The Journal of Chemical Physics

Machine learning of single molecule free energy surfaces and the impact of chemistry and environment upon structure and dynamics

Cite as: J. Chem. Phys. **142**, 105101 (2015); https://doi.org/10.1063/1.4914144
Submitted: 09 December 2014 • Accepted: 23 February 2015 • Published Online: 12 March 2015

Rachael A. Mansbach and Andrew L. Ferguson

Machine learning of single molecule free energy surfaces and the impact of chemistry and environment upon structure and dynamics

Rachael A. Mansbach[1] and Andrew L. Ferguson[2,a)]
[1]*Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*
[2]*Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*

(Received 9 December 2014; accepted 23 February 2015; published online 12 March 2015)

The conformational states explored by polymers and proteins can be controlled by environmental conditions (e.g., temperature, pressure, and solvent) and molecular chemistry (e.g., molecular weight and side chain identity). We introduce an approach employing the diffusion map nonlinear machine learning technique to recover single molecule free energy landscapes from molecular simulations, quantify changes to the landscape as a function of external conditions and molecular chemistry, and relate these changes to modifications of molecular structure and dynamics. In an application to an *n*-eicosane chain, we quantify the thermally accessible chain configurations as a function of temperature and solvent conditions. In an application to a family of polyglutamate-derivative homopeptides, we quantify helical stability as a function of side chain length, resolve the critical side chain length for the helix-coil transition, and expose the molecular mechanisms underpinning side chain-mediated helix stability. By quantifying single molecule responses through perturbations to the underlying free energy surface, our approach provides a quantitative bridge between experimentally controllable variables and microscopic molecular behavior, guiding and informing rational engineering of desirable molecular structure and function. © *2015 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4914144]

I. INTRODUCTION

The conformations adopted by polymers and proteins dictate their structural and functional properties.[1–4] The solubility of *n*-alkanes in water is intimately related to the collapse of the chain at sufficiently high molecular weights,[5] and the activity of biological enzymes is contingent on the structure and fluctuations of the active site.[6,7] The conformational ensemble explored by polymers and proteins can be controlled by chemical modification of the molecules themselves (e.g., side chain chemistry) or by changing environmental conditions (e.g., temperature, pressure, and solvent). For instance, hydrophobic polypeptides such as deca-alanine adopt α-helical conformations in water whereas deca-glutamate exists as a random coil.[8] Similarly, moving a polystyrene chain from a good solvent, such as xylene, to a poor solvent, such as methanol, results in collapse of the chain from a swollen, extended conformation to a collapsed globule.[9] A quantitative understanding of the impact of molecular chemistry and environmental conditions upon the microscopic conformations adopted by polymers and peptides is of fundamental interest in understanding the structural and functional properties of the chains and a prerequisite to the rational design of polymers and proteins with desired structural and functional properties.

The conformation of a molecule comprising N atoms exists in the $3N$-dimensional phase space defined by the Cartesian coordinates of the constituent atoms. Cooperative couplings between the degrees of freedom render the effective dimensionality (i.e., the number of variables required to effectively specify the conformational state of the molecule) far smaller than $3N$.[10–14] Two-dimensional descriptions have been calculated for dialanine[15] and the src homology domain,[13] and three-dimensional descriptions have been calculated for the antimicrobial peptide microcin J25[16] and *n*-alkane chains.[17] In a geometric sense, the cooperative couplings cause the molecule to explore only a restricted volume of the complete $3N$-dimensional phase space, and the shape and extent of this *intrinsic manifold*[14,17,18] can be systematically extracted using machine learning techniques.[10,13,14,19–21] Molecules containing more than a few atoms are expected to possess nonlinear intrinsic manifolds that cannot be easily discovered by linear approaches.[13,14,16,17] Diffusion maps are a powerful nonlinear dimensionality reduction technique that we, and others, have previously used to recover the intrinsic manifolds of polymers and peptides from molecular simulations.[14,16,17,22–28]

In this work, we apply diffusion maps to conformational ensembles constructed by aggregating molecular simulations of (i) *n*-alkanes at different temperatures and in different solvent conditions and (ii) peptides with different side chain chemistries. Provided that there is partial overlap between the molecular configurations sampled by each system, these "composite" diffusion maps can discover the intrinsic manifold spanning the multi-system ensemble and expose the impact of molecular chemistry and environmental conditions upon the accessible molecular conformations and pathways. By explicitly linking the impact of external control parameters (e.g., side chain chemistry, solvent environment, and temperature) and single molecule behavior (e.g., conformational

a)Author to whom correspondence should be addressed. Electronic mail: alf@illinois.edu

ensemble and structural stability), this approach can guide and inform how to tailor molecular chemistry and environment to obtain desirable structural and/or functional behaviors. In this work, we quantitate how solvent conditions and temperature influence the folding of an *n*-alkane chain and how side chain length mediates the helix-coil transition and stability of helical conformations of short peptides.

The goals of this work are threefold: (i) to present a new method to quantify the impact of molecular chemistry and environmental conditions on the microscopic structure of single molecules, (ii) to validate this methodology in the analysis of the well-studied, but conformationally rich, model of a solvated *n*-alkane chain, and (iii) to apply this approach to develop new molecular level insight and understanding into the tunable stabilization of helical peptides by side chain chemistry. The structure of this paper is as follows. In Sec. II, we provide details of our molecular simulation methodology and our composite diffusion map approach. In Sec. III A, we detail the application of our methodology to *n*-eicosane alkane chains to characterize the impact upon chain structure and folding at two temperatures and in three solvent environments. In Sec. III B, we apply our approach to a family of homomeric decapeptides composed of non-natural glutamate amino acid derivatives to demonstrate that the helix-coil transition and stability of the helical conformation may be controlled by tuning the length of the amino acid side chains. In Sec. IV, we present our conclusions and perspectives for future work.

II. THEORETICAL METHODS

A. Molecular simulations of *n*-eicosane

We conducted molecular dynamics simulations of *n*-eicosane ($C_{20}H_{42}$) in three solvent environments—neat phase, aqueous solution, and ideal gas—at two temperatures—298 K and 323 K—for a total of six systems. The ideal gas phase refers to an isolated chain that interacts only with itself. Neat and solvated phase molecular dynamics simulations were conducted using the GROMACS 4.6 simulation suite,[29] employing the united atom Transferable Potential for Phase Equilibria (TraPPE) model for the alkane chain[30] and the simple point charge (SPC) model for water.[31] Initial configurations were generated with the help of the PRODRG Server.[32] Simulations were conducted in the NPT ensemble at 298 K and 323 K at 1 bar, employing a Nosé-Hoover thermostat[33] and Parrinello-Rahman barostat.[34] The neat phase exists as a crystalline solid at 298 K and a liquid at 323 K. Electrostatic interactions in the aqueous phase were treated using particle mesh Ewald with a real-space cutoff of 1.4 nm and a 0.12 nm Fourier grid spacing,[35] and Lennard-Jones interactions shifted smoothly to zero at 1.4 nm. Equilibration runs of 10 ps for the aqueous phase and 10 ns and 90 ns for the 323 K and 298 K neat phases, respectively, were conducted. Production runs of 30 ns were performed, with configurations saved every 2 ps. Simulations of *n*-eicosane in the ideal gas phase were carried out using configurational-bias chain regrowth Monte Carlo[5,36,37] at 298 K and 323 K. Monte Carlo sampling was performed for 150 001 steps with coordinates saved every 10 steps. Complete details of our simulation methodology are provided in the supplementary material.[38]

B. Molecular simulations of polyglutamate-derivative peptides

We conducted molecular dynamics simulations of decamers of the five polyglutamate-derivatives poly(γ-(3-aminoethyl)-L-glutamate) ($PAGG_{10}$), poly(γ-(3-aminopropyl)-L-glutamate) ($PAPG_{10}$), poly(γ-(4-aminobutyl)-L-glutamate) ($PABG_{10}$), poly(γ-(5-aminopentanyl)-L-glutamate) ($PATG_{10}$), and poly(γ-(6-aminohexyl)-L-glutamate) ($PAHG_{10}$), plus polylysine (PL_{10}) as a control[8] (cf. Fig. 5). Initial peptide configurations were manually constructed assisted by the GlycoBioChem PRODRG2 Server[32] and the Bax Group PDB Utility Server (http://spin.niddk.nih.gov/bax/nmrserver/pdbutil). Simulations were conducted using the GROMACS 4.6 simulation suite,[29] employing the CHARMM27 force field for the peptides[39] and the TIP3P water model.[40] Simulations were conducted in the NPT ensemble at 298 K and 1 bar, employing a Nosé-Hoover thermostat[33] and Parrinello-Rahman barostat.[34] Electrostatic interactions were treated using particle mesh Ewald with a real-space cutoff of 1.2 nm and a 0.12 nm Fourier grid spacing,[35] and Lennard-Jones interactions shifted smoothly to zero at 1.2 nm. A 32 ns equilibration run was conducted for each of $PAPG_{10}$, $PABG_{10}$, $PATG_{10}$, and $PAHG_{10}$ and 92 ns runs for PL_{10} and $PAGG_{10}$. We then performed 28 ns production runs, saving simulation snapshots every 2 ps. Full details of our simulation methodology are presented in the supplementary material.[38]

C. Composite diffusion maps

The diffusion map[41–44] is a nonlinear manifold learning algorithm that has been profitably employed to discover low-dimensional descriptions for polymers and peptides within the $3N$-dimensional Cartesian phase space specifying the positions of the N atoms in the system.[14,16–18,22–24] Diffusion maps are more powerful and flexible than linear approaches (e.g., principal components analysis[45]) since they do not make linear approximations to the inherently nonlinear manifolds occupied by long polymers and peptides.[13,14,16,17] The low-dimensional embeddings discovered by diffusion maps expose the conformational space explored by the system, reveal the slow dynamical modes governing its evolution, and provide dynamically meaningful order parameters in which to construct dynamically meaningful free energy surfaces (FESs).[14]

In this work, we apply diffusion maps to molecular simulation trajectories following the methodology detailed in Refs. 14 and 17. In brief, given an ensemble of R system snapshots, we compute the symmetric R-by-R matrix, $\mathbf{P}$, where the matrix element P_{ij} corresponds to the pairwise distance between system snapshots i and j. Diffusion maps require that the scalar distance metric should serve as a good structural proxy for the dynamic proximity of system snapshots on short time scales.[14,17] Specifically, the distance measure should be well correlated with structural similarity

below some characteristic value of the measure, ϵ. As we shall see, the pairwise distances matrix will be convolved with a Gaussian kernel of bandwidth ϵ, effectively discarding elements containing distances much larger than the characteristic distance. An appropriate value of ϵ can be inferred from the data in a systematic manner.[46] We employ as a natural distance metric for biomolecular systems, which has worked well in previous studies, the root mean square distance (RMSD) between the atomic coordinates of pairs of system configurations that we translationally and rotationally align using the Kabsch algorithm.[14,16,17,47] In this work, our system snapshots comprise the (united) atom coordinates of single *n*-alkane or peptide chains. The environment of the chain in the *n*-eicosane simulations (i.e., water, other chains, and vacuum) is not explicitly considered in our distance measure, but its effects are implicitly captured in the configurations explored by the tagged molecule.[17] Explicit representation of solvent molecules in the distance metric in a spatially invariant manner is complicated by their inherent fungibility.[27,48] (We have recently proposed a distance metric based on graph matching between molecular clusters to surmount this difficulty and enable direct application of diffusion maps to multi-molecular phenomena.[49]) Similarly, by considering only the coordinates of the heavy backbone atoms of the peptide chains, the impact of the side chains upon the backbone configurational ensemble is implicitly captured. By constructing composite diffusion maps over the configurational ensembles harvested from multiple simulations under different environmental conditions for *n*-eicosane and side chain chemistries for the peptides, we resolve the impact of these factors upon the single molecule configurational ensembles and dynamical motions.

Having computed the R-by-R pairwise distances matrix, $\mathbf{P}$, we create a stochastic Markov matrix by convoluting the elements of the $\mathbf{P}$ matrix with a Gaussian kernel with bandwidth ϵ to form the matrix $\mathbf{A}$,

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

This soft thresholding of the pairwise distances attenuates large values—where the distance metric may not reliably reflect dynamic proximity—to retain only local distance information.[14,17,49] The Gaussian kernel is the infinitesimal generator of a diffusion process, and by forming this convolution, we model a random walk over the data points in the high-dimensional space.[44] It is by analyzing the spectral properties of this random walk that the diffusion map generates a low-dimensional embedding of the data.[43,44] We specify ϵ in an automated manner using the approach in Ref. 46.

We next compute the diagonal matrix $\mathbf{D}$ as the row sums of $\mathbf{A}$,

$$D_{ij} = \begin{cases} \sum_{k=1}^{R} A_{kj}, & \text{if } i = j, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

to form the right stochastic Markov matrix $\mathbf{M}$,

$$\mathbf{M} = \mathbf{D}^{-1}\mathbf{A}. \tag{3}$$

The elements M_{ij} may be interpreted as the transition probability from snapshot i to snapshot j, and $\mathbf{M}$ may therefore be regarded as defining a discrete random walk over the data.[43,44] The matrix $\mathbf{M}$ is closely related to the normalized graph Laplacian, which is a discrete approximation to the backward Fokker-Planck operator describing a diffusion process in the presence of potential wells.[17,42,46]

The right eigenvectors of the $\mathbf{M}$ matrix $\{\vec{\phi}_1, \vec{\phi}_2, \ldots, \vec{\phi}_R\}$, possessing associated eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_R$, are discrete approximations to the eigenfunctions of the Fokker-Planck equation,[43,44] which are identifiable as particular collective modes of the diffusion process over the data. By the Markov property of $\mathbf{M}$, $\lambda_1 = 1$ and $\vec{\phi}_1 = \vec{1}$ (i.e., the "all-ones" vector). Frequently, the eigenvalue spectrum exhibits a spectral gap separating a small number of lower order modes from the remainder of the spectrum.[14,16,17,23,27] The leading modes govern the long-time slow evolution of the system to which the higher order modes are effectively slaved.[14] Under the Mori-Zwanzig formalism,[50] the modes above the gap define a "slow subspace" for the diffusion process, constituting the low-dimensional intrinsic manifold to which the dynamical evolution of the system is effectively confined.[49]

For a spectral gap after the $(k+1)$th eigenvalue, the important slow dynamics of the system are contained with the first k non-trivial eigenvectors (recall that $\vec{\phi}_1 = \vec{1}$) motivating construction of the $k << 3N$-dimensional *diffusion map embedding* of the ith snapshot, $\vec{x}_i$, into the ith component of the top k non-trivial eigenvectors,[14,17,43,44]

$$\vec{x}_i \rightarrow (\vec{\phi}_2(i), \vec{\phi}_3(i), \ldots, \vec{\phi}_{k+1}(i)). \tag{4}$$

We have developed in-house C++ code to construct diffusion map embeddings that employs the implicitly restarted Arnoldi method implemented in the Parallel ARPACK libraries to compute the leading eigenvector/eigenvalue pairs of the $\mathbf{M}$ matrix.[51]

D. Free energy surfaces

Assuming that the simulation snapshots from which the diffusion map was constructed constitute a sample from an equilibrium distribution, FESs over the intrinsic manifold may be estimated from histograms over the diffusion map embedding.[17] Clearly the ensemble composed of data drawn from independent, non-interacting simulations under different conditions does not meet this criterion, but it is satisfied for points sampled from a single (sufficiently long) simulation. Specifically, $\beta F(\vec{\xi}) = -\ln\hat{P}(\vec{\xi}) + C$, where $\beta = 1/k_B T$, k_B is Boltzmann's constant, T is the temperature, $\vec{\xi}$ is a k-dimensional vector specifying a point on the intrinsic manifold spanned by the vectors $(\vec{\phi}_2, \vec{\phi}_3, \ldots, \vec{\phi}_{k+1})$, $F(\vec{\xi})$ is the free energy at $\vec{\xi}$, $\hat{P}(\vec{\xi})$ is a histogram approximation to the density of points on the manifold at $\vec{\xi}$, and C is an arbitrary additive constant. In the NPT ensemble, F is identifiable as the Gibbs free energy.

E. Visualization of diffusion map modes

A drawback of nonlinear dimensionality reduction approaches relative to linear techniques is the absence of an explicit mapping between the input variables (e.g., the atomic coordinates) and the variables parameterizing the low dimensional embeddings. This can make it challenging to identify a clear physical interpretation of the low-dimensional collective modes.[14,17,49] Protocols exist to systematically

search pools of candidate variables to find variable combinations approximating the nonlinear collective modes,[52,53] but the complexity of the identified combinations can themselves obscure physical interpretability. Indeed, the very existence of a simple physical interpretation for the modes underlying a complex dynamical process is not assured.[14,49] In the present work, we present a modification of a procedure detailed by Berry *et al.*[54] to visualize the collective modes discovered by the diffusion map (i.e., the eigenvectors $\{\vec{\phi}_l\}_{l=2}^{k+1}$) by extracting those configurations in the high-dimensional ambient space that contribute strongly to a particular mode. This procedure offers a transparent visual interpretation of the diffusion map modes in terms of dynamical motions in the original configurational space, providing an interpretive bridge between the high-dimensional space and the collective variables spanning the low-dimensional intrinsic manifold.

Following Berry *et al.*,[54] we form a projection of each of the top k non-trivial eigenvectors of the $\mathbf{M}$ matrix as

$$\vec{q}_l = \mathbf{D}\vec{\phi}_l, \quad l = 2 \ldots (k+1). \tag{5}$$

Since the main diagonal of the $\mathbf{D}$ matrix is the stationary distribution of the random walk over the data points,[17] the $\{\vec{q}_l\}_{l=2}^{k+1}$ constitute a scaling of the slow collective relaxations over the data by the equilibrium distribution. We collect the ensemble of R system snapshots in the R-by-$3N$ matrix $\mathbf{X}$, the rows of which are the $3N$-dimensional snapshots $\{\vec{x}_i\}_{i=1}^{R}$. The matrix-vector product of the lth R-by-1 scaled diffusion map mode, q_l, with the R-by-$3N$ snapshot matrix, $\mathbf{X}$, is the sum of the R snapshots weighted by q_l,

$$\Gamma_l = \vec{q}_l^{\,T}\mathbf{X} = \langle \vec{x}, \vec{q}_l \rangle = \sum_{i=1}^{R} \vec{x}_i \vec{q}_l(i). \tag{6}$$

The $3N$-dimensional vector Γ_l is a configuration in the original Cartesian space constructed as an average over the $\{\vec{x}_i\}_{i=1}^{R}$ snapshots comprising the ensemble, where each snapshot is weighted by the corresponding element of the scaled diffusion map mode q_l. Configurations in the ensemble that vary strongly with the lth diffusion map mode contribute much to the average, whereas those which are weakly associated with the mode contribute little.[54]

To better elucidate the dynamical motions associated with each diffusion map mode in configurational space, we visualize the sum on the right hand side of Eq. (6) by superposing the rotationally and translationally aligned configurations in the ensemble, $\{\vec{x}_i\}_{i=1}^{R}$, and color each configuration, i, according to its weight in the sum given by the ith element of $\vec{q}_l$. This representation provides an elegant means to clearly visualize the configurational motions associated with the collective modes discovered by the diffusion map.

III. RESULTS AND DISCUSSION

A. Thermodynamics and dynamical motions of *n*-eicosane as a function of solvent and temperature

N-alkanes are ubiquitous in nature as constituents of natural gas and oil, in the chemical and processing industries as fuels and lubricants, and in daily life as components of gasoline and petroleum jelly.[30,55–57] Their chemical simplicity enables relative ease in the calculation and understanding of their structure and properties but belies their rich conformational and thermodynamic behaviors.[5,17,58–60] N-alkanes may be considered the prototypical hydrophobic chain, and their structure and behavior in aqueous environments are of interest in understanding the impact of the hydrophobic effect upon protein structure and dynamics.[5,17,61–66] It is therefore of fundamental interest to characterize their structure and conformational dynamics in water and the neat melt or crystal.

A large number of previous investigations have been performed to unravel the structural properties of n-alkanes. Sun *et al.*[65] analyzed the preferred equilibrium conformations of n-octadecane in different environments, and several groups have conducted both theoretical[5,67,68] and experimental[69] studies of the molecular weight at which n-alkanes transition from an extended to a collapsed state in aqueous solvent. Athawale *et al.*[66] broadly studied conformational transitions, while Jorgensen *et al.* focused on the stability of conformations and structural conformations for n-butane[70] and Chakrabarty and Bagchi[58] did the same for n-alkanes of varying lengths. Mountain and Thirumalai studied the mechanism of urea denaturation.[63] There have also been many studies of alkane aqueous solubility,[71,72] hydrophobicity,[64,73–76] vapor-liquid equilibrium,[77–82] and hydration.[83]

We have previously employed diffusion maps to characterize the folding of n-octane (C_8H_{18}), n-hexadecane ($C_{16}H_{34}$), and n-tetracosane ($C_{24}H_{50}$) both in aqueous solution and in the ideal gas at 298 K at 1 bar.[17] We identified a three-dimensional intrinsic manifold spanned by collective variables that were well-correlated with the principal moments of the n-alkane chain gyration tensor. In the present study, we both use this system as a testing ground to prove our methodology and go beyond our previous work by using diffusion maps to quantify the impact of temperature and solvent upon chain structure and dynamics.

1. Composite diffusion maps

We constructed a composite diffusion map comprising the $6 \times 15\,001 = 90\,006$ snapshots of the n-eicosane chain harvested from each of the six simulations (i.e., neat phase, aqueous phase, and ideal gas at 298 K and 323 K). Using the approach in Ref. 46, we selected the bandwidth of the Gaussian kernel as $\epsilon = 1.83 \times 10^{-3}$. To estimate the dimensionality of the intrinsic manifold, and therefore the effective dimensionality of the system, we employ two independent approaches that we have profitably employed in previous studies:[17,49] the L-method of Salvador and Chan[84] and the plateau dimension of Sauer, Yorke, and Casdagli.[85] First, we applied the L-method[84] to identify in the eigenvalue spectrum in Fig. 1(a) a spectral gap after λ_4, implying an effective system dimensionality of ($k = 3$) and suggesting that we construct diffusion map embeddings in $\{\vec{\phi}_2, \vec{\phi}_3, \vec{\phi}_4\}$. Second, we computed the fractal dimension, d_{frac}, of the intrinsic manifold as a function of the number of eigenvectors, k, in the diffusion map embedding.[85,86] As illustrated in Fig. 1(b), d_{frac} saturates at the *plateau dimension*[85] of $d_{frac} \approx 4.3$, implying that the important dynamical motions of the system are contained within a five-dimensional

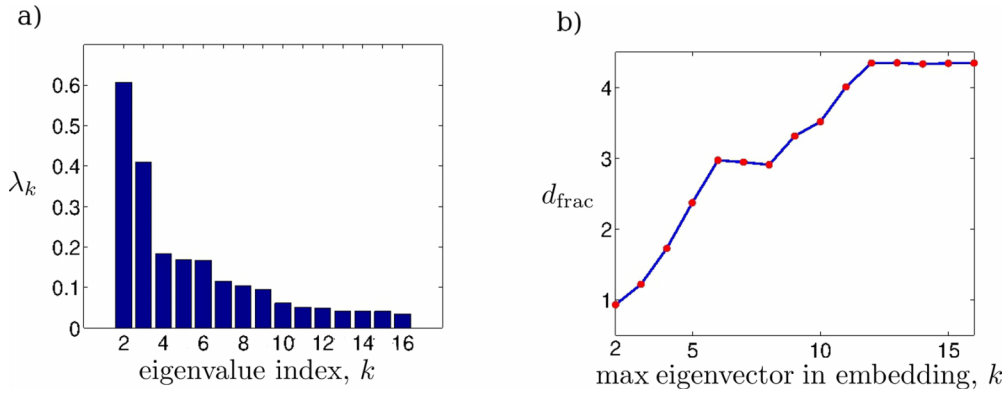

FIG. 1. Determination of the effective system dimensionality for the composite n-eicosane system. (a) The eigenvalue spectrum of the composite diffusion map in which the trivial $\lambda_1 = 1$ eigenvalue has been omitted for viewing clarity. Application of the L-method to the top 15 non-trivial eigenvalues identifies a spectral gap after λ_4, suggesting an effective dimensionality of the intrinsic manifold of $k = 3$. (b) The fractal dimension, d_{frac}, of the intrinsic manifold as a function of the number of eigenvectors incorporated into the diffusion map embedding plateaus at $d_{frac} \approx 4.3$, suggesting that the intrinsic manifold is contained within a ($k = 5$)-dimensional space.

space and suggesting that we construct ($k = 5$)-dimensional diffusion map embeddings. To be conservative, we adopt the larger of the k values identified by these two approaches and construct the $k = 5$ dimensional diffusion map embedding $\vec{x}_i \rightarrow (\vec{\phi}_2(i), \vec{\phi}_3(i), \vec{\phi}_4(i), \vec{\phi}_5(i), \vec{\phi}_6(i))$.

The eigenvectors spanning the diffusion map embedding are orthogonal by construction but can exhibit functional dependencies wherein multiple eigenvectors describe the same collective dynamical motion.[17] As we have previously observed, this is analogous to multivariate Fourier series wherein $\sin(x)$ and $\sin(2x)$ are orthogonal Fourier components identifiable as different harmonics describing the same direction in Cartesian space.[17] Analysis of the five non-trivial eigenvectors revealed that $\vec{\phi}_2$ and $\vec{\phi}_3$ were functionally dependent, defining an effectively one-dimensional manifold in their two-dimensional space (Fig. S1(a) in the supplementary material[38]). Following Ref. 17, we eliminate this redundancy by representing $\vec{\phi}_2$ and $\vec{\phi}_3$ by the arclength, $\vec{\alpha}_{23}$, of their one-dimensional manifold fitted by two piecewise continuous 2nd order polynomials. Further analysis revealed a second redundancy between $\vec{\phi}_4$ and $\vec{\alpha}_{23}$ that we eliminated by representing these two variables by their arclength, $\vec{\alpha}_{234}$, fitted by two piecewise continuous 3rd order polynomials (Fig. S1(b)[38]). We did not detect any further functional dependencies in the top five eigenvectors. Ultimately, this procedure allowed us to synthesize the three-dimensional diffusion map embedding $\vec{x}_i \rightarrow (\vec{\alpha}_{234}(i), \vec{\phi}_5(i), \vec{\phi}_6(i))$.

In Fig. 2, we present the composite diffusion map embedding of all 90 006 snapshots harvested from the six systems. Consistent with our previous work,[17] the three dimensions of the intrinsic manifold of n-eicosane ($C_{20}H_{42}$) are well-correlated with the principal moments of the gyration tensor of the hydrocarbon chain, $\{g_1, g_2, g_3\}$, physically interpretable as measures of the length of an instantaneous chain configuration along its longest, next longest, and shortest axes.[87] These moments present a useful interpretive "bridge" variable with which to help understand the collective configurational motions of the chain across the manifold. We project onto the embedding representative chain configurations visualized using Visual Molecular Dynamics (VMD).[88]

The first moment of the gyration tensor, g_1, is well correlated with α_{234}, indicating that this dimension of the intrinsic manifold corresponds to global chain collapse from an extended all-*trans* configuration to a hairpin or helix (Figs. 2(a) and 2(b)). The second moment, g_2, is correlated with ϕ_6, corresponding to the position of a bend in the chain (Figs. 2(c) and 2(d)). Finally, g_3 is associated with ϕ_5, describing the deviation of the chain configuration from planarity toward helical configurations (Figs. 2(e) and 2(f)). The mirror symmetries apparent in the manifold in ϕ_5 (bend near the head and tail) and ϕ_6 (right- and left-handed helices) are a consequence of the inherent head-tail and mirror indistinguishability of the n-eicosane chain. The emergence of these symmetries in the intrinsic manifold provides a good internal check that our simulations are sufficiently long to fully explore the thermally accessible configurational space. We have verified that differences in the free energy surface across the symmetry planes in both ϕ_5 and ϕ_6 are on the order of thermal noise, differing by $\sim k_B T$ (Fig. S2[38]).

In Figs. 2(g) and 2(h), we color the snapshots according to the simulation from which they were drawn—neat phase, aqueous solvent, or ideal gas at 298 K or 323 K. The partial overlap in the configurational ensembles drawn from these different environments enables the synthesis of a single unified diffusion map embedding, but it is clear that the imposition of different solvent conditions and temperatures restricts the thermally accessible configurational space to different regions of the intrinsic manifold. In Figs. S3 and S4, we present scatterplots of the embedding of each individual simulation within the unified intrinsic manifold and in Fig. S5 three-dimensional views of the composite scatterplot.[38]

2. Free energy surfaces

We present in Fig. 3 free energy surfaces for each n-eicosane system over the unified intrinsic manifold. In Fig. S6, we present an alternative viewing angle that better illustrates the two "wings" containing the helical chain configurations and in Fig. S2 the one-dimensional projections of the free energy into α_{234}, ϕ_5, and ϕ_6 with associated error bars.[38]

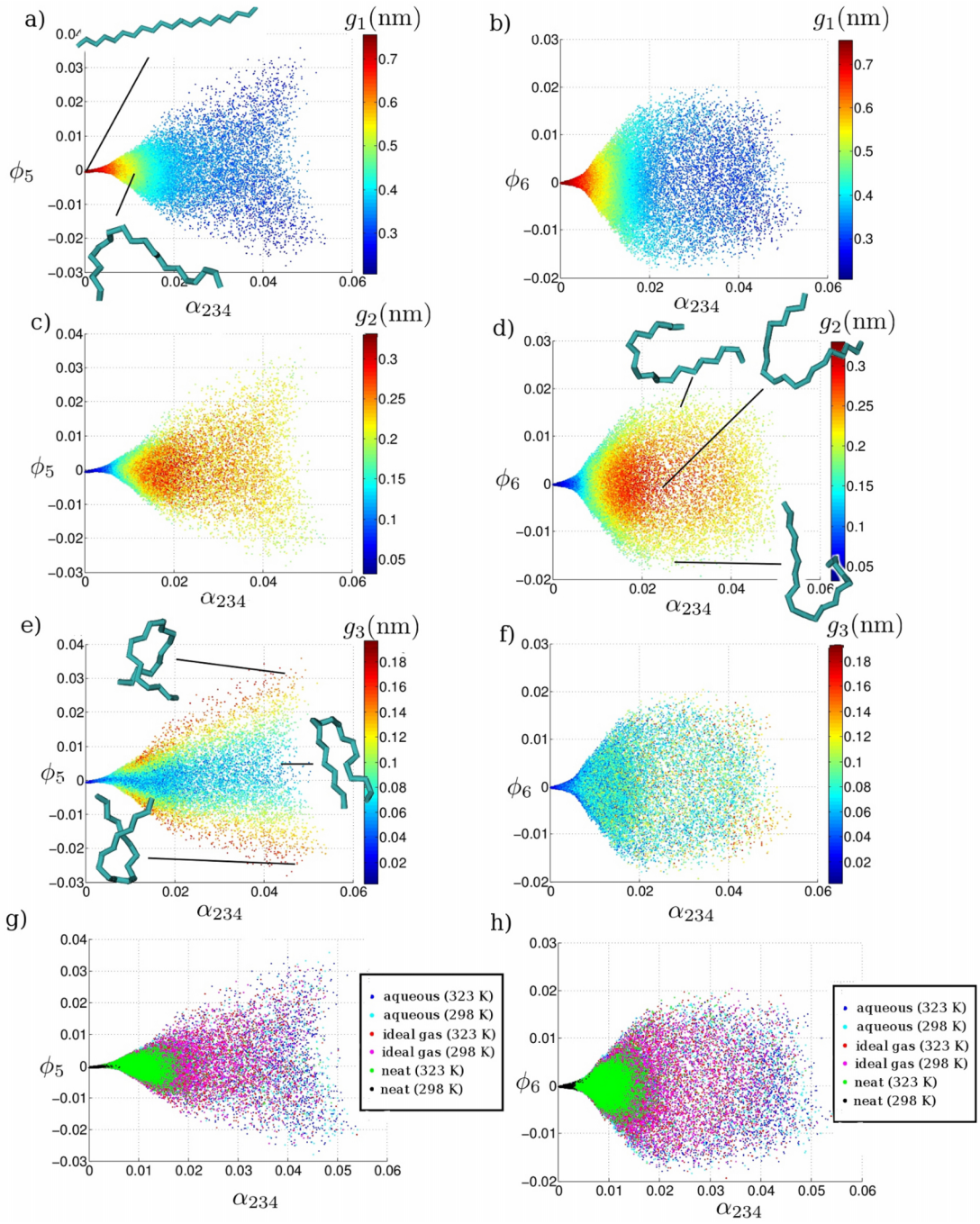

FIG. 2. Diffusion map embeddings into $[\alpha_{234}, \phi_5, \phi_6]$ of the 90 006 snapshots harvested from the six n-eicosane systems: the aqueous phase, the ideal gas phase, and the neat phase at 323 K and 298 K. Panels (a), (c), (e), and (g) display the two-dimensional elevation showing eigenvector ϕ_5 versus the arclength α_{234} parametrizing the eigenvectors ϕ_2, ϕ_3, and ϕ_4, whereas panels (b), (d), (f), and (g) show the ϕ_6 versus α_{234} elevation. Snapshots in panels (a) and (b) are colored by the value of the first moment of the gyration tensor, g_1, (c) and (d) by the second moment, g_2, (e) and (f) by the third moment, g_3, and (g) and (h) by the simulation from which the point was harvested. Representative snapshots are selected for visualization to illustrate the progression of molecular configurations along the manifold.

In effect, we have constructed low-dimensional free energy landscapes for each single simulation in a common basis set derived from multiple simulations. From these landscapes, we can resolve the regions of the manifold populated under different environmental conditions and quantify the relative stability of various chain configurations. We note that the regions of the free energy landscape too high in free energy to be explored by our unbiased molecular dynamics simulations could be sampled using accelerated sampling techniques such as umbrella sampling[89] or metadynamics[90] directly in the low-dimensional collective variables spanning the intrinsic manifold.[91]

As detailed in Refs. 14 and 17, an attractive feature of the diffusion map is that under the assumptions that the system can be well-described as a diffusion process on short time scales, and that the pairwise similarity metric is a good measure of short-time molecular motions, then the intrinsic manifold discovered by the diffusion map is dynamically

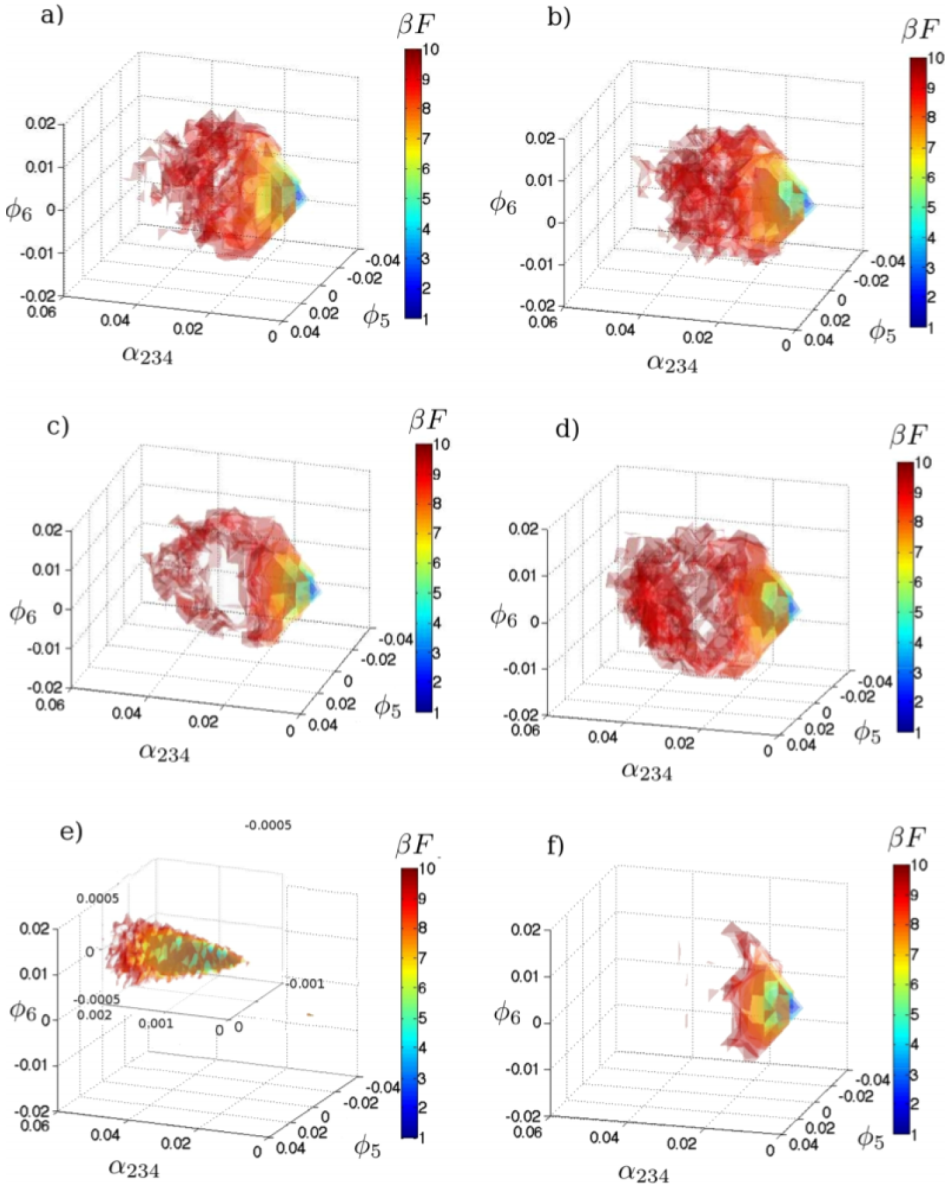

FIG. 3. Free energy surfaces for each of the six n-eicosane systems constructed over the intrinsic manifold in $[\alpha_{234}, \phi_5, \phi_6]$ presented in Fig. 2. Determined up to an additive constant, the free energy of the most populated bin for each system was arbitrarily shifted to $\beta F = 1.5$. Free energy isosurfaces in each panel are plotted in increments of $0.8\, k_B T$ (i.e., $\beta F = 1.5, 2.3, 3.1, \ldots$) for n-eicosane in (a) ideal gas at 298 K, (b) ideal gas at 323 K, (c) aqueous solution at 298 K, (d) aqueous solution at 323 K, (e) neat phase at 298 K, and (f) neat phase at 323 K. The viewing angle exposes the "donut" structure within the free energy surface in the aqueous phase, an alternative viewing angle that better exposes the "wings" containing helical chain configurations in the aqueous phase is presented in Fig. S6.[38] Panel (e) includes an inset displaying a zoomed in view of the free-energy surface of the neat phase at 298 K.

interpretable in the following sense: (1) Euclidean distances over the intrinsic manifold correspond to *diffusion distances* in the high-dimensional configurational space, measuring the time required for one configuration to dynamically evolve into another,[44,92] and (2) the eigenvectors spanning the intrinsic manifold are identifiable as the slow collective configurational motions governing the long-time evolution of the system to which the remaining fast degrees of freedom are effectively slaved.[44] Accordingly, pathways traced over the intrinsic manifold describe the evolution of the system in its slowest dynamical modes, revealing the microscopic molecular mechanisms underpinning the system dynamics. The RMSD distance employed in the present work is expected to satisfy both of these assumptions, conveying this dynamic interpretability to the free energy surfaces in Fig. 3.[14,17]

We now proceed to make some remarks and analysis of the n-eicosane free energy surfaces in Fig. 3, deferring a more detailed and comprehensive discussion to the supplementary material.[38] In the ideal gas phase, the chain populates almost the entire volume of the intrinsic manifold at both 298 K and 323 K (Figs. 3(a) and 3(b)). Upon immersion in water at 298 K (Fig. 3(c)), the aqueous solvent causes the chain to populate only a subset of the manifold, producing free energy profiles very similar to those we have previously reported for n-hexadecane and n-tetracosane.[17] The topography of these landscapes is attributable to the hydrophobic effect, which stabilizes compact chain configurations and induces a high wetting/dewetting free energy barrier responsible for the "donut" topology that we have previously reported.[17,73] The influence of hydrophobicity is mitigated at higher temperature,[93] reducing the size of the donut hole at 323 K (Fig. 3(d)) and making the free energy landscape more similar to that in the ideal gas. In the neat phase crystal at 298 K (Fig. 3(e)), chains exist almost exclusively in the all-*trans* conformation, populating only a tiny fraction of the intrinsic manifold. In the neat liquid at 323 K (Fig. 3(f)), the chain explores a somewhat greater configurational diversity of configurations, but collapsed configurations remain strongly disfavored, and in the absence of the hydrophobic effect, the donut hole is absent.

3. Visualization of dynamical modes

The chemical simplicity of *n*-eicosane chains allowed us to correlate the three dimensions of the diffusion map embedding with the principal moments of the chain gyration tensor, rendering physical interpretation of the intrinsic manifold relatively straightforward. Specifically, α_{234} is correlated with global chain collapse, ϕ_5 with the position of a bend in the chain, and ϕ_6 with chain helicity.[17] To more systematically interpret the slow collective modes identified by the diffusion map, we employed a procedure to expose in the original configuration space the simulation snapshots contributing most to each mode (Sec. II E). This technique is particularly valuable for more complex systems where physical bridge variables may not be readily apparent.

In Fig. 4, we visualize the dynamical motions associated with α_{234}, ϕ_5, and ϕ_6 for each of the three systems—ideal gas, aqueous phase, and neat phase—at 298 K. In the ideal gas and aqueous phase, the chain configurations varying most strongly with α_{234} reveal that motions over the intrinsic manifold in this direction correspond to the development of a symmetric bend in the middle of a planar chain (Figs. 4(a) and 4(d)). In the neat phase crystal, only elongated configurations are extracted (Fig. 4(g)), consistent with the existence of crystalline *n*-eicosane chains in their all-*trans* state. Taken together, these visualizations illustrate that the range of global symmetric folding motions sampled by the chain in the aqueous phase and ideal gas is comparable, whereas it is massively restricted in the neat phase. The motions in ϕ_5 in the ideal gas and aqueous phase show this collective mode to be associated with helical, out-of-plane twisting (Figs. 4(b) and 4(e)). This is precisely the motion executed by the chain as it moves into the high and low ϕ_5 wings of the intrinsic manifold containing the helical coils (Figs. S2(b) and S2(e)[38]). The neat phase trajectory contains very few configurations that contribute with large positive or negative q_l weights (Fig. 4(h)), indicating the absence of out of plane twisting motions for chains in the crystal (cf. Fig. S2(h)[38]). Finally, visualizations of ϕ_6 in the ideal gas show motions in this mode are associated with migration of the position of the bend in the chain (Fig. 4(c)). These motions correspond to the configurational evolution of the chain as it transits vertically over the intrinsic manifold in Fig. 3(a). The analogous visualization in the aqueous phase (Fig. 4(f)) contains more strongly kinked configurations that vary more strongly with ϕ_6 as evinced by their larger q_l weights. Hydrophobic destabilization of the symmetrically kinked chain configurations residing in the donut hole of the free energy landscape (cf. Sec. III A 2) forces the chain to adopt more severely asymmetrically kinked configurations residing in the exterior of the donut (Fig. 3(c)). The neat phase trajectory contains a small number of asymmetrically kinked configurations that resonate with the ϕ_6 mode (Fig. 4(i)), indicating that even in the crystal phase the thermal energy of the chains enables the development of small asymmetrically displaced kinks in the chain backbone.

Analogous visualizations for the systems at 323 K show the dynamical motions associated with each collective variable in the aqueous phase and ideal gas to be very similar to those at 298 K (Figs. S7(a)–S7(f)), as may have been anticipated by the relatively small impact of the elevated temperature upon the free energy landscapes of these two systems (Figs. S2(a)–S2(f)).[38] A close inspection of the α_{234} mode in the aqueous phase (Fig. S7(d)[38]) reveals that the population skews slightly towards more heavily folded configurations compared to that at 298 K (Fig. 4(d)), consistent with the elevated population of the donut hole of the free energy landscape at elevated temperature. The phase change from crystal at 298 K to liquid at 323 K experienced by

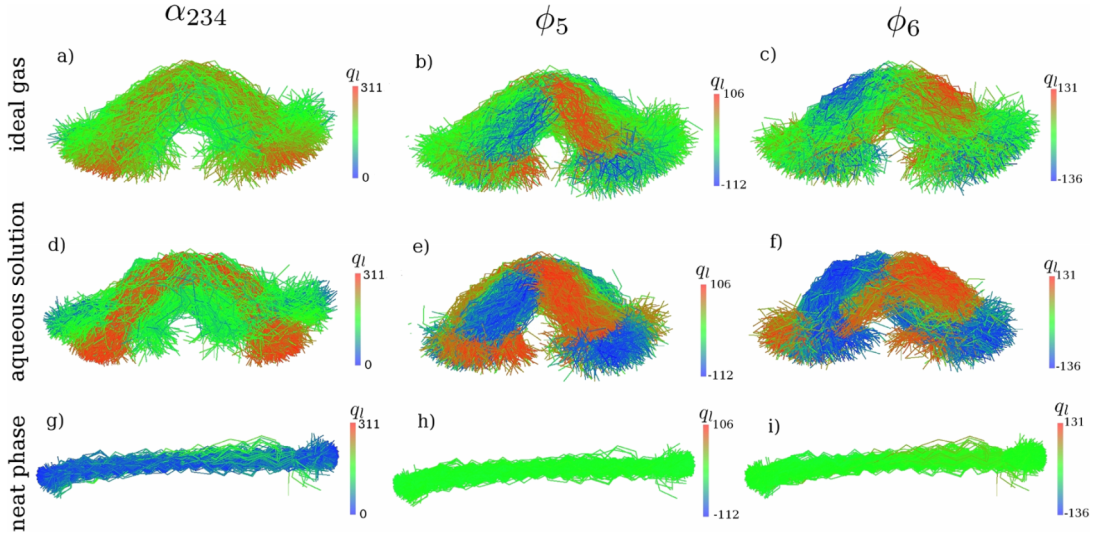

FIG. 4. Visualization of dynamical modes associated with the collective order parameters α_{234}, ϕ_5, and ϕ_6 spanning the unified intrinsic manifold for the *n*-eicosane systems at 298 K. Modes are visualized using the procedure detailed in Sec. II E to identify for each system the snapshots from the molecular simulation trajectory corresponding to the 500 most positive elements and 500 most negative elements of the $\vec{q}_l$ vector. These configurations vary most strongly with each of the collective order parameters and therefore represent the extremes of the configurational motions in the high-dimensional ambient space associated with each mode. For clarity of viewing, we isolate only the 1000 most extreme configurations and present the snapshots as line drawings colored according to the associated $\vec{q}_l$ weight. The top row—panels (a)–(c)—corresponds to the ideal gas, the middle row—panels (d)–(f)—to the aqueous phase, and the bottom row—panels (g)–(i)—to the neat phase. The left column—panels (a), (d), (g)—corresponds to the collective order parameter α_{234}, the middle column—panels (b), (e), (h)—to ϕ_5, and the right column—panels (c), (f), (i)—to ϕ_6.

the neat system is associated with a massive increase in the populated volume of the intrinsic manifold (Figs. 3(e) and 3(f)) and is manifested in these visualizations by a concomitant increase in the diversity and q_l weights of the extracted chain configurations (Figs. S7(g)–S7(i)[38]).

In sum, the visualizations we have presented here provide a configurational interpretation of the dynamical motions associated with α_{234}, ϕ_5, and ϕ_6 in the high dimensional original space. In previous work, we gained intuition into these modes by correlating them with physical "bridge" variables. Compared to this approach, the visualization procedure provides a more richly detailed interpretation, and furnishes a dynamical complement to the free energy landscapes by resolving the important configurational motions of the chain. Importantly, this visualization procedure requires neither painstaking inspection of representative chain configurations nor tedious correlation with a pool of candidate "bridge" variables. Further, it does not presuppose the existence of a correspondence of the nonlinear collective variables to a simple physical order parameter, since the visualization and interpretation are conducted within the high dimensional ambient space in which the system resides.

B. Thermodynamics and dynamical motions of polyglutamate-derivative peptide decamers as a function of side chain length

Having demonstrated our approach for the relatively simple and well-understood *n*-eicosane systems, we now employ composite diffusion maps to develop new insight into the more complex and poorly understood role of side chain chemistry on peptide structure and dynamics.[8] The α-helix is one of the most common and ubiquitous secondary structure elements in proteins[94] and has been exploited in numerous applications, including biologically active cell-penetrating peptides capable of delivering molecular cargoes,[95] non-toxic amyloid fibrils,[96] molecular sensors of membrane curvature,[97] protein-based hydrogels,[98] *de novo* designed peptides with specific enzymatic activity,[99] and in mediating protein-protein interactions.[100] Recently, Lu *et al.* engineered ultra-stable water-soluble α-helical peptides composed of non-natural amino acid residues with elongated hydrophobic side chains bearing a terminal positively charged amine group.[8] The terminal charges on the side chains maintained water solubility, while the stability of the α-helical conformation relative to the random coil was correlated with the length of the hydrophobic side chains. This trend was rationalized in terms of reduced electrostatic repulsion between the terminal charges, but the impact of side chain length upon peptide structure, and the relative interplay of hydrophobic and electrostatic interactions, remains poorly understood at a molecular level. Inspired by this work, it is the goal of the present study to employ molecular simulations and composite diffusion maps to establish a quantitative bridge between side chain length and peptide structure, unravel the molecular details of the observed trend in helical stability, and develop insight into the stabilization mechanism. It is our anticipation that these findings will develop new understanding of peptide stabilization by non-natural amino acids and help to guide peptide engineering and design.

1. Composite diffusion maps and visualization of dynamical modes

We constructed a composite diffusion map comprising the $6 \times 14\,001 = 84\,006$ snapshots of the carbon backbones of the five decameric homopeptides belonging to the family of non-natural polyglutamate derivatives studied by Lu *et al.*,[8] poly(γ-(3-aminoethyl)-L-glutamate) ($PAGG_{10}$), poly(γ-(3-aminopropyl)-L-glutamate) ($PAPG_{10}$), poly(γ-(4-aminobutyl)-L-glutamate) ($PABG_{10}$), poly(γ-(5-aminopentanyl)-L-glutamate) ($PATG_{10}$), and poly(γ-(6-aminohexyl)-L-glutamate) ($PAHG_{10}$), plus polylysine (PL_{10}), as a control peptide known to exist as a random coil with no helical content. The chemical structures of these peptides are illustrated in Fig. 5. We selected the bandwidth of the Gaussian kernel as $\epsilon = 4.98 \times 10^{-2}$ using the approach in Ref. 46. Application of the L-method[84] to the diffusion map eigenvalue spectrum illustrated in Fig. 6(a) identified a spectral gap after λ_4, implying that we construct ($k = 3$)-dimensional diffusion map embeddings. As illustrated in Fig. 6(b), the fractal dimension saturates at a plateau dimension[85] of $d_{frac} \approx 5$, indicating that the important dynamical motions of the system are contained within an approximately five-dimensional space. To be conservative, we construct the $k = 5$ dimensional diffusion map embedding $\vec{x}_i \rightarrow (\vec{\phi}_2(i), \vec{\phi}_3(i), \vec{\phi}_4(i), \vec{\phi}_5(i), \vec{\phi}_6(i))$. We observed no functional dependencies within these top five eigenvectors. We constructed diffusion maps by considering only the heavy backbone atoms of the peptides that are identical between all six systems, thereby implicitly treating the side chain chemistry as an external variable that influences the configurational ensemble adopted by the peptide backbone. We quantify the impact of the side chain chemistry within the low-dimensional projections synthesized by the composite diffusion map.

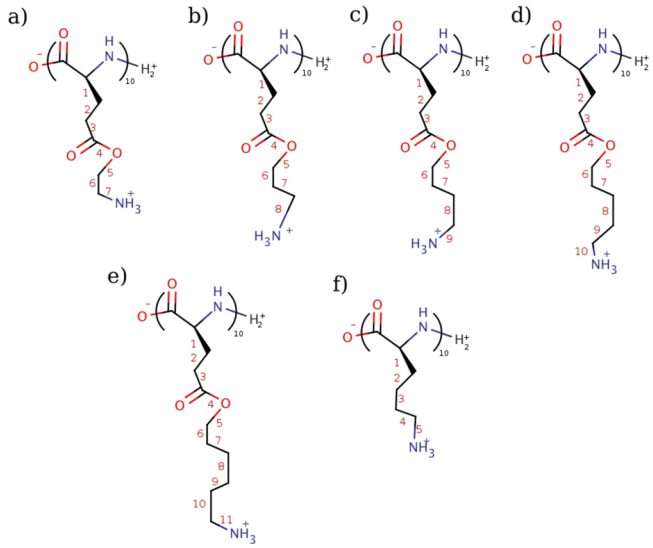

FIG. 5. Chemical structures of decamer homopeptides studied herein. (a) poly(γ-(3-aminoethyl)-L-glutamate) ($PAGG_{10}$), (b) poly(γ-(3-aminopropyl)-L-glutamate) ($PAPG_{10}$), (c) poly(γ-(4-aminobutyl)-L-glutamate) ($PABG_{10}$), (d) poly(γ-(5-aminopentanyl)-L-glutamate) ($PATG_{10}$), (e) poly (γ-(6-aminohexyl)-L-glutamate) ($PAHG_{10}$), and (f) polylysine (PL_{10}). Structures were produced using Marvin 14.7.7.0 (ChemAxon, 2014) (http://www.chemaxon.com).

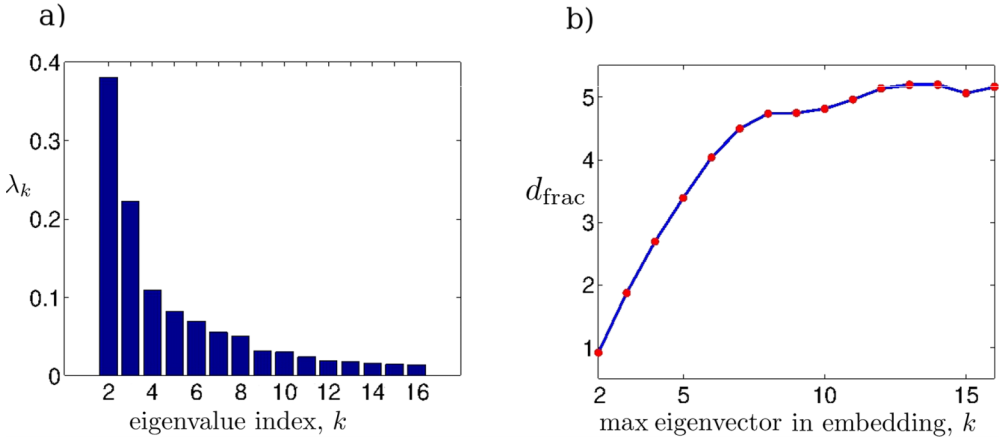

FIG. 6. Determination of the effective system dimensionality for the composite peptide system. (a) The eigenvalue spectrum of the composite diffusion map in which the trivial $\lambda_1 = 1$ eigenvalue has been omitted for viewing clarity. Application of the L-method to the top 15 non-trivial eigenvalues identifies a spectral gap after λ_4, suggesting an effective dimensionality of the intrinsic manifold of $k = 3$. (b) The fractal dimension, d_{frac}, of the intrinsic manifold as a function of the number of eigenvectors incorporated into the diffusion map embedding plateaus at $d_{frac} \approx 5$, suggesting that the intrinsic manifold is contained within a ($k = 5$)-dimensional space.

We present in Fig. 7 the composite diffusion map embedding of all 84 006 snapshots drawn from the six peptide systems into the leading three leading eigenvectors $\{\vec{\phi}_2, \vec{\phi}_3, \vec{\phi}_4\}$ together with visualizations of the peptide backbone at selected representative snapshots. Projections into $\{\vec{\phi}_2, \vec{\phi}_5, \vec{\phi}_6\}$ are presented in Fig. S8.[38] Although the population over the composite intrinsic manifold is bimodal, the diffusion map eigenvalue spectrum possesses a single unit eigenvalue (Fig. 6(a)) indicating that the Markov matrix constructed over the data defines a single unified diffusion process over the data. This assures that the eigenvectors spanning the manifold define a single, well-defined basis in which to project the simulation snapshots.

As illustrated in Fig. 7(f), the larger cloud contains exclusively PL_{10} and $PAGG_{10}$ molecules, whereas the smaller cloud comprises the peptides $PAPG_{10}$, $PABG_{10}$, $PATG_{10}$, and $PAHG_{10}$. Adopting as a useful bridge variable the root mean squared deviation of the backbone from an ideal α-helix, $RMSD_{helix}$, Fig. 7(a) reveals the smaller cloud to contain highly helical peptide configurations, with $RMSD_{helix} < 0.22$ nm from the idealized helix, whereas the larger cloud contains random coils with $RMSD_{helix} > 0.24$ nm. The composite diffusion map has clearly partitioned the six peptides into two classes: random coils (PL_{10} and $PAGG_{10}$) and α-helices ($PAPG_{10}$, $PABG_{10}$, $PATG_{10}$, and $PAHG_{10}$). That there is no overlap in the embedding of $PAGG_{10}$ with those of the peptides possessing longer side chains indicates that there are no peptide backbone conformations shared between the former and the latter ensembles. The discrete transition from a random coil state for $PAGG_{10}$ possessing a side chain of length 7 covalent bonds (Fig. 5(a)) to an α-helical state for $PAPG_{10}$ possessing a side chain of length 8 covalent bonds (Fig. 5(b)) reveals the existence of a critical side chain length beyond which the peptide folds into an α-helix.

The diffusion map identified five collective modes governing the long-time dynamical motions of the peptide family. The leading eigenvector, ϕ_2, is the primary variable discriminating between the two clouds, with the smaller, α-helical cloud existing at values of $\phi_2 < 3.9 \times 10^{-5}$ and the larger, random coil cloud at $\phi_2 > 8.9 \times 10^{-4}$. The collective dynamical motions associated with this variable govern the existence of the peptide backbone in one or other of these two conformational states, and—as the leading eigenvector in the embedding—proceed on the slowest time scales. As illustrated in Fig. 7(c), the first moment of the gyration tensor, g_1, is strongly correlated with ϕ_2 ($\rho_{\mathrm{Pearson}} = 0.81$, $p < 0.001$ (two-tailed Student's t-test)), with shorter lengths of the peptide along its longest axis corresponding to more tightly folded α-helical configurations. Deeper insight into the character of this dynamical mode is furnished by the visualization procedure detailed in Sec. II E. For the helical peptides, $PAPG_{10}$, $PABG_{10}$, $PATG_{10}$, and $PAHG_{10}$, the backbone configurations that resonate most strongly with ϕ_2 bound a narrow, structurally homogeneous range (Figs. 8(c)–8(f)). Consistent with the narrow span of the small α-helical cloud in ϕ_2, this indicates that there is very little dynamical motion of the helical peptides at equilibrium along this collective variable. In the case of the random coil peptides, PL_{10} and $PAGG_{10}$, there is a much greater diversity in the configurational ensemble that resonates with ϕ_2 (Figs. 8(a) and 8(b)). The dynamical motions of $PAGG_{10}$ correspond to the collapse from an extended, asymmetrically kinked to a collapsed, symmetrically kinked random coil. PL_{10}, on the other hand, executes motions taking it from extended configurations to those possessing some degree of backbone helicity, reminiscent of the early stages of folding into a helical conformation. The degree of helicity, however, remains weak with $RMSD_{helix} > 0.24$ nm and is insufficient to bridge the gap separating the larger, random coil cloud to the smaller, α-helical cloud.

As illustrated in Fig. 7(b), ϕ_4 is quite strongly correlated with the effective helical radius, r_{helix}, computed as the mean radial distance of the two-dimensional projection of the backbone C_α atom coordinates along the axis of the helix ($\rho_{\mathrm{Pearson}} = 0.64$, $p < 0.001$). For all peptides considered, our visualization procedure showed this collective mode to be associated with tightening of helix-like configurations to possess a smaller effective helical radius (Figs. 8(m)–8(r)).

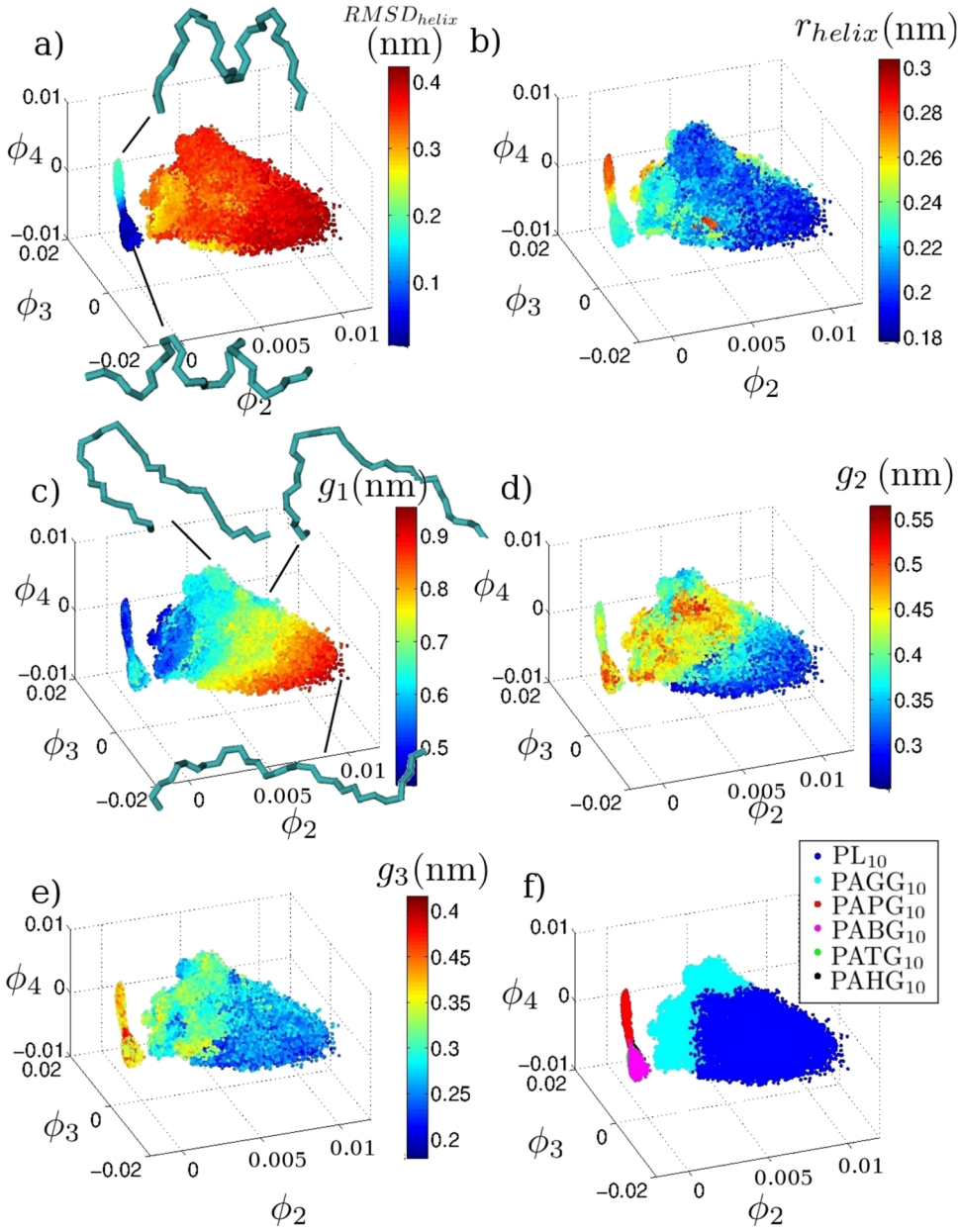

FIG. 7. Diffusion map embeddings into $[\phi_2, \phi_3, \phi_4]$ of the 84 006 snapshots harvested from the six peptides: PL_{10}, $PAGG_{10}$, $PAPG_{10}$, $PABG_{10}$, $PATG_{10}$, and $PAHG_{10}$. Snapshots are colored by (a) the root mean squared deviation of the backbone from an ideal α-helix, $RMSD_{helix}$, (b) the effective helical radius, r_{helix}, computed as the mean radial distance of the two-dimensional projection of the backbone C_α atom coordinates along the axis of the helix ($r_{helix}^{ideal} = 0.23$ nm), (c) the first moment of the gyration tensor, g_1, computed over the heavy backbone atoms of the peptide, (d) the second moment of the gyration tensor, g_2, (e) the third moment of the gyration tensor, g_3, and (f) by the peptide identity. Representative snapshots are selected for visualization to illustrate the progression of molecular configurations along the manifold.

We were unable to identify any physical bridge variables strongly correlated with ϕ_3, ϕ_5, or ϕ_6, but our visualization procedure clearly illuminates the collective backbone motions associated with these modes. As illustrated in Fig. 7(d), ϕ_3 is weakly correlated with g_2 ($\rho_{\text{Pearson}} = 0.37$, $p < 0.001$). Visualization of the dynamical modes reveals very little configurational diversity along the ϕ_3 axis for the peptides residing in the α-helical cloud (Figs. 8(i)–8(l)), consistent with the small width of the cloud in this dimension. In the case of the random coil peptides, PL_{10} and $PAGG_{10}$, this mode appears to correspond to the adoption of kinked β-hairpin-like configurations (Figs. 8(g) and 8(h)). Visualization of ϕ_5 and ϕ_6 reveals that the helical peptides again show very little configurational diversity in these modes (Figs. S9(c)–S9(f) and S9(i)–S9(l)), as may have been anticipated from the compact nature of the helical point cloud in these dimensions (Fig. S8).[38] In the case of the two random coils, however, our visualizations indicate that motions in ϕ_5 and ϕ_6 correspond to planar folding (Figs. S9(a) and S9(b)) and helical twisting (Figs. S9(g) and S9(h)) of the peptide backbone.[38]

2. Free energy surfaces

We present in Fig. 9 the free energy surfaces in $[\vec{\phi}_2, \vec{\phi}_3, \vec{\phi}_4]$ for each peptide over the unified intrinsic manifold. The free energy landscapes are dynamically interpretable in the sense that pathways over the free energy landscape correspond to the evolution of the system in its slow collective modes. Analogous free energy surfaces in $[\vec{\phi}_2, \vec{\phi}_5, \vec{\phi}_6]$ are presented in Fig. S10, and one-dimensional projections into $\{\vec{\phi}_i\}_{i=2}^{6}$ with associated error bars in Fig. S11.[38]

The single molecule free energy landscapes for the two random coils—PL_{10} and $PAGG_{10}$—span a large volume of the intrinsic manifold and contain a diverse ensemble of

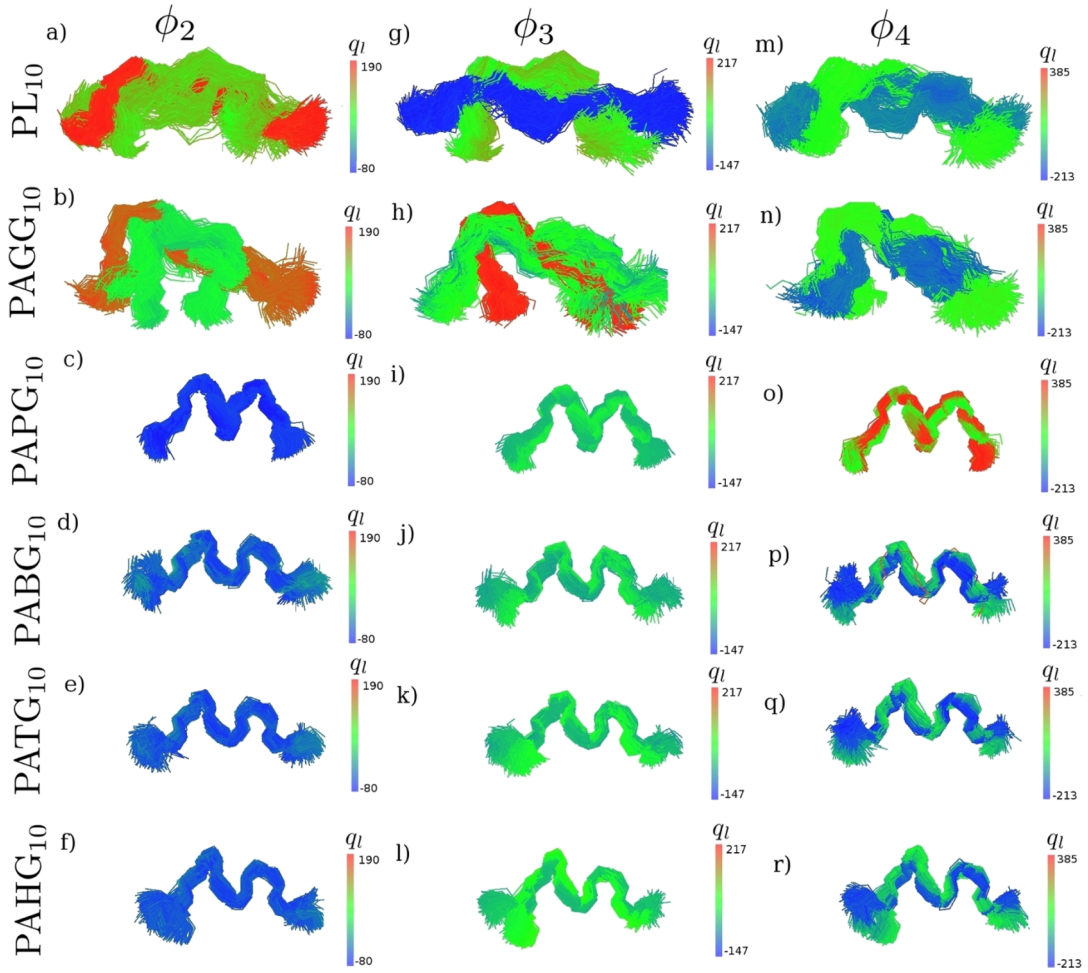

FIG. 8. Visualization of dynamical modes associated with the collective order parameters ϕ_2, ϕ_3, and ϕ_4 spanning the unified intrinsic manifold for the polypeptide systems. Modes are visualized in the manner described in the caption to Fig. 4. The first column—panels (a)–(f)—corresponds to the collective order parameter ϕ_2, the middle column—panels (g)–(l)—to ϕ_3, and the third column—panels (m)–(r)—to ϕ_4. The first row—panels (a), (g), (m)—corresponds to PL_{10}, the second row—panels (b), (h), (n)—to $PAGG_{10}$, the third row—panels (c), (i), (o)—to $PAPG_{10}$, the fourth row—panels (d), (j), (p)—to $PABG_{10}$, the fifth row—panels (e), (k), (q)—to $PATG_{10}$, and the final row—panels (f), (l), (r) to $PAHG_{10}$.

configurations lying close in free energy (Figs. 9(a), 9(b), S10(a), and S10(b)[38]). The free energy structure within these landscapes indicates that the peptides do not exist as truly "random" coils, but preferentially adopt particular weakly structured conformations.[101] PL_{10} exhibits a single, large free energy well in this projection, with the global free energy minimum at ($\phi_2 = 0.008, \phi_3 = -0.004, \phi_4 = 0.002$) containing extended conformations. In contrast, $PAGG_{10}$ possesses a deep global free energy minimum at ($\phi_2 = 0.002, \phi_3 = 0.005$, $\phi_4 = -0.003$), containing hairpin-like configurations that are bent out-of-plane, and a metastable basin at ($\phi_2 = 0.006$, $\phi_3 = 0.01, \phi_4 = 0.005$) corresponding to planar hairpins. A comparison of the structure of these two landscapes suggests that the shorter side chains of PL_{10} (Fig. 5(f)) cause the peptide backbone to favor more extended configurations, whereas the longer side chains of $PAGG_{10}$ (Fig. 5(a)) are sufficiently long to stabilize β-hairpin-like collapsed structures, but not α-helices. The free energy landscapes in Figs. S10(a) and S10(b) resolve the weakly structured configurations in ϕ_5 and ϕ_6.[38]

In contrast to the random coils, the free energy landscapes for the four helix forming peptides—$PAPG_{10}$, $PABG_{10}$, $PATG_{10}$, and $PAHG_{10}$—occupy only a very small fraction of the intrinsic manifold, existing as highly localized single-welled free energy funnels containing structurally homogeneous conformations (Figs. 9(c)–9(f) and S10(c)–S10(f)[38]). With increasing chain length (cf. Figs. 5(b)–5(e)), the free energy minimum progresses downwards in ϕ_4, corresponding to the adoption of increasingly ideal α-helical conformations (Figs. 7(a) and 7(b)). Of these four peptides, $PAPG_{10}$ possesses the shortest side chains, and the global minimum of its free energy landscape is centered on an α-helix with $RMSD_{helix}$ = 0.15 nm and r_{helix} = 0.26 nm. Its free energy landscape is somewhat more extended than those of the other three peptides, indicative of a greater conformational diversity and a less tightly folded and labile helix. With increasingly elongated side chains, $PABG_{10}$, $PATG_{10}$, and $PAHG_{10}$ possess more compact free energy landscapes centered on helices with $RMSD_{helix}$ = 0.019, 0.016, and 0.015 nm, respectively, and all of which possess r_{helix} = 0.23 nm. The very low $RMSD_{helix}$ values, together with the close approach of the effective helix radius to that of an ideal helix (r_{helix}^{ideal} = 0.23 nm), indicate that the side chain chemistry of these three peptides has caused the backbone to adopt very tight, nearly ideal helical conformations.

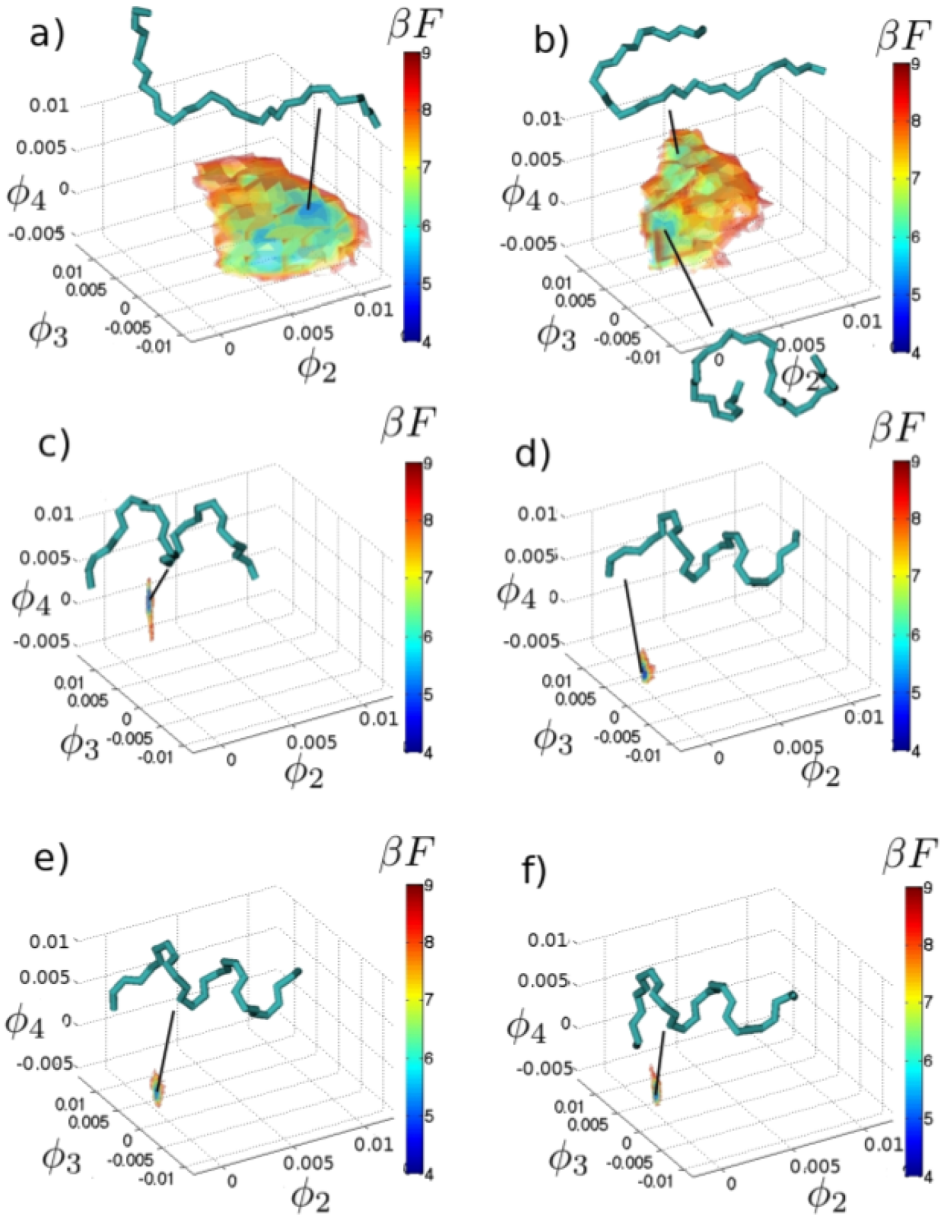

FIG. 9. Free energy surfaces for each of the six polypeptide systems constructed over the intrinsic manifold in $[\phi_2, \phi_3, \phi_4]$ (Fig. 7). Determined up to an additive constant, the free energy of the most populated bin for each system was arbitrarily shifted to $\beta F = 4.0$. Free energy isosurfaces in each panel are plotted in increments of 0.5 $k_B T$ (i.e., $\beta F = 4.0, 4.5, 5.0, \ldots$) for (a) PL_{10}, (b) $PAGG_{10}$, (c) $PAPG_{10}$, (d) $PABG_{10}$, (e) $PATG_{10}$, and (f) $PAHG_{10}$.

3. Mechanism of α-helix stabilization by polyglutamate-derivative side chains

In experimental studies of $PAPG_{57}$, $PATG_{50}$, and $PAHG_{57}$, Lu *et al.* found the stability of the α-helix relative to the random coil to be correlated with side chain length.[8] These researchers attributed this trend to a decrease in the electrostatic repulsion between the terminal charges, permitting the intrinsically hydrophobic side chains to condense around the peptide backbone and stabilize collapsed helical conformations. We test this hypothesis by first reporting trends in measures of helicity as a function of side chain length (Figs. 10(a)–10(d)) and then analyzing the structure of the peptide backbone, side chains, and solvent molecules to ascertain the molecular mechanism by which elongation of polyglutamate-derivative side chains favors α-helicity (Figs. 10(e)–10(g)).

In Fig. 10(a), we report the root mean squared deviation of the backbone from an ideal α-helix, $RMSD_{helix}$. PL_{10} and $PAGG_{10}$ exist as random coils, $PAPG_{10}$ is partially helical, and $PABG_{10}$, $PATG_{10}$, and $PAHG_{10}$ adopt essentially ideal helix configurations. The effective helical radius, r_{helix}, and helical twist, γ_{helix} presented in Figs. 10(b) and 10(c), show the same trend, with the radii and twists of $PABG_{10}$, $PATG_{10}$, and $PAHG_{10}$ close to those of an ideal α-helix, $r_{helix}^{ideal} = 0.23$ nm and $\gamma_{helix}^{ideal} = 100°$, while $PAPG_{10}$ forms a somewhat looser helix with $r_{helix} = 0.27$ nm and $\gamma_{helix} = 85°$. r_{helix} is defined as the radial distance of the two-dimensional projection of the backbone C_α coordinates along the helical axis, and γ_{helix} is the angle between residues in this projection. The application of these statistics to the PL_{10} and $PAGG_{10}$ random coils that do not exist as α-helices is, therefore, somewhat uninterpretable. The mean per residue molar ellipticity of the peptides at 222 nm, $[\theta]_{222\ nm}$, was calculated using DichroCalc to compute circular dichroism spectra from the peptide trajectories (http://comp.chem.nottingham.ac.uk/cgi-bin/dichrocalc/bin/getparams.cgi).[102] As illustrated in Fig. 10(d), our calculated values of $[\theta]_{222\ nm}$ are in qualitative agreement

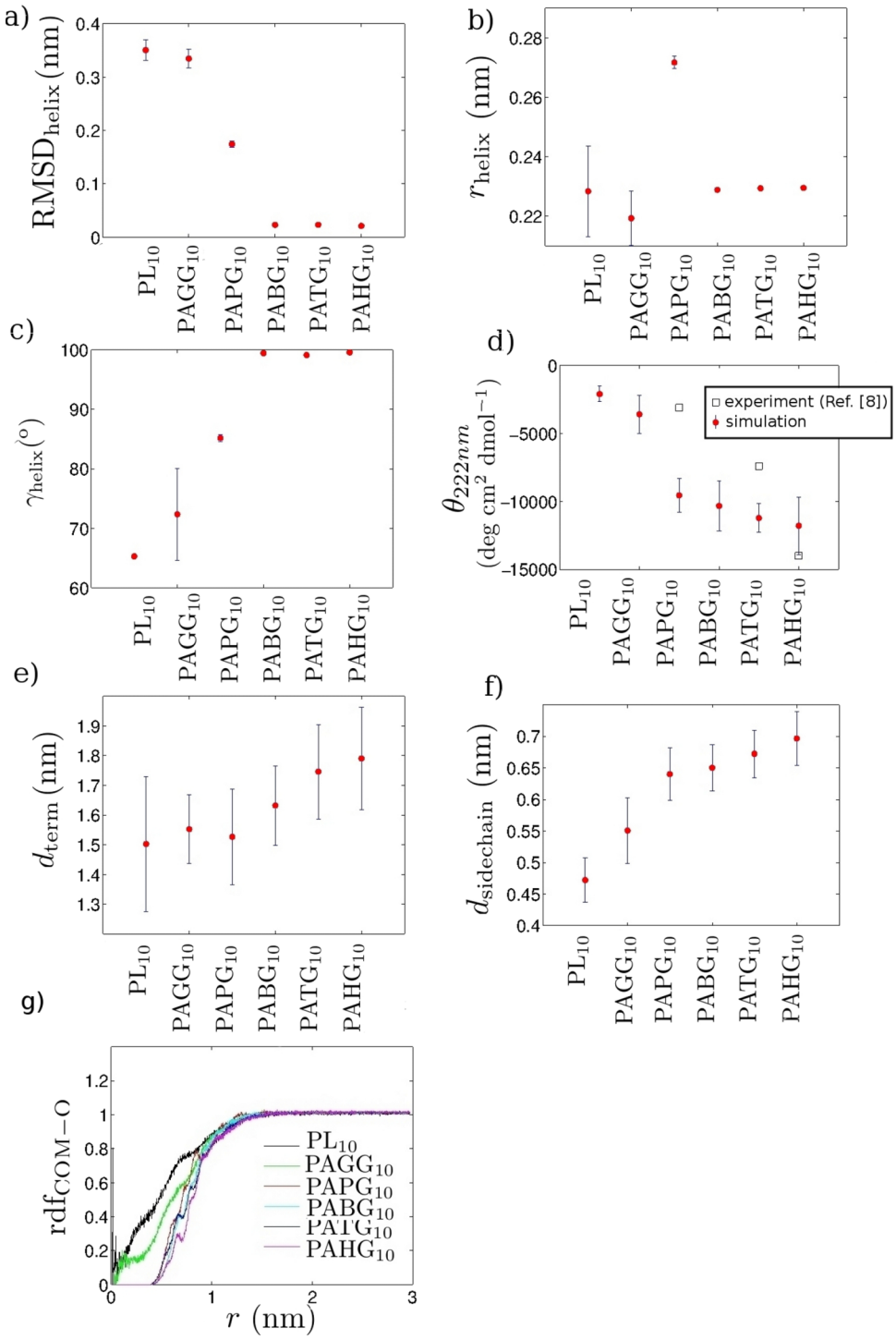

FIG. 10. Equilibrium structural metrics for the polyglutamate-derivative peptides and the polylysine control (cf. Fig. 5). Data points are plotted as red circles corresponding to the mean value computed over 28 ns molecular dynamics production runs, and error bars correspond to the associated standard error estimated by $n = 5$ block averaging. (a) The root mean squared deviation of the backbone from an ideal α-helix, $\mathrm{RMSD_{helix}}$. (b) The effective helical radius, $\mathrm{r_{helix}}$, computed as the mean radial distance of the two-dimensional projection of the backbone C_α atom coordinates along the axis of the helix ($\mathrm{r_{helix}^{ideal}} = 0.23$ nm). (c) The twist angle of the helix, γ_{helix}, defined as the angle between residues projected along the axis of the helix ($\gamma_{\mathrm{helix}}^{\mathrm{ideal}} = 100°$). (d) The mean per residue molar ellipticity at 222 nm, $[\theta]_{222\ \mathrm{nm}}$ (red circles) and the experimental measurements for $\mathrm{PAPG_{57}}$, $\mathrm{PATG_{50}}$, and $\mathrm{PAHG_{57}}$ reported by Lu *et al.* in Fig. 3(a) of Ref. 8 (black squares). (e) The mean distance between the charged termini of the side chains averaged over all $(10 \times 9)/2 = 45$ distinct pairs, d_{term}. (f) The mean distance of the side chain center of mass from the peptide center of mass averaged over all ten side chains, $d_{\mathrm{sidechain}}$. (g) The radial distribution function between the peptide backbone center of mass and the solvent O atoms, $\mathrm{rdf_{COM-O}}$. Non-zero values of $\mathrm{rdf_{COM-O}}$ (r = 0 nm) for the two random coils, $\mathrm{PL_{10}}$ and $\mathrm{PAGG_{10}}$, arise from the center of mass of loose peptide hairpins coinciding with solvent molecules between the peptide arms.

with those reported by Lu *et al.*,[8] showing a clear trend of increasing α-helicity (decreasing molar ellipticity) with side chain length. Part of the quantitative discrepancy is certainly attributable to the much longer peptides studied experimentally, 50 and 57-mers, compared to the 10-mers simulated herein.

The trends in $\mathrm{RMSD_{helix}}$, $\mathrm{r_{helix}}$, γ_{helix}, and $[\theta]_{222\ \mathrm{nm}}$ computed from our simulations are in good agreement with the findings of Lu *et al.* that increasing side chain length stabilizes α-helical configurations.[8] While there is clearly a large change in helical content between $\mathrm{PAGG_{10}}$ and $\mathrm{PAPG_{10}}$ along all three of these metrics, it is not possible to ascertain from these simple scalar measurements whether the coil to helix transition is continuous or discrete. Our composite diffusion map embeddings (cf. Fig. 7) show this unequivocally to be a discrete transition and reveal the existence of a critical side chain length. Specifically, elongating the side chain by a single C—C bond from 7 to 8 covalent bonds (i.e., $\mathrm{PAGG_{10}}$ to $\mathrm{PAPG_{10}}$) causes a qualitative shift in the conformational ensemble from one closely resembling the random coil configurations of $\mathrm{PL_{10}}$ to one which overlaps with the manifestly helical conformations of $\mathrm{PABG_{10}}$, $\mathrm{PATG_{10}}$, and

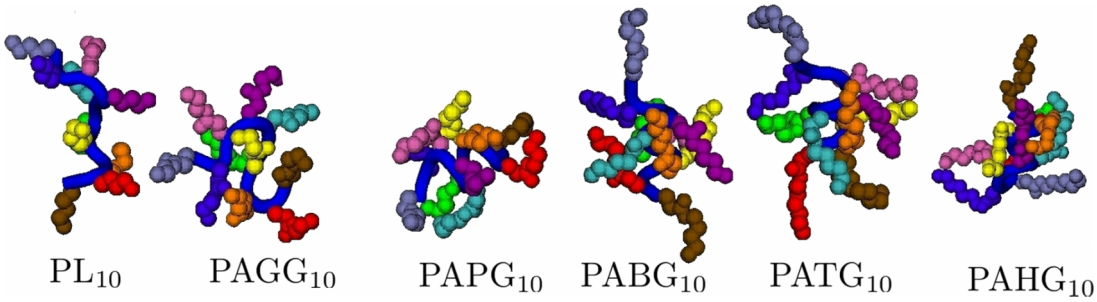

FIG. 11. Representative snapshots of the equilibrium configuration of each polyglutamate-derivative peptide and the polylysine control. The peptide backbone is illustrated by a blue ribbon and the heavy atoms of each of the ten side chains with different colored van der Waals spheres.

$PAHG_{10}$. Heartened by our good experimental agreement of our simulation results, we now proceed to probe the molecular mechanisms underlying this trend.

In Fig. 10(e), we report the mean distance between the charged termini of the side chains, d_{term}, as a function of side chain length. Within the standard error, d_{term} is approximately constant for PL_{10}, $PAGG_{10}$, and $PAPG_{10}$, before monotonically increasing between $PAPG_{10}$ and $PAHG_{10}$. This trend suggests a critical level of electrostatic repulsion between the positively charged side chain termini beyond which a compact helical configuration cannot be sustained. $PABG_{10}$, $PATG_{10}$, and $PAHG_{10}$ all exist as nearly ideal α-helices, with increasing side chain length resulting in increasing separation of the charged termini and reduction of electrostatic repulsion. The value of d_{term} for $PAPG_{10}$ lies on the same trend line as that of its longer cousins, but its markedly reduced helical content (cf. Figs. 10(a)–10(d)) suggests that at $d_{\text{term}} \approx 1.5$ nm, the electrostatic repulsion has reached a tipping point and has become sufficiently strong to destabilize the compact helix. Indeed, decreasing the side chain length to that of $PAGG_{10}$ or PL_{10} is accompanied by disruption of the helical structure in order to maintain $d_{\text{term}} \approx 1.5$ nm. Representative snapshots of the equilibrium structure of each peptide in Fig. 11 provide visual corroboration of this trend. These findings support the hypothesis of Lu *et al.*[8] that increasing side chain length mitigates electrostatic repulsion and stabilizes helix formation, and the existence of a critical chain length at which the electrostatic repulsion becomes too strong to sustain an α-helix.

In Fig. 10(f), we plot the mean distance from the peptide center of mass to the side chain center of mass, $d_{\text{sidechain}}$. Adding a single C—C bond of length 0.153 nm to the $PAGG_{10}$ random coil increases the distal protrusion of the side chain center of mass by 0.090 nm or about 59% of the covalent bond length. For α-helices—$PAPG_{10}$, $PABG_{10}$, $PATG_{10}$, and $PAHG_{10}$—side chain elongation increases $d_{\text{sidechain}}$ by less than 0.025 nm per C—C bond or approximately 16% of the bond length. This relatively gradual increase in $d_{\text{sidechain}}$ for the helical peptides is in contrast to the steeper trend in d_{term} of about 0.088 nm per C—C bond (Fig. 10(e)). The emergent physical picture is one in which the hydrophobic effect drives the side chains to adopt compact configurations wrapped closely around the peptide backbone to minimize disruption of the solvent hydrogen bonding network,[93] and do so in such a manner as to mitigate electrostatic repulsion between the positively charged termini (cf. Fig. 11).

In Fig. 10(g), we report the radial distribution functions between the peptide backbone center of mass and the solvent O atoms, $\text{rdf}_{\text{COM–O}}$. The coil to helix transition mediated by increasing the side chain length from 7 ($PAGG_{10}$) to 8 ($PAPG_{10}$) covalent bonds is accompanied by a marked reduction in solvent occupancy below 0.5 nm corresponding to exclusion of the solvent from the peptide core. Further elongation of the side chains has relatively minor impact upon the $\text{rdf}_{\text{COM–O}}$ profile, corroborating the physical picture in which elongation of the side chains for helical peptides results primarily in increased wrapping of the hydrophobic side chains around the peptide backbone with relatively little impact on the surrounding solvent (cf. Fig. 10(g)). These findings substantiate the hydrophobically driven stabilization mechanism of the α-helix, wherein the hydrophobic effect drives the side chains to shroud the peptide backbone from the solvent, driving the formation of hydrogen bonds between

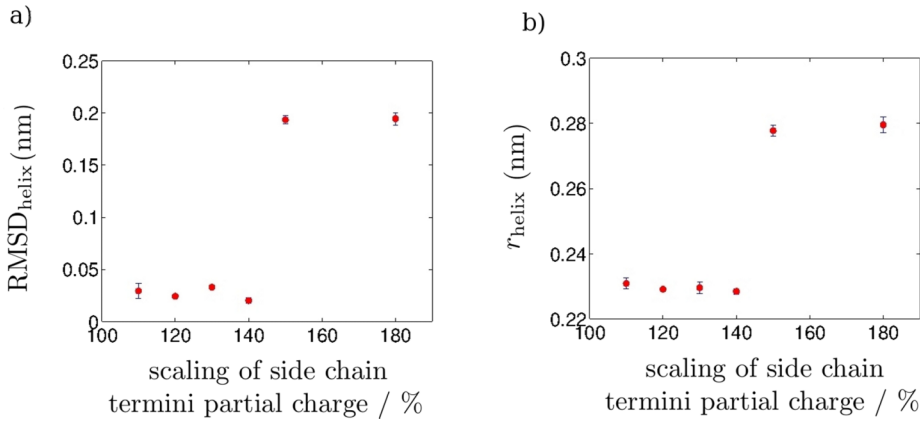

FIG. 12. Equilibrium structural metrics for modified $PATG_{10}$ peptides possessing artificially elevated partial charges on the atoms constituting the side chain terminal amino group. (a) The root mean squared deviation of the backbone from an ideal α-helix, $\text{RMSD}_{\text{helix}}$. (b) The effective helical radius, r_{helix}.

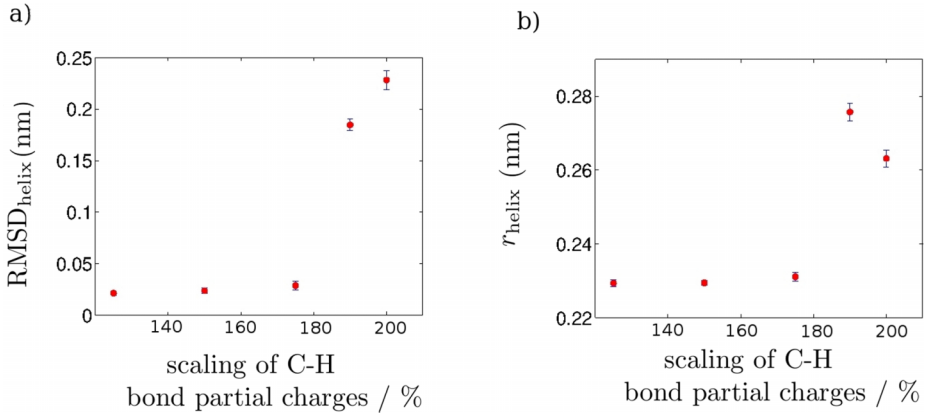

FIG. 13. Equilibrium structural metrics for modified $PATG_{10}$ peptides with artificially elevated side chain hydrophilicity by increasing the dipole moment of the terminal six C—H bonds on each side chain. (a) The root mean squared deviation of the backbone from an ideal α-helix, $RMSD_{helix}$. (b) The effective helical radius, r_{helix}.

the backbone C═O and N—H within the spiral of an α-helix. Such side chain-wrapped configurations are apparently inaccessible for side chains shorter than 9 covalent bonds due to disruption of the helix by electrostatic repulsion between the side chain termini (Fig. 10(e)). This mechanistic picture provides theoretical, molecular-level support for the experimental findings of Lu *et al.* that side chain hydrophobicity is critical for helix stabilization.[8]

To further test the hypothesis of Lu *et al.* that helical stability is governed by both electrostatic repulsion between the side chain termini and side chain hydrophobicity, we performed additional simulations of two peptide families constructed by (nonphysical) modifications of the $PATG_{10}$ molecule. To elucidate the effects of the charges on the end termini, we artificially increased the partial charges on the atoms constituting the terminal amino group on each $PATG_{10}$ side chain to 110%, 120%, 130%, 140%, 150%, and 180% of their natural values. Charge neutrality was maintained by a commensurate increase in the number of Cl^- counter-ions. We illustrate in Fig. 12 the response of $RMSD_{helix}$ and r_{helix} to the elevated charge, from which it is apparent that there exists a critical level of electrostatic repulsion above which the α-helix is destabilized. This is consistent with experimental findings that elevated surface charge density is correlated with reduced helical stability.[8,103] To assess the impact of side chain hydrophilicity, we artificially elevated side chain hydrophilicity by elevating the partial charges on the three CH_2 groups at the end of each $PATG_{10}$ side chain by 125%, 150%, 175%, 190%, and 200%. This effect of this operation is to increase the dipole moment of the terminal six C—H bonds, leaving the total system charge unchanged. As illustrated in Fig. 13, there also exists a critical value of side chain hydrophilicity, above which the hydrophobic effect becomes too weak to stabilize the helix.

IV. CONCLUSIONS

We have conducted molecular simulations of *n*-eicosane chains under different solvent environments and temperatures and a family of non-natural polyglutamate-derivative peptides with different side chain chemistries. By applying diffusion maps to the aggregated conformational ensemble of each of these two systems, we synthesized low-dimensional nonlinear embeddings to expose the impact of molecular chemistry and environmental conditions upon the single molecule free energy landscape and reveal their influence upon molecular behavior. Furthermore, these embeddings unveil the collective motions governing the slow dynamical modes of the molecule, revealing the important collective motions governing its long-time conformational dynamics. This approach provides a means to understand and inform the engineering of desirable single molecule structure and behavior by manipulating molecular chemistry and environmental conditions.

In our study of *n*-eicosane chains in the ideal gas, aqueous solution, and the neat phase at 298 K and 323 K, composite diffusion maps revealed the sub-regions of a three-dimensional intrinsic manifold populated by the chain under the various conditions. We find—with the exception of the crystalline solid—remarkably similar free energy landscapes and collective molecular motions for the chains under all conditions studied. Consistent with previous work, we find that the intrinsic dynamical motions of an isolated chain in the ideal gas are largely preserved upon immersion into aqueous solution, with relatively small perturbations of the ideal gas free energy landscape due to the influence of the water.[5] We also show that the dynamical motions in the hydrocarbon liquid are surprisingly similar to those in the ideal gas, although collapsed chain configurations are significantly less stable. In contrast, the thermally accessible conformational ensemble available to the chains in the hydrocarbon crystal is much more limited, populating only a small fraction of the complete intrinsic manifold localized around the fully extended all-*trans* conformation. Visualization of the dynamical modes associated with each dimension of the low-dimensional manifold extracted by diffusion maps facilitated interpretation of the collective dynamical motions associated with each mode without relying on the *a priori* availability of good physical "bridge" variables with which to correlate the low-dimensional collective order parameters. In sum, the low-dimensional free energy landscapes revealed by composite diffusion maps provide molecular insight into the equilibrium structure and dynamical motions of the *n*-eicosane chain and quantify the influence of temperature and solvent conditions upon single-molecule behavior. It would be of interest to explore the impact of additional environmental conditions (e.g., pressure, concentration, and flow) upon the behavior of hydrocarbon chains of different architectures and chemistries,

and these findings may have implications for the design of "switchable" polymers possessing the ability to rapidly change their molecular structure and function in response to an environmental trigger.

In an application of composite diffusion maps to a family of polyglutamate-derivative decamers, our approach discovered a five-dimensional intrinsic manifold containing the principal dynamical motions of the peptide family. Consistent with the experimental results of Lu *et al.*,[8] we found increasing stabilization of the α-helix with increasing side chain length. We also discovered the existence of a critical side chain length of 8 covalent bonds at which the peptide transitions from a random coil—$PAGG_{10}$ (7 covalent bonds)—to an α-helix—$PAPG_{10}$ (8), $PABG_{10}$ (9), $PATG_{10}$ (10), and $PAHG_{10}$ (11)—corresponding to a global structural change mediated by elongation of the side chain by a single C—C bond. Visualization of the dynamical modes of each peptide revealed a very restricted configurational range for the α-helices, moving along an essentially one-dimensional manifold limited to tightening and loosening of the helix. In contrast, the random coils explored a much richer and more diverse configurational space and preferentially occupied particular weakly structured conformational states. Analysis of the molecular-mechanisms underpinning this trend provides molecular-level support for the hypothesis of Lu *et al.* that reducing electrostatic repulsion between the charged side chain termini, coupled with intrinsic side chain hydrophobicity, drives the stabilization of helical configurations with increasing side chain length.[8] Specifically, we identify a critical level of side chain electrostatic repulsion at a side chain length of 8 covalent bonds ($PAPG_{10}$). Shorter side chains result in elevated repulsion that is alleviated by disruption of the back bone helicity to facilitate greater separation of the charged termini. Longer side chains mitigate this repulsion, facilitating hydrophobicity-induced condensation of the chains around the peptide backbone in a manner that maintains separation between the charged termini. Shrouded from the solvent by the hydrophobic side chains, the hydrogen bond donors and acceptors in the peptide backbone satisfy their electrostatic attractions by pairing up within an α-helix. This critical level of electrostatic repulsion underpins the discrete helix-coil transition unveiled by our composite diffusion maps and suggests a new principle for the rational design of marginally stable helices. In sum, we have provided theoretical corroboration of the stabilization helical hypothesis of Lu *et al.*,[8] revealed the molecular mechanisms underpinning this trend, and identified a critical polyglutamate-derivative side chain length underpinning the helix-coil transition. In future work, we will explore how these trends change for longer polypeptides, longer side chains, and different side chain chemistries. An improved understanding of the basic physical principles underpinning these trends enables new paradigms for the rational design of α-helical peptides with tunable and/or switchable stability.

ACKNOWLEDGMENTS

We thank Ziyuan Song and Dr. Jianjun Cheng for generous and fruitful discussions.

[1]D. Baker and A. Sali, Science **294**, 93 (2001).
[2]G. A. Petsko and D. Ringe, *Protein Structure and Function* (New Science Press, 2004).
[3]K. A. Dill and J. L. MacCallum, Science **338**, 1042 (2012).
[4]M. Rubinstein and R. Colby, *Polymers Physics* (Oxford University Press, Oxford, 2003), p. 113.
[5]A. L. Ferguson, P. G. Debenedetti, and A. Z. Panagiotopoulos, J. Phys. Chem. B **113**, 6405 (2009).
[6]S. Hammes-Schiffer and S. J. Benkovic, Annu. Rev. Biochem. **75**, 519 (2006).
[7]G. Bhabha, J. Lee, D. C. Ekiert, J. Gam, I. A. Wilson, H. J. Dyson, S. J. Benkovic, and P. E. Wright, Science **332**, 234 (2011).
[8]H. Lu, J. Wang, Y. Bai, J. W. Lang, S. Liu, Y. Lin, and J. Cheng, Nat. Commun. **2**, 206 (2011).
[9]A. Karim, S. Satija, J. Douglas, J. Ankner, and L. Fetters, Phys. Rev. Lett. **73**, 3407 (1994).
[10]A. E. García, Phys. Rev. Lett. **68**, 2696 (1992).
[11]A. Amadei, A. Linssen, and H. J. Berendsen, Proteins: Struct., Funct., Bioinf. **17**, 412 (1993).
[12]R. Hegger, A. Altis, P. H. Nguyen, and G. Stock, Phys. Rev. Lett. **98**, 028102 (2007).
[13]P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, Proc. Natl. Acad. Sci. U. S. A. **103**, 9885 (2006).
[14]A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti, Chem. Phys. Lett. **509**, 1 (2011).
[15]P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Natl. Acad. Sci. U. S. A. **97**, 5877 (2000).
[16]A. L. Ferguson, S. Zhang, I. Dikiy, A. Z. Panagiotopoulos, P. G. Debenedetti, and A. J. Link, Biophys. J. **99**, 3056 (2010).
[17]A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, Proc. Natl. Acad. Sci. U. S. A. **107**, 13597 (2010).
[18]R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, Multiscale Model. Simul. **7**, 842 (2008).
[19]T. Ichiye and M. Karplus, Proteins: Struct., Funct., Bioinf. **11**, 205 (1991).
[20]O. F. Lange and H. Grubmüller, Proteins: Struct., Funct., Bioinf. **70**, 1294 (2008).
[21]B. Schölkopf, A. Smola, and K. Müller, *International Conference on Artificial Neural Networks—ICANN*, Lecture Notes in Computer Science, edited by W. Gerstner, A. Germond, M. Hasler, and J. Nicoud (Springer, Berlin, 1997), Vol. 1327, pp. 583–588.
[22]A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, J. Chem. Phys. **134**, 135103 (2011).
[23]M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**, 124116 (2011).
[24]W. Zheng, M. A. Rohrdanz, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**, 144109 (2011).
[25]R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. U. S. A. **107**, 1088 (2010).
[26]A. Singer, R. Erban, I. G. Kevrekidis, and R. R. Coifman, Proc. Natl. Acad. Sci. U. S. A. **106**, 16090 (2009).
[27]B. E. Sonday, M. Haataja, and I. G. Kevrekidis, Phys. Rev. E **80**, 031102 (2009).
[28]P. Das, T. A. Frewen, I. G. Kevrekidis, and C. Clementi, "Think globally, move locally: Coarse graining of effective free energy surfaces," in *Coping with Complexity: Model Reduction and Data Analysis*, edited by A. Gorban and D. Roose (Springer, 2011), pp. 113–131.
[29]B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[30]M. G. Martin and J. I. Siepmann, J. Phys. Chem. B **102**, 2569 (1998).
[31]H. J. Berendsen, J. Postma, W. Van Gunsteren, and J. Hermans, "Interaction models for water in relation to protein hydration," in *Intermolecular Forces*, edited by B. Pullman (Springer, Dordrecht, Holland, 1981), pp. 331–342.
[32]A. W. Schuttelkopf and D. M. Van Aalten, Acta Crystallogr., Sect. D: Biol. Crystallogr. **60**, 1355 (2004).
[33]S. Nosé, J. Chem. Phys. **81**, 511 (1984).
[34]M. Parrinello and A. Rahman, J. Appl. Phys. **52**, 7182 (1981).
[35]U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).
[36]D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic Press, San Diego, 2002), Vol. 2.
[37]M. N. Rosenbluth and A. W. Rosenbluth, J. Chem. Phys. **23**, 356 (1955).
[38]See supplementary material at http://dx.doi.org/10.1063/1.4914144 for supplementary text and 11 supplementary figures.

[39]A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, J. Phys. Chem. B **102**, 3586 (1998).
[40]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[41]M. Belkin and P. Niyogi, Neural Comput. **15**, 1373 (2003).
[42]B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, Appl. Comput. Harmonic Anal. **21**, 113 (2006).
[43]R. R. Coifman and S. Lafon, Appl. Comput. Harmonic Anal. **21**, 5 (2006).
[44]R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proc. Natl. Acad. Sci. U. S. A. **102**, 7426 (2005).
[45]I. Jolliffe, *Principal Component Analysis* (John Wiley and Sons, 2005).
[46]R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, IEEE Trans. Image Process. **17**, 1891 (2008).
[47]W. Kabsch, Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **32**, 922 (1976).
[48]R. G. Littlejohn and M. Reinsch, Rev. Mod. Phys. **69**, 213 (1997).
[49]A. W. Long and A. L. Ferguson, J. Phys. Chem. B **118**, 4228 (2014).
[50]R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, New York City, 2001).
[51]K. J. Maschhoff and D. C. Sorensen, *Applied Parallel Computing Industrial Computation and Optimization* (Springer, 1996), pp. 478–486.
[52]B. Peters and B. L. Trout, J. Chem. Phys. **125**, 054108 (2006).
[53]A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769 (2005).
[54]T. Berry, J. Cressman, Z. Greguric-Ferencek, and T. Sauer, SIAM J. Appl. Dyn. Syst. **12**, 618 (2013).
[55]J. Brooks and J. Smith, Geochim. Cosmochim. Acta **33**, 1183 (1969).
[56]N. Chang, Z.-Y. Gu, H.-F. Wang, and X.-P. Yan, Anal. Chem. **83**, 7094 (2011).
[57]D. J. Abdallah and R. G. Weiss, Langmuir **16**, 352 (2000).
[58]S. Chakrabarty and B. Bagchi, J. Phys. Chem. B **113**, 8446 (2009).
[59]R. Ghosh, S. Banerjee, S. Chakrabarty, and B. Bagchi, J. Phys. Chem. B **115**, 7612 (2011).
[60]S. Chakrabarty and B. Bagchi, J. Chem. Phys. **133**, 214901 (2010).
[61]C. Tanford, J. Mol. Biol. **67**, 59 (1972).
[62]W. Kauzmann, Adv. Protein Chem. **14**, 1 (1959).
[63]R. D. Mountain and D. Thirumalai, J. Am. Chem. Soc. **125**, 1950 (2003).
[64]D. M. Huang and D. Chandler, J. Phys. Chem. B **106**, 2047 (2002).
[65]L. Sun, J. I. Siepmann, and M. R. Schure, J. Phys. Chem. B **110**, 10519 (2006).
[66]M. V. Athawale, G. Goel, T. Ghosh, T. M. Truskett, and S. Garde, Proc. Natl. Acad. Sci. U. S. A. **104**, 733 (2007).
[67]R. Underwood, J. Tomlinson-Phillips, and D. Ben-Amotz, J. Phys. Chem. B **114**, 8646 (2010).
[68]J. N. Byrd, R. J. Bartlett, and J. A. Montgomery, Jr., J. Phys. Chem. A **118**, 1706 (2014).
[69]N. O. Lüttschwager, T. N. Wassermann, R. A. Mata, and M. A. Suhm, Angew. Chem., Int. Ed. **52**, 463 (2013).
[70]W. L. Jorgensen, J. Chem. Phys. **77**, 5757 (1982).
[71]J. Tolls, J. van Dijk, E. J. Verbruggen, J. L. Hermens, B. Loeprecht, and G. Schüürmann, J. Phys. Chem. A **106**, 2760 (2002).
[72]A. Maączyński, M. Góral, B. Wiśniewska-Gocłowska, A. Skrzecz, and D. Shaw, Monatsh. Chem. **134**, 633 (2003).
[73]K. Lum, D. Chandler, and J. D. Weeks, J. Phys. Chem. B **103**, 4570 (1999).
[74]L. R. Pratt and D. Chandler, J. Chem. Phys. **67**, 3683 (2008).
[75]R. L. Baldwin, FEBS Lett. **587**, 1062 (2013).
[76]H. S. Ashbaugh, S. Garde, G. Hummer, E. W. Kaler, and M. E. Paulaitis, Biophys. J. **77**, 645 (1999).
[77]G. C. Boulougouris, J. R. Errington, I. G. Economou, A. Z. Panagiotopoulos, and D. N. Theodorou, J. Phys. Chem. B **104**, 4958 (2000).
[78]J. S. Chickos and W. Hanshaw, J. Chem. Eng. Data **49**, 77 (2004).
[79]J. J. de Pablo, M. Laso, and U. W. Suter, J. Chem. Phys. **96**, 6157 (1992).
[80]I. G. Economou, Fluid Phase Equilib. **183**, 259 (2001).
[81]J. R. Errington, G. C. Boulougouris, I. G. Economou, A. Z. Panagiotopoulos, and D. N. Theodorou, J. Phys. Chem. B **102**, 8865 (1998).
[82]P. V. Khadikar, D. Mandloi, A. V. Bajaj, and S. Joshi, Bioorg. Med. Chem. Lett. **13**, 419 (2003).
[83]A. V. Plyasunov and E. L. Shock, Geochim. Cosmochim. Acta **64**, 439 (2000).
[84]S. Salvador and P. Chan, *ICTAI '04 Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence* (IEEE, Boca Raton, Florida, 2004), pp. 576–584.
[85]T. Sauer, J. A. Yorke, and M. Casdagli, J. Stat. Phys. **65**, 579 (1991).
[86]P. Grassberger and I. Procaccia, Physica D **9**, 189 (1983).
[87]D. N. Theodorou and U. W. Suter, Macromolecules **18**, 1206 (1985).
[88]W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[89]G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[90]A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **99**, 12562 (2002).
[91]C. Laing, T. Frewen, and I. Kevrekidis, Nonlinearity **20**, 2127 (2007).
[92]B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis, in *Advances in Neural Information Processing Systems*, edited by Y. Weiss, B. Schölkopf, and J. Platt (MIT Press, Cambridge, MA, 2005), Vol. 18, pp. 955–962.
[93]N. T. Southall, K. A. Dill, and A. Haymet, J. Phys. Chem. B **106**, 521 (2002).
[94]C. Branden and J. Tooze, *Introduction to Protein Structure* (Garland, New York, 1991), Vol. 2.
[95]K. Usui, T. Kikuchi, M. Mie, E. Kobatake, and H. Mihara, Bioorg. Med. Chem. **21**, 2560 (2013).
[96]Y. Fezoui, D. M. Hartley, D. M. Walsh, D. J. Selkoe, J. J. Osterhout, and D. B. Teplow, Nat. Struct. Mol. Biol. **7**, 1095 (2000).
[97]G. Drin, J.-F. Casella, R. Gautier, T. Boehmer, T. U. Schwartz, and B. Antonny, Nat. Struct. Mol. Biol. **14**, 138 (2007).
[98]A. P. Nowak, V. Breedveld, L. Pakstis, B. Ozbas, D. J. Pine, D. Pochan, and T. J. Deming, Nature **417**, 424 (2002).
[99]S. C. Patel, L. H. Bradley, S. P. Jinadasa, and M. H. Hecht, Protein Sci. **18**, 1388 (2009).
[100]V. Azzarito, K. Long, N. S. Murphy, and A. J. Wilson, Nat. Chem. **5**, 161 (2013).
[101]G. A. Papoian, Proc. Natl. Acad. Sci. U. S. A. **105**, 14237 (2008).
[102]B. M. Bulheller and J. D. Hirst, Bioinformatics **25**, 539 (2009).
[103]H. Tang, L. Yin, K. H. Kim, and J. Cheng, Chem. Sci. **4**, 3839 (2013).