Recovery of Protein Folding Funnels from Single-Molecule Time Series by Delay Embeddings and Manifold Learning

Jiang Wang[†,¶] and Andrew L. Ferguson[*,‡]

†*Department of Physics, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, IL 61801, USA*

‡*Institute for Molecular Engineering, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA*

¶*Present address: Department of Chemistry, Rice University, P.O. Box 1892, MS 60 Houston, TX 77251, USA.*

E-mail: andrewferguson@uchicago.edu

Phone: (773) 702-3018

Abstract

The stability and folding of proteins is governed by the underlying single molecule free energy surface (smFES) mapping the free energy of the molecule as a function of configurational state. Ascertaining the smFES is of great value in understanding and engineering protein structure and function. By integrating tools from dynamical systems theory and nonlinear manifold learning, we describe an approach to reconstruct the multidimensional smFES for a protein from a time series in a single experimentally-measurable observable. We employ Takens' delay embeddings to project the time series into a high-dimensional space in which the projected dynamics are C^1-equivalent to the true system dynamics, and employ diffusion maps to recover a low-dimensional reconstruction of the smFES that is equivalent to the true smFES up to a smooth and invertible transformation. We validate the approach in molecular dynamics simulations of Trp-cage, Villin, and BBA to demonstrate that landscapes recovered from univariate time series in the head-to-tail distance are topologically identical – they precisely preserve the metastable states and folding pathways – and topographically approximate – the free energy barrier heights and well depths are approximately preserved – to the true landscapes determined from complete knowledge of all atomic coordinates. We go on to show that the reconstructed landscapes reliably predict temperature denaturation and identify point mutations and groups of mutations critical to folding. These results demonstrate that protein folding funnels can be reconstructed from experimentally-measurable time series and used to understand and engineer folding.

1 Introduction

The thermodynamic stability and folding pathways of proteins are governed by the underlying single molecule free energy surface (smFES) colloquially known as the folding funnel.[1,2] By quantifying protein stability and resolving low-free energy folding pathways, the smFES maps out the metastable states of the protein and the pathways between them, providing a wealth of understanding of protein structure, folding, and function.[1–7] Changes to the folding funnel as a function of the prevailing conditions (e.g., temperature, pressure, salt concentration) or mutations in the protein sequence can inform rational engineering of protein structure and function.[8–11] It is a primary objective in single molecule biophysics and protein folding and design to determine protein folding funnels.

The spatial location and configuration of a protein containing N atoms is uniquely specified by a $3N$-dimensional vector of Cartesian coordinates. Molecular dynamics calculations simulate the dynamical evolution of the protein through this high-dimensional configurational space.[12] Couplings between the molecular degrees of freedom mediated by bonded interactions, long-range electrostatic and dispersion interactions, and solvent-mediated hydrophobicity generically give rise to a small number of collective variables (CVs) describing the important large-scale / long-time motions of the protein to which the remaining degrees of freedom are effectively slaved.[13–16] Accordingly, the dynamical evolution of the protein is effectively restrained to a manifold of dimensionality far lower than the $3N$-dimensional space in which the dynamics are formulated.[6,7,7,14–21] The CVs spanning this low-dimensional intrinsic manifold present good order parameters with which to parameterize the smFES, since they naturally distinguish the various metastable states of the protein and are coincident with the important dynamical motions. Estimates of the CVs can be recovered from molecular simulation trajectories using dimensionality reduction techniques. Geometric dimensionality reduction techniques seek to determine a low-dimensional manifold residing within the high-dimensional simulation trajectory using linear[22–25] or nonlinear[26–34] approaches. Kinetic approaches instead seek to approximate the slowest eigenfunctions of the propagator

that evolves probability distributions over the molecular state space,[35] and can be formulated as a variational solution of a generalized eigenvalue problem.[35–41] Having identified CVs by such a method, molecular simulation data may then be projected onto these coordinates and the smFES estimated from the empirical distribution of points projected onto the low-dimensional intrinsic manifold.[6,7,17,39,42,43]

Molecular simulations, however, are subject to two primary sources of error: systematic errors introduced by the approximate nature of the classical mechanical force fields, and statistical errors introduced by incomplete sampling of configurational space.[44] Single molecule experiments are not subject to these deficiencies, but present their own difficulties since it is not currently possible to follow the coordinates of all atoms in a molecule as a function of time. X-ray crystallography and cryo-electron microscopy can resolve protein structures to essentially atomic resolution, but cannot track the protein dynamics. Fluorescent imaging techniques can probe protein dynamics by real-time optical tracking of conjugated fluorescent reporters.[45,46] Single molecule Förster resonance energy transfer (smFRET) operates as a "molecular yardstick" furnishing the distance between pairs of fluorophores grafted to particular locations in the protein.[46,47] Monitoring this distance provides coarse-grained information on the dynamical conformational changes of the protein, but is restricted to following a single (occasionally, a few) intramolecular distances and cannot furnish the location of all atomic coordinates as a function of time. How might one determine good CVs and recover estimates of the multidimensional smFES from experimental time series of a single molecular observable?

We have established a technique to achieve this goal by combining tools from dynamical systems theory and nonlinear manifold learning.[48] The approach appeals to Takens' Delay Embedding Theorem as a means to take time series in generic observables of a dynamical system and project them into a high-dimensional space within which the projected dynamics are related by a smooth, continuously differentiable function to the true dynamics of the system.[49–58] Manifold learning techniques are then used to extract from within this space

a low-dimensional manifold supporting a reconstruction of the smFES.[33,34,59] Takens' Theorem guarantees that the reconstructed manifold preserves all of the metastable states of the molecule and the transition pathways between them, but the height of the free energy barriers and depth of the free energy wells may be perturbed from their true values. In the context of protein folding, this approach provides a means to take univariate time series of single experimental observables (e.g., a smFRET intramolecular distance) and recover reconstructions of multidimensional protein folding funnels without ever requiring access to the atomic coordinates. In this work, we demonstrate empirically for a number of small proteins that the perturbation of the reconstructed smFES topography introduced by this procedure is relatively mild, and that the reconstructed landscapes can be used to understand and engineer protein stability and function. A schematic illustration of our approach is presented in Figure 1.

We have previously validated the approach in molecular dynamics simulations of a hydrophobic polymer chain to show the two-dimensional smFES reconstructed from knowledge of the dynamics of the chain head-to-tail distance to be topologically identical to the true landscape determined from analysis of the all-atom simulation trajectory.[48] It is the purpose of the present work to demonstrate the approach in realistic all-atom molecular dynamics simulations of small proteins, and assess the degree to which thermodynamic understanding may be extracted from the reconstructed folding funnels. Analysis of molecular simulation data allows us to validate the technique because the true smFES is available as a ground truth from analysis of the all-atom simulation trajectory. In the first part of this paper, we report the analysis of long folding trajectories of three small proteins Trp-cage, Villin, and BBA to demonstrate that our approach can faithfully recover topologically equivalent reconstructions of their protein folding funnels from univariate time series in the protein head-to-tail distance. We show the landscapes to be consistent with existing understanding, numerically validate the existence of the diffeomorphism asserted by Takens' Theorem, and demonstrate the degree of topographical perturbation of the free energy barrier heights and

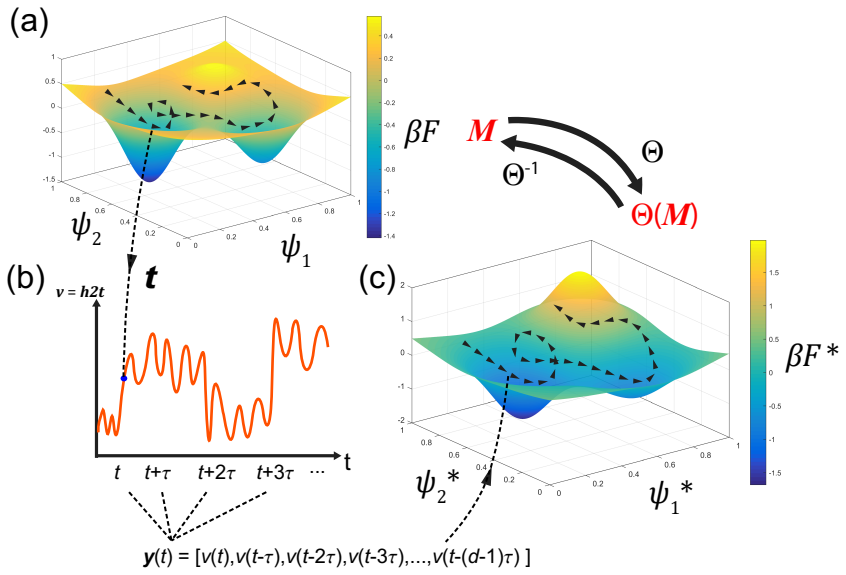


Figure 1: Schematic illustration of the smFES reconstruction approach. (a) The protein folding funnel is supported by a low-dimensional manifold M within the high-dimensional state space of all atomic coordinates. CVs parameterizing M can be ascertained by applying nonlinear manifold learning tools to molecular dynamics trajectories recording the dynamical evolution of all atomic coordinates. In this example, $\{\psi_1, \psi_2\}$ denote the CVs spanning the 2D manifold M and $\beta F(\psi_1, \psi_2)$ is the dimensionless free energy surface supported by M where $\beta = 1/k_B T$. (b) Tracking of an experimentally measurable coarse-grained observable of the system v furnishes a univariate time series that provides information on the dynamical motions of the protein, for example the intramolecular distance between two fluorophores measured by smFRET. Takens' Delay Embedding Theorem stipulates that the univariate time series can be projected into a high-dimensional space in which the dynamical evolution is C^1-equivalent to that of the true dynamics of the system being observed by forming d-dimensional delay embeddings $\vec{y}(t) = [v(t), v(t-\tau), v(t-2\tau), \ldots v(t-(d-1)\tau)]$. (c) Under some technical conditions on v, d, and τ discussed below, the dynamical evolution of the delay embedding $\vec{y}(t)$ maps out a manifold $\Theta(M)$ that is diffeomorphoic to M. In other words, Θ defines a smooth and invertible transformation between the true and reconstructed manifolds $\Theta : M \to \Theta(M)$. Consequently, $\Theta(M)$ is topologically equivalent to M in that it preserves the same metastable states and pathways between them, but may be topographically perturbed in that the free energy barrier heights and well depths may be perturbed. No theoretical bounds are available to limit the degree of this perturbation, but we provide empirical evidence in simulations of a number of small proteins that it is relatively mild. Applying nonlinear manifold learning to the delay embedding trajectories can identify the CVs $\{\psi_1^*, \psi_2^*\}$ spanning $\Theta(M)$ and supporting the reconstructed smFES $\beta F^*(\psi_1^*, \psi_2^*)$.

well depths to be relatively mild. In the second part of the paper, we analyze an ensemble of molecular simulations of Trp-cage at a variety of temperatures and containing different engineered mutations. The reconstructed landscapes reliably predict temperature denatu-

ration and accurately identify those mutations that have strong and weak influences on the stability of the native fold. This work demonstrates that protein folding landscapes may be reliably reconstructed from univariate time series in experimentally-accessible molecular observables, and the landscapes effectively used to understand and engineer protein folding.

2 Methods

2.1 Molecular dynamics simulations

Molecular simulation data analyzed in this work were derived from two sources. Long (> 100 μs) simulations of three fast-folding proteins obtained from D.E. Shaw Research (DESRES),[60] and an ensemble of simulations of Trp-cage at different temperatures and containing various mutations were conducted in-house.

DESRES simulations of Trp-cage, Villin, and BBA. Molecular dynamics simulations of Trp-cage, Villin, and BBA were conducted by D.E. Shaw Research as reported in Ref.[60] Simulations were conducted using the Desmond simulation package[61] running on the special purpose Anton supercomputer.[62] Systems were modeled using the CHARMM22* force field[63] and the modified TIP3P water model.[64,65] The N-termini, C-termini, Lys, Arg, Asp, and Glu residues were modeled as charged, and the His residues as neutral unless otherwise specified. Systems were equilibrated in the NPT ensemble and production runs conducted in the NVT ensemble. Temperature was maintained by a Nosé-Hoover thermostat with a time constant of 1 ps. Equations of motion were numerically integrated with a 2.5 fs time step. Lennard-Jones and short-range Coulomb interactions were treated with a 0.9-0.95 nm real space cutoff and long-range Coulombic forces treated with the Gaussian Split Ewald technique[66] over a 32×32×32 cubic grid. The 20-residue Trp-cage (DAYAQWLADGGPSSGRPPPS, PDB ID 2JOF), representing the thermostable K8A mutant of TC10b,[67] was simulated at 290 K for 208 μs. Simulations were conducted in 65 mM NaCl in a $\sim$3.7 nm cubic box of $\sim$1700 water molecules. We extracted the first 200

μs of the simulation trajectory for analysis. The 35-residue Villin (LSDEDFKAVFGMTR-SAFANLPLW(Nle)QQHL(Nle)KEKGLF, PDB ID 2F4K), representing the double norleucine mutant of the C-terminal fragment of the Villin headpiece,[68] was simulated at 360 K for 125 μs. Simulations were conducted in 40 mM NaCl in a ∼5.4 nm cubic box of ∼4400 water molecules, and the His residue is modeled as charged. We extracted the first 120 μs of the simulation trajectory for analysis. The 28-residue BBA (EQYTAKYKGRTFRNEKELRD-FIEKFKGR, PDB ID 1FME) was simulated at 325 K for 325 μs. Simulations were performed in a ∼4.7 nm cubic box of ∼3200 water molecules and four chloride ions to maintain charge neutrality. We extracted the first 200 μs of the simulation trajectory for analysis.

In-house simulations of Trp-cage. We also conducted in-house simulations of the 20-residue TC5b variant of the Trp-cage mini protein (NLYIQWLKDGGPSSGRPPPS, PDB ID 1L2Y).[69] Simulations were conducted using the OpenMM simulation suite.[70] The protein was modeled using the Amber99sb force field[71] and the water solvent modeled implicitly using the Amber99 GBSA-OBC model.[72] The Langevin equation of motion were numerically integrated with a 2 fs time step. A friction coefficient of 1 ps^{-1} was employed to maintain temperature at 380 K unless otherwise specified. Due to the use of implicit solvent, simulations were conducted in a formally infinite domain and no non-bonded cutoffs were employed. Bonds involving a hydrogen atom were constrained to a fixed length to improve the stability of the numerical integration. Each 1 μs simulation was conducted on a single NVIDIA GeForce GTX 770M GPU card to achieve execution speeds of ∼1100 ns/day. We performed 37 independent simulations partitioned into four classes: (I) 8 × simulations of wild-type TC5b Trp-cage at evenly-spaced temperatures spanning 300-440 K, (II) 4 × simulations of engineered Trp-cage mutants that were the subject of prior experimental study,[69] (III) 20 × alanine point mutants, and (IV) 5 × alanine tetrad mutants. An accounting of the 37 simulations is provided in Table 1.

2.2 Attractor reconstruction by Takens' delay embeddings

Takens' Delay Embedding Theorem provides a means to recover reconstructions of the geometry and topology of the phase space occupied by a dynamical system without having access to the dynamical evolution of all system degrees of freedom.[49–57] The theorem has been employed in diverse applications including climate modeling,[73] fishery forecasting,[58] dynamical mode decomposition of reaction-diffusion and liquid crystal growth,[74] and analysis of peptide dynamics.[75,76] In the context of molecular folding, we have previously shown how the theorem can be used to develop approximations for the smFES from measurements of one or more coarse-grained observables of the system.[48]

Let us consider a classical molecular system comprising N atoms that dynamically evolves under Newton's equations of motion within the $3N$-dimensional state space parameterized by all atomic coordinates. A simulation trajectory can be considered a ordered sequence of C simulation snapshots $\{\vec{r}_i\}_{i=1}^{C}$ describing the progression of the system through state space. Couplings between the degrees of freedom generically restrain the dynamics to occupy a k-dimensional intrinsic manifold $M \in \mathcal{R}^k \subset \mathcal{R}^{3N}$ with $k << 3N$.[6,7,7,14–21] This manifold can be recovered from $3N$-dimensional molecular simulation trajectories by applying manifold learning techniques to identify CVs spanning M and then estimating the smFES it supports from the distribution of projected points into these CVs.[6,7,17,39,42,43] We will refer to the smFES over M recovered from analysis of all-atom trajectories as the true smFES of the system.

Consider now a generic measurement function of a scalar observable of the system $v : \mathcal{R}^k \rightarrow \mathcal{R}$. In the context of protein folding, v may be an intramolecular distance between two fluorophores measured by smFRET.[46] Takens' Delay Embedding Theorem asserts that the state of the system is uniquely specified under the $d \geq (2k+1)$-dimensional delay embedding $\vec{y}(t) = [v(t), v(t-\tau), v(t-2\tau), \ldots v(t-(d-1)\tau)]$ that matches up each instantaneous scalar observation of the system $v(t)$ with $(d-1)$ past observations time-delayed by increments of τ. The theorem further asserts that the dynamics of $\vec{y}(t)$ are C^1-equivalent – related

by a smooth, continuously differentiable function – to the true all-atom dynamics and are restrained to a manifold $\Theta(M) \in \mathcal{R}^k \subset \mathcal{R}^d$ where $\Theta : M \to \Theta(M)$ defines a diffeomorphism – a smooth and invertible mapping – between the true and reconstructed manifolds.[48–57] The reconstructed manifold $\Theta(M)$ can be recovered from the time evolution of the delay embedding $\vec{y}(t)$ using the same manifold learning techniques as applied to the molecular simulation trajectories. We will refer to the smFES over $\Theta(M)$ recovered by appealing to Takens' delay embeddings of time series data as the reconstructed smFES of the system.

An important consequence of this theorem is that it provides a means to reconstruct protein folding funnels from time series measurements in experimentally accessible observables. The properties of the diffeomorphism $\Theta : M \to \Theta(M)$ relating the true and reconstructed manifolds are such that it cannot tear or restitch the manifold but may stretch and squash it.[49–52,56,77] Accordingly, the reconstructed manifold is guaranteed to be *topologically identical*, preserving the edges, continuity, and connectivity of the true manifold, and therefore maintain all of the metastable states of the molecule and the transition pathways between them. However, it may be *topographically perturbed*, such that the height of the free energy barriers and depth of the free energy wells are shifted away from their true values. We demonstrate in this work through empirical comparisons of M and $\Theta(M)$ that the topographical perturbations induced by the diffeomorphism are relatively mild, and that the reconstructed smFES preserves a high degree of quantitative interpretability.

We now consider a few technical, but important, aspects of the theorem that must be confronted in its practical application.

$\boxed{v}$ Taken's Theorem holds for any generic observable v of the system that is a function of all system degrees of freedom and does not contain any symmetries not present in the system.[49,57,78,79] Let us make three important observations about this condition. First, in applications to protein folding, the experimentally-accessible observable is typically expected to be some function of the protein coordinates such as an intramolecular distance. Depending on the choice of intramolecular distance, this observable may depend more or less strongly

on particular intramolecular degrees of freedom. Moreover, the dynamical evolution of the protein is also influenced by solvent degrees of freedom and the dynamics of deterministic or stochastic thermostats and barostats, applied constraints, and other external couplings. Accordingly, our application of Takens' Theorem is actually to observations of a subsystem – the portion of the protein governing the observed intramolecular distance – subject to a number of external couplings and forcings. The validity of this application is supported by generalizations of the theorem by Stark and co-workers who showed it to hold under very general conditions for (sub)systems subject to stochastic or deterministic forcing.[54,55]

Second, what if the chosen observable does contain symmetries not present in the system? For example, an intramolecular distance measured by smFRET is invariant to head-to-tail and mirror inversions of the molecule, whereas the molecule itself may not possess these symmetries. In this case, the reconstructed manifold $\Theta(M)$ will collapse out these symmetries and will not be diffeomorphic to M. It is necessary, therefore, to moderate our goal to instead seek a spatially symmetrized reconstruction of M that eliminates the symmetries present in the observable. This modification can be straightforwardly achieved by applying diffusion map manifold learning in a manner that mods out these spatial symmetries, and we describe how we achieve this in Section 2.3.[48] We also observe that it may be possible to lift the degeneracy by switching to an observable that does not contain these spurious symmetries, or by applying Takens' Theorem to multiple simultaneous observables that taken together eliminate these symmetries.[53]

Third, we adopt throughout this work the molecular head-to-tail distance as our experimentally-accessible time series and assert that this can, in principle, be measured by a technique such as smFRET. We appreciate that achieving this in practice entails significant experimental challenges, including the chemical conjugation of large fluorophores to small molecules and the dynamical perturbations they might induce, the recovery of long trajectories with high signal-to-noise ratios, sub-ms time resolution, the possibility of photobleaching, and the recording of reliable measurement of distances outside the 2-8 nm range.[46,47] It is the goal

of the present work to consider idealized smFRET measurements in order to validate the principles of our technique in long simulations of a number of small fast-folding proteins. This establishes the methodology in this idealized limit, and lays the groundwork for its future extension to real experimental data.

$\boxed{\tau}$ The theorem places no conditions on the delay time τ, but in practical applications a good choice of delay time is crucial in making best use of the data and producing high quality embeddings. We identify an appropriate value of τ using the approach of Fraser and Swinney, which calculates the mutual information between time-delayed pairs of the observable $(v(t), v(t-\tau))$.[80] The optimal value of τ is selected as that at the first minimum in the mutual information,[80] or – in the absence of a well-defined local minimum – the τ at which the mutual information decays to $1/e$ of its initial value.[57]

$\boxed{d}$ Takens' Theorem guarantees the existence of a diffeomorphism for delay embeddings of dimensionality $d \geq (2k+1)$. In practical applications, the dimensionality k of the intrinsic manifold may not be known *a priori*, and – although not assured – may exist for delay embeddings of lower dimensionality $k \leq d < (2k+1)$.[50,81] In practice, we use the method of Cao to ascertain the minimum delay embedding dimensionality at which the embedding becomes fully unfolded and there are no artificial intersections of the embedding or false nearest neighbors.[82,83]

Temporal symmetry breaking. A molecular system at thermodynamic equilibrium is constrained to obey time-reversal symmetry and detailed balance wherein every microscopic transition is equilibrated by the reverse process.[84,85] The construction of delay embeddings breaks time reversal symmetry by imposing a temporal ordering on the scalar observations. Specifically, a microscopic configuration $\vec{r}_A$ transitioning *to* a neighboring microstate $\vec{r}_B$ is distinguishable from the same microscopic configuration $\vec{r}_A$ that has just transitioned *from* $\vec{r}_B$. This distinguishability arises because the delay embedding vectors contain the history of the transition, and the delay embedding vectors for the forward and backward transitions are related by inversion of their elements. This temporal symmetry breaking

leads to an apparent violation of detailed balance within the reconstructed manifold $\Theta(M)$ since forward and backward transitions between any pair of microstates will not lie coincident within the manifold, but rather be separated across a plane of symmetry. We must eliminate the artificial temporal symmetry breaking in order to enforce detailed balance within the reconstructed manifold $\Theta(M)$ and achieve a diffeomorphism with the true manifold M where detailed balance is naturally enforced. We previously developed a means to eliminate this symmetry breaking by augmenting our scalar time series with its time reversed analog and identifying and eliminating these symmetry planes.[48] In this work, we introduce a simpler means to eliminate this temporal symmetry breaking in our application of diffusion maps, and we describe how we achieve this in Section 2.3. We note that elimination of this symmetry breaking is likely not warranted or desirable in the analysis of out-of-equilibrium systems.

Diffeomorphic bijection. The diffeomorphism $\Theta : M \to \Theta(M)$ asserts a smooth and invertible mapping between the two manifolds, and implies a bijection of each point in M to its image in $\Theta(M)$. For C observations of the molecular system $\{\vec{r}_i\}_{i=1}^{C}$, the manifold M will comprise C points, whereas $\Theta(M)$ typically contains $(C - d + 1)$ projections of the delay embedding vectors $\{\vec{y}_j\}_{j=1}^{(C-d+1)}$, where d is the dimensionality of the delay embedding. In order to establish a one-to-one mapping between the manifold, we adopt the convention that each delay vector projection on $\Theta(M)$ will be matched to the first of the d simulation snapshots from which the delay vector was constructed. (An alternative convention might choose to instead match the central, last, or mean simulation snapshot.) The initial $(d - 1)$ points in the simulation trajectory for which this average is undefined are discarded.

2.3 Diffusion maps manifold learning

Diffusion maps are an unsupervised nonlinear manifold learning technique that can be used to identify and extract low-dimensional manifolds latent within high-dimensional data sets.[17,33,34,59,86,87] The approach functions by constructing a random walk over the high-dimensional data set and performing a spectral decomposition of the resulting dynamics to

identify a low-dimensional surface to which the data are effectively restrained. In this work we use this approach to recover parameterizations of the intrinsic manifold M within the high-dimensional molecular simulation trajectories, and its image $\Theta(M)$ residing within the high-dimensional Takens' delay embeddings.

The first step in applying diffusion maps is to compute pairwise distances d_{ij} between each pair of points (i, j) in the high-dimensional data. In applying diffusion maps to the molecular simulation trajectories $\{\vec{r}_i\}_{i=1}^{C}$ we adopt the root mean squared deviation (RMSD) between the atomic coordinates of the C_α atoms of the translationally and rotationally aligned protein structures in each snapshot of the simulation trajectory that we efficiently compute using the Kabsh algorithm.[88] Furthermore, we also minimize this distance over head-to-tail inversion of the protein by inverting the indexing of the C_α atoms, and also mirror symmetry by minimizing over the original and mirror image of the C_α coordinates. This final step symmetrizes the manifold M with respect to these two inversions, modding out these transformations that cannot be distinguished by our choice of the head-to-tail distance as our observable and allowing $\Theta(M)$ to approximate the true manifold by application of Takens' Theorem to this observable. This choice of distance measure produces a *spatially symmetrized* representation of the manifold M. In applying diffusion maps to the delay embedding trajectories $\{\vec{y}_j\}_{j=1}^{(C-d+1)}$, we adopt the Euclidean distance between the the high-dimensional delay vectors $\vec{y}(t) = [v(t), v(t-\tau), v(t-2\tau), \ldots v(t-(d-1)\tau)]$ minimized under inversion of the ordering of the delay vector elements. In a similar way that head-to-tail and mirror inversion eliminates the spatial symmetries in the molecular configurations, this operation eliminates the spurious symmetry breaking introduced by the temporal ordering of the observables in the delay vector. This choice of distance measure produces a *temporally symmetrized* reconstructed manifold $\Theta(M)$.

After computing all pairwise distances, we convolute d_{ij} with a Gaussian kernel to form the matrix elements $A_{ij} = \exp(-d_{ij}^2/2\epsilon)$. The kernel bandwidth $\sqrt{\epsilon}$ defines the characteristic hop size of the random walk over the high-dimensional data set, which can be automati-

cally tuned to an appropriate value using the approach in Ref.[89] We then row-normalize $\mathbf{A}$ to obtain the Markov matrix $\mathbf{M} = \mathbf{D}^{-1}\mathbf{A}$, where $\mathbf{D}$ is a diagonal matrix with elements $D_{ii} = \sum_k A_{ik}$. The element M_{ij} of the right-stochastic Markov matrix $\mathbf{M}$ defines the hopping probability from state i to state j under the action of the discrete random walk in the high-dimensional space.[86] A spectral decomposition of $\mathbf{M}$ furnishes a set of eigenvalues $\lambda_1 = 1 \geq \lambda_2 \geq \ldots \geq \lambda_N$ and right eigenvectors $\{\psi_k\}_{k=1}^N$. The leading eigenvectors are discrete approximations to the leading eigenfunctions of the backward Fokker-Planck equation characterizing the slowest diffusion modes over the data.[33,86,89] A gap in the eigenvalue spectrum after λ_{k+1} informs a separation of time scales for the diffusion process and the identification of a k-dimensional intrinsic manifold within the high-dimensional space. After discarding the trivial top eigenvector $\psi_1 = \vec{1}$, the top k non-trivial eigenvectors serve as good CVs parameterizing the intrinsic manifold and inform the low-dimensional embedding,

$$\text{observation}_i \rightarrow \left(\vec{\psi}_2(i), \vec{\psi}_3(i), \ldots \vec{\psi}_{k+1}(i)\right). \tag{1}$$

The leading eigenvectors provide good embedding coordinates as the slowest relaxing modes of the random walk over the high-dimensional data, but the diffusion map does not provide an explicit mapping from the original coordinate space. In some cases it is possible to correlate the eigenvectors with physical variables by visualizing heat maps, performing linear correlation analyses, or screening pools of candidate physical variables.[90–92] It is not always possible, however, to identify a simple physical correspondence, and it is not surprising that the CVs emerging from the many-body interactions within complex dynamical systems may defy simple understanding or human intuition.[93] In the present work, this physical correspondence is welcome and useful, but not needed since all that we require of the CVs is to provide a good parameterization of the low-dimensional manifold.

Simple application of diffusion maps requires the calculation and storage of $N \times N$ pairwise distances, which can be computationally prohibitive for large N. In this work we employ

pivot diffusion maps (P-dMaps) as an algorithmic implementation that obviates the need to compute the full $N \times N$ matrix.[94] This approach computes the diffusion map embedding of $n << N$ pivot points spanning the manifold that are automatically selected on-the-fly and then subsequently projects in the remaining $(N - n)$ data points using the Nyström extension.[95–97] This technique drastically reduces the time complexity of applying diffusion maps from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \times n)$ with very little loss loss in accuracy.[94]

Large density variations over the intrinsic manifold can compromise the ability of the diffusion map embedding to construct a random walk that simultaneously has a sufficiently large step size to span the manifold but a sufficiently small step size to resolve important details within the high density regions of the space.[98–100] We resolve this difficulty by employing a density adaptive variant of diffusion maps that rescales pairwise distances as $d_{ij} \leftarrow d_{ij}^{\alpha}$ according to a tunable exponent $\alpha \in (0, 1]$.[100] This operation reduces the difference between large and small distances over the manifold while rigorously maintaining the triangle inequality. This mitigates the apparent density fluctuations over the manifold and produces superior diffusion map embeddings. Empirical tests suggest that an appropriate value of α may be estimated by reducing the apparent local density fluctuations to $\sim 10^2$.[100]

3 Results and Discussion

3.1 smFES reconstruction for Trp-cage, Villin, and BBA mini proteins

We first apply our approach to long molecular dynamics simulation trajectories of three proteins conducted by D.E. Shaw Research:[60] (i) the 20-residue Trp-cage (PDB ID 2JOF) fast-folding engineered mini protein that has been the subject of extensive experimental and computational study, (ii) the 35-residue Villin (PDB ID 2F4K) corresponding to the three-helix headpiece of an actin-binding protein, and (iii) the 28-residue BBA (PDB ID 1FME) designed FSD-EY protein containing a β-β-α native fold. Simulation details are provided in

Section 2.1. The goal of this study is to first recover estimates of the smFES by applying diffusion map manifold learning to the all-atom simulation trajectories (Section 2.3), and then compare these to reconstructions of the smFES determined by applying Takens' Theorem and diffusion maps to time series of the protein head-to-tail distance ($h2t$) between the first and last C_α atoms (Section 2.2). We extract this time series directly from the simulation trajectories under the presumption that such an intramolecular observable could, in principle, be measured by a technique such as smFRET.[46] We present in Figure 2 molecular images of the native state of each of the three proteins along with the $h2t$ time series extracted from each simulation trajectory.

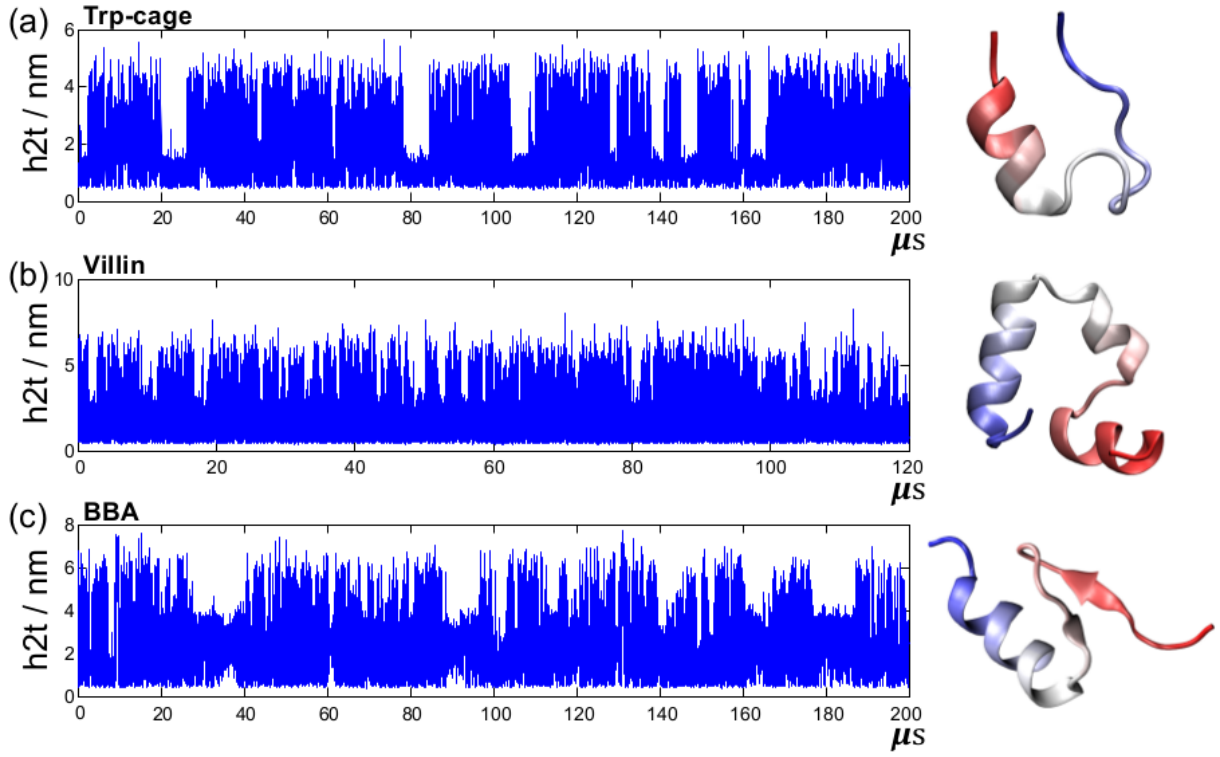


Figure 2: Time series in the head-to-tail distance ($h2t$) extracted from long molecular dynamics simulation trajectories for three mini-proteins (a) Trp-cage (PDB ID 2JOF), (b) Villin (PDB ID 2F4K), and (c) BBA (PDB ID 1FME).[60] Molecular models of the native state of each protein are also provided showing the secondary structural elements and shaded red-to-blue from the N-terminus to C-terminus. All molecular renderings in this work are constructed using VMD.[101]

3.1.1 Trp-cage (PDB ID 2JOF)

All-atom trajectories. We first apply diffusion maps to the 200 μs all-atom simulation trajectory of Trp-cage at $T = 290$ K to discover good CVs with which to parameterize the intrinsic manifold M supporting the true smFES. The 1,000,000 frame trajectory is sampled at 0.2 ns resolution, which we evenly subsample down to 100,000 frames for analysis. Distances between pairs of molecular configurations required in the application of diffusion maps are measured according to the RMSD between C_α atom coordinates minimized over translation, rotation, head-to-tail inversion, and mirror inversion. As detailed in Section 2.3, these symmetrizing operations are required to respect the fact that these transformations are not distinguishable to our choice of molecular observable $h2t$, and so we can only hope to recover a spatially symmetrized version of the smFES. We apply pivot diffusion maps with a pivot cutoff radius of 0.41 nm in the RMSD to identify 622 pivot points.[94] A diffusion map with bandwidth $\sqrt{\epsilon} = 1.0$ nm and density rescaling exponent $\alpha = 0.15$ produces an eigenvalue spectrum with a gap after λ_3 implying a 2D intrinsic manifold that can be parameterized by (ψ_2, ψ_3).

We present in Figure 3 the embedding of the all-atom simulation snapshots into the 2D intrinsic manifold M parameterized by these two CVs. Coloring each embedded point by the molecular radius of gyration R_g (Figure 3a) shows the top-left corner to contain highly extended configurations that flow first to the southeast and then to the northeast to occupy the collapsed configurations in the northeastern lobe. Coloring by the RMSD from the native state (Figure 3b) confirms that this lobe contains the native fold. Finally, restricting the RMSD to measure distance from the native N-terminal α-helix formed by residues 2-8 (RMSD-helix) shows that the lobe is distinguished from the rest of the embedding by folding of the N-terminal α-helix (Figure 3c).

We present in Figure 4 the smFES supported by the true intrinsic manifold M spanned by (ψ_2, ψ_3) estimated from histograms in the observed distribution of projected points over the manifold $\hat{P}(\boldsymbol{\psi})$ via the relation $\beta F(\boldsymbol{\psi}) = -\ln \hat{P}(\boldsymbol{\psi}) + C$, where $\boldsymbol{\psi}$ specifies the location

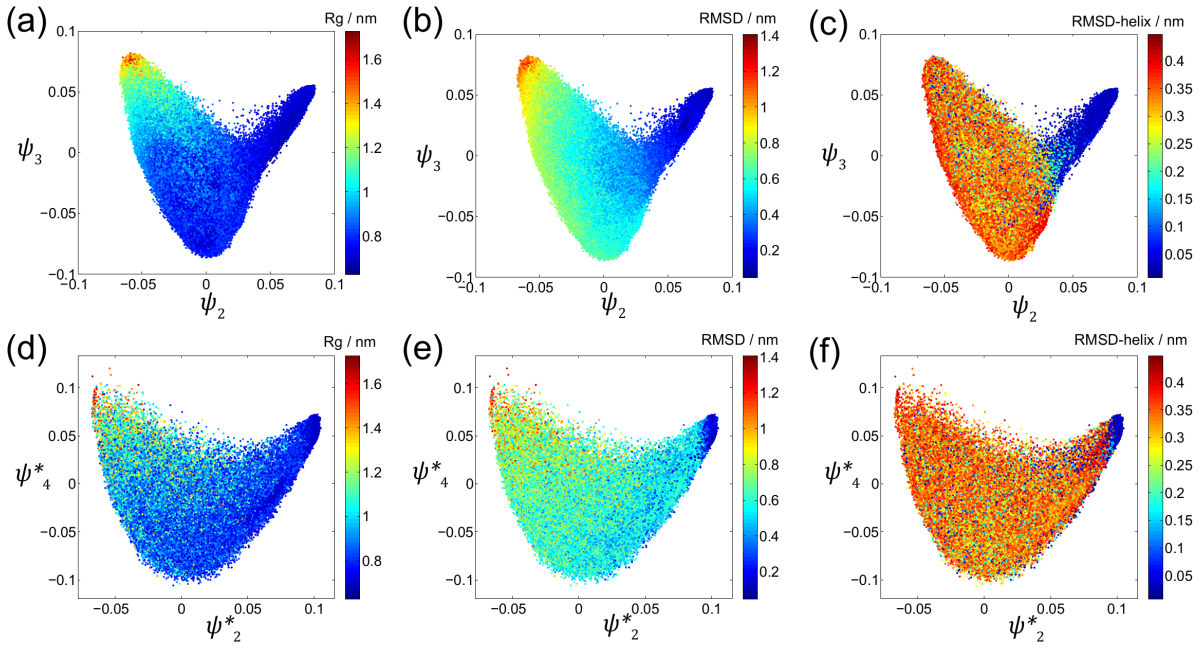


Figure 3: Representations of the 2D intrinsic manifold of Trp-cage (PDB ID 2JOF). (a-c) The true manifold M spanned by the CVs (ψ_2, ψ_3) discovered by application of diffusion maps to the all-atom simulation trajectory and colored by R_g, RMSD, and RMSD-helix. Each point represents a projection of an all-atom configuration observed in the molecular simulation trajectory. (d-e) The reconstructed manifold $\Theta(M)$ spanned by the CVs (ψ^*_2, ψ^*_4) discovered by construction of Takens' delay embeddings of the $h2t$ time series extracted from the simulation trajectory and subsequent application of diffusion maps. Each point represents a projection of a (d=50)-dimensional delay vector of $h2t$ values and is colored by the R_g, RMSD, and RMSD-helix corresponding to the first configuration in the delay vector.

on the manifold, $\beta = 1/k_BT$ where k_B is Boltzmann's constant and T is the simulation temperature, and C is an arbitrary additive constant reflecting our ignorance of the absolute free energy scale. The smFES over M illustrated in Figure 4b reveals the existence of two well-defined free energy wells corresponding to the native folded state **A** and a proximate metastable state **B**. Molecular renderings of representative molecular configurations within these basins are presented in Figure 4e. The native fold **A** resides at the global free energy minimum of the smFES and comprises a N-terminal α-helix, 3_{10}-helix, C-terminal polyproline tail, and well-defined hydrophobic core caging the Trp-6 side chain. The state **B** is a low-lying metastable state residing $\sim$0.5 k_BT higher in free energy than the native fold. It is structurally very similar to **A** with the exception that the 3_{10}-helix is unfolded, and is a

well-known structural intermediate identified in a number of prior studies.[102–105] These two low-lying states are connected by a narrow pass to a plateau lying $\sim$2.5 k_BT higher in free energy. This largely featureless plateau contains the unfolded ensemble, although some weak local minima corresponding to marginally metastable structures are apparent, one of which we have called out as configuration **C**.

The smFES is in good qualitative agreement with that recovered by Juraszek *et al.* employing the OPLS force field and SPC water model,[102] Kim *et al.* employing the Amber03w force field and TIP4P/2005 water,[103] and by us employing employing the Amber03 force field and implicit solvent.[104] In all cases the landscape exhibits well-defined low-lying minima near the native fold and a large unfolded ensemble. To facilitate comparison of the smFES with prior studies of Trp-cage, we present in Figure 4a the smFES reweighted into the two conventional order parameters (RMSD,RMSD-helix). The strong correlation between (ψ_2, ψ_3) and (RMSD,RMSD-helix) evinced in Figure 3 allows us to anticipate good preservation of the structure of the smFES in this projection, although the data-driven CVs identified by diffusion maps do a superior job in resolving structure within the unfolded ensemble. Kim *et al.* analyzed Trp-cage using diffusion maps[103] and we have previously employed artificial neural networks to similar effect.[104] In both instances a 2D intrinsic manifold was identified containing the two low-lying metastable **A** and **B** basins. However – in addition to the standard caveats regarding force fields and simulation protocols – we caution against too close comparisons with the present work due to our elimination of the spatial symmetries that mods out head-to-tail and mirror inversions.

h2t time series. We now recover the reconstructed manifold $\Theta(M)$ from a knowledge of only the time evolution of the single molecular observable by constructing high-dimensional Takens' delay embeddings and applying diffusion maps to extract the image of the low-dimensional manifold within this space. We generate the univariate time series in $h2t$ by tracking the value of $h2t$ in each of the 1,000,000 frames of the simulation trajectory (Figure 2a). Following the Takens' delay embedding protocol detailed in Section 2.2, we define an ap-

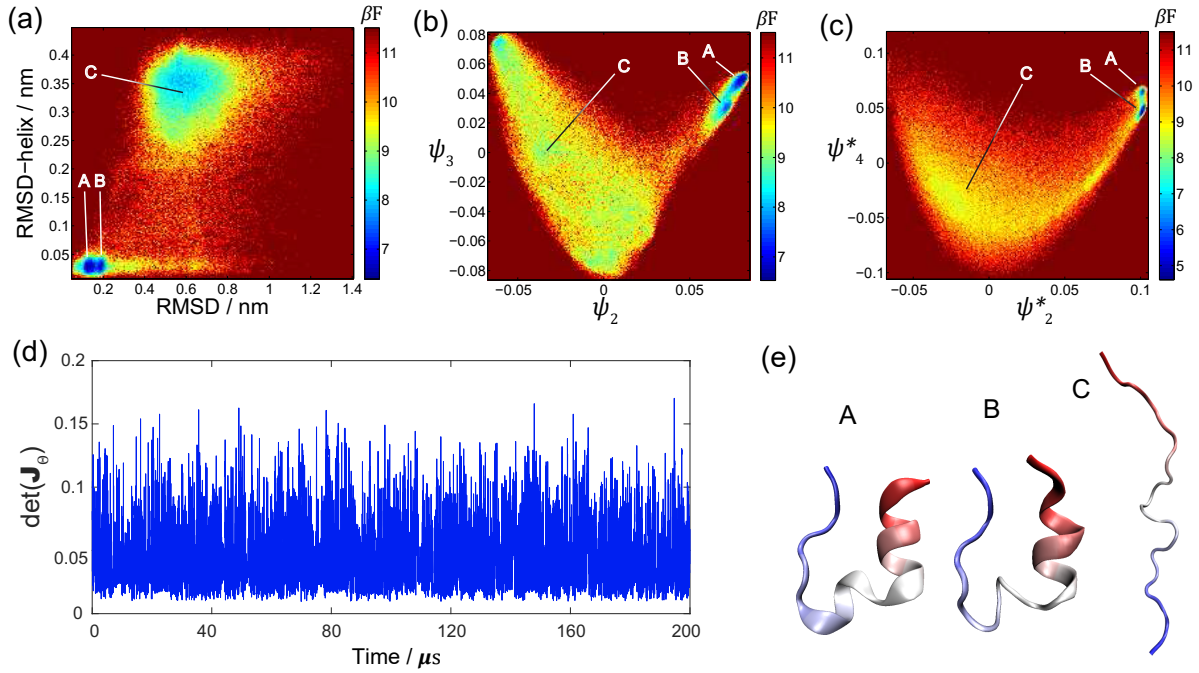


Figure 4: Single molecule free energy surfaces of Trp-cage (PDB ID 2JOF). (a) smFES spanned by the conventional CVs (RMSD,RMSD-helix). (b) smFES supported by the true intrinsic manifold M spanned by the CVs (ψ_2, ψ_3) identified by applying diffusion maps to the all-atom simulation trajectory. (c) smFES supported by the reconstructed manifold $\Theta(M)$ spanned by the CVs (ψ_2^*, ψ_4^*) identified by applying diffusion maps to delay embeddings of the $h2t$ time series. We report the Helmholtz free energy F dedimensionalized by $\beta = 1/k_B T$ where $T = 290$ K is the simulation temperature. (d) Determinant of the Jacobian $\mathbf{J}_\Theta$ for the forward transformation $\Theta : M \to \Theta(M)$ numerically evaluated at each point in the embedding and displayed as a function of elapsed simulation time. (e) Representative molecular snapshots at the locations identified in panels a-c.

propriate delay time $\tau = 0.6$ ns and delay embedding dimensionality of $d = 50$. Turning now to the 100,000 snapshot subsampled trajectory, we then construct delay embedding vectors employing these parameters for all configurations sufficiently advanced from the beginning of the simulation ($t > 30$ ns) so that all elements of the delay vector are defined. This results in the generation of 99,985 delay vectors populating a (d=50)-dimensional space. Pairwise distances between delay vectors are measured under a Euclidean metric minimized under inversion of the vector elements. As detailed in Section 2.3, this operation eliminates spurious symmetry breaking introduced into the temporal ordering of the delay vector elements. Application of pivot diffusion maps with a pivot cutoff radius of 8.5 nm in the delay vector

$h2t$ Euclidean distance identifies 515 pivot points. A diffusion map with bandwidth $\sqrt{\epsilon} =$ 1.45 nm, and density rescaling exponent $\alpha = 0.15$ produces an eigenvalue equation with a gap after λ_4^* indicating that embeddings should be constructed in $(\psi_2^*, \psi_3^*, \psi_4^*)$. (For clarity of exposition, we use an asterisk to distinguish quantities pertaining to the reconstructed smFES.) Analysis of these eigenvectors reveals that ψ_3^* is functionally correlated with ψ_2^* mapping out a 1D manifold in the (ψ_2^*, ψ_3^*) plane. Accordingly, ψ_3^* is effectively slaved to ψ_2^* and can be discarded from the low-dimensional embedding without loss of information on the structure of the low-dimensional manifold.[7,94]

We present in Figure 3d-f the projection of the delay vectors into the 2D reconstructed intrinsic manifold $\Theta(M)$ spanned by the CVs (ψ_2^*, ψ_4^*). The bijection between the true manifold M (Figure 3a-c) and the reconstructed manifold $\Theta(M)$ (Figure 3d-f) asserted by Takens' Theorem implies the existence of a nonlinear transformation between the CVs spanning the two spaces. Visual inspection of the upper and lower rows of Figure 3 suggests that ψ_2 is closely related to ψ_2^* and ψ_3 to ψ_4^*. This is confirmed by a linear correlation analysis that reveals Pearson correlation coefficients of $\rho(\psi_2, \psi_2^*) = 0.81$ ($p < 1 \times 10^{-10}$) and $\rho(\psi_3, \psi_4^*) =$ 0.44 ($p < 1 \times 10^{-10}$). There is also close correspondence in the gradations in R_g, RMSD, and RMSD-helix, although there is clearly some degree of non-uniform stretching and squashing. Further, the smFES supported by $\Theta(M)$ reported in Figure 4c preserves a similar topography to that over M reported in Figure 4b, demonstrating the existence of the native **A** and metastable **B** states and unfolded ensemble **C**.

The influence of the transformation $\Theta : M \to \Theta(M)$ is quantified through calculation of its Jacobian $\mathbf{J}_\Theta$ at each point on M. Verification that the Jacobian determinant $\det(\mathbf{J}_\Theta)$ does not pass through zero is, by the inverse function theorem, sufficient to validate the existence of the diffeomorphism asserted by Takens' Theorem.[106–108] The sign of the Jacobian determinant simply serves to specify whether orientation is preserved (positive) or inverted (negative). Takens' Theorem is silent, however, in placing any bounds on the degree to which the diffeomorphism may perturb the manifold through non-uniform compression and

dilation. This may be numerically evaluated *post hoc* by studying the magnitude of $\det(\mathbf{J}_\Theta)$ as an empirical measure of the degree of local stretching and squashing. A transformation for which the Jacobian determinant is of unit magnitude everywhere would be subject to no stretching or squashing and the two manifolds would be topologically and topographically identical. One for which the Jacobian determinant is a constant smaller (greater) than unit magnitude would be subject to uniform compression (dilation) everywhere, and the reconstructed manifold would be identical to the true manifold under a uniform rescaling. One for which the Jacobian determinant varies over several orders of magnitude is subject to large variations in local compression and/or dilation, thereby inducing large non-uniform perturbations in free energy barrier heights and well depths making interpretation of the smFES over the reconstructed manifold $\Theta(M)$ hopeless without knowledge of the invertible transformation $\Theta : M \to \Theta(M)$. Conversely, one for which the Jacobian determinant is restricted to vary over a small range implies relatively mild local variation in compression/dilation, and one might hope to glean semi-quantitative understanding of the true smFES even in the absence of the mapping. As we shall see, we find that all proteins studied in the present work fall into the latter case.

We numerically evaluate the Jacobian using the mesh-free approach detailed in Ref.[48] and report its value as a function of simulation time in Figure 4d. As detailed in Section 2.2 we establish the bijection required in this calculation by matching the first element of each d-dimensional delay vector to the corresponding all-atom simulation snapshot. That the determinants remain *single signed* numerically validates the existence of the diffeomorphism, but that they are not *single valued* is indicative of non-uniform stretching and squashing of the manifold under the transformation. Nevertheless, the determinant spans no more than about an order of magnitude over the range 0.01-0.15, which implies that the degree of non-uniform perturbation to the manifold is relatively mild. This is consistent with the qualitative analysis described above wherein the true and reconstructed manifolds are visually similar and preserve a similar topography of metastable states. This is a tantalizing

result since it implies that the topography of the smFES recovered from knowledge of only experimentally-accessible observables may provide a good semi-quantitative approximation to the true landscape.

We quantify the degree of topographical perturbation in two ways. First, we compute the depth of the global free energy minimum containing the native fold $\delta\beta F$ relative to the highest finite free energy resolved by the smFES. The stability of the native fold in the reconstructed landscape is $\delta\beta F^* = 6.20$, constituting a $\sim$44% overestimate compared to that computed over the true landscape $\delta\beta F = 4.29$. Although there is a sizable discrepancy between these values, it is remarkable that the value reconstructed from knowledge of only $h2t$ should provide so good an estimate within only a couple of k_BT. Second, we compute the Pearson correlation coefficient in the free energy assigned to each point within the true and reconstructed landscapes to identify a linear correlation of $\rho(\beta F, \beta F^*) = 0.53$ ($p < 1\times10^{-10}$), indicating a relatively strong linear relation between the reconstructed and true free energy values and suggesting that relative free energy differences over the reconstructed landscapes may possess useful interpretablity. We defer a deeper engagement of these issues to Section 3.2.

3.1.2 Villin (PDB ID 2F4K)

All-atom trajectories. In an analogous manner to Trp-cage, we apply diffusion maps to the 120 μs Villin simulation trajectory at $T = 360$ K that we uniformly subsample from 600,000 snapshots at 0.2 ns resolution down to 60,000 frames. Pivot diffusion maps are applied to the C_α atom coordinates under spatial symmetrization of RMSD pairwise distances with a pivot cutoff radius of 0.80 nm to identify 448 pivot points. A diffusion map with bandwidth $\sqrt{\epsilon} = 1.0$ nm and density rescaling exponent $\alpha = 0.7$ produces an eigenvalue spectrum with a gap after λ_3 implying construction of a 2D intrinsic manifold M spanned by (ψ_2, ψ_3).

The smFES in conventional CVs (RMSD,R_g) illustrated in Figure 5a is in good agreement with that supported by the true intrinsic manifold M displayed in Figure 5b. Both landscapes

evince a well-defined folded state **A** connected to an unfolded ensemble containing a diversity of configurations with varying degrees of native structure and packing within the three α helices, of which **B** is one example. These results are consistent with the $T = 360$ K landscape for native Villin (PDB ID 1YRF) recovered by Lei *et al.* by replica exchange molecular dynamics simulations employing the Amber03 force field and implicit solvent that resolved an equilibrium between the native fold and denatured ensemble.[109]

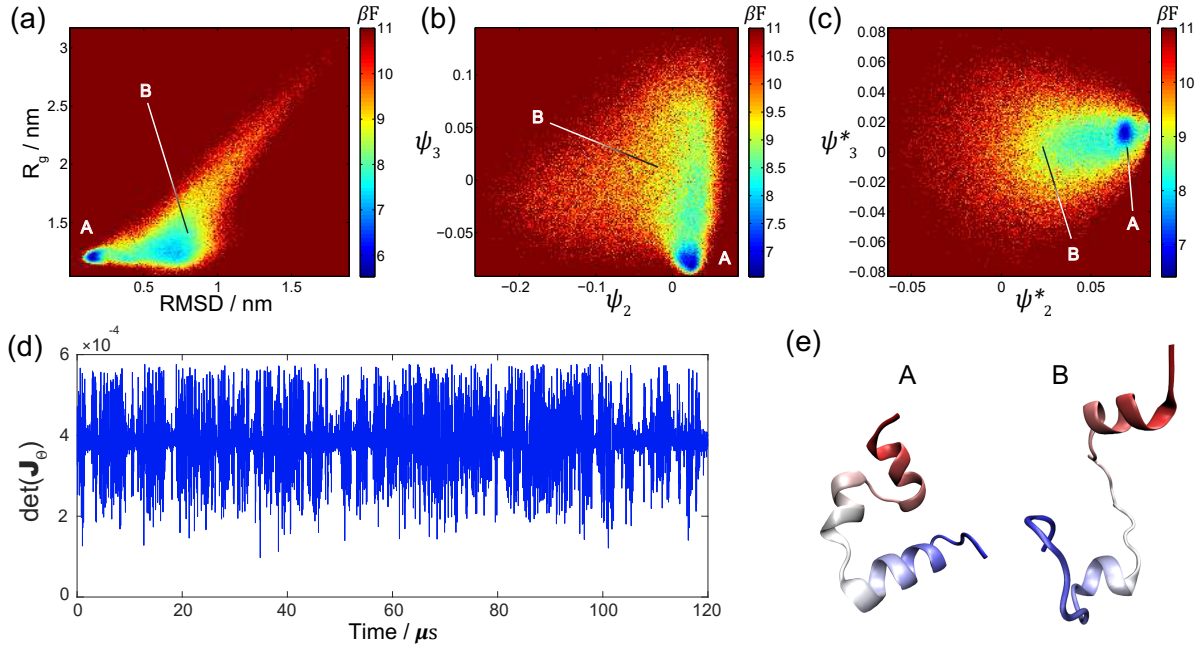


Figure 5: Single molecule free energy surfaces of Villin (PDB ID 2F4K). (a) smFES spanned by the conventional CVs (RMSD,R_g). (b) smFES supported by the true intrinsic manifold M spanned by the CVs (ψ_2, ψ_3) identified by applying diffusion maps to the all-atom simulation trajectory. (c) smFES supported by the reconstructed manifold $\Theta(M)$ spanned by the CVs (ψ_2^*, ψ_3^*) identified by applying diffusion maps to delay embeddings of the $h2t$ time series. We report the Helmholtz free energy F dedimensionalized by $\beta = 1/k_B T$ where $T = 360$ K is the simulation temperature. (d) Determinant of the Jacobian $\mathbf{J}_\Theta$ for the forward transformation $\Theta : M \to \Theta(M)$ numerically evaluated at each point in the embedding and displayed as a function of elapsed simulation time. (e) Representative molecular snapshots at the locations identified in panels a-c.

h2t time series. Takens' delay embeddings of the $h2t$ time series (Figure 2b) were constructed using a delay time of $\tau = 0.4$ ns and a delay embedding dimensionality of $d = 50$. Pivot diffusion maps were applied to the delay embedding vectors under temporal symmetrization of the Euclidean pairwise distances with a pivot cutoff of 8.0 nm to identify

1358 pivot points. A diffusion map with bandwidth $\sqrt{\epsilon} = 3.16$ nm, and density rescaling exponent $\alpha = 0.5$ produces an eigenvalue spectrum with a gap after λ_3^* implying that the reconstructed intrinsic manifold $\Theta(M)$ be constructed in (ψ_2^*, ψ_3^*).

Up to a trivial $\pi/2$ counter-clockwise rotation, the reconstructed smFES supported by $\Theta(M)$ in Figure 5c is visually similar to that supported by M presented in Figure 5b, with $\rho(\psi_2, \psi_3^*) = 0.33$ ($p < 1 \times 10^{-10}$) and $\rho(\psi_3, \psi_2^*) = -0.32$ ($p < 1 \times 10^{-10}$). Figure 5d illustrates that $\det(\mathbf{J}_\Theta)$ remains single signed – numerically validating the existence of the diffeomorphism – and spans a range of only about half an order of magnitude – indicative of a relatively uniform topographical perturbation over the manifold. This result is supported by good correspondence in the observed stability of the native state over the true and reconstructed manifolds of $\delta\beta F = 3.76$ and $\delta\beta F^* = 3.90$, with the reconstructed estimate in agreement with the true value within $\sim$4% error. Similarly, the the linear correlation between the free energies assigned to each point is also moderately strong at $\rho(\beta F, \beta F^*) = 0.46$ ($p < 1 \times 10^{-10}$).

3.1.3 BBA (PDB ID 1FME)

All-atom trajectories. The BBA simulation trajectory at $T = 325$ K comprising 1,000,000 snapshots at 0.2 ns resolution was uniformly subsampled to 100,000 frames. Pivot diffusion maps applied to the spatially symmetrized C_α pairwise RMSD distances with a pivot cutoff radius of 1.0 nm identified 596 pivots. A diffusion map with bandwidth $\sqrt{\epsilon} = 1.0$ nm and density rescaling exponent $\alpha = 0.15$ generated an eigenvalue spectrum with a gap after λ_3 implying construction of a 2D intrinsic manifold M spanned by (ψ_2, ψ_3).

The smFES in conventional CVs (RMSD,R_g) is presented in Figure 6a. In contrast to Trp-cage and Villin, the native fold **A** does not reside within a deep free energy well but is rather in competition with an equally stable partially folded ensemble **B** from which it is separated by a $\sim$1 k_BT free energy barrier. This is consistent with numerous prior studies that have reported the native fold of the BBA FSD-EY protein and related mutants to

be relatively unstable and in competition with partially folded intermediates.[60,110–113] The smFES over the true intrinsic manifold M in Figure 6b provides a superior embedding to that in the conventional CVs in that it better separates out the diversity of metastable states and reveals the native fold **A** to be just one of a number of nearly equi-stable structures including both partially folded **B** and fully unfolded **C**.

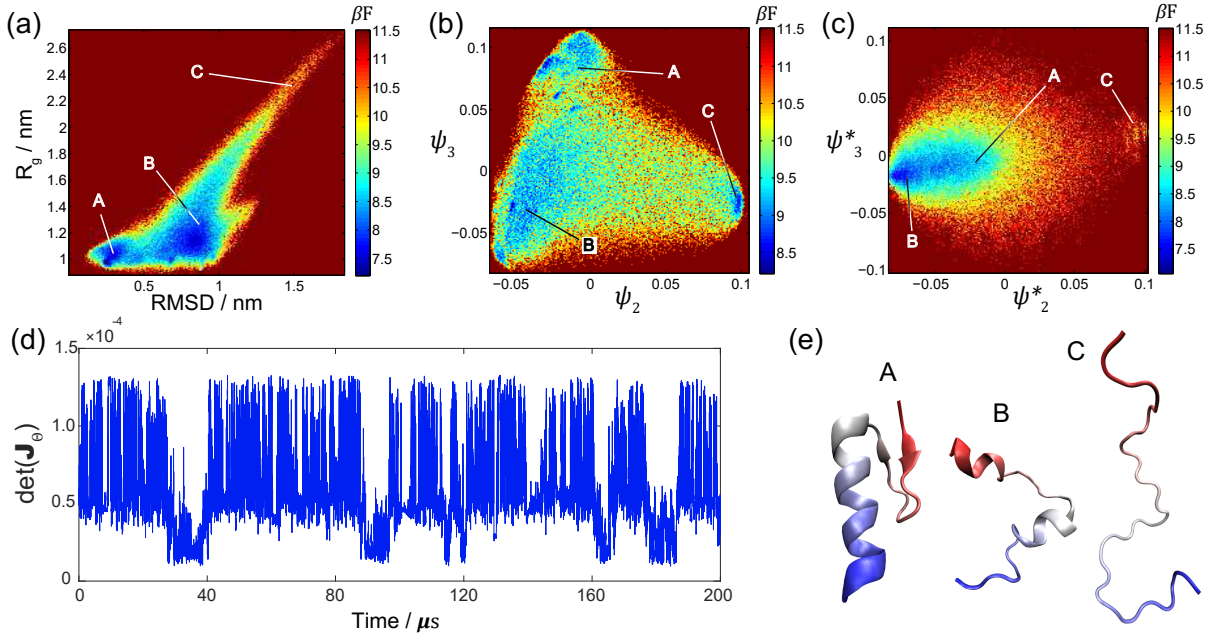


Figure 6: Single molecule free energy surfaces of BBA (PDB ID 1FME). (a) smFES spanned by the conventional CVs (RMSD,R_g). (b) smFES supported by the true intrinsic manifold M spanned by the CVs (ψ_2, ψ_3) identified by applying diffusion maps to the all-atom simulation trajectory. (c) smFES supported by the reconstructed manifold $\Theta(M)$ spanned by the CVs (ψ_2^*, ψ_3^*) identified by applying diffusion maps to delay embeddings of the $h2t$ time series. We report the Helmholtz free energy F dedimensionalized by $\beta = 1/k_B T$ where $T = 325$ K is the simulation temperature. (d) Determinant of the Jacobian $\mathbf{J}_\Theta$ for the forward transformation $\Theta : M \to \Theta(M)$ numerically evaluated at each point in the embedding and displayed as a function of elapsed simulation time. (e) Representative molecular snapshots at the locations identified in panels a-c.

h2t time series. Analysis of the $h2t$ time series (Figure 2c) identified $\tau = 0.8$ ns to be an appropriate delay time and $d = 50$ a suitable delay embedding. Pivot diffusion maps applied to the temporally symmetrized Euclidean pairwise distances between delay vectors with a pivot cutoff of 8.0 nm returned 840 pivot points. A diffusion map with bandwidth $\sqrt{\epsilon} = 3.16$ nm, and density rescaling exponent $\alpha = 0.5$ produced an eigenvalue spectrum

with a gap after λ_3^* indicating that the reconstructed intrinsic manifold $\Theta(M)$ should be parameterized by (ψ_2^*, ψ_3^*).

In contrast to Trp-cage and Villin – perhaps due to the presence of a diversity of nearly equally stable configurational states – the smFES over $\Theta(M)$ in Figure 6c is less visually similar to that over M in Figure 6b. Nevertheless, the Jacobian determinant $\det(\mathbf{J}_\Theta)$ is confirmed to be single signed, numerically verifying that M and $\Theta(M)$ are related by a diffeomorphism and that the smFES that they support are topologically identical. A linear correlation analysis reveals that $\rho(\psi_2, \psi_2^*) = 0.63$ ($p < 1 \times 10^{-10}$) and $\rho(\psi_3, \psi_3^*) = 0.04$ ($p < 1 \times 10^{-10}$), indicating that ψ_2 and ψ_2^* contain similar information content but that ψ_3 and ψ_3^* are linearly uncorrelated. We observe once again that $\det(\mathbf{J}_\Theta)$ spans only around an order of magnitude, limiting the non-uniformity in the topological perturbation exerted by the diffeomorphism over the manifold. As a result, we again find relatively good agreement between $\delta\beta F = 2.60$ and $\delta\beta F^* = 3.77$, and compute $\rho(\beta F, \beta F^*) = 0.56$ ($p < 1 \times 10^{-10}$).

3.2 Trp-cage smFES reconstruction, interpretation, and engineering

Analysis of the three proteins above demonstrated the capacity of our approach to reconstruct free energy landscapes from univariate time series in experimentally-measurable observables. The sign of the Jacobian determinant of the transformation between the true and reconstructed landscapes numerically validated the existence of the diffeomorphism that guarantees topological equivalence. Furthermore, the magnitude of the determinant was observed to span no more than about one order of magnitude, effectively bounding the degree of local deformation induced by the diffeomorphism. Accordingly, the barrier heights and well depths of the smFES over the reconstructed manifold $\Theta(M)$ may be expected to possess semi-quantitative interpretability. We now proceed to study the degree to which this is true by performing a comparative analysis of the changes in the true and reconstructed smFES for a Trp-cage variant subject to different temperatures and patterns of mutations. With a

view to ultimately applying our approach to experimental data for which the true smFES over M is unavailable, we seek to understand to what degree the reconstructed smFES over $\Theta(M)$ may be used to understand and engineer protein stability, folding, and ultimately function.

We study Trp-cage (PDB ID 1L2Y) in implicit solvent and conduct a total of 37 independent 1 μs simulations at a number of temperatures and under a number of engineered mutations (Section 2.1). A full accounting of the various simulations are provided in Table 1. Empirically, we observe numerous folding transitions at around $T = 380$ K and so choose to focus our investigations around this temperature. We note that the native Trp-cage studied in this portion of the work (PDB ID 1L2Y) differs from that studied in Section 3.1 (PDB ID 2JOF) by four point mutations in the N-terminal region. We first describe our recovery of the true and reconstructed smFES for the 37 simulations, and then assess to what extent folding may be interpreted and engineered through temperature and mutation from a knowledge of only the reconstructed landscapes.

3.2.1 Determination of composite true and reconstructed smFES

All-atom trajectories. We harvest 10,000 uniformly spaced frames from each of the 37 simulations to assemble a 370,000 frame composite ensemble to be analyzed by diffusion maps. By analyzing the union of molecular snapshots taken under all conditions studied, we ensure that the CVs recovered by diffusion maps span all regions of configurational space explored by the various simulations, and furnish a unified basis set with which to parameterize the underlying intrinsic manifold M.[114] Different systems are expected to populate the manifold differently and yield different free energy surfaces, but the use of a unified basis set in construction of M is crucial in drawing quantitative comparisons between the different systems. We represent each configuration as a vector of the 20 C_α atomic coordinates, enabling straightforward comparisons between snapshots drawn from different simulations. The only system requiring special treatment is simulation #11 comprising an engineered

Table 1: List of Trp-cage simulations. Four simulation classes were considered: (I) wild type at various temperatures, (II) engineered mutants,[69] (III) alanine scan, and (IV) alanine tetrads. Point mutations from the wild type TC5b Trp-cage (PDB ID 1L2Y)[69] are indicated in bold.

Index	Type	Name	T (K)	Sequence	$\delta\beta F$	$\delta\beta F^*$	$\frac{(\delta\beta F^* - \delta\beta F)}{\delta\beta F}/\%$	$\rho(\beta F, \beta F^*)$
1		300K	300	NLYIQ WLKDG GPSSG RPPPS	4.02	4.14	2.97	0.42
2		320K	320	NLYIQ WLKDG GPSSG RPPPS	3.92	4.20	7.12	0.44
3		340K	340	NLYIQ WLKDG GPSSG RPPPS	4.32	4.51	4.41	0.39
4	I	360K	360	NLYIQ WLKDG GPSSG RPPPS	4.24	4.44	4.78	0.47
5		380K	380	NLYIQ WLKDG GPSSG RPPPS	4.07	4.18	2.78	0.61
6		400K	400	NLYIQ WLKDG GPSSG RPPPS	3.54	3.56	0.51	0.52
7		420K	420	NLYIQ WLKDG GPSSG RPPPS	3.22	2.65	-17.82	0.31
8		440K	440	NLYIQ WLKDG GPSSG RPPPS	2.97	2.07	-30.14	0.04
9		TC3b	380	NL**FIE** WLK**N**G GPSSG **A**PPPS	3.60	3.84	6.73	0.51
10	II	TC4a	380	**DLFIE** WLK**N**G GPSSG RPPPS	3.82	4.09	7.25	0.58
11		TC4c	380	**KGLFIE** WLK**N**G GPSSG RPPPS	3.84	3.30	-14.16	0.52
12		TC5a	380	NL**F**IQ WLKDG GPSSG RPPPS	4.06	4.31	6.17	0.58
13		N1A	380	**A**LYIQ WLKDG GPSSG RPPPS	3.84	3.95	2.91	0.54
14		L2A	380	N**A**YIQ WLKDG GPSSG RPPPS	3.53	3.58	1.62	0.55
15		Y3A	380	NL**A**IQ WLKDG GPSSG RPPPS	3.81	4.16	9.25	0.56
16		I4A	380	NLY**A**Q WLKDG GPSSG RPPPS	4.12	4.41	7.13	0.59
17		Q5A	380	NLYI**A** WLKDG GPSSG RPPPS	3.98	4.41	10.88	0.56
18		W6A	380	NLYIQ **A**LKDG GPSSG RPPPS	2.53	1.95	-22.96	0.15
19		L7A	380	NLYIQ W**A**KDG GPSSG RPPPS	3.28	3.38	3.27	0.45
20		K8A	380	NLYIQ WL**A**DG GPSSG RPPPS	3.93	4.05	3.05	0.58
21		D9A	380	NLYIQ WLK**A**G GPSSG RPPPS	3.99	4.14	3.86	0.58
22		G10A	380	NLYIQ WLKD**A** GPSSG RPPPS	3.54	3.75	5.89	0.53
23	III	G11A	380	NLYIQ WLKDG **A**PSSG RPPPS	3.11	3.58	15.10	0.42
24		P12A	380	NLYIQ WLKDG G**A**SSG RPPPS	3.11	2.64	-15.24	0.31
25		S13A	380	NLYIQ WLKDG GP**A**SG RPPPS	4.14	4.40	6.21	0.60
26		S14A	380	NLYIQ WLKDG GPS**A**G RPPPS	3.66	3.84	4.80	0.56
27		G15A	380	NLYIQ WLKDG GPSS**A** RPPPS	3.16	3.22	1.96	0.44
28		R16A	380	NLYIQ WLKDG GPSSG **A**PPPS	3.62	3.76	3.78	0.49
29		P17A	380	NLYIQ WLKDG GPSSG R**A**PPS	3.64	3.74	2.75	0.56
30		P18A	380	NLYIQ WLKDG GPSSG RP**A**PS	3.33	3.33	0.00	0.52
31		P19A	380	NLYIQ WLKDG GPSSG RPP**A**S	3.30	3.02	-8.36	0.46
32		S20A	380	NLYIQ WLKDG GPSSG RPPP**A**	4.16	4.41	5.96	0.57
33		tetrad1	380	**AAAA**Q WLKDG GPSSG RPPPS	3.82	3.96	3.75	0.53
34		tetrad2	380	NLYI**A AAA**DG GPSSG RPPPS	2.48	2.01	-18.91	0.09
35	IV	tetrad3	380	NLYIQ WLK**AA AA**SSG RPPPS	3.16	2.56	-18.75	0.34
36		tetrad4	380	NLYIQ WLKDG GP**AAA A**PPS	2.94	2.56	-12.89	0.33
37		tetrad5	380	NLYIQ WLKDG GPSSG R**AAAA**	2.71	2.44	-9.81	0.08

mutant containing an addition N-terminal lysine that we simply neglect in our coordinate representation.

Pivot diffusion maps employing spatial symmetrization with a pivot cutoff of 0.38 nm in RMSD yielded 478 pivot points. A diffusion map with bandwidth $\sqrt{\epsilon} = 0.89$ nm and density rescaling exponent $\alpha = 1.0$ generated an eigenvalue spectrum with a gap after λ_3 implying that the true intrinsic manifold M is parameterized by the data-driven CVs (ψ_2, ψ_3).

We present in Figure 7 the embedding of the composite ensemble of all-atom configurations into the 2D intrinsic manifold M and colored by R_g (Figure 7a), RMSD (Figure 7b), and RMSD-helix (Figure 7c). The shape and gradations of the physical measures over the manifold are very similar to those for the manifold recovered above for the closely related Trp-cage mutant presented in Figure 3, although not identical due to differences in temperature, sequence, solvent, and force field (see Section 2.1).

h2t time series. We constructed delay embeddings of each of the 37 $h2t$ time series corresponding to each simulation, employing a delay time of $\tau = 0.02$ ns and a delay embedding dimensionality of $d = 20$. As for the all-atom analysis, we applied pivot diffusion maps to the composite ensemble of delay vectors over all 37 simulations in order to recover a unified basis set with which to parameterize the reconstructed intrinsic manifold $\Theta(M)$. This assures that there is a single well-defined transformation between each true and reconstructed smFES $\Theta : M \to \Theta(M)$ and all reconstructed smFES are supported by the same manifold. Importantly, this allows us to draw quantitative comparisons between the topography of the reconstructed smFES under different simulation conditions. Accordingly, even if the true smFES is unavailable and the transformation $\Theta : M \to \Theta(M)$ is unknown, we can still draw inferences about the effect of changing conditions by comparing the reconstructed landscapes within a common basis.

A pivot cutoff radius of 2.5 nm in the temporally symmetrized pairwise Euclidean distances identifies 970 pivot points, and a subsequent diffusion map with bandwidth $\sqrt{\epsilon} = 2.0$ nm and density rescaling exponent $\alpha = 0.25$ produces an eigenvalue spectrum with a gap

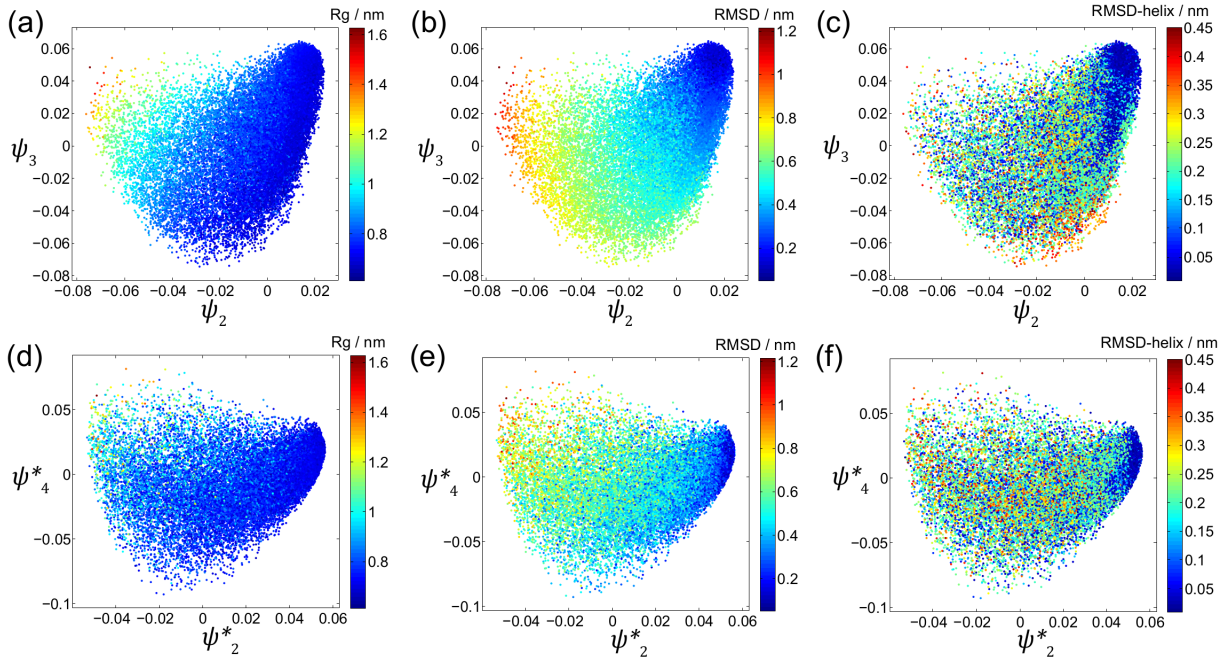


Figure 7: Representations of the 2D intrinsic manifold of Trp-cage (PDB ID 1L2Y). (a-c) The true manifold M spanned by the CVs (ψ_2, ψ_3) discovered by application of diffusion maps to the composite of 37 all-atom simulation trajectories and colored by R_g, RMSD, and RMSD-helix. Each point represents a projection of an all-atom configuration observed in the molecular simulation trajectory. (d-e) The reconstructed manifold $\Theta(M)$ spanned by the CVs (ψ_2^*, ψ_4^*) discovered by construction of Takens' delay embeddings of the $h2t$ time series extracted from each of the 37 simulation trajectories and subsequent application of diffusion maps. Each point represents a projection of a (d=50)-dimensional delay vector of $h2t$ values and is colored by the R_g, RMSD, and RMSD-helix corresponding to the first configuration in the delay vector.

after λ_4^*. As before, we found ψ_3^* to be effectively slaved to ψ_2^* mapping out a 1D manifold in the (ψ_2^*, ψ_3^*) plane and allowing us to parameterize the reconstructed intrinsic manifold $\Theta(M)$ in (ψ_2^*, ψ_4^*) without loss of information.[7,94]

Embeddings of the composite ensemble of delay vectors into the 2D reconstructed manifold $\Theta(M)$ colored by R_g, RMSD, and RMSD-helix (Figure 7d-f) show clear visual similarity to the all-atom embeddings into the true manifold M (Figure 7a-c). A Pearson correlation analysis confirms the existence of strong $\rho(\psi_2, \psi_2^*) = 0.81$ $(p < 1 \times 10^{-10})$ and moderate $\rho(\psi_3, \psi_4^*) = 0.29$ $(p < 1 \times 10^{-10})$ linear correlation.

3.2.2 The reconstructed smFES approximate the true smFES

For each of the 37 simulations we estimate the smFES supported by the true manifold M from histograms in the projected distribution of all-atom simulation snapshots into (ψ_2, ψ_3) via the relation $\beta F(\psi_2, \psi_3) = -\ln \hat{P}(\psi_2, \psi_3) + C$, where $\beta = 1/k_B T$ is evaluated at $T = 380$ K, and C is an arbitrary additive constant. Each independent simulation produces an independent smFES, but importantly these are all constructed over a common parameterization of M. We specify the value of C in each of the 37 different landscapes by asserting equality of the highest finite free energy observed. By an analogous procedure, we estimate the smFES supported by the reconstructed manifold $\Theta(M)$ from histograms in the projected distribution of delay vectors into (ψ_2^*, ψ_4^*) and using the relation $\beta F(\psi_2^*, \psi_4^*) = -\ln \hat{P}(\psi_2^*, \psi_4^*) + C^*$, where $\beta = 1/k_B T$ is evaluated at $T = 380$ K.

In subsequent sections we investigate the degree to which changes in the topography of the reconstructed smFES over $\Theta(M)$ as a function of temperature or mutations reflect changes to the true smFES over M. If we find good correspondence between these changes, then this provides support for our conjecture that protein folding landscapes recovered from experimental time series might be used to understand and engineer protein structure, stability, and folding. Before presenting the free energy surfaces themselves, we first assess the degree to which we might expect this correspondence to hold.

First, we numerically evaluate $\det(\mathbf{J}_\Theta)$ to confirm that it remains single signed and validate the existence of the diffeomorphism (Figure 8a). We observe that its magnitude varies only over approximately an order of magnitude, lending empirical support that the variation in the degree of local perturbation induced in the reconstructed manifold is relatively tightly bounded. We therefore expect that the topography of the true smFES may be approximately preserved within the reconstructed smFES and is therefore interpretable. Second, we report in Table 1 the measured stability of the native fold in the true $\delta\beta F$ and reconstructed $\delta\beta F^*$ landscapes, which shows the latter to approximate the former within a maximum error of $\sim$30%. These quantities exhibit a strong linear correlation of $\rho(\delta\beta F, \delta\beta F^*)$

$= 0.94$ ($p < 1 \times 10^{-10}$) and their relationship is well fit by a linear least squares fit of $(\delta\beta F^*) = 1.47(\delta\beta F) - 1.70$ ($R^2 = 0.89$) (Figure 8b). The slope close to unity and intercept close to zero indicates that native fold stabilities estimated from the reconstructed folding funnel are in semi-quantitative agreement with those measured over the true landscape. Third, the correlation in free energy assigned to each point within the pairs of true and reconstructed landscapes tends to exhibit moderately strong linear correlation, with a mean value of $\overline{\rho(\beta F, \beta F^*)} = 0.45$ over the 37 simulations. The particular values of $\rho(\beta F, \beta F^*)$ are reported in Table 1. We find that higher native state stabilities $\delta\beta F \gtrsim 3.0$ (Region B in Figure 8c) tend to possess stronger correlation coefficients, whereas shallower free energy minima have weaker correlation (Region A). Region A comprises Simulations #8, #18, #34, and #37 that correspond to high temperatures or mutations that strongly destabilize the native state and cause the protein to delocalize over the folding landscape. These observations suggest that the reconstructed smFES is a better approximation for the true smFES under conditions where the protein possesses a moderately stable native fold than when it is delocalized over a rather flat free energy landscape, and we should be wary of over-interpreting the reconstructed landscape in the latter case.

3.2.3 The reconstructed smFES accurately tracks temperature denaturation

The smFES for wild type Trp cage at $T = 320$ K, 380 K, and 420 K are illustrated in Figure 9. The landscapes for all eight temperatures considered (Simulations #1-#8 in Table 1) are presented in Figure S1 in the Supporting Information. For each temperature we project the smFES into the conventional variables (RMSD,RMSD-helix), the true manifold M spanned by (ψ_2,ψ_3), and the reconstructed manifold $\Theta(M)$ spanned by (ψ_2^*,ψ_4^*). The projection into conventional CVs is provided to facilitate comparison with prior work.

Under the force field and implicit solvent model employed in this work we observed multiple folding and unfolding events over the 1 μs trajectory at $T = 380$ K wherein Trp-cage rapidly switches between the two states **A** and **C**, where **A** is the native state (Figure 9j) and

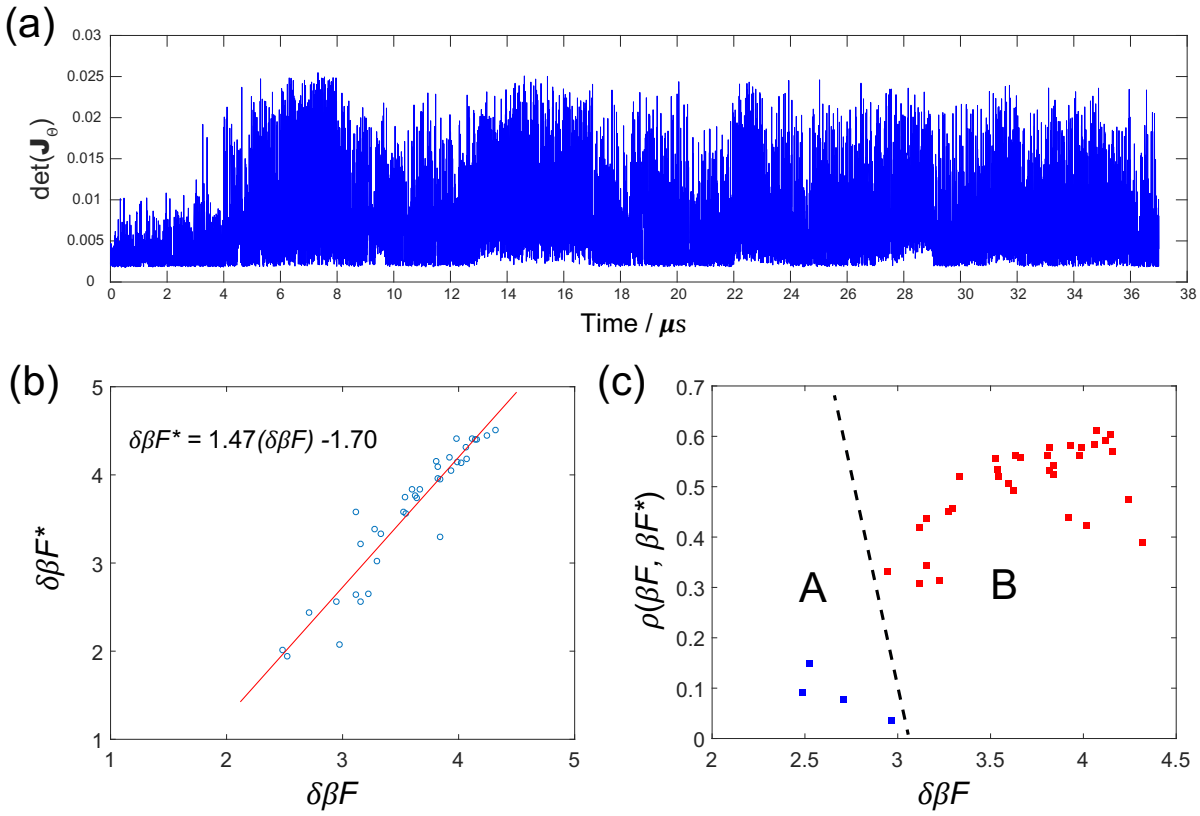


Figure 8: Numerical analysis of diffeomorphism between the Trp-cage (PDB ID 1L2Y) true M and reconstructed $\Theta(M)$ manifolds and the degree of induced topographical perturbation. (a) Determinant of the Jacobian $\mathbf{J}_\Theta$ for the forward transformation $\Theta : M \to \Theta(M)$. For representational clarity we concatenate the 37 independent 1 μs simulations (Table 1) into a single trajectory and represent $\det(\mathbf{J}_\Theta)$ at each point as a function of time. That $\det(\mathbf{J}_\Theta)$ is single signed verifies the existence of the diffeomorphism, and that it varies in magnitude over approximately only one order of magnitude bounds the range of local compression or dilation induced by the transformation. (b) Parity plot of the stability of the native fold in the smFES supported by the true $\delta\beta F$ and reconstructed $\delta\beta F^*$ manifolds. The data exhibit a Pearson correlation coefficient of $\rho(\delta\beta F, \delta\beta F^*) = 0.94$ and the red line indicates the least squares best fit line $(\delta\beta F^*) = 1.47(\delta\beta F) - 1.70$. (c) Scatter plot of native state stability $\delta\beta F$ against linear correlation coefficient between the free energy assigned to each configuration over the smFES over the true and reconstructed manifolds $\rho(\beta F, \beta F^*)$. The correlation coefficient $\rho(\beta F, \beta F^*)$ tends to be stronger, and therefore maximizes the interpretability of the reconstructed smFES, for higher native state stabilities (Region B, $\delta\beta F \gtrsim 3.0$) and diminishes for shallower native free energy wells (Region A, $\delta\beta F \lesssim 3.0$).

C possesses a partially unfolded N-terminal α-helix (Figure 9l). This dynamic equilibrium produces a folding funnel centered on the native state but with substantial exploration of nearby non-native configurations (Figure 9k,m,n): State **B** contains a partially unfolded 3_{10}

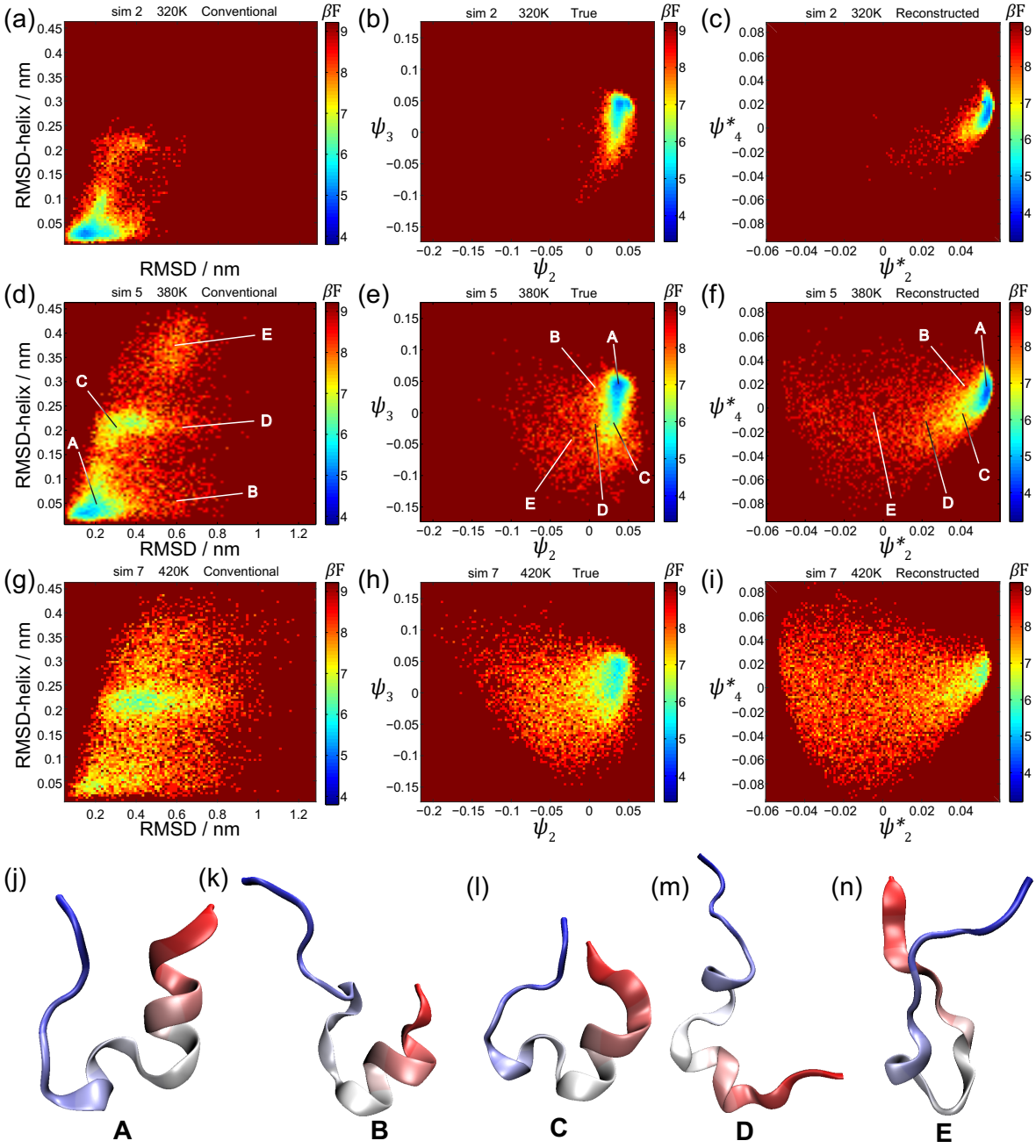


Figure 9: Single molecule free energy surfaces of wild type Trp-cage (PDB ID 1L2Y) as a function of temperature (Simulations #2, #5, and #7 in Table 1). smFES supported by conventional CVs (left), the true intrinsic manifold M (center) and reconstructed manifold $\Theta(M)$ (right) at (a-c) $T = 320$ K, (d-f) $T = 380$ K, and (g-i) $T = 420$ K. We report the Helmholtz free energy F dedimensionalized by $\beta = 1/k_B T$ at $T = 380$ K. (j-n) Representative molecular snapshots corresponding to the locations identified in panels (d-f).

helix, in **D** both the N-terminal α-helix and 3_{10} helix are partially unfolded, and in **E** the N-terminal α-helix and 3_{10} helix are essentially completely unfolded. The depth of the native well measured on the reconstructed manifold is $\delta\beta F^* = 4.18$ in very good agreement with that measured on the true landscape of $\delta\beta F = 4.07$, where we report free energy in units of $\beta = 1/k_B T$ at $T = 380$ K. The reconstructed smFES show that lowering the temperature to $T = 320$ K causes the protein to retreat into the native free energy well but negligible change in the depth of the native well to $\delta\beta F^* = 4.20$ (Figure 9c), whereas increasing the temperature to $T = 420$ K greatly destabilizes the native fold to $\delta\beta F^* = 2.65$ and induces the protein to delocalize over the folding funnel (Figure 9i). Of course, these trends are not unexpected and are generic features of natural proteins with a well-defined native fold. What is remarkable is that the observed trends in the shape and topography of a folding funnel reconstructed from the knowledge of the time evolution of a single coarse-grained observable should map so well to those over the true landscape that requires knowledge of all atomic coordinates. The temperature-dependent shape and size of the the true smFES (Figure 9b,e,h) are in good agreement with its reconstructed image, and the reconstructed smFES predictions for the low and high temperature native fold stabilities are in reasonably good agreement with the true values – $\delta\beta F = 3.92$ and $\delta\beta F^* = 4.20$ at $T = 320$ K and $\delta\beta F = 3.22$ and $\delta\beta F^* = 2.65$ at $T = 420$ K – within only 7.1% and 18% error.

3.2.4 The reconstructed smFES predicts the stability of engineered mutants

The TC5b Trp-cage variant that we adopt as our wild type sequence was experimentally engineered by Andersen and co-workers as a truncated mutant of a poorly folded 39-residue peptide exendin-4.[69] We now consider four more of their engineered mutants – TC3b, TC4a, TC4c, and TC5a (Simulations #9-#12 in Table 1) – to determine whether the reconstructed smFES can correctly identify their relative stabilities. Experimentally, TC5b (our wild type) and TC5a are found to be of comparable stability, and TC3b and TC4c to be of comparable and lower stability.[69] Stability measures for TC4a in water are not reported.[69]

Despite our use of an approximate implicit solvent model, the smFES over the true manifold M recapitulate these experimental trends (Figure 10b,h,k, Figure 9e), revealing native state stabilities of $\delta\beta F = 4.07$ and 4.06 for TC5b and TC5a, and $\delta\beta F = 3.84$ and 3.60 for TC4c and TC3b (Table 1). Remarkably, the reconstructed smFES illuminate very similar trends, showing the TC4c and TC3b funnels to be wider and shallower than those of TC5b and TC5a, and reporting native state stabilities of $\delta\beta F^* = 4.18$ and 4.31 for TC5b and TC5a, and $\delta\beta F^* = 3.30$ and 3.84 for TC4c and TC3b. Although the reconstructed smFES inverts the order of stability within each pair, the errors in the stability predictions are better than 14% and correspond to just a fraction of a k_BT. Finally, the reconstructed and true native stabilities of TC4a are $\delta\beta F = 3.82$ and $\delta\beta F^* = 4.09$, which agree to within 7% error. This analysis demonstrates the capacity of the reconstructed landscapes to provide semi-quantitative predictions of the stability and topographical character of the smFES for engineered protein mutants in agreement with the true smFES and experimental data.

3.2.5 The reconstructed smFES identifies significant and insignificant Ala point mutations

Alanine scanning is a technique that probes wild type residue contributions to protein structure or function by making all alanine point mutations.[115] We perform a computational version of this technique to determine whether the reconstructed smFES can reliably identify those point mutants that substantially perturb the free energy surface from those that are relatively silent and leave it largely unaffected. We illustrate in Figure 11 the smFES for four selected alanine point mutations Q5A, W6A, P12A, and S13A. The complete set of landscapes for all 20 point mutants is presented in Figure S2 in the Supporting Information.

The reconstructed smFES under the Q5A and S13A mutations (Figure 11c,l) are very similar to that for the wild type (Figure 9f), suggesting that alanine mutations at these positions do not strongly affect the folding landscape. This prediction is confirmed by the smFES over the true manifold that shows very little perturbation under the two point mutations

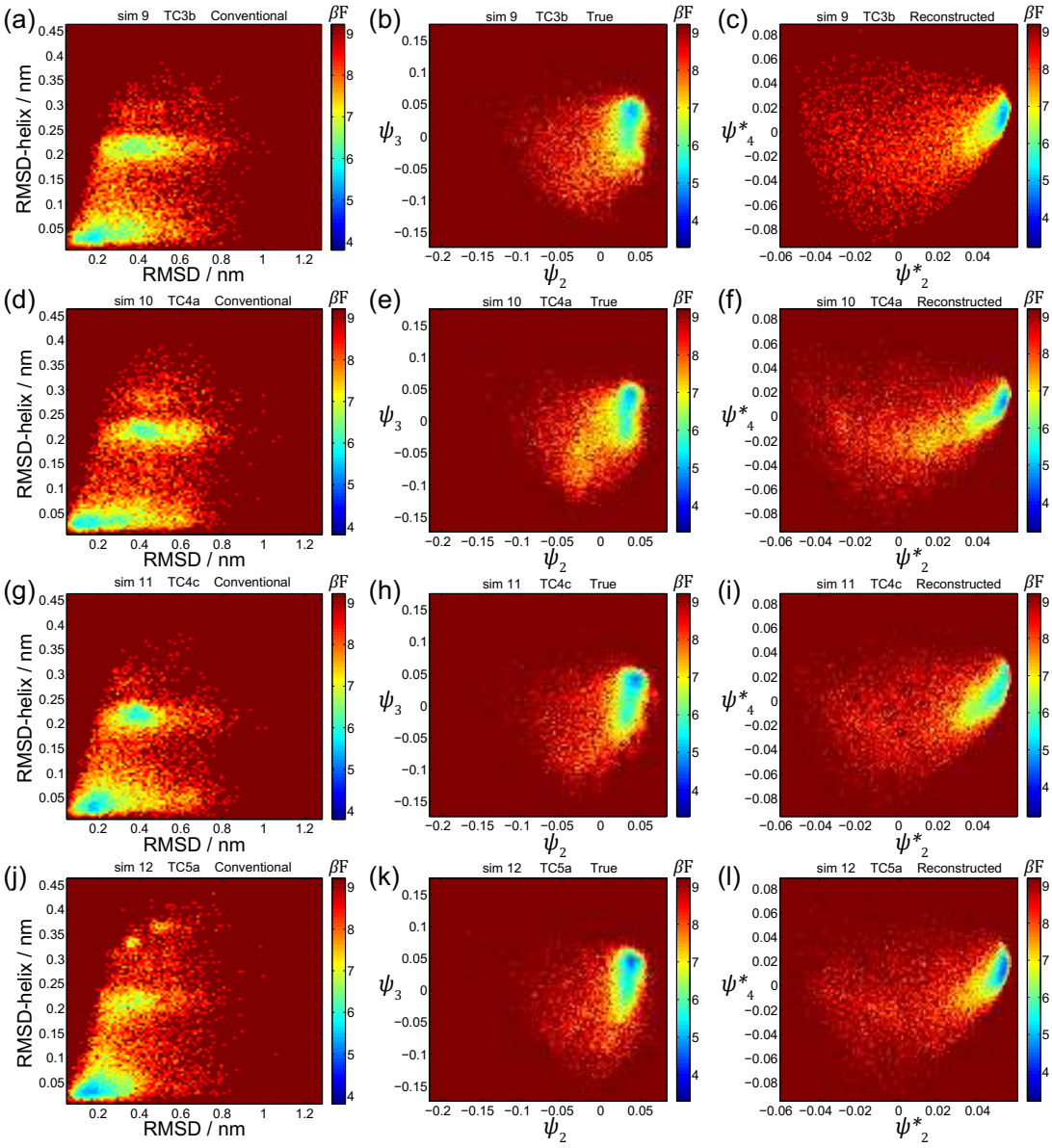


Figure 10: Single molecule free energy surfaces of experimentally engineered mutants of Trp-cage (PDB ID 1L2Y) (Simulations #9-#12 in Table 1). smFES supported by conventional CVs (left), the true intrinsic manifold M (center) and reconstructed manifold $\Theta(M)$ (right) for mutants (a-c) TC3b, (d-f) TC4a, (g-i) TC4c, and (j-l) TC5a. We report the Helmholtz free energy F dedimensionalized by $\beta = 1/k_B T$ at $T = 380$ K.

(Figure 11b,k, Figure 9e). There is also good quantitative agreement between the predicted small changes in native fold stabilities according to the reconstructed smFES $\delta\beta F^* = 4.18$ (wt), 4.41 (Q5A), and 4.40 (S13A) relative to those from the true smFES $\delta\beta F = 4.07$ (wt), 3.98 (Q5A), and 4.14 (S13A). Conversely, the reconstructed smFES exhibits large changes

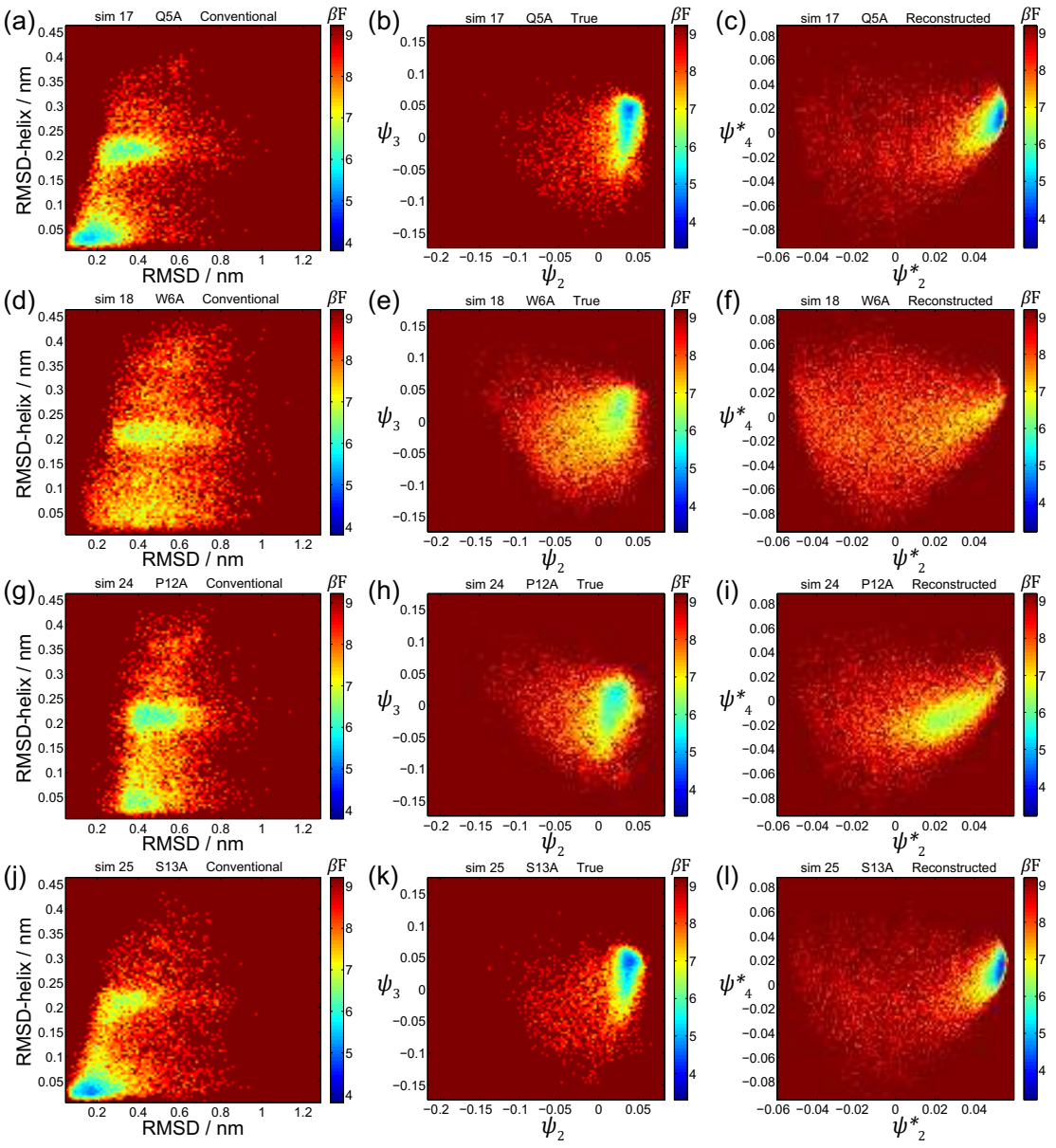


Figure 11: Single molecule free energy surfaces of selected alanine point mutants of Trp-cage (PDB ID 1L2Y) (Simulations #17, #18, #24, and #25 in Table 1). smFES supported by conventional CVs (left), the true intrinsic manifold M (center) and reconstructed manifold $\Theta(M)$ (right) for point mutants (a-c) Q5A, (d-f) W6A, (g-i) P12A, and (j-l) S13A. We report the Helmholtz free energy F dedimensionalized by $\beta = 1/k_B T$ at $T = 380$ K.

in the smFES under the W6A and P12A mutations (Figure 11f,i) wherein the native free energy well is destroyed, the weak residual minimum shifted away from the location of the wild type native fold, and the configurational ensemble delocalized over the folding funnel. These trends are in excellent accord with the observed changes over the true smFES (Fig-

ure 11e,h). The destabilization of the native state measured over the reconstructed smFES $\delta\beta F^* = 1.95$ (W6A) and 2.64 (P12A) again track those over the true smFES $\delta\beta F = 2.53$ (W6A), and 3.11 (P12A). In addition to the W6A and P12A mutations, the reconstructed smFES also predicts the L2A, L7A, G11A, G15A, P18A, and P19A to destabilize the native state by more than 0.5 k_BT, failing to identify only the G10A mutation identified by applying the same criterion to the true smFES. Accordingly, the reconstructed landscape has identified with high accuracy those Ala point mutations that do and do not substantially impact native state stability. The Trp-cage hydrophobic core comprises the Trp-6 side chain "caged" by encircling Tyr-3, Leu-7, Gly-11, Pro-12, Pro-18, and Pro-19 side chains, and many of the residues identified as significant by the alanine scan can be understood to disrupt formation of the hydrophobic core.[67,69,116] Finally, we note that the linear correlation between the true and reconstructed free energy values is moderately strong for all alanine point mutations $\rho(\beta F, \beta F^*) > 0.4$ with the exception of the two most destabilizing mutants W6A ($\rho(\beta F, \beta F^*) = 0.15$) and P12A ($\rho(\beta F, \beta F^*) = 0.31$) where abrogation of the native well and flattening of the landscape is observed to diminish the topographic fidelity of the reconstructed smFES.

3.2.6 The reconstructed smFES identifies significant and insignificant Ala tetrad mutations

As a more extreme example of alanine scanning we also perform alanine tetrad scans in which we mutate the five tetrads of contiguous residues 1-4, 5-8, 9-12, 13-16, 17-20 to alanine. The reconstructed and true smFES are again in very good agreement (Figure 12). The reconstructed smFES correctly predicts that all of the tetrad mutations substantially destabilize the native fold with the exception of Tetrad 1 in positions 1-4. This can be understood since these mutations are restricted to the N-terminal helix that can still form under alanine substitutions and which does not strongly participate in forming the hydrophobic core of the native fold. Three of the other tetrads mutate away the Trp-6 (Tetrad 2), Gly-11 and

Pro-12 (Tetrad 3), and Pro-18 and Pro-19 (Tetrad 5) residues that are critical for formation of the hydrophobic core.[67] The remaining Tetrad 4 replaces Ser-13, Ser-14, Gly-15, and Arg-16 with four contiguous Ala residues, which form a short α-helix that disrupts the native tertiary structure. The native state stabilities predicted over the reconstructed landscape $\delta\beta F^* = 3.96$ (Tetrad 1), 2.01 (Tetrad 2), 2.56 (Tetrad 3), 2.56 (Tetrad 4), and 2.44 (Tetrad 5) are again in quite good agreement with those over the true smFES $\delta\beta F = 3.82$ (Tetrad 1), 2.48 (Tetrad 2), 3.16 (Tetrad 3), 2.94 (Tetrad 4), and 2.71 (Tetrad 5). Tetrad 1 preserves a stable native fold and maintains reasonably strong linear correlation in the free energy $\rho(\beta F, \beta F^*) = 0.53$, but flattening of the landscape drives the linear correlation below 0.35 for the remaining four tetrads. These results show that knowledge of only the $h2t$ time series is sufficient to reconstruct a smFES that can accurately predict the effect of multiple mutations upon the stability and topography of the folding funnel.

4 Conclusions

We have demonstrated an integration of nonlinear manifold learning techniques with Takens' Delay Embedding Theorem to reconstruct single molecule free energy surfaces (smFES) for protein folding from univariate time series recording the dynamical evolution of the head-to-tail distance of the molecule. We have previously validated the approach in molecular dynamics simulations of a hydrocarbon chain,[48] and the present work represents its first application to realistic molecular simulations of protein folding. Intramolecular distances are experimentally accessible observables that can be followed by experimental techniques such as single molecule Förster resonance energy transfer (smFRET). Subject to the elimination of spatial symmetries in the observation variable and remediation of spurious temporal symmetry breaking in the delay embeddings, Takens' Theorem guarantees that the reconstructed smFES is diffeomorphic – related by a smooth and invertible transformation – to the true smFES determined from complete knowledge of the dynamical evolution of all molecular

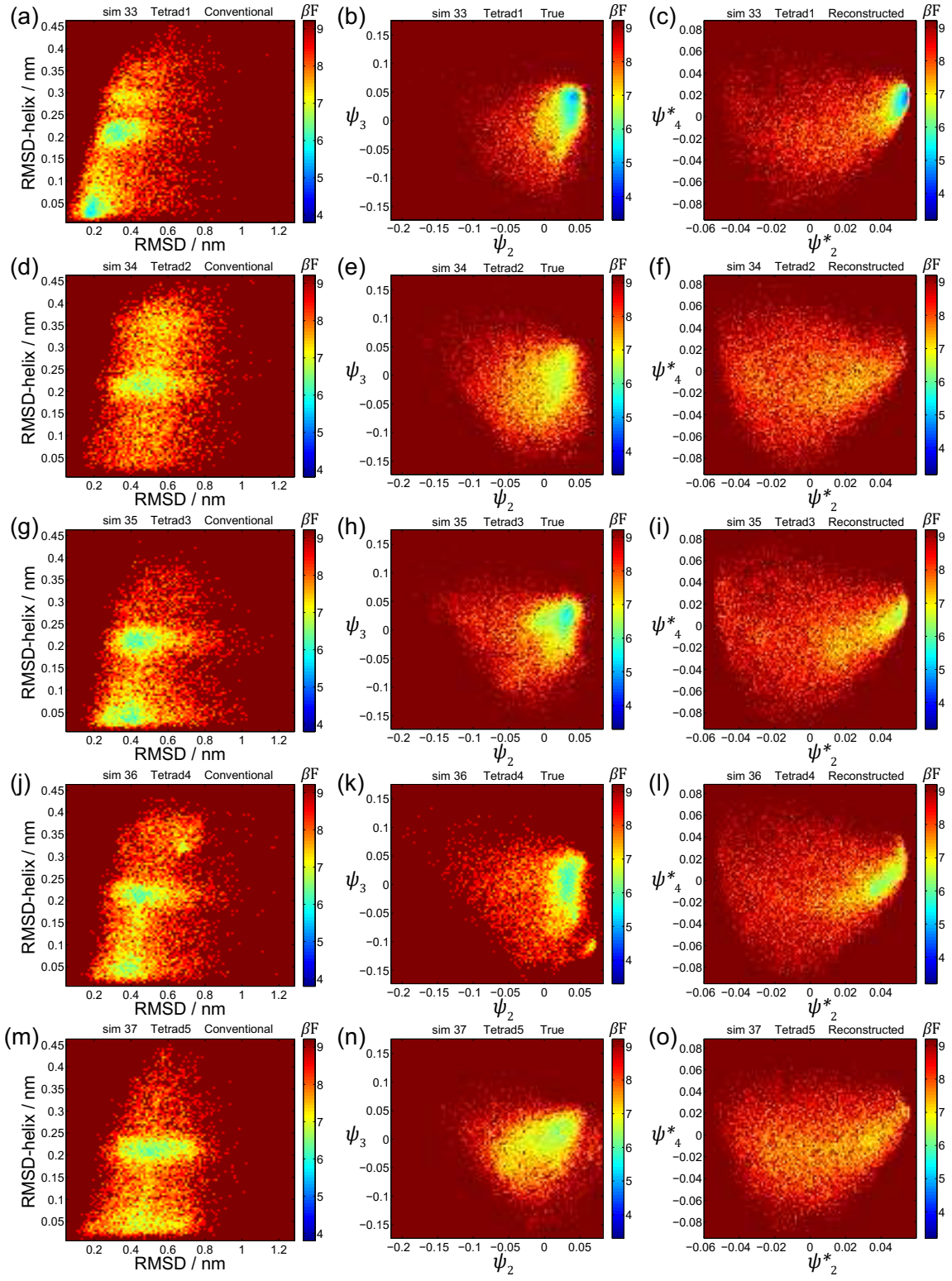


Figure 12: Single molecule free energy surfaces of selected alanine tetrad mutants of Trp-cage (PDB ID 1L2Y) (Simulations #33-#37 in Table 1). smFES supported by conventional CVs (left), the true intrinsic manifold M (center) and reconstructed manifold $\Theta(M)$ (right) for alanine tetrad mutants inserted into positions (a-c) 1-4, (d-f) 5-8, (g-i) 9-12, (j-l) 13-16, and (m-o) 17-20. We report the Helmholtz free energy F dedimensionalized by $\beta = 1/k_B T$ at $T = 380$ K.

degrees of freedom.

Molecular simulation trajectories give us access to both the true and reconstructed landscapes, and we demonstrate that the latter do indeed provide topologically equivalent representations of the former in long simulations of three small proteins Trp-cage, Villin, and BBA in explicit solvent. We further show in an ensemble of simulations of Trp-cage in implicit solvent, that the reconstructed landscapes reliably recapitulate topographical changes to the true landscape as a function of temperature or the introduction of amino acid mutations. We numerically compute the local Jacobian of the transformation between the true and reconstructed landscapes to the validate the existence of the diffeomorphism and quantify the topographical perturbation. Theoretical bounds on the local variation are not known, but our empirical analyses show that for the proteins studied here and the head-to-tail distance as a univariate observable, the degree of perturbation lies within approximately one order of magnitude. This empirical result bounds the degree of topographical perturbation of the reconstructed smFES and allows us to semi-quantitatively interpret the free energy barrier heights and well depths. This demonstrates, in principle, the potential to use protein folding landscapes reconstructed from time series in experimentally accessible observables to understand and engineer protein stability and folding.

This work reconstructs protein folding funnels from synthetic, idealized smFRET time series at arbitrarily high time resolution, subject to no measurement noise, and without regard to any concerns of photobleaching or perturbation of the underlying molecular motions by bulky dye molecules. Applications to real experimental data must confront these issues, but the present work establishes proof of principle in this idealized limit. In future work, we will engage these matters by artificially corrupting molecular simulation time series with shot noise, reducing the time resolution, directly simulating the dye molecules, and exploring the use of more and different molecular observables. Further, this work has provided empirical evidence that the degree of topographical perturbation may be bounded, and we are working to place theoretical bounds on the range of local variation in compression and dilation.

5 Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at `http://pubs.acs.org`. Figures provide additional single molecule free energy surfaces for Trp-cage (PDB ID 1L2Y) at various temperatures and subjected to alanine scan point mutations.

6 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. DMS-1841810. We are grateful to D.E. Shaw Research (DESRES) for sharing the Trp-cage, Villin, and BBA simulation trajectories.

References

(1) Dill, K. A.; Chan, H. S. From Levinthal to Pathways to Funnels. *Nat. Struct. Biol.* **1997**, *4*, 10–19.

(2) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of Protein Folding: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545–600.

(3) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Proteins: Struct. Func. Bioinf.* **1995**, *21*, 167–195.

(4) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. The Protein Folding Problem. *Annu. Rev. Biophys.* **2008**, *37*, 289–316.

(5) Dill, K. A.; MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science* **2012**, *338*, 1042–1046.

(6) Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. Low-Dimensional, Free-Energy Landscapes of Protein-Folding Reactions by Nonlinear Dimensionality Reduction. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 9885–9890.

(7) Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. Systematic Determination of Order Parameters for Chain Dynamics Using Diffusion Maps. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 13597–13602.

(8) Pande, V. S.; Grosberg, A. Y.; Tanaka, T. Heteropolymer Freezing and Design: Towards Physical Models of Protein Folding. *Rev. Mod. Phys.* **2000**, *72*, 259–314.

(9) Huang, P.-S.; Boyken, S. E.; Baker, D. The Coming of Age of De Novo Protein Design. *Nature* **2016**, *537*, 320–327.

(10) Lu, P.; Min, D.; DiMaio, F.; Wei, K. Y.; Vahey, M. D.; Boyken, S. E.; Chen, Z.; Fallas, J. A.; Ueda, G.; Sheffler, W. et al. Accurate Computational Design of Multipass Transmembrane Proteins. *Science* **2018**, *359*, 1042–1046.

(11) Alberstein, R.; Suzuki, Y.; Paesani, F.; Tezcan, F. A. Engineering the Entropy-Driven Free-Energy Landscape of a Dynamic Nanoporous Protein Assembly. *Nat. Chem.* **2018**, *10*, 732–739.

(12) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*; Academic Press, 2001.

(13) Ichiye, T.; Karplus, M. Collective Motions in Proteins: A Covariance Analysis of Atomic Fluctuations in Molecular Dynamics and Normal Mode Simulations. *Proteins: Struct. Func. Bioinf.* **1991**, *11*, 205–217.

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

(15) Amadei, A.; Linssen, A.; Berendsen, H. J. Essential Dynamics of Proteins. *Proteins: Struct. Func. Bioinf.* **1993**, *17*, 412–425.

(16) Hegger, R.; Altis, A.; Nguyen, P. H.; Stock, G. How Complex is the Dynamics of Peptide Folding? *Phys. Rev. Lett.* **2007**, *98*, 028102.

(17) Ferguson, A. L.; Panagiotopoulos, A. Z.; Kevrekidis, I. G.; Debenedetti, P. G. Nonlinear Dimensionality Reduction in Molecular Simulation: The Diffusion Map Approach. *Chem. Phys. Lett.* **2011**, *509*, 1–11.

(18) Zhuravlev, P. I.; Materese, C. K.; Papoian, G. A. Deconstructing the Native State: Energy Landscapes, Function, and Dynamics of Globular Proteins. *J. Phys. Chem. B* **2009**, *113*, 8800–8812.

(19) Plaku, E.; Stamati, H.; Clementi, C.; Kavraki, L. E. Fast and Reliable Analysis of Molecular Motion Using Proximity Relations and Dimensionality Reduction. *Proteins: Struct. Func. Bioinf.* **2007**, *67*, 897–907.

(20) Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. Integrating Diffusion Maps with Umbrella Sampling: Application to Alanine Dipeptide. *J. Chem. Phys.* **2011**, *134*, 135103.

(21) Ferguson, A. L.; Zhang, S.; Dikiy, I.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Link, J. A. An Experimental and Computational Investigation of Spontaneous Lasso Formation in Microcin J25. *Biophys. J.* **2010**, *99*, 3056–3065.

(22) Jolliffe, I. *Principal Component Analysis*; Wiley Online Library, 2002.

(23) Chatterjee, A. An Introduction to the Proper Orthogonal Decomposition. *Curr. Sci.* **2000**, 808–817.

(24) Liang, Y.; Lee, H.; Lim, S.; Lin, W.; Lee, K.; Wu, C. Proper Orthogonal Decomposition and its Applications—Part I: Theory. *J. Sound Vib.* **2002**, *252*, 527–544.

(25) Jolliffe, I. T. *Principal Component Analysis*; Springer, 1986; pp 115–128.

(26) Roweis, S. T.; Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **2000**, *290*, 2323–2326.

(27) Zhang, Z.; Wang, J. MLLE: Modified Locally Linear Embedding Using Multiple Weights. Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference. Cambridge, 2007; pp 1593–1600.

(28) Tenenbaum, J. B.; De Silva, V.; Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **2000**, *290*, 2319–2323.

(29) Schölkopf, B.; Smola, A.; Müller, K.-R. Kernel Principal Component Analysis. Artificial Neural Networks — ICANN'97: 7th International Conference Lausanne, Switzerland, October 8–10, 1997 Proceeedings. Berlin Heidelberg, 1997; pp 583–588.

(30) Kramer, M. A. Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. *AIChE J.* **1991**, *37*, 233–243.

(31) Nguyen, P. H. Complexity of Free Energy Landscapes of Peptides Revealed by Nonlinear Principal Component Analysis. *Proteins: Struct. Func. Bioinf.* **2006**, *65*, 898–913.

(32) Scholz, M.; Fraunholz, M.; Selbig, J. In *Principal Manifolds for Data Visualization and Dimension Reduction*; Gorban, A. N., Kegl, B., Wunsch, D. C., Zinovyev, A., Eds.; Lecture Notes in Computational Science and Engineering 58; Springer: Berlin Heidelberg, 2008; pp 44–67.

(33) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. Geometric Diffusions as a Tool for Harmonic Analysis and Structure Definition of Data: Diffusion Maps. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7426–7431.

(34) Coifman, R. R.; Lafon, S. Diffusion Maps. *Appl. Comput. Harm. Anal.* **2006**, *21*, 5–30.

(35) Noé, F.; Nuske, F. A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems. *Multiscale Model. Simul.* **2013**, *11*, 635–655.

(36) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theor. Comput.* **2013**, *9*, 2000–2009.

(37) Pérez-Hernández, G.; Noé, F. Hierarchical Time-Lagged Independent Component Analysis: Computing Slow Modes and Reaction Coordinates for Large Molecular Systems. *J. Chem. Theor. Comput.* **2016**, *12*, 6118–6129.

(38) Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noé, F. Variational Approach to Molecular Kinetics. *J. Chem. Theor. Comput.* **2014**, *10*, 1739–1752.

(39) Boninsegna, L.; Gobbo, G.; Noé, F.; Clementi, C. Investigating Molecular Kinetics by Variationally Optimized Diffusion Maps. *J. Chem. Theor. Comput.* **2015**, *11*, 5947–5960.

(40) Molgedey, L.; Schuster, H. G. Separation of a Mixture of Independent Signals Using Time Delayed Correlations. *Phys. Rev. Lett.* **1994**, *72*, 3634.

(41) Blaschke, T.; Berkes, P.; Wiskott, L. What is the Relation Between Slow Feature Analysis and Independent Component Analysis? *Neural Comput.* **2006**, *18*, 2495–2508.

(42) Hernández, C. X.; Wayment-Steele, H. K.; Sultan, M. M.; Husic, B. E.; Pande, V. S. Variational Encoding of Complex Dynamics. *Phys. Rev. E* **2018**, *97*, 062412.

(43) Ferguson, A. L. BayesWHAM: A Bayesian Approach for Free Energy Estimation, Reweighting, and Uncertainty Quantification in the Weighted Histogram Analysis Method. *J. Comput. Chem.* **2017**, *38*, 1583–1605.

(44) Karplus, M.; Petsko, G. A. Molecular Dynamics Simulations in Biology. *Nature* **1990**, *347*, 631–639.

(45) Chang, J. C.; Rosenthal, S. J. *Biomedical Nanotechnology*; Springer, 2011; pp 51–62.

(46) Roy, R.; Hohng, S.; Ha, T. A Practical Guide to Single-Molecule FRET. *Nat. Methods* **2008**, *5*, 507–516.

(47) Zerze, G. H.; Best, R. B.; Mittal, J. Modest Influence of FRET Chromophores on the Properties of Unfolded Proteins. *Biophys. J.* **2014**, *107*, 1654–1660.

(48) Wang, J.; Ferguson, A. L. Nonlinear Reconstruction of Single-Molecule Free-Energy Surfaces from Univariate Time Series. *Phys. Rev. E* **2016**, *93*, 032412.

(49) Takens, F. Detecting Strange Attractors in Turbulence. *Dynamical Systems and Turbulence* **1981**, *898*, 366–381.

(50) Sauer, T.; Yorke, J. A.; Casdagli, M. Embedology. *J. Stat. Phys.* **1991**, *65*, 579–616.

(51) Packard, N.; Crutchfield, J.; Farmer, J.; Shaw, R. Geometry from a Time Series. *Phys. Rev. Lett.* **1980**, *45*, 712–716.

(52) Broomhead, D. S.; King, G. P. Extracting Qualitative Dynamics from Experimental Data. *Physica D.* **1986**, *20*, 217–236.

(53) Cao, L.; Mees, A.; Judd, K. Dynamics From Multivariate Time Series. *Physica D.* **1998**, *121*, 75–88.

(54) Stark, J. Delay Embeddings for Forced Systems. I. Deterministic Forcing. *J. Nonlin. Sci.* **1999**, *9*, 255–332.

(55) Stark, J.; Broomhead, D. S.; Davies, M.; Huke, J. Delay Embeddings for Forced Systems. II. Stochastic Forcing. *J. Nonlin. Sci.* **2003**, *13*, 519–577.

(56) Vialar, T. *Complex and Chaotic Nonlinear Dynamics*; Springer, 2009.

(57) Kantz, H.; Schreiber, T. *Nonlinear Time Series Analysis*; Cambridge University Press, 2004.

(58) Ye, H.; Beamish, R. J.; Glaser, S. M.; Grant, S. C. H.; Hsieh, C.-H.; Richards, L. J.; Schnute, J. T.; Sugihara, G. Equation-Free Mechanistic Ecosystem Forecasting Using Empirical Dynamic Modeling. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, E1569–E1576.

(59) Coifman, R. R.; Kevrekidis, I. G.; Lafon, S.; Maggioni, M.; Nadler, B. Diffusion Maps, Reduction Coordinates, and Low Dimensional Representation of Stochastic Systems. *Multiscale Model. Simul.* **2008**, *7*, 842–864.

(60) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334*, 517–520.

(61) Bowers, K. J.; Chow, D. E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D. et al. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. SC 2006 Conference, Proceedings of the ACM/IEEE. 2006; pp 43–43.

(62) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. et al. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341–346.

(63) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How Robust are Protein Folding Simulations With Respect to Force Field Parameterization? *Biophys. J.* **2011**, *100*, L47–L49.

(64) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(65) MacKerell Jr, A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S. et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(66) Shan, Y.; Klepeis, J. L.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Gaussian Split Ewald: A Fast Ewald Mesh Method for Molecular Simulation. *J. Chem. Phys.* **2005**, *122*, 054101.

(67) Barua, B.; Lin, J. C.; Williams, V. D.; Kummler, P.; Neidigh, J. W.; Andersen, N. H. The Trp-Cage: Optimizing the Stability of a Globular Miniprotein. *Protein Eng. Des. Sel.* **2008**, *21*, 171–185.

(68) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. Sub-Microsecond Protein Folding. *J. Molec. Biol.* **2006**, *359*, 546–553.

(69) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. Designing a 20-Residue Protein. *Nat. Struct. Biol.* **2002**, *9*, 425–430.

(70) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D. et al. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLOS Comput. Biol.* **2017**, *13*, e1005659.

(71) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins: Struct. Func. Bioinf.* **2006**, *65*, 712–725.

(72) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes With a Modified Generalized Born Model. *Proteins: Struct. Func. Bioinf.* **2004**, *55*, 383–394.

(73) Giannakis, D.; Majda, A. J. Nonlinear Laplacian Spectral Analysis for Time Series With Intermittency and Low-Frequency Variability. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 2222–2227.

(74) Berry, T.; Cressman, J.; Greguric-Ferencek, Z.; Sauer, T. Time-Scale Separation from Diffusion-Mapped Delay Coordinates. *SIAM J. Appl. Dyn. Syst.* **2013**, *12*, 618–649.

(75) Villani, V.; Tamburro, A. M.; Zaldivar Comenges, J. M. Conformational Chaos and Biomolecular Instability in Aqueous Solution. *J. Chem. Soc., Perkin Trans. 2* **2000**, *13*, 2177–2184.

(76) Villani, V.; Zaldivar Comenges, J. M. Analysis of Biomolecular Chaos in Aqueous Solution. *Theor. Chem. Acc.* **2000**, *104*, 290–295.

(77) Hashemian, B.; Arroyo, M. Topological Obstructions in the Way of Data-Driven Collective Variables. *J. Chem. Phys.* **2015**, *142*, 044102.

(78) Cross, D. J.; Gilmore, R. Differential Embedding of the Lorenz Attractor. *Phys. Rev. E* **2010**, *81*, 066220.

(79) Letellier, C.; Gouesbet, G. Topological Characterization of Reconstructed Attractors Modding Out Symmetries. *Journal de Physique II* **1996**, *6*, 1615–1638.

(80) Fraser, A. M.; Swinney, H. L. Independent Coordinates for Strange Attractors from Mutual Information. *Phys. Rev. A* **1986**, *33*, 1134–1140.

(81) Letellier, C.; Maquet, J.; Le Sceller, L.; Gouesbet, G.; Aguirre, L. On the Non-Equivalence of Observables in Phase-Space Reconstructions from Recorded Time Series. *J. Phys. A* **1998**, *31*, 7913.

(82) Cao, L. Practical Method for Determining the Minimum Embedding Dimension of a Scalar Time Series. *Physica D.* **1997**, *110*, 43–50.

(83) Kennel, M. B.; Brown, R.; Abarbanel, H. D. Determining Embedding Dimension for Phase-Space Reconstruction Using a Geometrical Construction. *Phys. Rev. A* **1992**, *45*, 3403.

(84) Zwanzig, R. *Nonequilibrium Statistical Mechanics*; Oxford University Press, 2001.

(85) Tolman, R. C. *The Principles of Statistical Mechanics*; Courier Corporation, 1979.

(86) Nadler, B.; Lafon, S.; Coifman, R. R.; Kevrekidis, I. G. *Advances in Neural Information Processing Systems 18: Proceedings of the 2005 Conference (Neural Information Processing)*; The MIT Press, 2006; pp 955–962.

(87) Nadler, B.; Lafon, S.; Coifman, R. R.; Kevrekidis, I. G. Diffusion Maps, Spectral Clustering and Reaction Coordinates of Dynamical Systems. *Appl. Comput. Harm. Anal.* **2006**, *21*, 113–127.

(88) Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr. A* **1976**, *32*, 922–923.

(89) Coifman, R. R.; Shkolnisky, Y.; Sigworth, F. J.; Singer, A. Graph Laplacian Tomography From Unknown Random Projections. *IEEE Trans. Image Process.* **2008**, *17*, 1891–1899.

(90) Ma, A.; Dinner, A. R. Automatic Method for Identifying Reaction Coordinates in Complex Systems. *J. Phys. Chem. B* **2005**, *109*, 6769–6779.

(91) Peters, B.; Trout, B. L. Obtaining Reaction Coordinates by Likelihood Maximization. *J. Chem. Phys.* **2006**, *125*, 054108.

(92) Peters, B.; Beckham, G. T.; Trout, B. L. Extensions to the Likelihood Maximization Approach for Finding Reaction Coordinates. *J. Chem. Phys.* **2007**, *127*, 034109.

(93) Wang, J.; Ferguson, A. L. Nonlinear Machine Learning in Simulations of Soft and Biological Materials. *Mol. Sim.* **2017**, *44*, 1090–1107.

(94) Wang, J.; Ferguson, A. L. A Study of the Morphology, Dynamics, and Folding Pathways of Ring Polymers with Supramolecular Topological Constraints Using Molecular Simulation and Nonlinear Manifold Learning. *Macromol.* **2018**, *51*, 598–616.

(95) Long, A. W.; Ferguson, A. L. Landmark Diffusion Maps (L-dMaps): Accelerated Manifold Learning Out-of-Sample Extension. *Appl. Comput. Harm. Anal.* **2017**, in press (doi: 10.1016/j.acha.2017.08.004).

(96) Sonday, B. E.; Haataja, M.; Kevrekidis, I. G. Coarse-Graining the Dynamics of a Driven Interface in the Presence of Mobile Impurities: Effective Description via Diffusion Maps. *Phys. Rev. E* **2009**, *80*, 031102.

(97) Laing, C. R.; Frewen, T. A.; Kevrekidis, I. G. Coarse-Grained Dynamics of an Activity Bump in a Neural Field Model. *Nonlinearity* **2007**, *20*, 2127–2146.

(98) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. Determination of Reaction Coordinates via Locally Scaled Diffusion Map. *J. Chem. Phys.* **2011**, *134*, 03B624.

(99) Stamati, H.; Clementi, C.; Kavraki, L. E. Application of Nonlinear Dimensionality Reduction to Characterize the Conformational Landscape of Small Peptides. *Proteins: Struct. Func. Bioinf.* **2010**, *78*, 223–235.

(100) Wang, J.; Gayatri, M. A.; Ferguson, A. L. Mesoscale Simulation and Machine Learning of Asphaltene Aggregation Phase Behavior and Molecular Assembly Landscapes. *J. Phys. Chem. B* **2017**, *121*, 4923–4944.

(101) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph. Model.* **1996**, *14*, 33–38.

(102) Juraszek, J.; Bolhuis, P. Sampling the Multiple Folding Mechanisms of Trp-Cage in Explicit Solvent. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 15859–15864.

(103) Kim, S. B.; Dsilva, C. J.; Kevrekidis, I. G.; Debenedetti, P. G. Systematic Characterization of Protein Folding Pathways Using Diffusion Maps: Application to Trp-cage Miniprotein. *J. Chem. Phys.* **2015**, *142*, 085101.

(104) Chen, W.; Tan, A. R.; Ferguson, A. L. Collective Variable Discovery and Enhanced Sampling Using Autoencoders: Innovations in Network Architecture and Error Function Design. *J. Chem. Phys.* **2018**, *149*, 072312.

(105) Chen, W.; Ferguson, A. L. Molecular Enhanced Sampling With Autoencoders: On-the-Fly Nonlinear Collective Variable Discovery and Accelerated Free Energy Landscape Exploration. *J. Comput. Chem.* **2018**, *39*, 2079–2102.

(106) Letellier, C.; Aguirre, L.; Maquet, J. How the Choice of the Observable May Influence the Analysis of Nonlinear Dynamical Systems. *Commun. Nonlin. Sci.* **2006**, *11*, 555–576.

(107) Letellier, C.; Aguirre, L. A. Investigating Nonlinear Dynamics from Time Series: The Influence of Symmetries and the Choice of Observables. *Chaos* **2002**, *12*, 549–558.

(108) Kass, R. E.; Vos, P. W. *Geometrical Foundations of Asymptotic Inference*; John Wiley & Sons, 2011; Chapter Appendix A: Diffeomorphisms and the Inverse Function Theorem, pp 300–303.

(109) Lei, H.; Wu, C.; Liu, H.; Duan, Y. Folding Free-Energy Landscape of Villin Headpiece Subdomain From Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4925–4930.

(110) Sarisky, C. A.; Mayo, S. L. The $\beta\beta\alpha$ Fold: Explorations in Sequence Space. *J. Molec. Biol.* **2001**, *307*, 1411–1418.

(111) Li, W.; Zhang, J.; Wang, W. Understanding the Folding and Stability of a Zinc Finger-Based Full Sequence Design Protein With Replica Exchange Molecular Dynamics Simulations. *Proteins: Struct. Func. Bioinf.* **2007**, *67*, 338–349.

(112) Mohanty, S.; Hansmann, U. H. E. Folding of a Miniprotein With Mixed Fold. *J. Chem. Phys.* **2007**, *127*, 035102.

(113) Lei, H.; Wang, Z.-X.; Wu, C.; Duan, Y. Dual Folding Pathways of an α/β Protein From All-Atom Ab Initio Folding Simulations. *J. Chem. Phys.* **2009**, *131*, 165105.

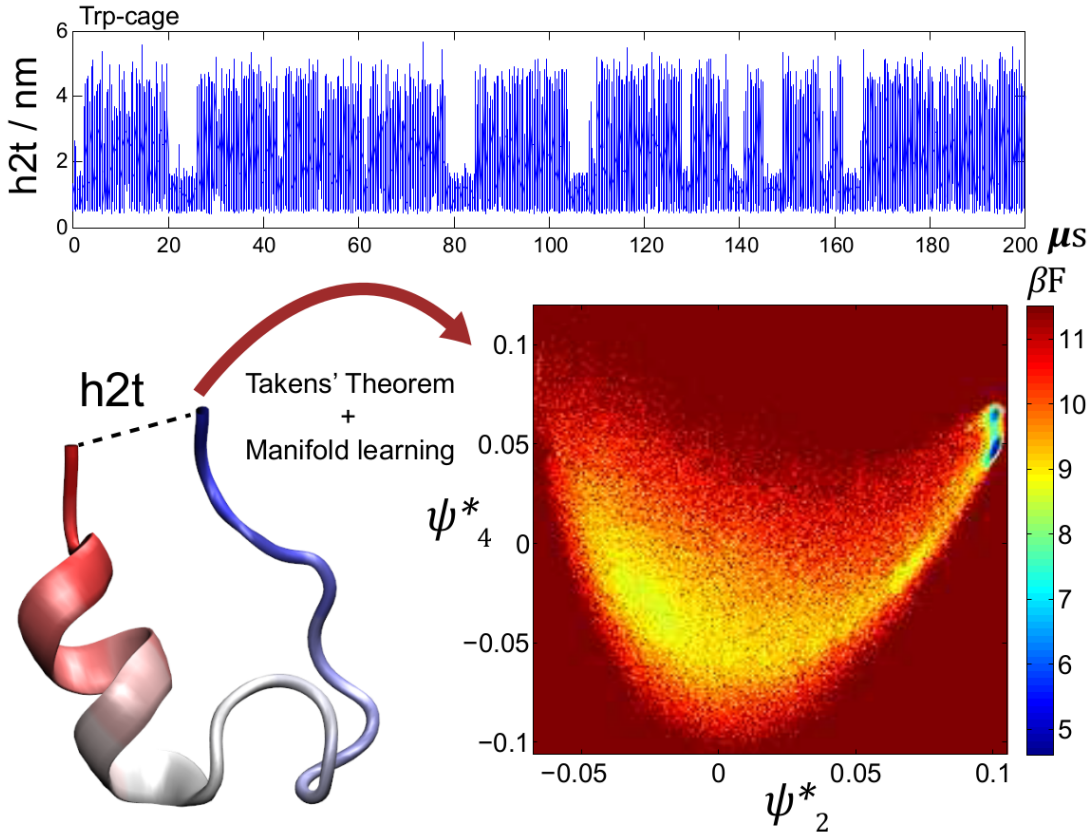
(114) Mansbach, R. A.; Ferguson, A. L. Machine Learning of Single Molecule Free Energy Surfaces and the Impact of Chemistry and Environment upon Structure and Dynamics. *J. Chem. Phys.* **2015**, *142*, 105101.

(115) Morrison, K. L.; Weiss, G. A. Combinatorial Alanine-Scanning. *Curr. Opin. Chem. Biol.* **2001**, *5*, 302–307.

(116) Herman, R. E.; Badders, D.; Fuller, M.; Makienko, E. G.; Houston, M. E.; Quay, S. C.;

Johnson, P. H. The Trp Cage Motif as a Scaffold for the Display of a Randomized Peptide Library on Bacteriophage T7. *J. Biol. Chem.* **2007**, *282*, 9813–9824.

TOC Graphic


Trp-cage
h2t / nm
0
2
4
6
0
20
40
60
80
100
120
140
160
180
200
μs
h2t
Takens' Theorem
+
Manifold learning
βF
11
10
9
8
7
6
5
0.1
0.05
0
−0.05
−0.1
ψ*4
−0.05
0
0.05
0.1
ψ*2