describe the slow modes of the dynamical evolution of $y(t)$ over $\Theta(M)$, which Takens' theorem asserts is C^1-equivalent to the dynamical evolution of the molecular system on M [32]. As illustrated in Fig. 9(b), the reconstructed intrinsic manifold produced by embedding the 9963 delay vectors into this three-dimensional space exists as a two-dimensional surface resembling a potato chip. Application of h-NLPCA confirms this assessment, showing 99.79% of the variance in the data to reside within the top two nonlinear principal components [Fig. 9(c)], permitting the projection of $\Theta(M)$ into $[\vec{\Upsilon}_1^*,\vec{\Upsilon}_2^*]$ with essentially no loss of information. We present in Fig. 10(a) the projection of $\Theta(M)$ into $[\vec{\phi}_2^*,\vec{\phi}_3^*,\vec{\phi}_4^*]$, and in Fig. 10(b) its projection into $[\vec{\Upsilon}_1^*,\vec{\Upsilon}_2^*]$.

D. Temporally symmetrized smFES from delay embeddings

A necessary condition for the existence of a diffeomorphism between M and $\Theta(M)$ is that the manifolds possess

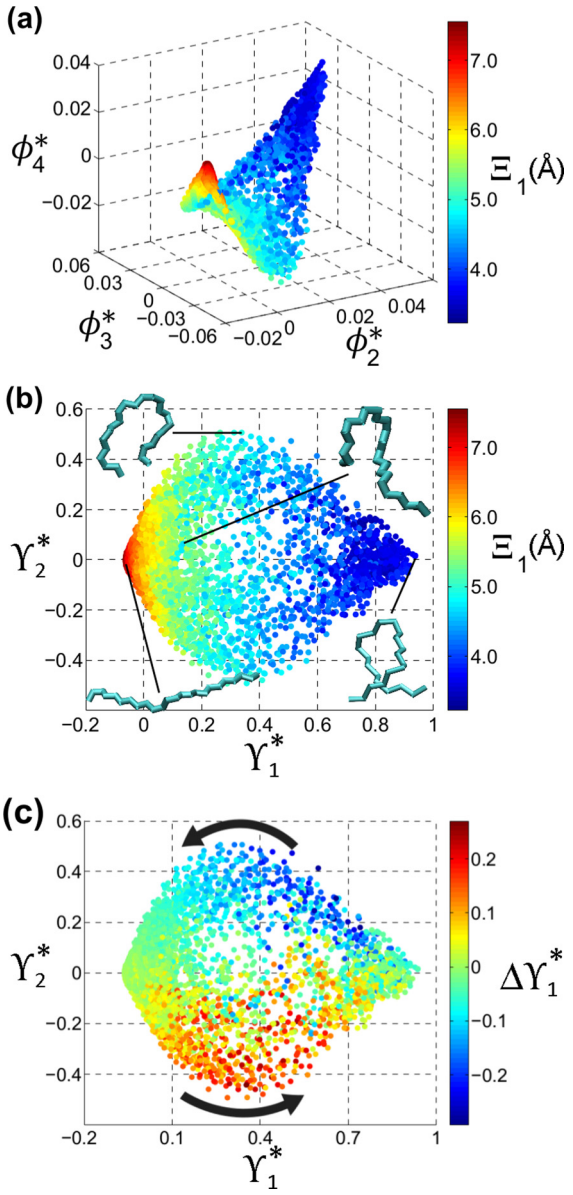

FIG. 10. Recovery of the reconstructed intrinsic manifold by the application of diffusion maps and h-NLPCA to the 20-dimensional Takens' delay embedding of the ℓ scalar time series. (a) Projection of the 9963 delay embedding vectors into the top three collective modes $[\vec{\phi}_2^*,\vec{\phi}_3^*,\vec{\phi}_4^*]$ identified by diffusion maps. Points are colored by the first principal moment of the gyration tensor averaged over the 20 molecular configurations constituting each delay embedding vector, Ξ_1. (b) Projection of the reconstructed intrinsic manifold in panel (a) into the top two nonlinear principal components $[\vec{\Upsilon}_1^*,\vec{\Upsilon}_2^*]$ recovered from the application of h-NLPCA to the diffusion map embedding. High (low) Ξ_1, extended (collapsed) chain configurations lie at low (high) values of Υ_1^*. In the delay embedding vectors selected for visualization, the 10th of the 20 configurations constituting the delay embedding is visualized. (c) Reproduction of panel (b) with points colored by the change in Υ_1^* between consecutive delay embedding vector projections, $\Delta\Upsilon_1^*(t_i) = \Upsilon_1^*(t_i+\tau) - \Upsilon_1^*(t_i)$. The delay embedding breaks the symmetry of Newton's equations of motion such that collapse and extension pathways are embedded into different regions of the reconstructed intrinsic manifold. Chain collapse proceeds as indicated by the lower arrow, and extension by the upper, resulting in a net counterclockwise flow.

the same dimensionality. It is encouraging, therefore, that the sequential application of diffusion maps and h-NLPCA furnishes $M \in \mathbb{R}^2$ from the molecular simulation trajectory in $\mathbb{R}^{72}$ [Figs. 7(a)–7(c)], and $\Theta(M) \in \mathbb{R}^2$ from the Takens' delay embedding of ℓ in $\mathbb{R}^{20}$ [Fig. 10(b)]. Inspection of these two manifolds, however, reveals that $\Theta(M)$ possesses a reflection symmetry across the Υ_1^* axis that is absent in M, suggesting that the two manifolds are not topologically equivalent. Indeed the determinant of the Jacobian of the coordinate transformation between the two manifolds is not single-signed (Methods Summary), confirming the absence of a diffeomorphism [52,53,64]. What is the root of this apparent contradiction to Takens' theorem?

As illustrated in Fig. 10(b), delay embedding vectors residing at low values of Υ_1^* correspond to extended chain configurations with large values of Ξ_1, whereas those at high values of Υ_1^* correspond to collapsed hairpins and helices with small Ξ_1. By computing the change in Υ_1^* between successive delay embedding vectors, the origin of the reflection symmetry in $\Theta(M)$ is revealed. In Fig. 10(c), we color each point in $\Theta(M)$ by $\Delta\Upsilon_1^*(t_i) = \Upsilon_1^*(t_i+\tau) - \Upsilon_1^*(t_i)$. The process of chain collapse from low to high Υ_1^* (high to low Ξ_1) is indicated by the lower black arrow, corresponding to progression along the lower half of $\Theta(M)$ passing through negative values of Υ_2^*. The reverse process, chain extension from high to low Υ_1^* (low to high Ξ_1), is indicated by the upper black arrow, and corresponds to progression along the upper half of $\Theta(M)$ passing through positive values of Υ_2^*. Accordingly, the dynamical evolution of the delay embedding defined by Eq. (1) produces a counterclockwise flow around $\Theta(M)$ as the chain collapses and extends.

The existence of separate pathways for chain collapse and extension stands in apparent contradiction to the expectation that a classical molecular system in thermodynamic equilibrium should obey detailed balance and exhibit microscopic reversibility [5,65]. In other words, it should not be possible to tell from the observation of a single molecular configuration whether the chain is in the process of collapse or extension, and the sequence of configurations in collapse and extension pathways should be coincident in the intrinsic manifold. This expectation is borne out in the intrinsic manifold, M, recovered from the molecular simulation trajectory where each data point corresponds to a single observation of the system, but not for that recovered from the delay embedding, $\Theta(M)$, where each point corresponds to 20 successive observations. The critical difference is that the construction of a delay embedding as a sequence of measurements breaks the time reversibility of Newton's equations of motion, such that it is possible to ascertain from the series of measurements whether the chain is in the process of collapsing or extending. Specifically, the delay embedding of a particular chain configuration in a collapse pathway will comprise 20 measurements of ℓ decreasing in value, whereas the delay embedding of an identical chain configuration undergoing extension will comprise 20 measurements of ℓ increasing in value. Accordingly, two otherwise identical chain configurations are necessarily embedded in different coordinates in the delay space. This symmetry breaking induced by the delay embedding separates the collapse and extension

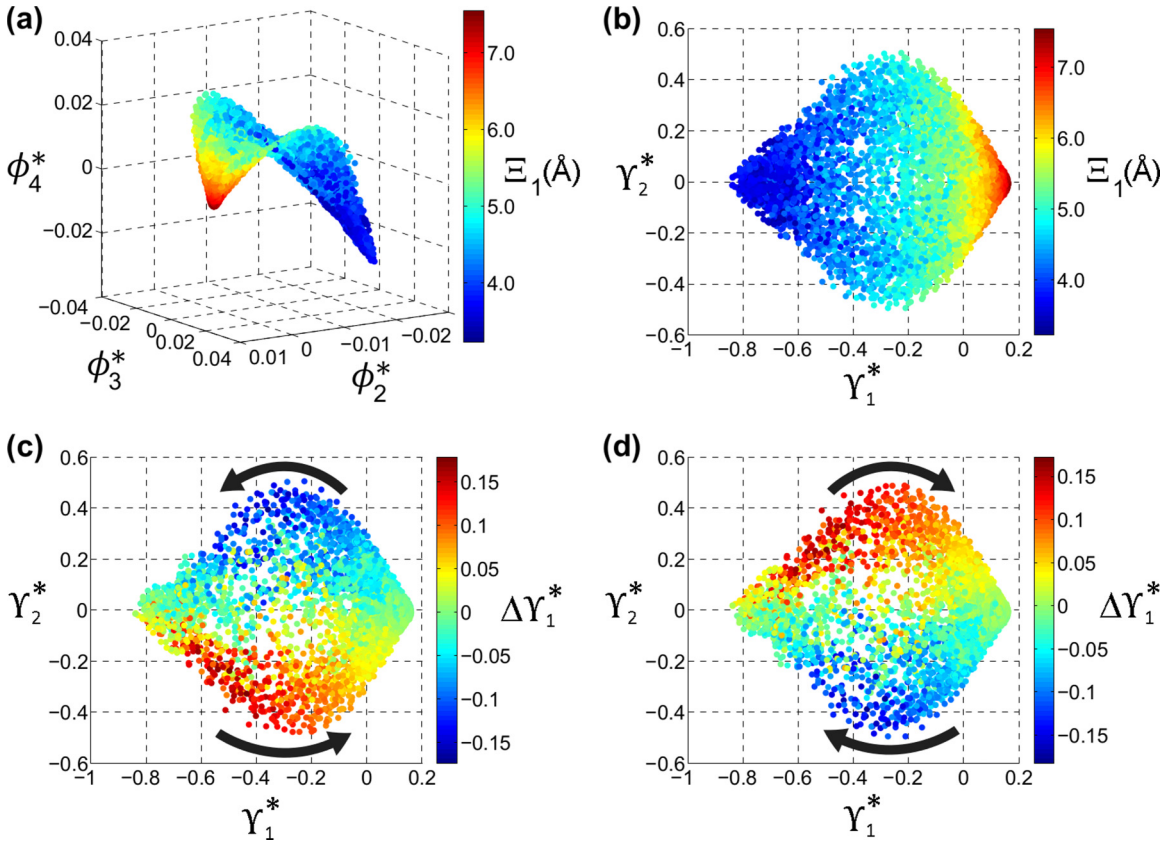

FIG. 11. Recovery of the temporally augmented reconstructed intrinsic manifold by the application of diffusion maps and h-NLPCA to the 20-dimensional Takens' delay embedding of the ℓ scalar time series produced by concatenating the delay embeddings resulting from the forward and backward simulations. (a) Projection of the 19 926 delay embedding vectors into the top three collective modes $[\vec{\phi}_2^*,\vec{\phi}_3^*,\vec{\phi}_4^*]$ identified by diffusion maps. (b) Projection of the reconstructed intrinsic manifold in panel (a) into the top two nonlinear principal components $[\vec{\Upsilon}_1^*,\vec{\Upsilon}_2^*]$ recovered from the application of h-NLPCA to the diffusion map embedding. (c) Reproduction of panel (b) visualizing only the 9963 data points derived from the forward trajectory colored by $\Delta\Upsilon_1^*(t_i) = \Upsilon_1^*(t_i+\tau) - \Upsilon_1^*(t_i)$. (d) Reproduction of panel (b) visualizing only the 9963 data points derived from the backward trajectory colored by $\Delta\Upsilon_1^*(t_i)$.

pathways over the reconstructed intrinsic manifold and gives rise to the observed reflection symmetry across the Υ_1^* axis in Fig. 10(b).

We will now describe a procedure to eliminate the temporal symmetry breaking artificially introduced by the delay embedding. Given a trajectory from a dynamical system known to obey a particular symmetry, additional trajectories can be generated "for free" by applying the symmetry operation to the observed trajectory [66–68]. Our molecular simulation evolves according to Newton's equations of motion, the time reversibility of which make the time-reversed simulation an equally valid system trajectory. Following Refs. [66,67], we can double our data by concatenating the forward and reverse trajectories, and then exploit the fact that our system is in thermodynamic equilibrium to appeal to detailed balance to retain the collapse and extension pathways that are coincident on the reconstructed manifold.

Specifically, we take the $K' = 9963$ delay vectors defined by Eq. (1), and invert the order of the elements to generate the delay vector produced by the time-reversed simulation, $\vec{v}(t_i) = [\ell(t_i+(d-1)\tau),\ldots,\ell(t_i+\tau),\ell(t_i)]$, such that $\{\vec{v}(t_i)\}_{i=1}^{K'}$ defines the reconstructed intrinsic manifold of the backwards trajectory. We augment the delay embedding generated from the forward simulation trajectory with that produced by the backward trajectory to generate a combined ensemble of $2K' = 19\,926$ points, $\{\vec{y}(t_i),\vec{v}(t_i)\}_{i=1}^{K'}$, and apply diffusion maps to extract the three-dimensional embedding into $[\vec{\phi}_2^*,\vec{\phi}_3^*,\vec{\phi}_4^*]$ in Fig. 11(a). Application of h-NLPCA reveals 99.89% of the variance to reside in the top two nonlinear principal components, allowing us to generate the two-dimensional projection into $[\vec{\Upsilon}_1^*,\vec{\Upsilon}_2^*]$ in Fig. 11(b). This object is the augmented reconstructed intrinsic manifold recovered from the delay embedding generated from the combined forward and backward simulation trajectories.

Every forward delay vector, $\vec{y}(t_i)$, possesses a backwards partner, $\vec{v}(t_i)$, containing the same values of ℓ in reverse order. By analyzing these 9963 pairs, we find that each member of the pair is embedded with identical values of Υ_1^*, but their Υ_2^* coordinates differ in sign, such that $\vec{y}(t_i) \mapsto [\Upsilon_1^*,\Upsilon_2^*]$ and $\vec{v}(t_i) \mapsto [\Upsilon_1^*,-\Upsilon_2^*]$ (Fig. 12). Moreover, the flow over the manifold of the forward points, $\vec{y}(t_i)$, is counterclockwise [Fig. 11(c)], whereas that of the backwards points, $\vec{v}(t_i)$, is clockwise [Fig. 11(d)]. In sum, upon reversing the arrow of time, delay embedding vectors extracted during a collapse event containing successively smaller values of ℓ with a negative value of Υ_2^*, have become observations from an

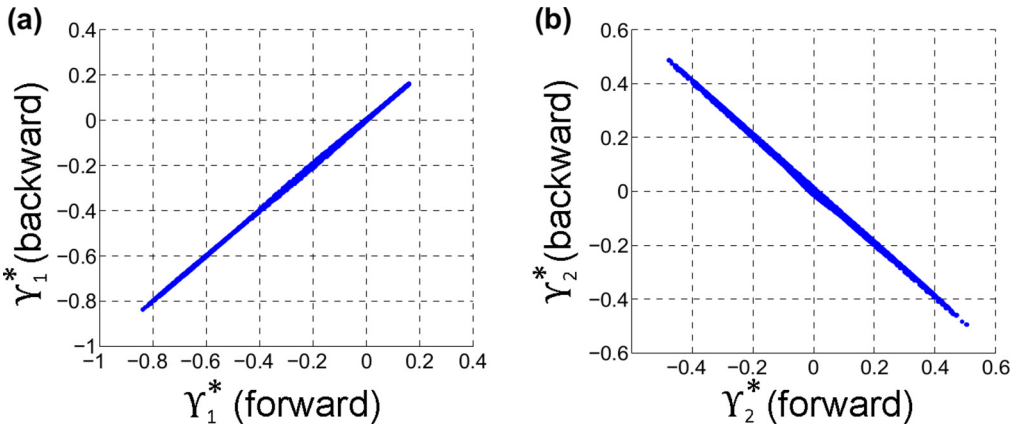

FIG. 12. Comparison of the $[\vec{\Upsilon}_1^*, \vec{\Upsilon}_2^*]$ coordinates of the forward, $\{\vec{y}(t_i)\}$, and backward, $\{\vec{v}(t_i)\}$, delay embeddings in the reconstructed manifold in Fig. 11(b). The embedding of each forward delay vector, $\vec{y}(t_i)$, and its backward delay vector partner, $\vec{v}(t_i)$, possess (a) identical values of Υ_1^*, and (b) sign-inverted values of Υ_2^*, such that $\vec{y}(t_i) \mapsto [\Upsilon_1^*, \Upsilon_2^*]$ and $\vec{v}(t_i) \mapsto [\Upsilon_1^*, -\Upsilon_2^*]$.

extension event containing successively larger values of ℓ with a positive value of Υ_2^*.

For a system at thermodynamic equilibrium, detailed balance asserts that every elementary process is equilibrated by its reverse process [65]. In the present case, each elementary step along a collapse pathway should be balanced by the reverse step along an extension pathway, and so the collapse and extension pathways must be coincident on the intrinsic manifold. We enforce detailed balance and eliminate the temporal symmetry breaking caused by the delay embedding by taking our reconstructed manifold from the combined forward and backward delay embeddings, and retaining from each pair $\{\vec{y}(t_i), \vec{v}(t_i)\}$ the one possessing the larger value of Υ_1^*. This procedure reduces our data back down to K' points, and

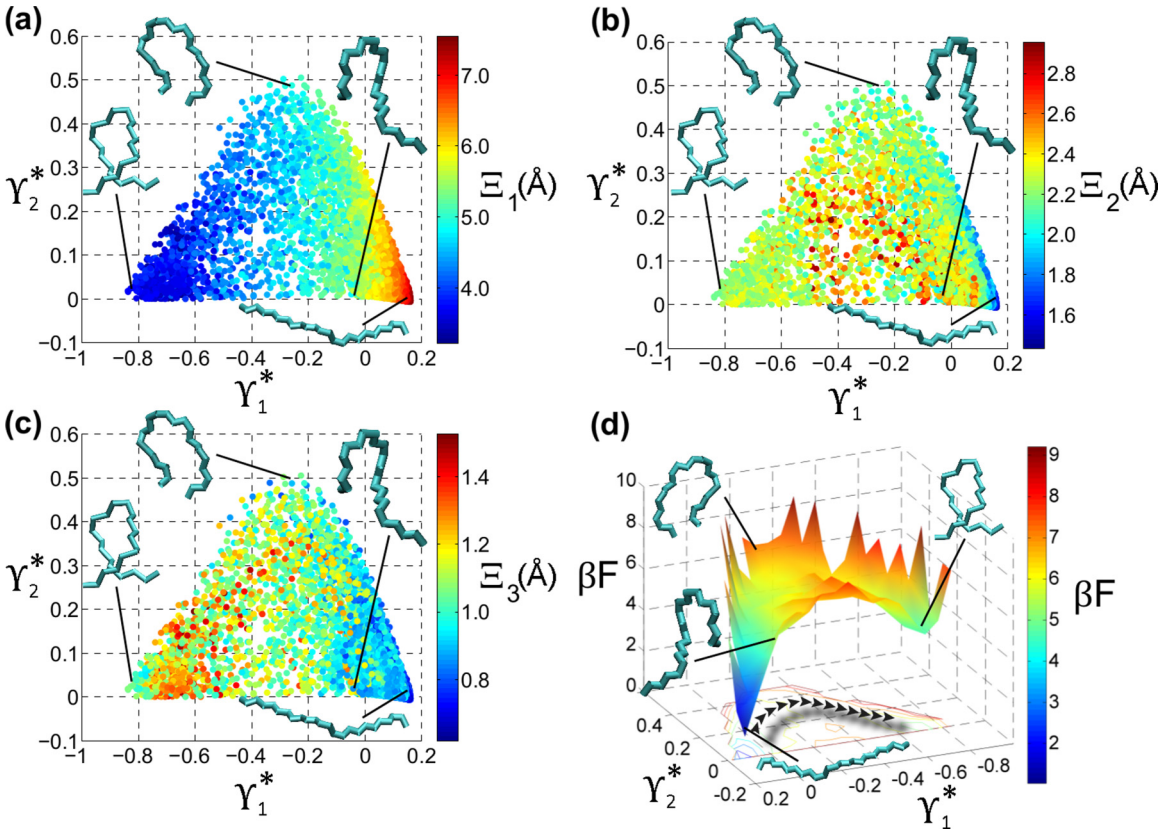

FIG. 13. Embedding of the temporally symmetrized ℓ delay embedding into the top two nonlinear principal components $[\vec{\Upsilon}_1^*, \vec{\Upsilon}_2^*]$ identified by sequential application of diffusion maps and h-NLPCA. Projection of the 9963 delay embedding vectors colored by (a) Ξ_1, (b) Ξ_2, and (c) Ξ_3. (d) The smFES $F(\vec{\Upsilon}_1^*, \vec{\Upsilon}_2^*)$ over which the "kink-and-slide" collapse pathway is indicated by chevrons. In the delay embedding vectors selected for visualization, the 10th of the 20 configurations constituting the delay embedding is visualized.

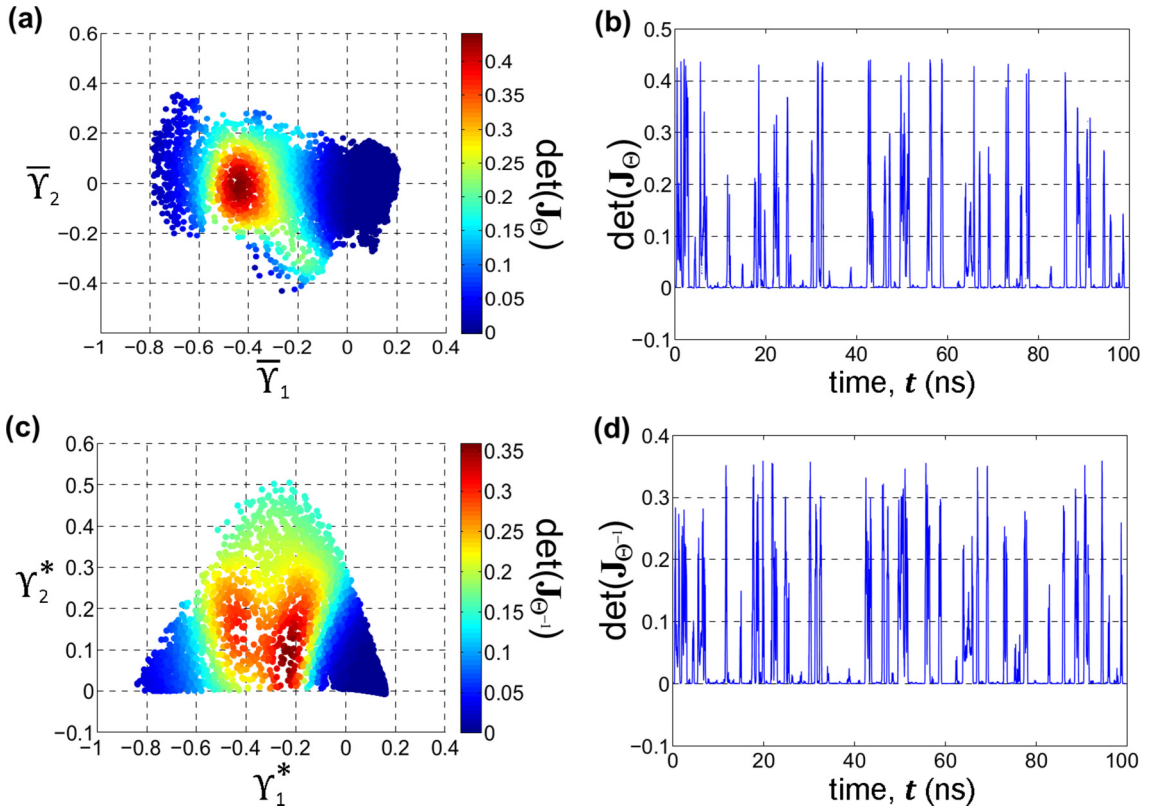

FIG. 14. Empirical validation of the existence of a diffeomorphism between the intrinsic manifolds recovered from the atomistic simulation trajectory and the ℓ delay embedding. (a) Determinant of the Jacobian, $\det(\mathbf{J}_\Theta)$, of the forward mapping, $\Theta : M \to \Theta(M)$, at each point on the 20-point averaged intrinsic manifold, M. (b) $\det(\mathbf{J}_\Theta)(t)$ for each 20-point averaged point on M expanded out as a linear time series for clarity of viewing. (c) Determinant of the Jacobian, $\det(\mathbf{J}_{\Theta^{-1}})$, of the reverse mapping, $\Theta^{-1} : \Theta(M) \to M$, at each point on the reconstructed manifold, $\Theta(M)$. (d) $\det(\mathbf{J}_{\Theta^{-1}})(t)$ for each point on $\Theta(M)$ expanded as a linear time series. Projected points with fewer then 40 neighbors within the bandwidth of the Gaussian kernel were not displayed due to insufficiently many neighbors to return a robust estimation of the Jacobian matrix elements. That both $\det(\mathbf{J}_\Theta)$ and $\det(\mathbf{J}_{\Theta^{-1}})$ remain single-signed verifies that the two manifolds are diffeomorphic.

its net effect is simple: the embedding of the chain extension events observed in the forward simulation trajectory—the lower half of the manifold in Fig. 11(c)—are reflected across the Υ_1^* axis to lie coincident with, but in the opposite sense to, the chain collapse events in the upper half of the manifold.

We note that the same result could have been approximately achieved by simply reflecting Fig. 10(b) across the Υ_1^* axis without going through the process of creating the combined forward and backward delay embedding. However, for finite data there is no guarantee that the forward trajectory alone will possess a symmetry plane exactly coincident with this axis. In contrast, augmenting the data with its reverse delay embedding provides each point with a temporally symmetric partner and offers a systematic procedure to unambiguously detect any temporal symmetries and guarantee a precise plane of reflection. For example, the intrinsic manifold in Fig. 11(b) possesses a single symmetry plane along the Υ_1^* axis, revealing the existence of precisely one temporal symmetry and providing a means to remove it by performing the reflection. We anticipate that this protocol will prove particularly useful in applications to systems possessing multiple stable states and/or higher-dimensional intrinsic manifolds.

We illustrate in Figs. 13(a)–13(c) the resultant temporally symmetrized intrinsic manifold, $\Theta(M)$. To aid in visual interpretation of the landscapes, we superpose onto selected points in the delay embedding the 10th of the 20 configurations constituting the delay vector. This configurational information is available from our molecular simulation trajectories, but would typically be unavailable in an application to experimental data. In practice, the values of the physical observable constituting the delay embedding vectors can reveal coarse-grained features of the molecule as it moves over the reconstructed landscape. For example, in the present case a knowledge of the head-to-tail extent permits identification of the extended, partially collapsed, and fully collapsed states of the molecule and the folding pathway connecting them. The distribution of $\{\xi_1,\xi_2,\xi_3\}$ over the spatially symmetrized manifold, M, recovered from the atomistic trajectory in Figs. 7(a)–7(c), is visually consistent with that of $\{\Xi_1,\Xi_2,\Xi_3\}$ over the temporally symmetrized manifold, $\Theta(M)$, recovered from the delay embeddings in Figs. 13(a)–13(c). The topology and topography of the smFES over M [Fig. 7(d)] and $\Theta(M)$ [Fig. 13(d)] also appear similar, possessing a single global free energy minimum corresponding to extended chain configurations connected by a "kink-and-slide" pathway to a shallow local minimum containing the collapsed helical coils. In the next section we will verify that the two landscapes are topologically identical.

E. Topological and geometric equivalence of smFES

Takens' theorem asserts that the manifolds M and $\Theta(M)$ supporting the spatially and temporally symmetrized smFES in Figs. 7(d) and 13(d) should be topologically equivalent, such that one may be continuously and smoothly transformed into the other. Mathematically, the two manifolds are related by a diffeomorphism such that Θ is an invertible mapping, and Θ and Θ^{-1} are both smooth. By the inverse function theorem, if there exists a one-to-one correspondence between points on the manifolds, and the Jacobian determinant of the coordinate transformation relating the two manifolds does not change sign, then the manifolds are globally diffeomorphic [52,53,64,69]. By computing the Jacobian transformation between the manifolds we will empirically confirm the existence of this diffeomorphism and verify the topological equivalence of the smFES recovered from the ℓ delay embedding to that computed from a complete knowledge of all molecular degrees of freedom.

The delay embedding defined by Eq. (1) maps 20 ℓ observations of the molecular system in real space into a single point in the delay embedding, $\vec{y}(t_i) = [\ell(t_i), \ell(t_i + \tau), \ldots, \ell(t_i + 19\tau)]$. Moreover, this embedding results in 9963 points in the delay embedding compared to 10 001 in the simulation trajectory. In order to draw a one-to-one correspondence between the points composing M and $\Theta(M)$, we define the following mapping,

$$\overrightarrow{[\vec{p}(t_i), \vec{p}(t_i + \tau), \ldots, \vec{p}(t_i + 19\tau)]} = \vec{P}(t_i) \mapsto \vec{p}^{*}(t_i), \quad (2)$$

where $\vec{p}(t_i) = [\vec{\Upsilon}_1(t_i), \vec{\Upsilon}_2(t_i)]$ represents the coordinates on the manifold M of the chain configuration extracted from the molecular simulation at time t_i, $\vec{P}(t_i)$ is the average of the 20 coordinates $[\vec{p}(t_i), \vec{p}(t_i + \tau), \ldots, \vec{p}(t_i + 19\tau)]$, and $\vec{p}^{*}(t_i) = [\vec{\Upsilon}_1^{*}(t_j), \vec{\Upsilon}_2^{*}(t_j)]$ represents the coordinates on the manifold $\Theta(M)$ of the delay embedding corresponding to the 20 observations of the head-to-tail chain distance $[\ell(t_i), \ell(t_i + \tau), \ldots, \ell(t_i + 19\tau)]$. The terminal snapshots from the molecular simulation trajectory for which $\vec{P}(t_i)$ is undefined are removed from M. In this manner we define an unambiguous mapping between 9963 points in M and $\Theta(M)$.

Having defined the one-to-one mapping, we now numerically compute for every point the Jacobians, $\mathbf{J}_{\Theta}$ and $\mathbf{J}_{\Theta^{-1}}$, of the forward, $\Theta : M \to \Theta(M)$, and reverse, $\Theta^{-1} : \Theta(M) \to M$, mappings (Methods Summary). In Figs. 14(a)–14(b) we illustrate the Jacobian determinant of the forward transformation, $\det(\mathbf{J}_{\Theta})$, at each point on the 20-point averaged manifold M, and in Figs. 14(c)–14(d) that of the reverse transformation, $\det(\mathbf{J}_{\Theta^{-1}})$, at each point on $\Theta(M)$. The magnitude of the Jacobian determinant gives the factor by which the local region is scaled under the transformation, and the sign indicates whether or not the orientation is preserved. That $\det(\mathbf{J}_{\Theta})$ and $\det(\mathbf{J}_{\Theta^{-1}})$ remain single-signed over their respective manifolds indicates that a smooth and invertible transformation exists at each point on the manifold, providing empirical validation that the manifolds are diffeomorphic.

III. CONCLUSIONS

We have integrated delay embeddings with nonlinear dimensionality reduction techniques to recover from molecular simulations a representation of the single-molecule free-energy surface of an n-tetracosane chain in water from measurements of only the head-to-tail distance of the chain. Subject to the elimination of spatial symmetries associated with our choice of the measurement observable, and temporal symmetry breaking induced by the delay embedding, we have verified that the smFES recovered in this manner is geometrically and topologically equivalent to that recovered from a trajectory in which the temporal evolution of all molecular degrees of freedom are known.

This work demonstrates that topologically equivalent representations of single-molecule free-energy surfaces can be extracted from the analysis of univariate time series, laying the foundations for the inference of biomolecular folding landscapes directly from experimental measurements. Much work, however, remains to be done. We considered the idealized case of a simple homopolymer chain for which the spatial symmetries to be eliminated given our choice of measurement observable were clear, and which possessed a single temporal symmetry. Furthermore, we analyzed a continuous, noise-free 100 ns time series with 10 ps resolution. Delay embeddings of short, noisy, low-resolution, and temporally disjoint experimental smFRET trajectories will require careful processing, and the impact of these factors upon the resultant smFES remains to be ascertained. In future work, we plan to (i) extend our study to molecular simulations of peptides and proteins, (ii) explore multichannel measurements of several observables, none of which may, in itself, be generic, (iii) examine the impact of the temporal resolution of the time series on the reconstruction fidelity, and (iv) confront the influence of noise by artificially contaminating our simulated scalar time series to lay empirical bounds on tolerable signal-to-noise ratios. Finally, Takens' theorem asserts the existence of a diffeomorphism between the true smFES and that recovered from delay embeddings, but the transformation itself is not supplied. Although the topology of the landscape is maintained, interpretation of its topography (i.e., the height of the free-energy wells and barriers) under the action of an unknown Jacobian presents a challenge. We are currently working to place limits on the degree of stretching/squashing of the smFES under the diffeomorphic transformation induced by the delay embedding under different choices of physical observable and delay embedding parameters both theoretically, using tools from real analysis and probability theory, and empirically, by conducting molecular simulations of more biologically realistic systems.

IV. METHODS SUMMARY

Our theoretical and computational methods are summarized below. Full details of the molecular simulations, phase space reconstruction, dimensionality reduction, and diffeomorphism validation are provided in Appendix A.

Molecular simulations. Molecular dynamics simulations were conducted using the GROMACS 4.6 simulation suite [70] employing the TraPPE potential [71] for n-tetracosane and the SPC model of water [72]. The PRODRG2 server assisted in the construction of chain topologies [73]. Lennard-Jones interactions were shifted smoothly to zero at 1.4 nm,

and Lorentz-Berthelot combining rules used to determine dispersion interactions between unlike atoms [74]. Electrostatic interactions were treated using particle mesh Ewald with a real-space cutoff of 1.4 nm and a 0.12 nm reciprocal-space grid spacing [75]. Simulations were maintained at 298 K and 1 bar using a Nosé-Hoover thermostat [76] and an isotropic Parrinello-Rahman barostat [77]. Equations of motion were integrated using the leap-frog algorithm [78] with a 2 fs time step, and the system equilibrated for 1 ns before performing a 100 ns production run. System configurations were saved every 10 ps.

Phase space reconstruction. Given a dynamical system that evolves over a k-dimensional manifold M, and a univariate time series in a generic measurement function $v : \mathbb{R}^k \to \mathbb{R}$, $\{v(t)\}_{t=0}^{T}$, Takens' theorem asserts that the state of the system is uniquely specified by a $d \geqslant (2k+1)$-dimensional delay embedding $\vec{y}(t) = \Theta(v(t)) = [v(t), v(t+\tau), v(t+2\tau), \ldots, v(t+(d-1)\tau)]$, where $\Theta : M \to \Theta(M)$ is a diffeomorphism, defining an invertible function mapping the manifold M to a geometrically and topological equivalent embedding, $\Theta(M)$, in the d-dimensional Euclidean space [29–31,33,34]. In practice, $k < d < (2k+1)$ can be sufficient to uniquely specify the system state [30,60]. Employing the head-to-tail distance, ℓ, of the chain as our univariate time series, we use the mutual information approach of Fraser and Swinney [61] to select an appropriate delay time of $\tau = 20$ ps, and the false nearest neighbors approach of Cao [62] to select an appropriate delay embedding dimensionality of $d = 20$.

Dimensionality reduction. We apply diffusion maps [39,40] followed by hierarchical nonlinear principal components analysis (h-NLPCA) [44,45] to (i) the 72-dimensional molecular dynamics trajectories of Cartesian coordinates of the n-tetracosane united atoms to extract the intrinsic manifold $M \in \mathbb{R}^2$, and (ii) the 20-dimensional delay embeddings of ℓ to extract the reconstructed intrinsic manifold $\Theta(M) \in \mathbb{R}^2$. The diffusion map is a nonlinear machine learning approach to extract low-dimensional nonlinear manifolds resident within high-dimensional spaces [3,39]. By performing a spectral analysis of a discrete random walk over K high-dimensional observations in $\mathbb{R}^D$, the diffusion map infers a low-dimensional mapping into $\mathbb{R}^k$ with $k < D \ll K$: observation$_i \mapsto [\vec{\phi}_2(i), \vec{\phi}_3(i), \ldots, \vec{\phi}_{k+1}(i)]$. The $\{\vec{\phi}_j\}_{j=1}^{K}$ constitute the eigenvectors of the discrete random walk, with associated eigenvalues $\{\lambda_j\}_{j=1}^{K}$. By the nature of the random walk, the top pair is trivial ($\vec{\phi}_1 = \vec{1}$, $\lambda_1 = 1$). A gap in the eigenvalue spectrum defines an appropriate number of eigenvectors, k, to incorporate in the embedding. We have previously shown that diffusion map embedding can contain functional dependencies between the embedding variables $\{\vec{\phi}_j\}_{j=2}^{k+1}$ [3]. We employ h-NLPCA [44,45] to identify and eliminate such dependencies and achieve lower-dimensional representations of M and $\Theta(M)$ beyond that attainable by diffusion maps alone.

Diffeomorphism validation. Takens' theorem asserts that $\Theta : M \in \mathbb{R}^2 \to \Theta(M) \in \mathbb{R}^2$ is a diffeomorphism, such that M and $\Theta(M)$ are geometrically and topologically equivalent manifolds related by a smooth and invertible transformation [29–31,33,34]. By the inverse function theorem, a global diffeomorphism exists if the mapping $\Theta : M \to \Theta(M)$ is bijective and its Jacobian determinant, $\det(\mathbf{J}_\Theta)$, does not pass through zero [52,64]. Using a mesh-free approach based on a smoothed-particle hydrodynamics formulation to estimate the partial derivatives constituting the elements of $\mathbf{J}_\Theta$ [79,80], we empirically verify the existence of this global diffeomorphism, proving that the single-molecule free-energy surface over $\Theta(M)$ is geometrically and topologically equivalent to that over M.

ACKNOWLEDGMENTS

We thank Prof. Haw Yang for generous discussions and encouragement at the inception of this work, and Prof. Lee DeVille and Prof. Seppe Kuehn for fruitful discussions and useful suggestions. This work was partially supported by a grant from the Initiative for Mathematical Sciences and Engineering at the University of Illinois at Urbana-Champaign.

APPENDIX A: MATERIALS AND METHODS

1. Molecular dynamics simulations of *n*-tetracosane

Following Ref. [3], we performed molecular dynamics simulations of a coarse-grained n-tetracosane ($C_{24}H_{50}$) chain in water using the GROMACS 4.6 simulation suite [70]. Initial chain configurations were constructed using the GlycoBioChem PRODRG2 server [73,84] and modeled using the TraPPE potential [71], which represents each CH_2 and CH_3 group as a single united atom. Accordingly, the chain configuration is completely specified by a $(3\times24 = 72)$-dimensional vector specifying the Cartesian coordinates of each united atom. Simulations were initialized by placing the chain in a $5\times5\times5$ nm cubic box with periodic boundary conditions, and solvating to a density of 1.0 g/cm^3 by 4117 water molecules modeled by the SPC potential [72]. High-energy overlaps were removed by steepest descent energy minimization to eliminate forces exceeding 2000 kJ/mol nm. Lennard-Jones interactions were shifted smoothly to zero at 1.4 nm, and Lorentz-Berthelot combining rules used to determine dispersion interactions between unlike atoms [74]. The 5 nm cubic box size was sufficiently large that the n-alkane chain did not interact with itself through the periodic boundary, even in a fully extended all-*trans* configuration. Electrostatic interactions were treated using particle mesh Ewald (PME) with a real-space cutoff of 1.4 nm and a 0.12 nm reciprocal-space grid spacing [75]. Simulations were conducted in the *NPT* ensemble at 298 K and 1 bar using a Nosé-Hoover thermostat [76] and an isotropic Parrinello-Rahman barostat [77]. Equations of motion were numerically integrated using the leap-frog algorithm [78] employing a 2 fs time step. As required by the TraPPE and SPC potentials, bond lengths were fixed to their equilibrium values using the LINCS algorithm [86]. The system was subjected to a 1 ns equilibration run, after which time the temperature, pressure, and energy had all attained stable average values, before conducting a 100 ns production run. System configurations were saved every 10 ps to generate a molecular dynamics trajectory comprising 10 001 snapshots.

2. Phase space reconstruction

A molecular system can be considered—provided that electronic degrees of freedom are not relevant on the time scales of interest—as a dynamical system evolving according to the laws of classical mechanics. The phase space of the system defines the ensemble of accessible system states in the high-dimensional space spanned by the Cartesian coordinates of all atoms in the system. As we shall describe below, cooperative couplings between molecular degrees of freedom can cause the accessible phase space to define a relatively low-dimensional structure that can be extracted by the application of nonlinear manifold learning to molecular simulation trajectories [3]. Experimentally, we typically do not have a complete knowledge of all system degrees of freedom, instead possessing measurements of a small number of experimental observables. Attractor reconstruction methods seek to infer the geometry and topology of the phase space of the dynamical system from few observables, typically without requiring any knowledge of the underlying governing equations. These approaches are made possible by two theorems. The Whitney embedding theorem [87] states that $(2k+1)$ independent measurements of a k-dimensional dynamical system unambiguously specify the system state. More precisely, the mapping from the k-dimensional manifold, M, upon which the system evolves into the $(2k+1)$ Euclidean space, U, is surjective and structure preserving, defining an *embedding* of the manifold. An embedding is a smooth and invertible map, Θ, such that $\Theta(M)$ is a geometrically and topologically equivalent "realization" of M in the space U [32]. In other words, $\Theta(M) \in U$ is a fully unfolded image of M, with each point on M uniquely located on $\Theta(M)$, and $\Theta(M)$ is a reconstruction of the phase space of the system [58]. Takens' delay embedding theorem [29–31,33,34] builds on Whitney to show that the embedding can be constructed from a single measurement function, $v : \mathbb{R}^k \to \mathbb{R}$, that produces a univariate time series, $v(t)_{t=0}^{T}$—or its discrete analog, $\{v(t_i)\}_{i=1}^{K}$, where K is the number of evenly spaced time points—by forming a *delay embedding*,

$$\begin{aligned}\vec{y}(t) &= \Theta(v(t)) \\ &= [v(t), v(t+\tau), v(t+2\tau), \ldots, v(t+(d-1)\tau)], \end{aligned} \quad \text{(A1)}$$

where τ is the delay time between successive system observations and d is the delay embedding dimensionality. The theorem guarantees that if $v(t)$ is a generic observable (i.e., a function that depends on all system degrees of freedom and contains no symmetries that are not present in the system being observed [50]) and $d \geqslant (2k+1)$ then (i) the $\vec{y}(t)$ define an embedding, and therefore a reconstruction, of the k-dimensional phase space of the dynamical system [58], (ii) the dynamical evolution of the system on M is C^1-equivalent to that on $\Theta(M)$, and (iii) $\Theta : M \to \Theta(M)$ is a diffeomorphism, an invertible function mapping M to $\Theta(M)$ such that both Θ and Θ^{-1} are smooth [32], and implying the geometrical and topological equivalence of M and $\Theta(M)$ such that one may be smoothly and invertibly transformed into the other. The existence of a diffeomorphism is not assured, but possible nonetheless, for $k < d < (2k+1)$ [30,60]. The great value of Takens' theorem is that, in principal, it permits attractor reconstruction from a single system measurement. In practice, it can be challenging to determine appropriate values of τ and d, and to confront issues of sampling noise, finite data, and weak dependence of the observable on one, or more, system degrees of freedom [32,58].

In this work, we adopt as our univariate system observable in which to construct delay embeddings the head-to-tail distance, ℓ, of the n-tetracosane chain. This observable constitutes an observable that can be, in principal, experimentally measured by single-molecule FRET [21]. As we discuss in the main text, this measurement function does not satisfy the criterion of a generic observable, since it is invariant to two symmetries of the n-alkane chain [50]: (i) head-to-tail inversion, and (ii) mirror symmetry. In other words, this observable cannot distinguish (i) the head-to-tail directionality of the chain, or (ii) the right- or left-handedness of chiral chain conformations. We confront this difficulty by removing these symmetries in the space of the real space chain dynamics such that the delay embedding can provide a reconstruction of the spatially symmetrized phase space.

Takens' theorem holds for any value of the delay time, τ, but, in practice, finite trajectories and sampling noise make the quality of the reconstruction strongly dependent on the choice of τ [61]. Following Fraser and Swinney [61], we select an appropriate value of τ by computing the mutual information, I, between measurements of the head-to-tail distance at times t and $(t+\tau)$,

$$I(\ell(t),\ell(t+\tau)) = \sum_t P(\ell(t),\ell(t+\tau)) \log_2 \left(\frac{P(\ell(t),\ell(t+\tau))}{P(\ell(t))P(\ell(t+\tau))} \right), \quad \text{(A2)}$$

where $P(\ell(t),\ell(t+\tau))$ is the joint probability distribution function for $\ell(t)$ and $\ell(t+\tau)$, and $P(\ell(t))$ is the probability distribution function for $\ell(t)$. The values of ℓ observed over the course of the simulation lie within the range 0.3537–2.8845 nm, and we estimate the probability distributions using a bin size of 0.5 nm. As shown in Fig. 15(a), $I(\ell(t),\ell(t+\tau))$ monotonically decreases with τ as knowledge of the value of $\ell(t)$ becomes progressively less informative of $\ell(t+\tau)$. Fraser and Swinney suggest as a good delay time the value of τ corresponding to the first local minimum in the mutual information [61]. In the absence of a minimum, we instead follow Kantz and Schreiber to select the value of τ at which the mutual information falls to $1/e$ of its initial value [49], leading us to select $\tau = 20$ ps.

Having chosen a suitable delay time, we use the approach of Cao [62] based on the false nearest neighbors method of Kennel *et al.* [63] to determine an appropriate delay embedding dimensionality, d, as the minimum dimensionality at which the reconstructed phase space becomes fully unfolded [cf. Eq. (A1)] [62]. Too low a delay embedding dimensionality causes points far apart on the original manifold, M, to be artificially proximate in the reconstructed manifold, $\Theta(M)$, due to self-intersections of an incompletely unfolded reconstruction image. When the nearest neighbors of each point in the reconstruction no longer change with increasing embedding dimensionality, the attractor is fully unfolded. The minimum value of d at which this behavior is observed is an appropriate choice of embedding dimensionality. Cao defined

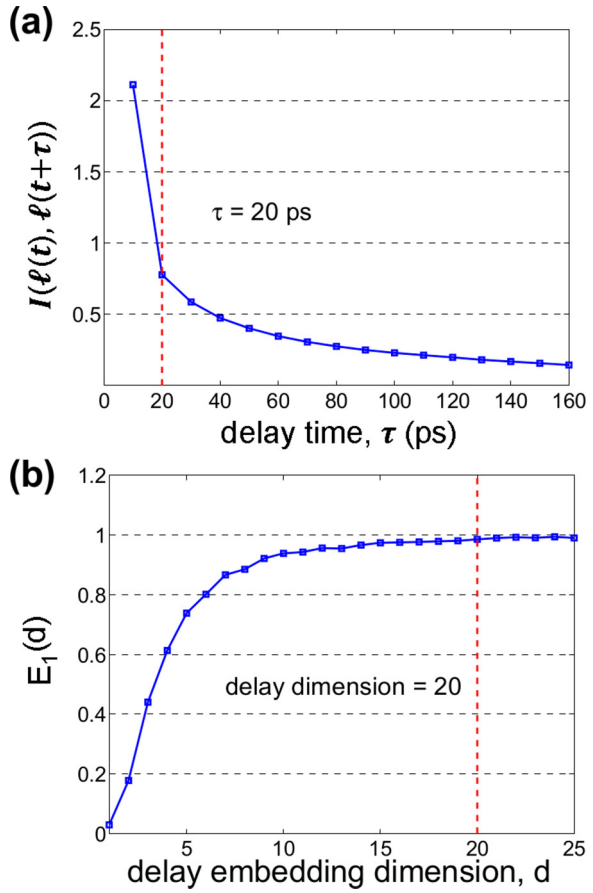

FIG. 15. Empirical selection of delay time, τ, and delay embedding dimensionality, d. (a) The mutual information in the ℓ signal. We choose $\tau = 20$ ps corresponding to the delay time at which $I(\ell(t),\ell(t+\tau))$ drops to $1/e$ of its initial maximum, identifying the minimum period of time beyond which subsequent measurements of ℓ contain significant "new" information about the system. (b) The characteristic function $E_1(d)$, at a delay time of $\tau = 20$ ps, measures the number of false nearest neighbors of points in the reconstructed embedding as a function of the delay embedding dimensionality. Once the phase space reconstruction has been fully unfolded, $E_1(d)$ saturates to unity, motivating our choice of $d = 20$.

a characteristic function, $E_1(d)$ (see Ref. [62] for details), that provides a measure of the number of false nearest neighbors as a function of d, and—in the case of deterministic processes—saturates at unity once the phase space reconstruction is fully unfolded [35,62]. As illustrated in Fig. 15(b), we find $E_1(d)$ to reach saturation beyond a delay embedding dimensionality of $\sim$20, motivating us to choose $d = 20$.

3. Diffusion maps

The diffusion map [39,40,81,82] is a nonlinear manifold learning technique that has been previously employed by ourselves and others to infer low-dimensional parametrizations of the free-energy surface for polymers, biomolecules, and colloids [3,13,14,16,17,88]. Linear approaches, such as principal components analysis (PCA) [89], are restricted to seek low-dimensional hyperplanes parametrizing the data in the high-dimensional space. Nonlinear approaches can discover convoluted and curvilinear manifolds [6], which is of particular value in applications to polymers and macromolecules possessing complex couplings (e.g., covalent bonds, dispersion interactions, the hydrophobic effect) between their degrees of freedom [3,4,6,14]. In this work, we employ diffusion maps to discover low-dimensional nonlinear parametrizations within (i) 72-dimensional molecular dynamics simulation trajectories recording the Cartesian coordinates of each united atom in a n-tetracosane chain, and (ii) 20-dimensional Takens' delay embeddings of the head-to-tail distance of the chain recorded over the course of the simulation trajectory.

We have previously described the application of diffusion maps to molecular simulations in Refs. [3,6]. In brief, given an ensemble of K observations in D-dimensional space, we first compute the $K \times K$ pairwise distances matrix, $\mathbf{P}$, the elements P_{ij} of which hold the pairwise distances between observations i and j. In the application of diffusion maps directly to the molecular dynamics simulation of n-tetracosane, the observations correspond to the 72-dimensional vectors recording the Cartesian coordinates of the 24 united atoms. Following previous studies, we adopt as our distance metric the root mean squared distance (RMSD) between the united atom coordinates of pairs of configurations translationally and rotationally aligned using the Kabsch algorithm [3,6,14,90]. Our molecular dynamics simulations explicitly represent the solvent molecules surrounding the n-tetracosane chain, but it is a challenge to explicitly incorporate solvent degrees of freedom into the application of diffusion maps due to the identical and fungible nature of solvent molecules [69,91]. Instead, we implicitly capture the impact of the solvent degrees of freedom through their influence on the configurational ensemble sampled by the chain over the course of the simulation [3,6]. We have recently proposed a means to explicitly incorporate many-body effects into the mapping [16], but this methodology has yet been applied to realistic molecular systems. In the application of diffusion maps to delay embeddings of the head-to-tail distance, ℓ, of the chain, the observations correspond to 20-dimensional vectors recording 20 sequential observations of the chain length. In this case we adopt the Euclidean to measure pairwise distances between the vectors.

In the next step, we form the matrix $\mathbf{A}$ by convoluting the elements of the pairwise distances matrix $\mathbf{P}$ with a Gaussian kernel of bandwidth ϵ,

$$A_{ij} = \exp\left(-P_{ij}^2/2\epsilon\right), \quad i,j = 1,\ldots,K. \tag{A3}$$

An appropriate bandwidth is systematically defined using the procedure detailed in Ref. [83]. The diagonal matrix, $\mathbf{D}$, is computed from the row sums of $\mathbf{A}$,

$$D_{ii} = \sum_{j=1}^{K} A_{ij}, \quad i = 1,\ldots,K, \tag{A4}$$

and the right-stochastic Markov matrix $\mathbf{M}$ formed as the matrix product,

$$\mathbf{M} = \mathbf{D}^{-1}\mathbf{A}, \tag{A5}$$

defines a discrete random walk over the observations with a characteristic step size of ϵ [39]. By analyzing the spectral properties of this process we can discover low-dimensional

structures within the high-dimensional data [3,39,40]. Specifically, $\mathbf{M}$ diagonalizes as [3],

$$\mathbf{M} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Psi}^T, \tag{A6}$$

where $\mathbf{\Lambda}$ is a diagonal matrix holding the eigenvalues, $\lambda_1 = 1 \geqslant \lambda_2 \geqslant \ldots \lambda_K$, with associated left, $\mathbf{\Psi} = \{\vec{\psi}_i\}_{i=1}^K$, and right, $\mathbf{\Phi} = \{\vec{\phi}_i\}_{i=1}^K$, column eigenvectors, which form a biorthogonal set, $\mathbf{\Psi}^T\mathbf{\Phi} = \mathbf{1}$ [3]. $\lambda_1 = 1$ and $\vec{\phi}_1 = \vec{1}$ by the Markov property. By expanding an arbitrary initial probability distribution over the data into the basis of the left eigenvectors, $\vec{p}_0 = \sum_{j=1}^K \alpha_j \vec{\psi}_j$, the distribution after k steps of the discrete diffusion process can be written as $\vec{p}_k = \sum_{j=1}^K \alpha_j \lambda_j^k \vec{\psi}_j$. From this expression $\vec{\psi}_1$ (with $\lambda_1 = 1$) is identifiable as the equilibrium distribution, and higher eigenvectors as transient modes with increasingly faster relaxation times.

If the system admits a description as a diffusion process, a gap within the eigenvalue spectrum can be identified after λ_{k+1}. The slow relaxations of the distribution over the data are defined by the leading $(k+1)$ eigenvectors, to which the remaining modes are effectively slaved [3,39,83]. Geometrically, the leading eigenvectors define a low-dimensional subspace, known as the intrinsic manifold, to which the data are effectively restrained. The *diffusion map* defines the embedding of the ith observation into the ith component of the leading k nontrivial right eigenvectors of $\mathbf{M}$,

$$\text{observation}_i \mapsto [\vec{\phi}_2(i), \vec{\phi}_3(i), \ldots, \vec{\phi}_{k+1}(i)], \tag{A7}$$

where $\vec{\phi}_1$ is dropped as the trivial all-ones vector. For $k < D \ll K$, the diffusion map achieves dimensionality reduction by defining a projection of our K observations onto a low-dimensional (nonlinear) intrinsic manifold discovered within the high-dimensional space. Under the assumptions that the system dynamics are well approximated by a diffusion process, and the pairwise distance metric is a good measure of short-time diffusive motions, then the diffusion map embedding possesses two valuable properties: (i) Euclidean distances in the embedding are equivalent to *diffusion distances* in the original space, defining the time required for the system to evolve from one state to another, and (ii) the eigenvectors spanning the diffusion map embedding are identifiable as the slow dynamical modes governing the long-time evolution of the system [3,6,39,82].

The diffusion map defines a projection of the data onto an intrinsic manifold in $\mathbb{R}^k$. By compiling histograms of the observed distribution of points over the manifold, $\hat{P}(\{\vec{\phi}_i\}_{i=2}^{k+1})$, the free-energy profile over the surface is estimated as $F(\{\vec{\phi}_i\}_{i=2}^{k+1}) = -k_B T \ln \hat{P}(\{\vec{\phi}_i\}_{i=2}^{k+1}) + C$, where k_B is Boltzmann's constant, T is the temperature, and C is an arbitrary constant. This hypersurface in $\mathbb{R}^{k+1}$ defined by the application of diffusion maps to the molecular simulation trajectory is the smFES [3,6,14]. Appealing to Takens' theorem, an equivalent representation of the smFES, related by a smooth and invertible transformation, is obtained by the application of diffusion maps to delay embeddings of the molecular head-to-tail distance.

4. Hierarchical nonlinear principal components analysis

As we have previously reported, although the eigenvectors, $\{\vec{\phi}_i\}_{i=1}^k$, spanning the diffusion map embedding are orthogonal, two (or more) eigenvectors can correspond to the same dynamical mode of the system [3]. We have previously drawn the analogy with multivariate Fourier series wherein $\sin(x)$ and $\sin(2x)$ are orthogonal Fourier components oriented in the same spatial direction [3]. Such dependencies are detectable as approximately one-dimensional projections of the manifold in particular eigenvector pairs, and we have eliminated this redundancy by successively replacing functionally dependent pairs of eigenvectors by the arclength of the one-dimensional curve mapped out by the projection of the data into their subspace [3]. This procedure is valuable in providing further dimensionality reduction beyond that furnished by the diffusion map by elimination of redundancies between collective variables in the low-dimensional embedding.

In this work, we adopt a more sophisticated approach to eliminate these functional dependencies using hierarchical nonlinear principal components analysis (h-NLPCA) approach developed by Scholz and Vigário [44,45]. This approach offers several benefits over the replacement of redundant eigenvector pairs by their arclength in that it can be applied to simultaneously eliminate multidimensional and nonlinear functional dependencies (i.e., to recover a q-dimensional surface in the subspace of p eigenvectors, where $p > q$), is straightforward to apply in an automated fashion, and explicitly quantifies the degree of information loss in the dimensionality reduction through the fraction of variance explained.

The h-NLPCA algorithm may be considered a nonlinear analog of standard principal components analysis (PCA) [89] in that it seeks to infer a hierarchically ordered set of (nonlinear) principal components in the sense that the top q components explain the maximum possible variance within a q-dimensional nonlinear projection [45]. This hierarchical dimensionality reduction is achieved using a multilayer perceptron with an autoassociative topology, commonly known as an autoencoder [45]. The topology of the h-NLPCA autoencoder is illustrated in Fig. 16. For a p-dimensional data set, we establish p nodes in the input, bottleneck, and output layers, and $r > p$ nodes in the mapping and demapping layers. The input, bottleneck, and output layers employ linear activation functions, while mapping and demapping layers employ the nonlinear tanh activation function [45]. In this work, we are interested in identifying and eliminating eigenvector redundancies in three-dimensional diffusion map embeddings, so we set $p = 3$ and $r = 8$.

The network is trained (i.e., the parameters of the linear and nonlinear activation functions are tuned) to perform the identity mapping by enforcing that the output data, $\mathbf{x}'$, approximate the input data, $\mathbf{x}$, by minimizing the hierarchical error, $E_H = \sum_{i=1}^p E_i$, where $E_i = \frac{1}{2}||\mathbf{x}^{'(i)} - \mathbf{x}||_2^2$ is the squared reconstruction error of the network employing nodes $\{1,2,\ldots,i\}$ in the bottleneck layer, and $\mathbf{x}'^{(i)}$ are the output reconstructions of the input data produced by the network employing nodes $\{1,2,\ldots,i\}$ in the bottleneck layer. The parameters of the activation functions in the network are tuned to minimize E_H using conjugate gradient descent [44,45,92]. Following Ref. [44], the minimization is regularized by

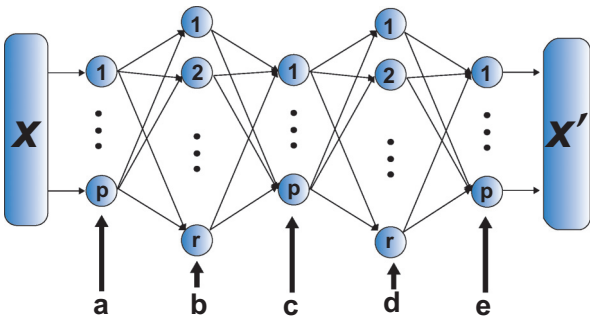

FIG. 16. Topology of the autoassociative multilayer perceptron (autoencoder) used to perform h-NLPCA. The input (a), bottleneck (c), and output (e) layers each possess a number of nodes, p, equal to the dimensionality of the input data, $\mathbf{x}$. The mapping (b) and demapping (d) layers contain $r > p$ nodes. The input, bottleneck, and output layers employ linear activation functions, while mapping and demapping layers employ the nonlinear tanh activation function. The autoencoder is trained to perform the identify mapping (i.e., the output of the neural network, $\mathbf{x}'$, is equal to its input, $\mathbf{x}$) by minimizing the hierarchical error, $E_H = \sum_{i=1}^{p} E_i$, using conjugate gradient descent.

applying an L_2 penalty to penalize large network weights and stabilize training of the autoencoder network parameters. Network construction and training is performed using the open-source "Nonlinear PCA toolbox for Matlab" developed by Scholz [45,85].

Dimensionality reduction $\mathbb{R}^p \to \mathbb{R}^q$ is achieved by projecting the p-dimensional input observation x into the output values of the first $\{1,2,\ldots,q\}$ nodes in the bottleneck layer. This dimensionality reduction is inherently hierarchical, since minimization of E_H guarantees that the squared reconstruction error of the nonlinear projection into the q-dimensional subspace is minimized subject to minimization of the squared reconstruction error in all $(i = 1,2,\ldots,q-1)$-dimensional subspaces [45]. We choose an appropriate value of q by searching for a gap in the spectrum of the fraction of variance explained as a function of the dimensionality of the nonlinear projection.

5. Empirical validation of diffeomorphism

Takens' embedding theorem asserts that for $d \geqslant (2k+1)$ and $v : \mathbb{R}^k \to \mathbb{R}$ a generic measurement function of a k-dimensional dynamical system, the delay embedding $\Theta : M \to \Theta(M)$, where $\Theta(v(t)) = [v(t), v(t+\tau), v(t+2\tau), \ldots, v(t+(d-1)\tau)]$, is a diffeomorphism (i.e., an invertible function such that the function, Θ, and its inverse, Θ^{-1}, are smooth) mapping the k-dimensional manifold M to a submanifold $\Theta(M)$ of the d-dimensional Euclidean space U [32]. We recover the manifold, M, by applying diffusion maps to the molecular dynamics simulation trajectory of the n-tetracosane chain, and its image, $\Theta(M)$, by applying them to the delay embedding of the head-to-tail distance of the chain.

By the inverse function theorem, if the determinant of the Jacobian, $\mathbf{J}_\Theta$, of the mapping $\Theta : \mathbb{R}^k \to \mathbb{R}^k$ does not change sign over the manifold M (i.e., does not pass through zero) and is bijective [i.e., there is a one-to-one correspondence between points on M and $\Theta(M)$], then at each point over the manifolds there exists an invertible map, and M and $\Theta(M)$ are globally diffeomorphic [52,53,64]. In the language of control theory, the system is *observable* from $v(t)$ since its delay embedding projection onto $\Theta(M)$ unambiguously specifies the system state on M [52]. Defining the $k \times k$ Jacobian matrix as

$$\begin{aligned}\mathbf{J}_\Theta(z_1,\ldots,z_k) &= \frac{\partial(F_1,\ldots,F_k)}{\partial(z_1,\ldots,z_k)} \\ &= \begin{bmatrix} \nabla_{\vec{z}} F_1 \\ \vdots \\ \nabla_{\vec{z}} F_k \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial F_1}{\partial z_1} & \cdots & \frac{\partial F_1}{\partial z_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_k}{\partial z_1} & \cdots & \frac{\partial F_k}{\partial z_k} \end{bmatrix},\end{aligned} \tag{A8}$$

where $\vec{z} = [z_1,\ldots,z_k]$ defines a point on M and $\vec{F} = [F_1,\ldots,F_k]$ the corresponding point on $\Theta(M)$ under the mapping. We will show that the two manifolds we recover by diffusion maps are diffeomorphic [i.e., $\Theta(M)$ can be obtained by a continuous and smooth transformation of M, and vice versa] by demonstrating that $\det(\mathbf{J}_\Theta)$ remains single-signed over M and, equivalently, $\det(\mathbf{J}_{\Theta^{-1}})$ remains single-signed over $\Theta(M)$.

In order to compute the elements of $\mathbf{J}_\Theta$, we must draw a one-to-one correspondence between the points defining the representations of the intrinsic manifold inferred from the simulation trajectory, M, and delay embedding, $\Theta(M)$. By Eq. (A1), each point in $\Theta(M)$ comprises d simulation snapshots, leading us to define the following mapping,

$$\begin{aligned}&\overrightarrow{[\vec{p}(t_i), \vec{p}(t_i+\tau), \ldots, \vec{p}(t_i+(d-1)\tau)]} \\ &\quad = \vec{P}(t_i) = \vec{z}(t_i) \mapsto \vec{p}^*(t_i) = \vec{F}(t_i),\end{aligned} \tag{A9}$$

where $\vec{p}(t_i)$ are the coordinates on the manifold M of the chain configuration extracted from the molecular simulation at time t_i, $\vec{P}(t_i)$ is the average of the d coordinates $[\vec{p}(t_i), \vec{p}(t_i+\tau), \ldots, \vec{p}(t_i+(d-1)\tau)]$, and $\vec{p}^*(t_i)$ are the coordinates on the manifold $\Theta(M)$ of the delay embedding corresponding to the d observations of the head-to-tail chain distance $[\ell(t_i), \ell(t_i+\tau), \ldots, \ell(t_i+(d-1)\tau)]$. The terminal snapshots from the molecular simulation trajectory for which $\vec{P}(t_i)$ is undefined are removed from M. In this manner we define an unambiguous mapping between points in M and $\Theta(M)$.

We employ this bijection to compute the elements of $\mathbf{J}_\Theta(\vec{z})$ over the manifold M. We evaluate the partial derivatives constituting the matrix elements of the spatially dependent Jacobian matrix using a mesh-free method to estimate partial derivatives based on a formulation used in smoothed-particle hydrodynamics (SPH) that it is more robust to noise than simple finite difference estimators [79,80]. In this approach, the value of a quantity ζ at any point $\vec{z}$—not necessarily coincident with a projection of any particular observation j—on the k-dimensional manifold M is expressed as a kernel-weighted sum of the value of ζ at all projected observations, $\{\vec{z}^{(j)}\}_{j=1}^{K}$, over the manifold,

$$\zeta(\vec{z}) = \frac{\sum_j \zeta(\vec{z}^{(j)}) W(|\vec{z} - \vec{z}^{(j)}|)}{\sum_j W(|\vec{z} - \vec{z}^{(j)}|)}, \tag{A10}$$

where $W(|\vec{z}-\vec{z}^{(j)}|)$ is a kernel function for which we adopt a k-dimensional Gaussian,

$$W(|\vec{z}-\vec{z}^{(j)}|)=\exp\left[-\tfrac{1}{2}(\vec{z}-\vec{z}^{(j)})\cdot\Sigma^{-2}\cdot(\vec{z}-\vec{z}^{(j)})\right], \tag{A11}$$

where $\Sigma=\mathrm{diag}(\sigma_1,\sigma_2,\ldots,\sigma_k)$ is a $k\times k$ diagonal matrix of the standard deviations of the Gaussian in each dimension. We adopt an isotropic Gaussian kernel such that $\sigma=\sigma_1=\sigma_2=\ldots=\sigma_k$. The spatial derivative $\nabla_{\vec{z}}\zeta$ is straightforwardly obtained as

$$\begin{aligned}\nabla_{\vec{z}}\zeta &= \frac{\sum_j \zeta(\vec{z}^{(j)})\nabla_{\vec{z}}W(|\vec{z}-\vec{z}^{(j)}|)}{\sum_j W(|\vec{z}-\vec{z}^{(j)}|)}-\frac{\sum_j \nabla_{\vec{z}}W(|\vec{z}-\vec{z}^{(j)}|)\cdot\sum_j \zeta(\vec{z}^{(j)})W(|\vec{z}-\vec{z}^{(j)}|)}{\left[\sum_j W(|\vec{z}-\vec{z}^{(j)}|)\right]^2}\\ &= \frac{\sum_j(\vec{z}-\vec{z}^{(j)})\cdot\Sigma^{-2}W(|\vec{z}-\vec{z}^{(j)}|)\cdot\sum_j \zeta(\vec{z}^{(j)})W(|\vec{z}-\vec{z}^{(j)}|)}{\left[\sum_j W(|\vec{z}-\vec{z}^{(j)}|)\right]^2}-\frac{\sum_j(\vec{z}-\vec{z}^{(j)})\cdot\Sigma^{-2}\zeta(\vec{z}^{(j)})W(|\vec{z}-\vec{z}^{(j)}|)}{\sum_j W(|\vec{z}-\vec{z}^{(j)}|)}.\end{aligned} \tag{A12}$$

We compute from this expression the rows of $\mathbf{J}_\Theta(\vec{z})$ in Eq. (A8) at each projected observation, $\vec{z}^{(j)}$, by setting $\zeta=\{F_q\}_{q=1}^k$ and $\vec{z}=\{\vec{z}^{(j)}\}_{j=1}^K$.

The value of σ in the Gaussian kernel [Eq. (A11)] controls the characteristic "smoothing length" over which neighboring points contribute to the estimate of $\zeta(\vec{z})$. To assure that our results are robust to the choice of this parameter, we define the function, $R(\sigma,\Delta\sigma)$, measuring the relative change in $\det(\mathbf{J}_\Theta)$ averaged over all projected observations $\{\vec{z}^{(j)}\}_{j=1}^K$ as a function of the kernel bandwidth σ and a perturbation $\Delta\sigma$,

$$R(\sigma;\Delta\sigma)=\frac{\sum_{j=1}^K|\det[\mathbf{J}_\Theta(\vec{z}^{(j)};\sigma+\Delta\sigma)]-\det[\mathbf{J}_\Theta(\vec{z}^{(j)};\sigma)]|}{\sum_{j=1}^K|\det[\mathbf{J}_\Theta(\vec{z}^{(j)};\sigma)]|}, \tag{A13}$$

where we take absolute values to eliminate any fortuitous cancellation of positive and negative deviations.

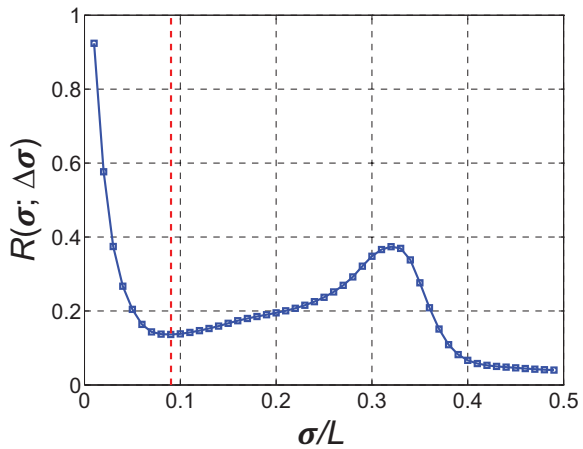

FIG. 17. Selection of Gaussian kernel bandwidth σ in mesh-free Jacobian estimation for the mapping Θ. $R(\sigma;\Delta\sigma)$ measures the relative change in the Jacobian determinant, $\det[\mathbf{J}_\Theta(\vec{z}^{(j)};\sigma)]$, averaged over all observations, $\{\vec{z}^{(j)}\}_{j=1}^K$, under small perturbations to the bandwidth of $\Delta\sigma/L=0.01$. $L=1.0$ is the characteristic size of the manifold M in Fig. 14(a). $R(\sigma;\Delta\sigma)$ reaches a local minimum at $\sigma/L=0.09$ corresponding to a balance between incorporating sufficiently many neighbors into the estimator to robustly evaluate partial derivates, but not so many as to incorporate irrelevant nonlocal information.

In Fig. 17 we plot $R(\sigma;\Delta\sigma)$ for $\sigma/L=[0.01{:}0.01{:}0.50]$ with $\Delta\sigma/L=0.01$, where $L=1.0$ is the largest distance between any two points over the manifold M illustrated in Fig. 14(a). At small values of the kernel bandwidth ($\sigma/L<0.05$), $\det[\mathbf{J}_\Theta(\vec{z}^{(j)};\sigma)]$ changes rapidly with σ as evinced by large values of $R(\sigma;\Delta\sigma)$. This is attributable to noisy estimates of the partial derivatives due to the inclusion of insufficiently many neighbors into the estimator in Eq. (A10). At larger bandwidths ($0.05<\sigma/L<0.25$), $R(\sigma;\Delta\sigma)$ approximately plateaus. Moving to higher bandwidths ($0.25<\sigma/L<0.30$), $R(\sigma;\Delta\sigma)$ increases as the bandwidth becomes so large that nonlocal information irrelevant to the partial derivative estimation is incorporated into the estimator. Finally, for $\sigma/L>0.40$, $R(\sigma;\Delta\sigma)$ approaches zero as essentially all points in the embedding are incorporated into the estimator. The plateau region at $0.05<\sigma/L<0.25$ represents a balance between incorporating sufficiently many neighbors to robustly compute the partial derivatives, and not so many as to incorporate irrelevant nonlocal information into the estimator. That $R(\sigma;\Delta\sigma)$ does not reach zero in this region can be understood as a steady change in $\det[\mathbf{J}_\Theta(\vec{z}^{(j)};\sigma)]$ with σ due to the elevated smoothing of the data associated with the incorporation of more neighbors into the estimator. Accordingly, we choose as our bandwidth $\sigma/L=0.09$, corresponding to the weak local minimum in the plateau region.

In Fig. 18 we plot $R(\sigma;\Delta\sigma)$ corresponding to the Jacobian, $\mathbf{J}_{\Theta^{-1}}$, of the reverse mapping $\Theta^{-1}:\mathbb{R}^k\to\mathbb{R}^k$, from which we select a bandwidth of $\sigma/L=0.11$ for $\mathbf{J}_{\Theta^{-1}}(\vec{F})$.

APPENDIX B: TAKENS' RECONSTRUCTION OF TWO TOY SYSTEMS

The capacity of Takens' theorem to recover multidimensional manifolds containing the dynamics of a multidimensional time series from observations of a single scalar system observable can be unintuitive. Given the possibly alien nature of these ideas, we present below applications of Takens' theorem to demonstrate the recovery from univariate time series of the multidimensional landscapes of two simple but nontrivial toy systems, one deterministic—the Lorenz model—and one stochastic—two-dimensional Brownian motion in a three-well potential.

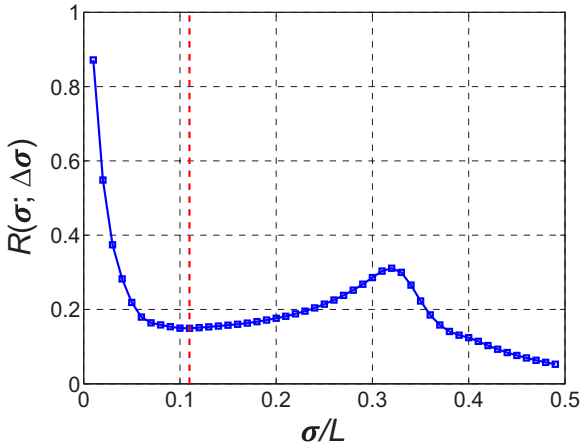

FIG. 18. Selection of Gaussian kernel bandwidth σ in mesh-free Jacobian estimation for the mapping Θ^{-1}. $R(\sigma;\Delta\sigma)$ measures the relative change in the Jacobian determinant, $\det[\mathbf{J}_{\Theta^{-1}}(\vec{F}^{(j)};\sigma)]$, averaged over all observations, $\{\vec{F}^{(j)}\}_{j=1}^{K}$, under small perturbations to the bandwidth of $\Delta\sigma/L = 0.01$. $L = 1.0$ is the characteristic size of the manifold $\Theta(M)$ in Fig. 14(c). $R(\sigma;\Delta\sigma)$ reaches a local minimum at $\sigma/L = 0.11$ corresponding to a balance between incorporating sufficiently many neighbors into the estimator to robustly evaluate partial derivates, but not so many as to incorporate irrelevant nonlocal information.

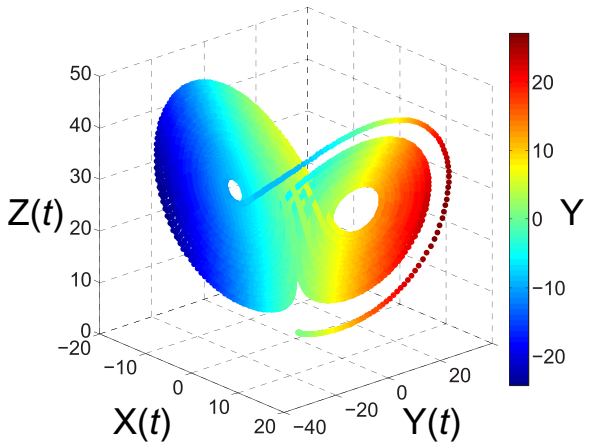

FIG. 19. The Lorenz attractor is a fractal object of correlation dimension (2.05 ± 0.01) [94] within the three-dimensional phase space of the Lorenz model [Eq. (B1)]. Points represent samples from a single numerical trajectory of the Lorenz system and are colored according to the value of their y coordinate. We will employ Takens' theorem to demonstrate that topologically equivalent representations of this attractor can be reconstructed from univariate time series in a single system observable.

1. Reconstruction of the Lorenz attractor

The (dimensionless) Lorenz model [93] is defined by a set of three coupled ordinary differential equations defining trajectories of $\{x,y,z\} \in \mathbb{R}^3$,

$$\begin{aligned} \frac{dx}{dt} &= \sigma(y-x), \\ \frac{dy}{dt} &= x(\rho - z) - y, \\ \frac{dz}{dt} &= xy - \beta z, \end{aligned} \tag{B1}$$

where σ, ρ, and β are constants. Following the original formulation by Lorenz, we choose $\sigma = 10$, $\rho = 28$, $\beta = 8/3$ [93]. Under these conditions, the dynamics of the system are chaotic (i.e., exhibit sensitive dependence on initial conditions). The chaotic trajectories of $\{x,y,z\} \in \mathbb{R}^3$ define a low-dimensional fractal *attractor* of correlation dimension (2.05 ± 0.01) [94], constituting the subset of three-dimensional phase space—the *intrinsic manifold*—to which the system will evolve from an arbitrary initial condition. We can draw an analogy to molecular systems wherein the intrinsic manifold defines the low-dimensional surface in a high-dimensional Cartesian coordinate space to which the molecular motions are effectively restrained.

We illustrate in Fig. 19 a numerical trajectory of the Lorenz system commencing from the initial point $(x,y,z) = (0.00,1.00,1.05)$ over the range $t = [0,100]$. Points are displayed at intervals of $\Delta t = 0.0049$. Numerical integration was performed in MATLAB using the ode45 integrator with a tolerance of 1×10^{-6} [95]. The trajectory evolves to a closed subset of the phase space resembling a butterfly—the eponymous Lorenz attractor [93]. It is the goal of this demonstration to show that we can use Takens' theorem to recover a topologically equivalent reconstruction of the Lorenz attractor from delay embeddings of a univariate time series in a single system observable. The Lorenz model is a canonical test problem in dynamical systems theory, and reconstruction of the Lorenz attractor has been considered in many previous works [32,50,51,53,62,94,96].

Takens' delay embedding theorem [29–34] provides a means to reconstruct a *topologically equivalent* realization, $\Theta(M)$, of the *intrinsic manifold*, M, from a scalar time series in a single observable. The theorem requires that the observable is *generic*, and *does not contain any symmetries that are not present in the system*. Takens' theorem proceeds by constructing a high-dimensional *delay embedding* of the scalar time series to reconstruct a representation of the intrinsic manifold, $\Theta(M)$, over which the dynamical evolution of the system is C^1*-equivalent* to that over the intrinsic manifold, M, in the original space. Before proceeding, let us examine the precise meaning of the italicized terms.

(1) The *intrinsic manifold* is the low-dimensional subset of the high-dimensional phase space within which the system dynamics are restricted to reside, typically as a result of couplings between system degrees of freedom. In the Lorenz system, this is the Lorenz attractor.

(2) *Topologically equivalent* means that the relative arrangement and connectivity of the geometric features of the original intrinsic manifold, M, are preserved in its reconstruction, $\Theta(M)$. In the present case of the Lorenz attractor, for example, we should expect the Takens' reconstruction to properly preserve the two wings of the butterfly and the holes in their centers. Takens' theorem does not, however, claim that the reconstruction will preserve the *topography* of the intrinsic manifold, so that the manifold may be stretched and squashed, and the distribution of points over its surface densified or rarefied. In the Lorenz system, we should expect that the

butterfly wings may be bent and contorted in the Takens' reconstruction.

(3) *Generic observable* means a scalar measurement of the system dynamics that is a function of all degrees of freedom. In principle, an arbitrary function will suffice, but in practice, reconstruction may fail if the observable depends only very weakly on some degrees of freedom. In the Lorenz model, a generic observable would be any function of all three coordinates $f(x,y,z)$. As we show below, since the x coordinate is coupled to the evolution of y and z, so $f(x,y,z) = x(y,z)$ constitutes a generic observable.

(4) That the generic observable *does not contain any symmetries that are not present in the system* means that the observable should not be symmetric in its arguments (i.e., the system degrees of freedom) in any manner that the system is not. As can be checked from Eq. (B1), the Lorenz system is invariant under the transformation $(x,y,z) \to (-x,-y,z)$, indicating that it possesses a symmetry in x-y that is visually manifested in the two butterfly wings [53]. The variable z is unchanged under the action of this symmetry, and so cannot resolve the two wings [53]. As previously explored in Refs. [51,53,54], we should expect that Takens' delay embedding in the observable $f(x,y,z) = z(x,y)$ will not respect this topological symmetry, and will collapse together the two wings of the butterfly. This reconstructed attractor will *not*, therefore, be topologically equivalent to the Lorenz attractor.

(5) A *delay embedding* is the procedure by which discrete scalar time series $\{x(t_i)\}_{i=1}^{K}$ is converted into a d-dimensional vector time series $\{\vec{y}(t_i)\}_{i=1}^{K'}$ under the following operation,

$$\vec{y}(t_i) = [x(t_i), x(t_i - \tau), \ldots, x(t_i - (d-1)\tau)], \qquad \text{(B2)}$$

where τ is the delay time. This operation has the effect of representing the state of the system at any particular time instant as a d-dimensional vector of evenly spaced observations of x recording the past history of the system. In this work, we choose to work with future rather than past embeddings [i.e., replacing $\tau \to (-\tau)$], which represents the system state by its future trajectory in x. Systematic means exist to choose appropriate values of d [62,63] and τ [61].

(6) That the dynamical evolution is C^1-*equivalent* to that in the original space means that the manner in which the dynamics of the system evolve in the reconstructed intrinsic manifold, $\Theta(M)$, residing within the d-dimensional space constructed from the Takens' delay embedding is related by a continuous and smooth (i.e., at least once differentiable) function whose inverse is also continuous and smooth, to the dynamic evolution in the original intrinsic manifold, M. In the case of the Lorenz system, the manner in which the chaotic Lorenz trajectories orbit around the Lorenz attractor—a figure eight flow around the butterfly wings—can be mapped by a smooth and invertible function to the manner in which the trajectories in the Takens' delay embedding orbit around the reconstructed attractor.

Using the methodology detailed in the main text—Methods Summary: Phase space reconstruction—and presented in more detail above—Appendix A2: Phase space reconstruction—we will now proceed to construct Takens' delay embeddings to generate reconstructions of the Lorenz attractor in Fig. 19 from univariate time series in a single system observable.

Observable $= x(t)$. We first consider the Lorenz variable x as our univariate observable, producing a univariate time series $x(t)$. This is a generic observable of the system, since the evolution of x is coupled to that of y and z [i.e., $f(x,y,z) = x(y,z)$; cf. Eq. (B1)]. Given our scalar time series $\{x(t_i)\}_{i=1}^{K}$ measured at intervals of $\Delta t = 0.0049$ over the course of the time horizon $t = [0,100]$, Takens' theorem prescribes that we construct the delay embedding,

$$\vec{y}(t_i) = [x(t_i), x(t_i + \tau), \ldots, x(t_i + (d-1)\tau)], \qquad \text{(B3)}$$

where τ is the delay time between successive system observations, d is the delay embedding dimensionality, and the projected time series $\{\vec{y}(t_i)\}_{i=1}^{K'}$ defines the reconstructed intrinsic manifold $\Theta(M) \in \mathbb{R}^d$. Empirical tools exist to select τ and d. We use the mutual information approach of Fraser and Swinney to choose $\tau = 0.429$ [61], and the approach of Cao [62] based on the false nearest neighbors method of Kennel *et al.* [63] to select $d = 3$.

We present in Fig. 20 the ensemble of $\{\vec{y}(t_i)\} \in \mathbb{R}^3$ synthesized by our delay embedding of the scalar time series $\{x(t_i)\}$. Takens' theorem asserts that this embedding constitutes a topologically equivalent reconstruction of the Lorenz attractor, and that the dynamical flow of the points over the manifold is C^1-equivalent to that over the original attractor. The relatively simple and low-dimensional nature of the Lorenz system allows this to be confirmed from visual inspection, from which it is apparent that the reconstruction reproduces the two wings, the holes in their centers, and the figure eight flow of points over the manifold.

Observable $= z(t)$. We now consider the Lorenz variable z as our observable, producing a univariate time series $z(t)$. Following an identical approach to that above, we construct Takens' delay embeddings with $\tau = 0.146$ [61], and $d = 3$ [62]. As discussed above, in this case we should *not* expect the reconstruction to preserve the topology of the Lorenz

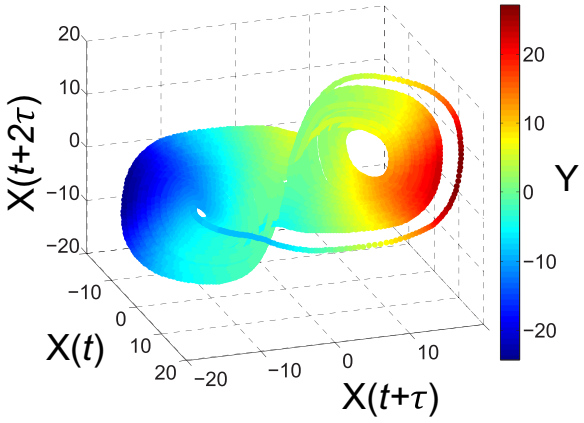

FIG. 20. Reconstruction of the Lorenz attractor from three-dimensional Takens' delay embeddings of the scalar time series $x(t)$. To assist in comparisons with Fig. 19, each point $\vec{y}(t_i) = (x(t_i), x(t_i + \tau), x(t_i + 2\tau))$, where $\tau = 0.429$, is colored according to the y coordinate of the center point, $y(t_i + \tau)$. By Takens' theorem, this reconstruction is topologically identical to the Lorenz attractor in Fig. 19.

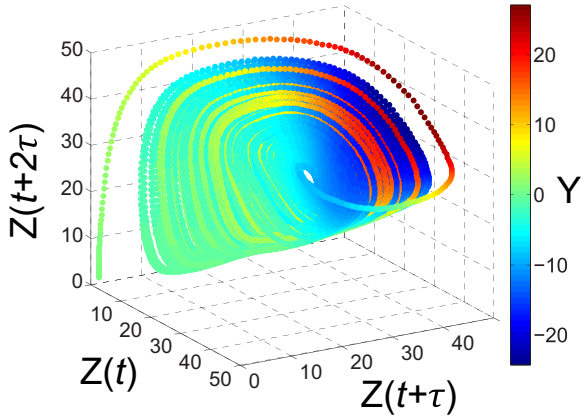

FIG. 21. Reconstruction of the Lorenz attractor from three-dimensional Takens' delay embeddings of the scalar time series $z(t)$. To assist in comparisons with Fig. 19, each point $\vec{y}(t_i) = (x(t_i), x(t_i+\tau), x(t_i+2\tau))$, where $\tau = 0.146$, is colored according to the y coordinate of the center point, $y(t_i+\tau)$. The observable z contains a spurious symmetry that is not present in the Lorenz system such that it cannot distinguish between the two butterfly wings. This violates a key assumption for the success of Takens' theorem causing it to fail, and the reconstructed attractor is not topologically equivalent to the Lorenz attractor in Fig. 19.

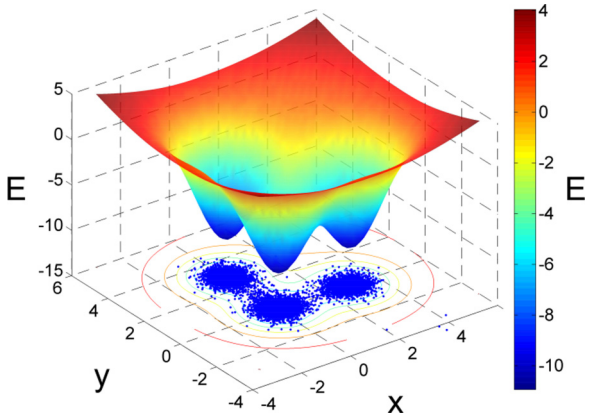

FIG. 22. Two-dimensional Brownian dynamics in a three-well potential energy landscape. The point cloud corresponds to the 10 000 snapshots of the particle location recorded over the course of the numerical Brownian dynamics simulation.

attractor since the observable z does contain a symmetry not present in the system, in that it cannot distinguish the two butterfly wings. The topological inequivalence is apparent from the three-dimensional delay embedding presented in Fig. 21, where the reconstruction has collapsed together the two butterfly wings. This negative example demonstrates the importance of using a scalar observable that does not contain any spurious symmetries not present in the system. This condition turns out to be of central importance in our application of Takens' theorem to the $C_{24}H_{50}$ n-alkane chain in the main text.

2. Reconstruction of the potential energy landscape of a Brownian particle

As a second example, we consider the application of Takens' theorem to reconstruct a two-dimensional potential energy landscape from univariate measurements of the dynamics of a Brownian point particle within the potential. Since a point particle possesses no configurational entropy, the potential energy $E(\vec{r})$ is identical to the free energy $F(\vec{r})$, where $\vec{r}$ is the coordinates of the particle on the potential surface. In contrast, the polymer considered in the main text does contain configurational entropy, so in that case it is the free-energy landscape that we seek to recover.

The motion of a Brownian particle with constant and isotropic diffusivity, D, in an external potential, $E(\vec{r})$, is given by the equation of motion,

$$\dot{\vec{r}}(t) = -D\nabla E(\vec{r}) + \sqrt{2D}\xi(t), \tag{B4}$$

where $\xi(t)$ is a stationary Gaussian process with $\langle \xi(t) \rangle = 0$ and $\langle \xi(t)\xi(t') \rangle = \delta(t-t')$ [97]. For clarity of exposition, we consider the motion in dimensionless form such that the particle position $\vec{r}$, time t, and potential E are all dimensionless. (The dimensional equation of motion can be placed in dimensionless form by scaling it with characteristic values for time, length, and energy.) We consider the potential energy landscape illustrated by the surface in Fig. 22, comprising three identical isotropic Gaussian wells and a long ranged quadratic restraining potential to prevent the particle from drifting off to infinity. Mathematically, the dimensionless potential energy surface is given by

$$E(x,y) = \frac{1}{2}\kappa(\vec{r}-\vec{v}_i)^T(\vec{r}-\vec{v}_i) + \sum_{i=1}^{3} \frac{\alpha}{\sqrt{(2\pi)^2|\Sigma|}} \exp\left[\frac{1}{2}(\vec{r}-\vec{\mu}_i)^T\Sigma^{-1}(\vec{r}-\vec{\mu}_i)\right], \tag{B5}$$

where $\vec{r} = (x,y)$, the quadratic restraining potential, is centered at $\vec{v} = (1.5,1.5)$ and possesses a spring constant $\kappa = 0.2$, and the Gaussian wells of depth $\alpha = (-70)$ and unit variance $\Sigma = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ are centered at $\vec{\mu}_1 = (0.0,0.0)$, $\vec{\mu}_2 = (0.3,0.0)$, and $\vec{\mu}_3 = (0.0,0.3)$.

We adopt a diffusivity of $D = 0.04$, and initially locate the particle at the origin $\vec{r}_0 = (0,0)$. We simulate the trajectory of the Brownian particle over the time horizon $t = [0,200,000]$ by numerically integrating the equation of motion using the Ermak-McCammon equation employing a time step $\Delta t = 0.2$ [97]. We sample the particle location every 100 steps. The point cloud defining the location of the particle at each of these 10 000 time points is projected over the potential energy landscape in Fig. 22.

Given complete knowledge of the location of the Brownian particle at each time point [i.e., its (x,y) coordinates] we can compute an approximation for the underlying (dimensionless) potential energy landscape can be estimated from the observed probability distribution of snapshots in the point cloud, $\hat{P}(x,y)$, using the statistical mechanical relationship, $E(x,y) = -\ln\hat{P}(x,y) + C$, where C is an arbitrary constant. The approximation generated from the 10 000 samples of the (x,y) system coordinates recorded over the course of the simulation is presented in Fig. 23. It is the goal of

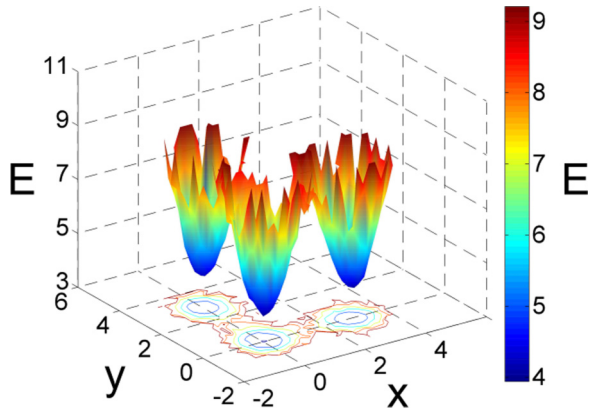

FIG. 23. Potential energy surface estimated from 10 000 samples of the system coordinates (x,y) recorded over the course of the Brownian dynamics simulation.

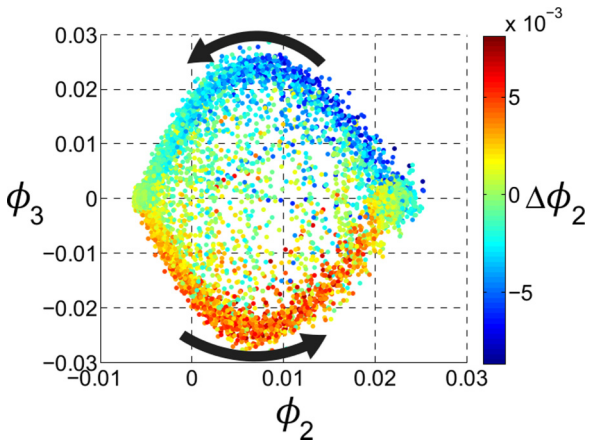

FIG. 24. Projection of the 20-dimensional delay embedding of the scalar time series in $x(t)$ extracted from the Brownian dynamics simulations into the top two collective modes $[\vec{\phi}_2,\vec{\phi}_3]$ identified by diffusion maps. Points are colored by the change in ϕ_2 between consecutive delay embedding vector projections, $\Delta\phi_2(t_i) = \phi_2(t_i+\tau) - \phi_2(t_i)$.

this example to show that we can apply Takens' theorem to recover a topologically equivalent representation of the potential energy surface in Fig. 23 from time series in a single system observable.

Before proceeding, we note that since the Brownian dynamics proceed over a two-dimensional intrinsic manifold in a two-dimensional phase space, there is no need to apply dimensionality reduction algorithms to the simulation trajectory to synthesize a low-dimensional projection of the intrinsic manifold residing within a high-dimensional ambient space. This stands in contrast to our analysis of the dynamics of a polymer chain considered in the main text, where we employ nonlinear dimensionality reduction to extract a three-dimensional intrinsic manifold from a 72-dimensional coordinate space.

Observable $= x(t)$. We first consider the case that we have access to only the x coordinate of the particle dynamics. The dynamical evolution of x and y are coupled through the potential energy landscape, $E(x,y)$, meaning that the evolution of x depends also on y, and x is a generic observable of the system [i.e., $f(x,y) = x(y)$; cf. Fig. 22]. This observable does, however, contain a spurious symmetry not present in the system, since it cannot distinguish the two Gaussian wells located at $(x,y) = (0.0,0.0)$ and $(0.0, 0.3)$. As is visually apparent from Fig. 22, these wells collapse together under projection onto the x axis, and we should *not* expect to be able to reconstruct a topologically equivalent potential energy landscape from Takens' delay embeddings of $x(t)$. This is precisely analogous to the case of the z observable in the Lorenz system described above.

Following the same methodology as described for the Lorenz attractor above, we construct a Takens' delay embedding in the scalar time series $\{x(t_i)\}_{i=1}^{K}$ measured at equally spaced intervals of $\Delta t = 0.2$ over the course of the time horizon $t = [0,200,000]$ [cf. Eq. (B3)], employing the mutual information approach of Fraser and Swinney to choose $\tau = 0.4$ [61], and the nearest neighbors approach of Cao [62] to select $d = 20$. Takens' theorem asserts that the dimensionality of the reconstructed intrinsic manifold must be the same as that of the original intrinsic manifold, such that the two-dimensional reconstruction lies latent within the 20-dimensional delay embedding space. Using the approach detailed in the main text—Methods Summary: Dimensionality reduction—and presented in more detail above—Appendix A3: Diffusion maps, and Appendix A4: Hierarchical nonlinear principal components analysis—we use nonlinear dimensionality reduction to identify and extract the two-dimensional reconstructed manifold. Indeed, application of diffusion maps to the 20-dimensional delay embedding identifies a two-dimensional embedding into the top two collective modes $[\vec{\phi}_2,\vec{\phi}_3]$ that we present in Fig. 24.

It is clear from visual inspection of the embedding that there exists an axis of reflection symmetry along $\phi_3 = 0$. What is the source of this symmetry? Coloring the points over the two-dimensional manifold by the change in $\Delta\phi_2(t_i) = \phi_2(t_i+\tau) - \phi_2(t_i)$ reveals a counterclockwise flow over the manifold such that transitions of the system from the left side to the right side of the reconstructed manifold progress by the lower pathway, whereas transitions from the right to left progress by the upper. The existence of separate pathways for transitions of the Brownian particle from left to right and right to left contradicts the expectation that a system in thermodynamic equilibrium should obey detailed balance and exhibit microscopic reversibility [5,65]. Accordingly, it should not be possible to tell from a single observation of the particle coordinates whether it is moving to the right or left. This expectation is met for the point cloud over the original intrinsic manifold (Fig. 22), but not for that recovered from the delay embedding (Fig. 24). The difference is that the delay embedding vectors are formed from a series of successive measurements of the system, from which it is possible to ascertain whether the particle is in the process of moving right or moving left. Specifically, the delay embedding of a particle moving along a particular pathway towards the right will comprise a sequence of measurements of x increasing in value, whereas that for a particle moving along the same pathway but towards the left will comprise a sequence of x measurements decreasing in value. Accordingly, a particle

located at coordinates $(\tilde{x},\tilde{y})$ and moving to the right will be embedded in a different location in delay space from a particle at $(\tilde{x},\tilde{y})$ but moving to the left. In effect, the construction of a delay embedding has broken the symmetry of the underlying Newtonian mechanics, leading to the observed reflection symmetry in Fig. 24. As we shall see, the well on the left side of the reconstructed intrinsic manifold at $(\phi_2,\phi_3) = (-0.005,0)$ corresponds to the collapsing together of the two potential wells at $(x,y) = (0.0,0.0)$ and (0.0, 0.3), and the one on the right at at $(\phi_2,\phi_3) = (0.02,0)$ the the remaining well at $(x,y) = (0.3,0.0)$. We do not see an analogous top-bottom symmetry in addition to the left-right just described due to the inability of the observable x to distinguish the two Gaussian wells at $(x,y) = (0.0,0.0)$ and (0.0, 0.3) (cf. Fig. 23).

For clarity of exposition, we eliminate the spurious symmetry introduced by the delay embedding by projecting the points embedded into the lower portion of the attractor into the upper portion by reflecting them through the observed axis of symmetry along $\phi_3 = 0$. We observe analogous spurious temporal symmetries in our reconstruction of the intrinsic manifold of the $C_{24}H_{50}$ chain in the main text, and describe therein a more rigorous approach to eliminate this symmetry in Results and Discussion: Temporally symmetrized smFES from delay embeddings.

Having removed the spurious symmetry, we now proceed to reconstruct the potential energy landscape over the reconstructed intrinsic manifold using the relationship $E(\phi_2,\phi_3) = -\ln\hat{P}(\phi_2,\phi_3) + C$, where C is an arbitrary constant. As illustrated in Fig. 25, we see that the reconstructed potential energy landscape $E(\phi_2,\phi_3)$ contains only two basins compared to the three basins in the original potential $E(x,y)$, and is therefore *not* topologically equivalent. The reason, of course, is the spurious symmetry in the observable x from which the delay embeddings were constructed that cannot distinguish between the two Gaussian wells at $(x,y) = (0.0,0.0)$ and (0.0, 0.3). Indeed, by inspecting the x values of the points in the delay embedding, we have verified that the deeper basin on the left corresponds to the collapsing together of the two wells at at $(x,y) = (0.0,0.0)$ and (0.0, 0.3), and that on the right to the single well at $(x,y) = (0.3,0.0)$.

Observable $= x(t) - y(t)$. We now consider as our generic observable a linear combination of x and y that does not contain any spurious spatial symmetries. The expectation is that this observable should permit the recovery of topologically equivalent reconstructions of the potential energy landscape of the Brownian particle using Takens' theorem. Adopting as our scalar observable $f(x,y) = (x - y)$, we followed precisely the same protocol as above to construct $d = 20$-dimensional delay embeddings with a delay time of $\tau = 0.4$. Diffusion maps recover a three-dimensional manifold from within the 20-dimensional delay space spanned by the top three collective variables $[\phi_2,\phi_3,\phi_4]$. Analysis revealed ϕ_2 and ϕ_3 to be functionally dependent, defining a one-dimensional manifold in the space of these two variables. We eliminated this redundancy using hierarchical nonlinear principal components analysis (h-NLPCA) to extract this one-dimensional manifold that we term Γ_{23}. The combined application of diffusion maps and h-NLPCA allows us to generate a two-dimensional reconstruction of the intrinsic manifold in $[\Gamma_{23},\phi_4]$. After removing the spurious temporal symmetry in ϕ_4 introduced by the delay embedding (cf. Fig. 24), we generated the reconstructed potential energy landscape illustrated in Fig. 26. In this case, our reconstruction is a topologically equivalent reconstruction of the original landscape (Fig. 23), reproducing its structure and geometry. Analysis of the (x,y) coordinates corresponding to the points in the delay embedding reveals the well at $(\Gamma_{23},\phi_4) = (0.2,0.0)$ corresponds to that at $(x,y) = (0.0,3.0)$, that at $(\Gamma_{23},\phi_4) = (0.6,0.0)$ to $(x,y) = (0.0,0.0)$, and that at $(\Gamma_{23},\phi_4) = (1.2,0.0)$ to $(x,y) = (3.0,0.0)$. The low-energy pathways linking neighboring wells in the reconstructed potential energy surface reproduce their topological adjacency in the original potential energy landscape, the higher energy pathway linking the left and right wells corresponding to rarely observed transitions between the $(x,y) = (0.0,0.3)$ to $(x,y) = (0.3,0.0)$ wells that do not become trapped in the intervening $(x,y) = (0.0,0.0)$ basin.

By collecting time series in a generic observable that does not contain any symmetries not present in the system, we have used Takens' theorem to successfully reconstruct topologically equivalent representations of the original potential energy

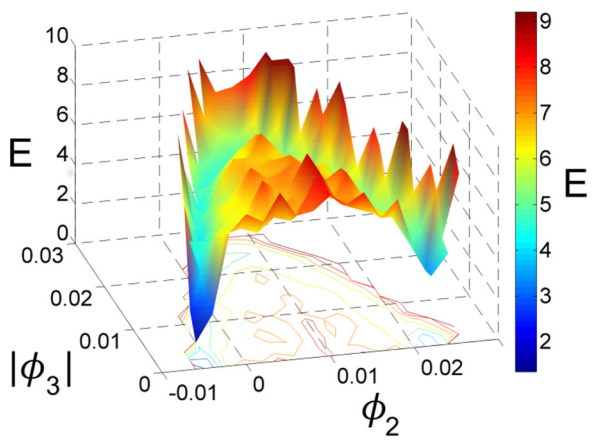

FIG. 25. Potential energy surface over the two-dimensional reconstructed intrinsic manifold generated from delay embeddings of the scalar time series in $x(t)$.

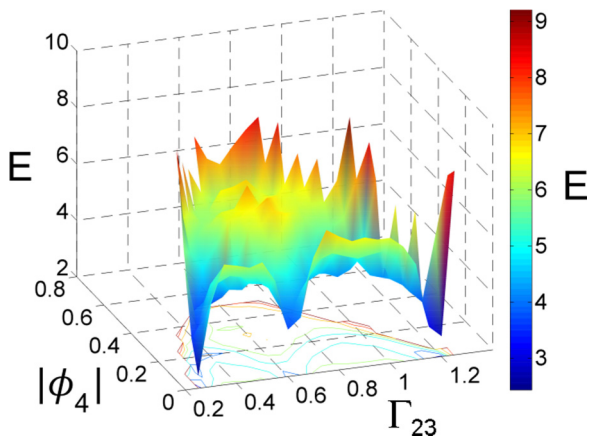

FIG. 26. Potential energy surface over the two-dimensional reconstructed intrinsic manifold generated from delay embeddings of the scalar time series in $x(t) - y(t)$.

landscape containing the Brownian particle. As specified by the theorem, we should not expect the *topography* of the original landscape to be preserved, such that the reconstructed intrinsic manifold may be a stretched and squashed version of the original manifold that nonetheless preserves its topological geometry and connectivity. Indeed, the shape of the reconstructed potential energy landscape in Fig. 26 is clearly different from that of the original landscape in Fig. 23, while still maintaining the topology.

Observable $= 4 \times \sin[x(t) - 1.5] + 3 \times \cos[y(t)]$. As a final example, we consider as our system observable a highly nonlinear system observable $f(x,y) = 4 \times \sin[x(t) - 1.5] + 3 \times \cos[y(t)]$ that does not contain any symmetries not present in the system. Following an identical procedure to that above, we recover the two-dimensional reconstructed energy landscape in Fig. 27. As is visually apparent, this reconstruction is a topologically equivalent reconstruction of the original potential energy landscape, containing the three energy wells and the transition paths between them. As might be anticipated from the complexity of the nonlinear scalar observable, the topography of the landscape is substantially perturbed compared to that recovered from the linear observable $f(x,y) = (x - y)$, although the topology is once again equivalent.

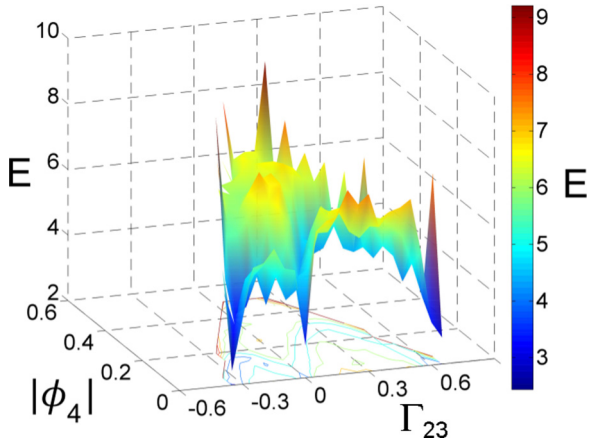

FIG. 27. Reconstructed potential energy landscape over the two-dimensional reconstructed intrinsic manifold generated from delay embeddings of the scalar time series in $4 \times \sin[x(t) - 1.5] + 3 \times \cos[y(t)]$.

[1] K. A. Dill and H. S. Chan, From Levinthal to pathways to funnels, Nat. Struct. Biol. **4**, 10 (1997).

[2] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Theory of protein folding: The energy landscape perspective, Annu. Rev. Phys. Chem. **48**, 545 (1997).

[3] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, Systematic determination of order parameters for chain dynamics using diffusion maps, Proc. Natl. Acad. Sci. USA **107**, 13597 (2010).

[4] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction, Proc. Natl. Acad. Sci. USA **103**, 9885 (2006).

[5] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, Oxford, 2001).

[6] A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti, Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach, Chem. Phys. Lett. **509**, 1 (2011).

[7] R. Coifman, I. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems, Multis. Model. Simul. **7**, 842 (2008).

[8] A. E. García, Large-Amplitude Nonlinear Motions in Proteins, Phys. Rev. Lett. **68**, 2696 (1992).

[9] P. I. Zhuravlev, C. K. Materese, and G. A. Papoian, Deconstructing the native state: Energy landscapes, function, and dynamics of globular proteins, J. Phys. Chem. B **113**, 8800 (2009).

[10] A. Amadei, A. Linssen, and H. J. Berendsen, Essential dynamics of proteins, Proteins: Struct., Funct., Genet. **17**, 412 (1993).

[11] E. Plaku, H. Stamati, C. Clementi, and L. E. Kavraki, Fast and reliable analysis of molecular motion using proximity relations and dimensionality reduction, Proteins: Struct., Funct., Genet. **67**, 897 (2007).

[12] R. Hegger, A. Altis, P. H. Nguyen, and G. Stock, How Complex Is the Dynamics of Peptide Folding? Phys. Rev. Lett. **98**, 028102 (2007).

[13] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, Integrating diffusion maps with umbrella sampling: Application to alanine dipeptide, J. Chem. Phys. **134**, 135103 (2011).

[14] A. L. Ferguson, S. Zhang, I. Dikiy, A. Z. Panagiotopoulos, P. G. Debenedetti, and J. A. Link, An experimental and computational investigation of spontaneous lasso formation in microcin J25, Biophys. J. **99**, 3056 (2010).

[15] P. G. Bolhuis, C. Dellago, and D. Chandler, Reaction coordinates of biomolecular isomerization, Proc. Natl. Acad. Sci. USA **97**, 5877 (2000).

[16] A. W. Long and A. L. Ferguson, Nonlinear machine learning of patchy colloid self-assembly pathways and mechanisms, J. Phys. Chem. B **118**, 4228 (2014).

[17] W. Zheng, M. A. Rohrdanz, M. Maggioni, and C. Clementi, Polymer reversal rate calculated via locally scaled diffusion map, J. Chem. Phys. **134**, 144109 (2011).

[18] A. Mukherjee, R. Lavery, B. Bagchi, and J. T. Hynes, On the molecular mechanism of drug intercalation into DNA: A simulation study of the intercalation pathway, free energy, and DNA structural changes, J. Am. Chem. Soc. **130**, 9747 (2008).

[19] D. Shukla, Y. Meng, B. Roux, and V. S. Pande, Activation pathway of src kinase reveals intermediate states as targets for drug design, Nat. Commun. **5**, 3397 (2014).

[20] J. Guan, B. Wang, and S. Granick, Automated single-molecule imaging to track DNA shape, Langmuir **27**, 6149 (2011).

[21] R. Roy, S. Hohng, and T. Ha, A practical guide to single-molecule FRET, Nat. Methods **5**, 507 (2008).

[22] G. H. Zerze, R. B. Best, and J. Mittal, Modest influence of FRET chromophores on the properties of unfolded proteins, Biophys. J. **107**, 1654 (2014).

[23] S. A. McKinney, C. Joo, and T. Ha, Analysis of single-molecule FRET trajectories using hidden Markov modeling, Biophys. J. **91**, 1941 (2006).
[24] J. P. Crutchfield and K. Young, Inferring Statistical Complexity, Phys. Rev. Lett. **63**, 105 (1989).
[25] C. R. Shalizi and J. P. Crutchfield, Computational mechanics: Pattern and prediction, structure and simplicity, J. Stat. Phys. **104**, 817 (2001).
[26] C.-B. Li, H. Yang, and T. Komatsuzaki, Multiscale complex network of protein conformational fluctuations in single-molecule time series, Proc. Natl. Acad. Sci. USA **105**, 536 (2008).
[27] K. R. Haas, H. Yang, and J.-W. Chu, Expectation-maximization of the potential of mean force and diffusion coefficient in Langevin dynamics from single molecule FRET data photon by photon, J. Phys. Chem. B **117**, 15591 (2013).
[28] L. A. Aguirre and C. Letellier, Modeling nonlinear dynamics and chaos: A review, Math. Probl. Eng. **2009**, 238960 (2009).
[29] F. Takens, Detecting strange attractors in turbulence, in *Proceedings of a Symposium on Dynamical Systems and Turbulence Warwick 1980*, edited by D. Rand and L.-S. Young, Series Lecture Notes in Mathematics (Springer, Berlin Heidelberg, 1981), Vol. 898, pp. 366–381.
[30] T. Sauer, J. A. Yorke, and M. Casdagli, Embedology, J. Stat. Phys. **65**, 579 (1991).
[31] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, Geometry from a Time Series, Phys. Rev. Lett. **45**, 712 (1980).
[32] D. S. Broomhead and G. P. King, Extracting qualitative dynamics from experimental data, Phys. D (Amsterdam, Neth.) **20**, 217 (1986).
[33] J. Stark, Delay embeddings for forced systems. I. Deterministic forcing, J. Nonlinear Sci. **9**, 255 (1999).
[34] J. Stark, D. S. Broomhead, M. Davies, and J. Huke, Delay embeddings for forced systems. II. Stochastic forcing, J. Nonlinear Sci. **13**, 519 (2003).
[35] V. Villani and J. M. Zaldivar Comenges, Analysis of biomolecular chaos in aqueous solution, Theor. Chem. Acc. **104**, 290 (2000).
[36] V. Villani, A. M. Tamburro, and J. M. Zaldivar Comenges, Conformational chaos and biomolecular instability in aqueous solution, J. Chem. Soc. Perkin Trans. 2 **13**, 2177 (2000).
[37] D. Giannakis and A. J. Majda, Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability, Proc. Natl. Acad. Sci. USA **109**, 2222 (2012).
[38] T. Berry, J. Cressman, Z. Greguric-Ferencek, and T. Sauer, Time-scale separation from diffusion-mapped delay coordinates, SIAM J. Appl. Dyn. Syst. **12**, 618 (2013).
[39] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, Proc. Natl. Acad. Sci. USA **102**, 7426 (2005).
[40] R. R. Coifman and S. Lafon, Diffusion maps, Appl. Comput. Harmonic Anal. **21**, 5 (2006).
[41] M. V. Athawale, G. Goel, T. Ghosh, T. M. Truskett, and S. Garde, Effects of length scales and attractions on the collapse of hydrophobic polymers in water, Proc. Natl. Acad. Sci. USA **104**, 733 (2007).
[42] D. Chandler, Interfaces and the driving force of hydrophobic assembly, Nature (London) **437**, 640 (2005).
[43] T. F. Miller, E. Vanden-Eijnden, and D. Chandler, Solvent coarse-graining and the string method applied to the hydrophobic collapse of a hydrated chain, Proc. Natl. Acad. Sci. USA **104**, 14559 (2007).
[44] M. Scholz and R. Vigário, Nonlinear PCA: A new hierarchical approach, in *10th European Symposium on Artificial Neural Networks (ESANN)*, edited by M. Verleysen (ESANN, Louvain-la-Neuve, Belgium, 2002), pp. 439–444.
[45] M. Scholz, M. Fraunholz, and J. Selbig, Nonlinear principal component analysis: Neural network models and applications, in *Principal Manifolds for Data Visualization and Dimension Reduction*, edited by A. N. Gorban, B. Kegl, D. C. Wunsch, and A. Zinovyev, No. 58 (Springer, Berlin, 2008), pp. 44–67.
[46] C. J. Dsilva, R. Talmon, R. R. Coifman, and I. G. Kevrekidis, Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study [Appl. Comput. Harmonic Anal. (to be published, 2015)].
[47] W. Humphrey, A. Dalke, and K. Schulten, VMD: visual molecular dynamics, J. Mol. Graphics **14**, 33 (1996).
[48] D. N. Theodorou and U. W. Suter, Shape of unperturbed linear polymers: Polypropylene, Macromolecules **18**, 1206 (1985).
[49] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, 2nd ed. (Cambridge University Press, Cambridge, 2005).
[50] D. J. Cross and R. Gilmore, Differential embedding of the Lorenz attractor, Phys. Rev. E **81**, 066220 (2010).
[51] C. Letellier and G. Gouesbet, Topological characterization of reconstructed attractors modding out symmetries, J. Phys. II **6**, 1615 (1996).
[52] C. Letellier, L. Aguirre, and J. Maquet, How the choice of the observable may influence the analysis of nonlinear dynamical systems, Commun. Nonlinear Sci. Numer. Simul. **11**, 555 (2006).
[53] C. Letellier and L. A. Aguirre, Investigating nonlinear dynamics from time series: The influence of symmetries and the choice of observables, Chaos: An Interdisciplinary J. Nonlinear Sci. **12**, 549 (2002).
[54] L. Cao, A. Mees, and K. Judd, Dynamics from multivariate time series, Phys. D (Amsterdam, Neth.) **121**, 75 (1998).
[55] B. Hashemian and M. Arroyo, Topological obstructions in the way of data-driven collective variables, J. Chem. Phys. **142**, 044102 (2015).
[56] E. Kranakis, *Advances in Network Analysis and Its Applications* (Springer, Berlin Heidelberg, 2013).
[57] K. B. Howell, *Ordinary Differential Equations: An Introduction to the Fundamentals* (CRC Press, Boca Raton, 2016).
[58] T. D. Sauer, Attractor reconstruction, Scholarpedia **1**, 1727 (2006).
[59] C. R. Shalizi, Methods and techniques of complex systems science: An overview, *Complex Systems Science in Biomedicine* (Springer, New York, 2006).
[60] C. Letellier, J. Maquet, L. Le Sceller, G. Gouesbet, and L. Aguirre, On the non-equivalence of observables in phase-space reconstructions from recorded time series, J. Phys. A: Math. Gen. **31**, 7913 (1998).
[61] A. M. Fraser and H. L. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A **33**, 1134 (1986).
[62] L. Cao, Practical method for determining the minimum embedding dimension of a scalar time series, Phys. D (Amsterdam, Neth.) **110**, 43 (1997).
[63] M. B. Kennel, R. Brown, and H. D. Abarbanel, Determining embedding dimension for phase-space reconstruction

using a geometrical construction, Phys. Rev. A **45**, 3403 (1992).
[64] R. E. Kass and P. W. Vos, in *Geometrical Foundations of Asymptotic Inference* (John Wiley & Sons, New York, 2011), pp. 300–303.
[65] R. C. Tolman, *The Principles of Statistical Mechanics* (Oxford University Press, Oxford, 1938).
[66] L. Sirovich, Turbulence and the dynamics of coherent structures. Part I: Coherent structures, Q. Appl. Math. **45** (1987).
[67] P. Holmes, J. L. Lumley, G. Berkooz, and C. W. Rowley, Turbulence, coherent structures, dynamical systems and symmetry, in *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, 2nd ed. (Cambridge University Press, Cambridge, 2012), Chap. 3, p. 83.
[68] N. Aubry, W.-Y. Lian, and E. S. Titi, Preserving symmetries in the proper orthogonal decomposition, SIAM J. Sci. Comput. **14**, 483 (1993).
[69] B. E. Sonday, M. Haataja, and I. G. Kevrekidis, Coarse-graining the dynamics of a driven interface in the presence of mobile impurities: Effective description via diffusion maps, Phys. Rev. E **80**, 031102 (2009).
[70] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, GROMACS: Fast, flexible, and free, J. Comput. Chem. **26**, 1701 (2005).
[71] M. G. Martin and J. I. Siepmann, Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes, J. Phys. Chem. B **102**, 2569 (1998).
[72] H. Berendsen, J. Postma, W. van Gunsteren, and J. Hermans, Interaction models for water in relation to protein hydration, in *Intermolecular Forces*, edited by B. Pullman (Reidel, Dordrecht, 1981), p. 331.
[73] A. W. Schüttelkopf and D. M. F. van Aalten, PRODRG: A tool for high-throughput crystallography of protein-ligand complexes, Acta Crystallogr., Sect. D **60**, 1355 (2004).
[74] M. P. Allen and D. J. Tildesley, *Computer Simulations of Liquids* (Oxford University Press, Oxford, 1989).
[75] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, A smooth particle mesh Ewald method, J. Chem. Phys. **103**, 8577 (1995).
[76] S. Nosé, A unified formulation of the constant temperature molecular dynamics methods, J. Chem. Phys. **81**, 511 (1984).
[77] M. Parrinello and A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method, J. Appl. Phys. **52**, 7182 (1981).
[78] R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles* (Taylor & Francis, New York, 1988).
[79] R. A. Gingold and J. J. Monaghan, Smoothed particle hydrodynamics: Theory and application to non-spherical stars, Mon. Not. R. Astron. Soc. **181**, 375 (1977).
[80] G. R. Liu and D. Karamanlidis, Mesh free methods: Moving beyond the finite element method, Appl. Mech. Rev. **56**, B17 (2003).
[81] M. Belkin and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. **15**, 1373 (2003).
[82] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, Diffusion maps, spectral clustering and reaction coordinates of dynamical systems, Appl. Comput. Harmonic Anal. **21**, 113 (2006).
[83] R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, Graph Laplacian tomography from unknown random projections, IEEE Trans. Image Proc. **17**, 1891 (2008).
[84] Available at http://davapc1.bioch.dundee.ac.uk/cgi-bin/prodrg.
[85] Available for free download at http://www.nlpca.org/matlab.html.
[86] B. Hess, H. Bekker, H. J. Berendsen, and J. G. Fraaije, LINCS: A linear constraint solver for molecular simulations, J. Comput. Chem. **18**, 1463 (1997).
[87] H. Whitney, Differentiable manifolds, Ann. Math. **37**, 645 (1936).
[88] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, Determination of reaction coordinates via locally scaled diffusion map, J. Chem. Phys. **134**, 124116 (2011).
[89] I. Jolliffe, *Principal Component Analysis*, 2nd ed. (Springer, New York, 2002).
[90] W. Kabsch, A solution for the best rotation to relate two sets of vectors, Acta Crystallogr., Sect. A **32**, 922 (1976).
[91] R. G. Littlejohn and M. Reinsch, Gauge fields in the separation of rotations and internal motions in the n-body problem, Rev. Mod. Phys. **69**, 213 (1997).
[92] M. Hestenes and E. Stiefel, Methods of conjugate gradients for solving linear systems, J. Res. Natl. Bureau Standards **49**, 409 (1952).
[93] E. Lorenz, Deterministic nonperiodic flow, J. Atmos. Sci. **20**, 130 (1963).
[94] P. Grassberger and I. Procaccia, Measuring the strangeness of strange attractors, Phys. D (Amsterdam, Neth.) **9**, 189 (1983).
[95] MATLAB, Ver. 7.10.0 (R2010a), MathWorks, Inc., Natick, MA, 2010.
[96] L. M. Pecora, L. Moniz, J. Nichols, and T. L. Carroll, A unified approach to attractor reconstruction, Chaos: An Interdisciplinary J. Nonlinear Sci. **17**, 013110 (2007).
[97] T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide* (Springer Science & Business Media, New York, 2010).