Integrating diffusion maps with umbrella sampling: Application to alanine dipeptide

Andrew L. Ferguson,[1,a)] Athanassios Z. Panagiotopoulos,[1] Pablo G. Debenedetti,[1] and Ioannis G. Kevrekidis[2]
[1]*Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey 08544, USA*
[2]*Department of Chemical and Biological Engineering and Program in Applied and Computational Mathematics, Princeton University, Princeton, New Jersey 08544, USA*

(Received 1 October 2010; accepted 16 March 2011; published online 7 April 2011)

Nonlinear dimensionality reduction techniques can be applied to molecular simulation trajectories to systematically extract a small number of variables with which to parametrize the important dynamical motions of the system. For molecular systems exhibiting free energy barriers exceeding a few $k_B T$, inadequate sampling of the barrier regions between stable or metastable basins can lead to a poor global characterization of the free energy landscape. We present an adaptation of a nonlinear dimensionality reduction technique known as the *diffusion map* that extends its applicability to biased umbrella sampling simulation trajectories in which restraining potentials are employed to drive the system into high free energy regions and improve sampling of phase space. We then propose a bootstrapped approach to iteratively discover good low-dimensional parametrizations by interleaving successive rounds of umbrella sampling and diffusion mapping, and we illustrate the technique through a study of alanine dipeptide in explicit solvent. © *2011 American Institute of Physics*. [doi:10.1063/1.3574394]

I. INTRODUCTION

Molecular simulations have demonstrated that the fundamental dynamical motions of biophysical systems may frequently be parametrized by a small number of collective variables.[1–7] The existence of such low-dimensional descriptions may be attributed to couplings between the degrees of freedom,[3–9] resulting in a small number of slowly evolving variables that govern the dynamics and to which the remaining fast degrees of freedom (or their statistics) are slaved. Such a description leads naturally to the modeling of biophysical systems as diffusion processes, in which a set of stochastic differential equations may be formulated in the slow variables, with the fast degrees of freedom represented as thermal noise.[9–12] From a geometric perspective, the low-dimensional manifold parametrized by the slow variables may be visualized as a (possibly highly convoluted) surface in phase space, which we term the "intrinsic manifold"[9] and upon which the system effectively evolves.

The systematic development of low-dimensional descriptions of biophysical systems can provide insight into the important underlying motions, which may be related to molecular function, folding or activity, or simply provide a deeper understanding of the conformational states sampled by the system. Furthermore, extraction of a small number of variables is of use in targeting simulations to the region of phase space of interest and informing inexpensive low-dimensional simulations in the slow variables.

In recent years, among other efforts, Ma and Dinner identified three order parameters with which to describe the $C_7^{\mathrm{eq}} \leftrightarrow \alpha_R$ isomerization of alanine dipeptide,[2] Clementi and co-workers determined two-dimensional parametrizations for a coarse-grained models of a β-hairpin[7] and the src homology 3 domain,[6] and we developed three-dimensional descriptions of solvated n-alkane chains[9] and of a 21-residue peptide, pro-microcin J25.[13] The development of low-dimensional parametrizations may be frustrated, however, for systems exhibiting high free energy barriers. Simulations of such systems can become trapped in a local free energy well, rarely escaping to adjacent states over the high free energy barriers surrounding the basin. Even if the simulation does escape to a neighboring well, the barriers themselves are infrequently sampled by such "reactive" trajectories and remain poorly characterized. Dimensionality reduction techniques may permit the synthesis of good low-dimensional descriptions of the important dynamics *within* well-sampled free energy basins, but poor sampling of the intervening barrier regions typically precludes the construction of globally valid parametrizations.

Given a molecular simulation trapped in a local free energy minimum, we would like to induce the system to escape the metastable basin and explore the surrounding phase space by driving it over the bounding free energy barriers. A standard approach employs biased sampling in which harmonic restraining potentials are applied to a set of independent simulations to restrict each trajectory to a particular region of phase space in the vicinity of the local minimum. The configurations explored by the various biased simulations provide good sampling of the region around the local minimum, including high free energy barrier regions. By explicitly removing the

[a)] Author to whom correspondence should be addressed. Electronic mail: aferguso@mit.edu. Tel.: (617) 252-1744. FAX: (617) 253-2272. Present address: Department of Chemical Engineering, MIT, E19-550, Cambridge, Massachusetts 02139, USA.

biases associated with each run, an unbiased free energy surface (FES) may be constructed over the entire region of phase space explored by the simulations. The application of biasing potentials to improve sampling of phase space and the subsequent aggregation of these runs to synthesize an unbiased free energy surface is a well-established molecular simulation technique known as umbrella sampling.[14]

Since the biasing potentials used to restrain the simulation to a particular region of phase space must be formulated as a function of the atomic coordinates of the system, we require good collective variables with which to parametrize the location and extent of the free energy basin within which the simulation is trapped. In practice, the success of umbrella sampling is contingent on the availability of such "good" variables, which may often be associated with the slow dynamical motions of the system between its stable and metastable states.[6,15] Such variables are rarely available *a priori*, but may be systematically determined by data mining simulation trajectories using dimensionality reduction techniques.

Building on this premise, in this work we develop a scheme to systematically and efficiently explore the thermally accessible phase space available to a system. Briefly, given an unbiased simulation trajectory trapped in a local free energy minimum, we propose that dimensionality reduction techniques be applied to determine good collective variables in which to then extend umbrella sampling beyond the free energy well and into the surrounding phase space. The bias associated with each simulation trajectory can subsequently be explicitly removed to construct the unbiased FES over the phase space explored by the biased simulations. The newly explored regions of phase space may be better parametrized by variables other than those in which the original umbrella sampling simulations were performed. Accordingly, we would like to apply dimensionality reduction to the biased ensemble of system configurations to determine new variables in which to conduct the second round of umbrella sampling. By repeating this process, we have conceived an iterative strategy to progressively explore larger regions of phase space by applying dimensionality reduction techniques to biased simulation data in order to determine good variables in which to conduct subsequent rounds of biased simulation.

Clearly, a prerequisite for the application of this iterative strategy is that the dimensionality reduction technique employed must be applicable to biased simulation data. While a number of dimensionality reduction techniques may be adapted to biased data sets, a class of nonlinear dimensionality reduction approaches known as manifold learning techniques, provide a means to extract *global* order parameters characterizing the important dynamical motions of the system over large regions of phase space.[16–18] Compared to linear dimensionality reduction approaches, manifold learning techniques are expected to identify a far smaller number of variables with which to adequately describe the important motions of the system, and therefore simplify the application of umbrella sampling by providing a more parsimonious low-dimensional description.[5,7] In this work, we present an adaptation of one such technique known as the *diffusion map*[16,19,20] to operate on biased ensembles of simulation configurations and permit its incorporation into an iterative strategy for the efficient exploration of phase space for systems possessing high free energy barriers.

The structure of this paper is as follows. In Sec. II, we outline the diffusion map technique[16,19,20] and present an adaptation of this approach to operate on biased simulation data. We then present an iterative method for the systematic identification of low-dimensional parametrizations for systems exhibiting high free energy barriers by interleaving umbrella sampling simulations and applications of the diffusion map. In Sec. III, we demonstrate the methodology in an application to solvated alanine dipeptide, which is a standard test system for novel simulation methodologies and is known to possess free energy barriers of several k_BT between its metastable states.[21] Finally, in Sec. IV we present our conclusions.

II. METHODS

A. Dimensionality reduction: The diffusion map

The diffusion map[16,19,20] is a nonlinear dimensionality reduction technique which we have applied to simulations of *n*-alkanes[9] and a 21-residue peptide[13] in explicit solvent. Other researchers have employed this approach to analyze simulations of networks of chemical reactions,[22] the mobility of a defect at an interface,[12] and gene regulatory networks.[23] Isomap,[17] ScIMAP,[6] and local linear embedding (LLE)[18] are other nonlinear dimensionality reduction approaches, which share certain commonalities with the diffusion map, and have also been applied to biophysical simulations.[6,7,24,25]

A molecular simulation trajectory may be considered a succession of $3R$-dimensional snapshots recording the coordinates of the R atoms in the system. Dimensionality reduction techniques aim to embed the trajectory into a k-dimensional space, $k \ll 3R$, where the order parameters spanning the space are good descriptors of the dynamic transitions between system states. Application of the diffusion map approach to simulation trajectories has been previously described by ourselves and others.[9,12,13,23,26,27] The underlying premise of the technique is that the global structure of the intrinsic manifold of the system—the surface in phase space upon which the data lie—may be reconstructed by "integration" of the local distances between neighboring data points.[19] The first step defines a scalar-valued similarity metric with which to compute distances between all pairs of simulation snapshots. This metric should capture a system's short time diffusive motions—in this case arising from thermal fluctuations of the constituent atoms and molecules—and in this regard need only be a locally meaningful measure. In the present work, we select the translationally and rotationally minimized root mean square deviation (RMSD) between the peptide atomic coordinates and denote the distance between snapshots i and j as RMSD_{ij}. As demonstrated in our prior work,[9] while this measure explicitly disregards all solvent degrees of freedom, the solvent interaction influences the structural conformations adopted by the peptide and its influence is therefore incorporated into the sampled peptide configurations.

For a trajectory, or composite of multiple trajectories, comprising N snapshots, the distances between all snapshot pairs are stored in a symmetric N-by-N matrix. A Gaussian kernel is then applied to each element of the matrix to synthesize the **A** matrix,

$$A_{ij} = \exp(-(\mathrm{RMSD}_{ij})^2/2\epsilon) \qquad i, j = 1 \ldots N, \tag{1}$$

where $\epsilon > 0$ is the bandwidth of the kernel, specifying the characteristic radius within which RMSD_{ij} is considered a meaningful similarity metric. Following Coifman *et al.*,[26] the range of appropriate ϵ values lie within the linear region of a log–log plot of $\sum_{i,j=1}^{N} A_{ij}$ vs ϵ, where the slope provides an estimate of the effective system dimensionality.

The row sums of **A** are stored in the main diagonal of the diagonal matrix **D**,

$$D_{ij} = \begin{cases} \sum_{q=1}^{N} A_{iq}, & \text{if } i = j \qquad i, j = 1 \ldots N, \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

from which the **A** matrix is row normalized to generate the right stochastic Markov transition matrix **M**,

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

We denote the N eigenvalues of **M** in nonascending order as $\{\lambda_i\}_{i=1}^{N}$, $\lambda_1 = 1 \geq \lambda_2 \geq \ldots \lambda_N$, and the corresponding eigenvectors as $\{\vec{\phi}_i\}_{i=1}^{N}$. The Markov property of **M** results in a trivial top eigenvalue of unity, $\lambda_1 = 1$, with associated "all-ones" eigenvector, $\vec{\phi}_1 = \vec{1}$.

The so-called k-dimensional *diffusion map* is the mapping of the ith snapshot of the simulation trajectory into the ith components of the top k nontrivial eigenvectors,[9,16,19,27]

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

In our terminology, we denote this mapping as an "embedding into the top k eigenvectors."[9,13] The number of eigenvectors to employ in the mapping, and hence the dimensionality of the embedding, may be determined by a gap in the spectrum of eigenvalues $\{\lambda_i\}_{i=1}^{N}$.[16,27] The components of the eigenvectors are the "diffusion map order parameters" identified by the approach.

Linear dimensionality reduction techniques such as principal component analysis (PCA)[3–5,8,13,28,29] typically explicitly furnish the transformation between the input variables and the low-dimensional embedding.[13] This, as we shall see in Sec. II B 3, is a very attractive feature which is typically not shared by nonlinear dimensionality reduction approaches. Indeed, the correspondence of the diffusion map order parameters to physical variables is not furnished by the approach, and may only be determined *a posteriori* by correlating the eigenvector components with candidate physical variables.[9,12] More elegant means to determine such correspondence is an active area of research.

Under the dual assumptions that the biophysical system may be well-modeled as a diffusion process in the sense described in Sec. I, and that the pairwise similarity metric captures the short time diffusive motions of the system, the diffusion map embedding of the system is a dynamically meaningful reconstruction of the intrinsic manifold. Pathways over the manifold are associated with motions of the system in its fundamental dynamical motions, and order parameters describing the paths are good descriptors of these modes.[9,19,23,26,27] Although these assumptions are expected to hold generally for biophysical systems, in the event that they do not, then the identified order parameters may not be associated with the underlying dynamical modes. Nevertheless, these variables still provide good separation of the local free energy minima of the system and offer a useful low-dimensional parametrization of its evolution from one state to another.[9]

Du *et al.* define the *transition coordinate* as the component of the motion between two states that possesses the largest relaxation time; in contrast, the *reaction coordinate* is defined in its transition state theoretical sense as the precise minimum free energy pathway linking the two states.[15] Accordingly, while motions in the degrees of freedom orthogonal to the transition coordinate may be large, it nevertheless serves as a useful order parameter with which to describe transitions of the system from one state to another. Applying these concepts to the current work, the order parameters furnished by the diffusion map serve as a good parametrization of the *transition coordinates* between the states of the system. Verification of the stronger condition that the *transition coordinate* is also good *reaction coordinate*, would require a separate evaluation of the committor probabilities along the putative coordinate using a technique such as transition path sampling.[9,30]

B. Integration of umbrella sampling into the diffusion map approach

The synthesis of a globally valid low-dimensional description—by the diffusion map or any other dimensionality reduction technique—requires that the distribution of snapshots in phase space be sufficiently dense that no group of snapshots is disconnected from the "bulk" of the data. For example, if two distinct free energy basins are well-sampled by the simulation while the barrier region between them is not, then dimensionality reduction will typically result in two distinct low-dimensional descriptions which are locally valid within their respective wells, but will fail to provide a satisfactory global parametrization of the two basins and the intervening transition region.

To develop a description of the data as a single unified diffusion process, the diffusion map approach requires that a succession of snapshots separated by hops of length $\mathcal{O}(\epsilon)$ or less in the pairwise similarity metric exists between every pair of snapshots in the trajectory. While, in principle, ϵ can be made sufficiently large to span the gaps resulting from poor sampling, this results in low-resolution diffusion map embeddings that inadequately separate the snapshots in well-sampled regions of phase space.[13] In prior work, we have demonstrated the use of a "deislanding" approach to discard disconnected groups of snapshots, thereby permitting the successful application of the diffusion map approach at the expense of eliminating a portion of the trajectory.[13]

By driving the simulation through the application of biasing potentials, umbrella sampling offers a means both to improve sampling of the high free energy regions of phase space and to facilitate the production of fully-connected data sets.[14,31] The success of the umbrella sampling approach depends crucially on the availability of good variables in which to construct the biasing potentials and drive the system. Ergo, the diffusion map approach provides a means to systematically furnish good variables in which to conduct umbrella sampling, and umbrella sampling provides the means to generate well-connected data sets to which the diffusion map may be effectively applied.

In the subsections which follow, we first describe the rudiments of umbrella sampling and the weighted histogram analysis technique required for the development of our methodology. Second, we present an extension of the diffusion map approach applicable to biased simulation trajectories. Finally, we discuss an iterative diffusion mapping /umbrella sampling protocol which combines these two methodologies as a means to systematically extract global order parameters for systems containing high free energy barriers.

1. Umbrella sampling and WHAM

Consider an N-dimensional umbrella sampling simulation in a set of putative umbrella variables, $\vec{\psi}$, which themselves are some function of the coordinates of the atoms constituting the system. The umbrella sampling may most easily conducted by partitioning the phase space over a N-dimensional grid, and running an independent simulation at each of the v vertices, where a harmonic restraining potential is applied to constrain the simulation to remain in the vicinity of the vertex. Good coverage of phase space is achieved by specifying the grid spacing and harmonic spring constants to permit some overlap between simulation trajectories at neighboring grid points, while adequately restraining each simulation around its own vertex to enforce good sampling of high free energy regions.[31] The weighted histogram analysis method (WHAM) introduced by Ferrenberg and Swendsen,[32] and refined by Kumar *et al.*,[33] seeks to optimally combine a set of biased simulation trajectories into a single unbiased probability distribution in the umbrella variables. This is achieved by self-consistently solving the multidimensional WHAM equations,[34]

$$P(\vec{\psi}) = \frac{\sum_{i=1}^{v} m_i P(\vec{\psi})_i^{\text{biased}}}{\sum_{j=1}^{v} m_j e^{+\beta(F_j - w_j(\vec{\psi}))}}, \tag{5}$$

$$e^{-\beta F_i} = \int e^{-\beta w_i(\vec{\psi})} P(\vec{\psi})\, \mathrm{d}\vec{\psi}, \tag{6}$$

where $P(\vec{\psi})_i^{\text{biased}}$ is the raw (biased) probability distribution function determined from run i as a function of the umbrella variables $\vec{\psi}$, $P(\vec{\psi})$ is the optimal estimate of the overall unbiased probability distribution function, m_i is the number of snapshots in run i, $w_i(\vec{\psi})$ is the restraining potential applied to run i, F_i is the free energy shift associated with run i, and $\beta = 1/k_B T$. The free energy shifts were all initially set to zero, and Eqs. (5) and (6) iteratively solved for $P(\vec{\psi})$ and $\{F_i\}_{i=1}^{v}$ using an in-house multidimensional solver. Since the values of the shifts are only meaningful to within an additive constant, F_1 was set to zero and the iteration terminated when the largest change in any $\{F_i\}_{i=2}^{v}$ value between successive iterations was less than the specified tolerance of 0.01 $k_B T$.

With $P(\vec{\psi})$ in hand, the free energy surface parametrized by the umbrella variables is given, to within an arbitrary constant, as

$$\beta G(\vec{\psi}) = -\ln P(\vec{\psi}) + \text{const.}, \tag{7}$$

where G denotes the Gibbs free energy, as is appropriate for this work in which umbrella sampling was conducted in the isothermal–isobaric (NPT) ensemble. The relative probabilities of two points lying on the FES at locations $\vec{\psi}_1$ and $\vec{\psi}_2$ is given by the exponential of the free energy difference between them,

$$\frac{P(\vec{\psi}_2)}{P(\vec{\psi}_1)} = e^{-\beta[G(\vec{\psi}_2) - G(\vec{\psi}_1)]}. \tag{8}$$

From $\{F_i\}_{i=1}^{v}$, the probability distribution function in order parameters other than the umbrella variables may be computed as

$$P(\vec{\xi}) = \frac{1}{m} \sum_{i=1}^{v} e^{-\beta F_i} \sum_{k=1}^{m_i} e^{+\beta w_i(\vec{\psi}[k])}\, \delta(\vec{\xi}_i[k] - \vec{\xi}), \tag{9}$$

where $\vec{\xi}$ is a vector of arbitrary order parameters each of which is some function of the atomic coordinates of the system, $P(\vec{\xi})$ is the probability distribution function as a function of $\vec{\xi}$, $m = \sum_{i=1}^{v} m_i$, $w_i(\vec{\psi}[k])$ is the value of the biasing potential in the kth snapshot of run i, $\vec{\xi}_i[k]$ is the value of $\vec{\xi}$ in the kth snapshot of run i, and δ denotes the Dirac delta function. The FES parametrized by $\vec{\xi}$ may be computed by the application of Eq. (7) to Eq. (9).

Equation (9) implies that, in principle, the probability distribution function may be reparametrized in any combination of order parameters, $\vec{\xi}$. However, whereas we expect to possess good sampling along the variables in which the system was driven by the application of biasing potentials, the exploration of phase space along coordinates orthogonal to these directions is driven only by spontaneous thermal fluctuations. In practice, therefore, insufficient statistics may be accumulated along arbitrary variables to permit an adequate parametrization of the FES in those coordinates.

2. Application of diffusion maps to umbrella sampling trajectories

The application of the diffusion map approach to an unbiased molecular dynamics simulation trajectory returns eigenvectors which are discrete approximations to the eigenfunctions of a continuous Fokker–Planck process over the data in the limit of $N \rightarrow \infty$ and $\epsilon \rightarrow 0$.[9] (In variants of the approach, the **A** matrix may be pretreated to result in diffusion map eigenvectors corresponding to different continuous space generators, other than diffusion.[19,20,27]) The Fokker–Planck equation describes the evolution of a probability density in the presence of potential wells, and in the present case describes the evolution of the density over the data set, where the depth of the potential wells is related to the local sampling density of phase space.[9] In this manner, the application of the diffusion map approach to unbiased simulation trajectories generates low-dimensional embeddings which preserve the distribution of data points over the intrinsic manifold, and therefore conserve the geometry of the free energy wells explored by the simulation.

Equation (8) prescribes that for every system configuration sampled at the top of a barrier of height ΔG in an unbiased trajectory, there will be approximately $e^{+\beta\Delta G}$ such configurations sampled at the bottom. For even modestly sized barriers, this dictates that trajectories which are sufficiently long to adequately sample the barrier regions and produce a fully-connected data set, will contain a very large number of snapshots. In principle, the diffusion map approach may be applied to arbitrarily long trajectories, but in practice it is limited by matrix storage requirements. For example, approximately 1 TB of RAM would be required to hold the elements of the **M** matrix in single precision during the eigenvector computation for a trajectory containing 500 000 snapshots. Furthermore, nonuniform subsampling an unbiased trajectory to give good coverage of the phase space in a sufficiently small number of snapshots by preferentially discarding snapshots from well-sampled basins and retaining those from sparsely sampled barriers, results in eigenvectors which no longer correspond to discrete approximations of the Fokker–Planck eigenfunctions, and therefore do not preserve the geometry of the free energy landscape over the intrinsic manifold. A similar remark applies to trajectories sampled from any biased distribution, such as those resulting from umbrella sampling simulations.

We now derive an adaptation of the diffusion map approach to permit its application to biased simulation data. We formulate the adaptation with an application to umbrella sampling data in mind, but it is generally applicable to any biased data set for which the underlying unbiased FES is computable. We shall demonstrate that the diffusion map embedding, synthesized by the application of the adapted technique to a set of umbrella sampling simulations in the umbrella variables $\vec{\psi}$, is essentially identical to that which would have resulted from the application of the original diffusion map approach to an unbiased trajectory—of sufficient length to explore the same phase space as that sampled by the umbrella simulations—over the FES parametrized by $\vec{\psi}$. This correspondence implicitly assumes that $\vec{\psi}$ provides a good parametrization of the FES, and the molecular system does not exhibit significant motions in coordinates orthogonal to this parametrization. The determination of an appropriate set of umbrella variables $\vec{\psi}$, and the validation of this assumption will be addressed in Sec. II B 3.

Consider a set of v umbrella sampling trajectories driven in a set of order parameters $\vec{\psi}$, for which the solution of the WHAM equations has synthesized an effective FES as a function of $\vec{\psi}$ [Eqs. (5), (6), and (7)]. The total number of snapshots contained in the v trajectories is $m = \sum_{i=1}^{v} m_i$, where m_i is the number of snapshots in trajectory i. The Boltzmann weight of each of the m snapshots lying on the FES parametrized by the umbrella variables $\vec{\psi}$ is given by

$$P(\vec{\psi}[i]) = Ce^{-\beta G(\vec{\psi}[i])} = \frac{e^{-\beta G(\vec{\psi}[i])}}{\sum_{i=1}^{v} e^{-\beta G(\vec{\psi}[i])}}, \qquad (10)$$

where $\vec{\psi}[i]$ is the value of the umbrella variables associated with snapshot i, $G(\vec{\psi}[i])$ is the value of the free energy on the FES parametrized by $\vec{\psi}$ at the location $\vec{\psi}[i]$, and where we have chosen the arbitrary multiplicative constant $C = \sum_{i=1}^{v} e^{-\beta G(\vec{\psi}[i])}$ to generate a normalized probability distribution. Considering the ensemble of m snapshots as a *discrete state space* distributed over the region of phase space explored by the umbrella sampling runs, we define the integer *multiplicity* of snapshot i as the expected number of times the snapshot would be visited in a sample of size s over the discrete state space,

$$\hat{c}_i = \text{round}[sP(\vec{\psi}[i])], \qquad (11)$$

where $\hat{c}_i$ is rounded to the nearest integer. Snapshots for which $\hat{c}_i = 0$ are removed from the data set, leaving m' snapshots which are re-indexed accordingly. The parameter s may be made arbitrarily large so as to result in arbitrarily few instances of $\hat{c}_i = 0$, such that sufficiently many snapshots from the high free energy barrier regions will be retained within the remaining set of m' snapshots to result in a fully connected ensemble. We define fully connected in the aforementioned sense that any snapshot may be reached from any other by a series of hops of length $\mathcal{O}(\epsilon)$ or less in the pairwise similarity metric between intervening snapshots, where ϵ is sufficiently small to resolve the important features of the underlying FES.

If the set of m – or more precisely m', since the $(m-m')$ snapshots with $\hat{c}_i = 0$ are discarded from the data set—snapshots is too large for efficient matrix storage, the ensemble of m snapshots harvested from the umbrella sampling runs may be subsampled without replacement prior to the computation of snapshot multiplicities. Since the multiplicity of each snapshot is calculated according to its free energy on the FES computed by the self-consistent solution of the WHAM equations over all m snapshots, snapshot multiplicities are computed in exactly the same manner [i.e., using Eqs. (10) and (11)] irrespective of whether or not subsampling is performed. Correspondingly, matrix storage costs may be reduced by subsampling the set of m snapshots according to any $\vec{\psi}$ space distribution that provides adequate coverage of the high free energy barrier regions and results in a sufficiently well-connected ensemble of snapshots.

Similar to the original formulation of the diffusion map approach described in Sec. II A, we now construct a set of m'-by-m' matrices from the m' snapshots,

$$\hat{C}_{ij} = \begin{cases} \hat{c}_i, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}, \qquad (12)$$

$$\hat{A}_{ij} = \exp(-(\text{RMSD}_{ij})^2/2\epsilon), \qquad (13)$$

$$\hat{D}_{ij} = \begin{cases} \hat{A}(i,\cdot) \cdot diag(\hat{\mathbf{C}}), & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}, \qquad (14)$$

$$\hat{\mathbf{M}} = \hat{\mathbf{D}}^{-1}\hat{\mathbf{A}}, \tag{15}$$

for $i, j = 1 \ldots m'$, where $\hat{A}(i, \cdot)$ is the ith row of $\hat{\mathbf{A}}$, $\cdot$ denotes the dot product and $diag(\hat{\mathbf{C}})$ is the main diagonal of $\hat{\mathbf{C}}$. The snapshot multiplicities are explicitly accounted for in the row normalization of the $\hat{\mathbf{A}}$ matrix by the $\hat{\mathbf{D}}$ matrix.

We now proceed to the heart of the adaptation by generating the set of snapshots formed by explicitly replicating each of the m' snapshots by its multiplicity, resulting in an ensemble of $\tilde{m} = \sum_{i=1}^{m'} \hat{c}_i$ snapshots. As we shall discuss further later, this ensemble may be considered an unbiased sampling of the discrete state space, or equivalently a collection of snapshots resulting from the evolution of an unbiased trajectory over the FES parametrized by $\vec{\psi}$. The application of the original diffusion map approach (Sec. II A) to this ensemble of $\tilde{m}$ snapshots would embed the snapshots in the top eigenvectors resulting from the solution of the eigenvalue problem,

$$\mathbf{M}\vec{\phi}_i = \lambda_i \vec{\phi}_i, \tag{16}$$

where $\mathbf{M}$ is the Markov transition matrix constructed over all $\tilde{m}$ snapshots. Motivated by the fact that repeated snapshots are embedded into identical locations, we now demonstrate that this embedding may be constructed by considering a much smaller eigenvalue problem consisting of only the m' unrepeated snapshots.

Without loss of generality, the row and column indices of the $\mathbf{M}$ matrix in Eq. (16) may be arranged such that repeated snapshots are grouped contiguously, giving the matrix and its eigenvectors the block structure illustrated in Fig. 1. All of the elements in each block of $\mathbf{M}$ and $\vec{\phi}_i$ are identical, motivating the construction of a new matrix-vector pair in which each block is collapsed into a single element. In the case of the $\tilde{m}$-by-$\tilde{m}$ $\mathbf{M}$ matrix, this procedure yields precisely the m'-by-m' $\hat{\mathbf{M}}$ matrix. For the $\vec{\phi}_i$ vector of length $\tilde{m}$, this results in the a vector of length m' which we denote $\vec{\zeta}_i$.

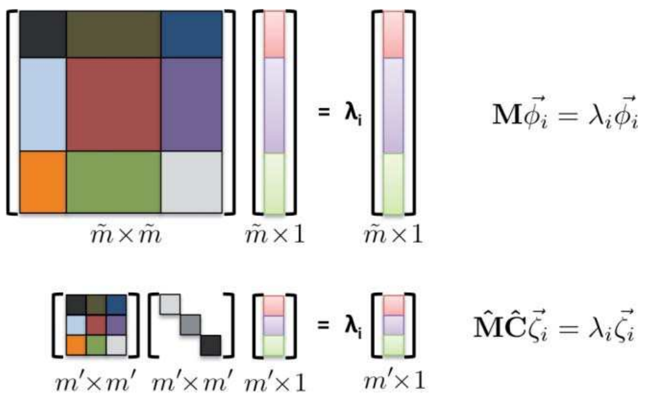
$= \lambda_i$
$\mathbf{M}\vec{\phi}_i = \lambda_i\vec{\phi}_i$
$\tilde{m}\times\tilde{m}$
$\tilde{m}\times 1$
$\tilde{m}\times 1$
$= \lambda_i$
$\hat{\mathbf{M}}\hat{\mathbf{C}}\vec{\zeta}_i = \lambda_i\vec{\zeta}_i$
$m'\times m'$
$m'\times m'$
$m'\times 1$
$m'\times 1$

FIG. 1. Multiplicity within an ensemble of $\tilde{m}$ snapshots leads to a block structure in the $\tilde{m}$-by-$\tilde{m}$ $\mathbf{M}$ matrix and its eigenvectors $\vec{\phi}_i$. The value of all elements within a block are identical. As illustrated for the case of $m' = 3$ independent snapshots, the block structure may be exploited to construct the m'-by-m' $\hat{\mathbf{M}}$ matrix and diagonal $\hat{\mathbf{C}}$ matrix. The eigenvalues of $\hat{\mathbf{M}}\hat{\mathbf{C}}$ are identical to the nonzero eigenvalues of $\mathbf{M}$, and the eigenvectors $\vec{\zeta}_i$ are identical to $\vec{\phi}_i$ with repeated elements removed.

Exploiting the block structure, Eq. (16) may be rewritten as

$$\begin{aligned}
\lambda_i \vec{\phi}_i[k] &= \sum_{j=1}^{\tilde{m}} M(k, j)\vec{\phi}_i[j] \\
&= M(k, 1)\vec{\phi}_i[1] + M(k, 2)\vec{\phi}_i[2] \\
&\quad + \cdots + M(k, \tilde{m})\vec{\phi}_i[\tilde{m}] \\
&= \sum_{\text{block}=1}^{m'} \hat{c}_{\text{block}} M(k, \text{block})\vec{\phi}_i[\text{block}],
\end{aligned} \tag{17}$$

where the sum in the final line is indexed over the column blocks of the $\mathbf{M}$ matrix and the row blocks of the $\vec{\phi}_i$ vector. $\hat{c}_{\text{block}}$ denotes the multiplicity of the snapshot associated with the current block, $M(k, \text{block})$ represents the identical value of all the elements in this block of the $\mathbf{M}$ matrix, and $\vec{\phi}_i[\text{block}]$ is the identical value of the elements in the corresponding block of the $\vec{\phi}_i$ vector. This operation may therefore be conceived as collapsing the columnwise block structure of the $\mathbf{M}$ matrix. Now proceeding to collapse the rowwise block structure, we index over the elements of $\vec{\zeta}_i$ rather than those of $\vec{\phi}_i$,

$$\begin{aligned}
\lambda_i \vec{\zeta}_i[k] &= \sum_{j=1}^{m'} \hat{c}(j)\hat{M}(k, j)\vec{\zeta}_i[j] \\
&= \sum_{j=1}^{m'} \hat{C}(j, j)\hat{M}(k, j)\vec{\zeta}_i[j],
\end{aligned} \tag{18}$$

$$\Rightarrow \hat{\mathbf{M}}\hat{\mathbf{C}}\vec{\zeta}_i = \lambda_i \vec{\zeta}_i. \tag{19}$$

The identity of $\vec{\zeta}_i$ as the vector in which each block of $\vec{\phi}_i$ is collapsed into a single element, is enforced by the normalization condition,

$$\vec{\phi}_i \cdot \vec{\phi}_i = \hat{\mathbf{C}}\vec{\zeta}_i \cdot \vec{\zeta}_i = 1. \tag{20}$$

Due to repeated rows, the $\tilde{m}$-by-$\tilde{m}$ $\mathbf{M}$ matrix is rank deficient, possessing m' nonzero eigenvalues. The m' eigenvalues of $\hat{\mathbf{M}}\hat{\mathbf{C}}$ [Eq. (19)] are identical to the m' non-zero eigenvalues of $\mathbf{M}$ [Eq. (16)].

The adapted diffusion map embedding of the m' snapshots proceeds in an analogous manner to that described by Eq. (4), in which the ith snapshot is mapped into the ith components of the top k nontrivial eigenvectors,

$$\text{snapshot}_i \mapsto (\vec{\zeta}_2(i), \vec{\zeta}_3(i), \ldots \vec{\zeta}_{k+1}(i)), \tag{21}$$

where an appropriate value of k may be determined by a gap in the eigenvalue spectrum.

Since identical snapshots are embedded into identical locations, the embedding of the m' independent snapshots by this adaptation of the diffusion map approach [Eq. (21)] is identical to the embedding of the $\tilde{m} = \sum_{i=1}^{m'} \hat{c}_i$ snapshots by the original diffusion map approach, where each snapshot is explicitly replicated by its multiplicity [Eq. (4)]. The multiplicities of the snapshots in the embedding of the m' snapshots are of use in the construction of appropriately weighted free energy surfaces.

Conceptually, it may be useful to consider of the set of $\tilde{m}$ snapshots as having been generated from a sufficiently long unbiased trajectory over the discrete state space of m' snapshots drawn from the umbrella sampling simulations. If the variables in which the umbrella sampling was conducted completely parametrize the slow dynamical motions of the system—an assumption addressed in Sec. II B 3—then in the limit of infinitely dense umbrella sampling (i.e., $m' \rightarrow \infty$), the set of $\tilde{m}$ snapshots may be considered to have been drawn from the evolution of a hypothetical continuous dynamical trajectory over the unbiased FES computed from the umbrella sampling data. Considering this FES as a reconstruction of the intrinsic manifold of the molecular system, the trajectory effectively describes the evolution of the molecular system in its fundamental dynamical motions. Conceiving of the original $\tilde{m}$-by-$\tilde{m}$ eigenvalue problem as the application of the original diffusion map approach to this hypothetical trajectory, we have demonstrated that this is related to a far smaller, and therefore less computationally expensive, m'-by-m' eigenvalue problem which results in an identical low-dimensional embedding of the data.

To summarize this section, we have presented an adaptation of the diffusion map approach to permit its application to biased simulation data. By reconstructing the unbiased, underlying FES from the biased data, we appropriately reweighted each of the data points to generate a new ensemble, which may be thought of as resulting from an unbiased trajectory over the FES. We then reformulated the diffusion map approach to permit its application to this new ensemble in a computationally efficient manner. This adaptation of the diffusion map approach facilitates its application to biased simulations of systems possessing high free energy barriers, where unbiased sampling would typically result in poor characterization of the barrier regions and difficulties in the construction of globally valid low-dimensional parametrizations.

3. *The umbrella bootstrapped diffusion map approach*

A key assumption of the methodology presented in Sec. II B 2 is that the variables in which the umbrella sampling is conducted, and which parametrize the FES resulting from the solution of the WHAM equations, are good variables with which to characterize the important dynamical motions of the system. Specifically, it is assumed that there are no additional slow variables along which the system exhibits significant motions that are projected out in the construction of the FES. This assumption is critical for the appropriate reweighting of the snapshots in the reformulated eigenvalue problem encapsulated in Eq. (19), and failure of this assumption can lead to diffusion map embeddings which do not correspond to the intrinsic manifold of the system.

A set of variables parametrizing the slow, dynamical motions of the system, and in which to perform the umbrella sampling, is rarely known *a priori*. Indeed the ultimate goal of the methodology presented herein is to systematically identify good global order parameters with which to parametrize these modes. Accordingly, we propose a bootstrapped approach in which we alternate between rounds of umbrella sampling and diffusion mapping until we converge upon an appropriate set of order parameters. This process is illustrated in Fig. 2.

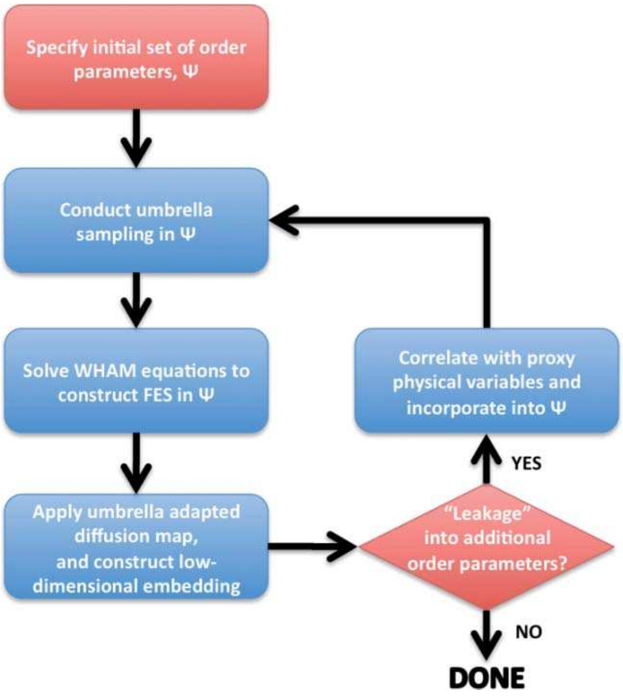
Specify initial set of order parameters, Ψ
Conduct umbrella sampling in Ψ
Solve WHAM equations to construct FES in Ψ
Apply umbrella adapted diffusion map, and construct low-dimensional embedding
"Leakage" into additional order parameters?
YES
Correlate with proxy physical variables and incorporate into Ψ
NO
DONE

FIG. 2. Flowchart illustrating the umbrella bootstrapped diffusion map approach.

The protocol commences by specifying a set of starting variables, $\vec{\psi}^{\,0}$, in which to conduct the first round of umbrella sampling. To be implemented in the umbrella sampling simulations, these variables must be some function of the atomic coordinates of the system, but need not be particularly good descriptors of the underlying motions of the system since this procedure will guide us toward such a set. While such variables may be determined heuristically from intuition, or from prior knowledge of the system, a more systematic approach would be to apply dimensionality reduction—PCA, for example—to an initial, short, unbiased molecular simulation trajectory.

We then perform the first round of umbrella sampling in $\vec{\psi}^{\,0}$, solve the multidimensional WHAM equations to synthesize the FES parametrized by $\vec{\psi}^{\,0}$, and apply the adaptation of the diffusion map approach outlined in Sec. II B 2 to the data. By construction, the system will exhibit good sampling along the umbrella variables in which the system was driven, with spontaneous thermal fluctuations allowing the system to explore coordinates orthogonal to these variables. If any of these orthogonal coordinates describe a slowly evolving mode of the system, and provided the free energy barriers are not too high, the simulation will naturally "leak" along these directions. The existence of such "leakage" will be manifest as the emergence of pathways in the diffusion map embedding which are described by additional order parameters beyond those in which the umbrella sampling was conducted. By *correlating the emergent order parameters with physical variables*, the diffusion map approach suggests additional physical variables along which to extend the exploration of phase space by their incorporation into successive rounds of umbrella sampling.

Furthermore, inspection of the FES parametrized by $\vec{\psi}^{\,0}$, and construction and inspection of the FES parametrized by $\vec{\psi}^{\,0}$ plus any emergent order parameters using Eq. (9), can reveal whether the initial choice of order parameters was a good one. For example, suppose the initial order parameters in $\vec{\psi}^{\,0}$ describe fast dynamical motions which are effectively slaved to the slow modes of the system. In performing the umbrella sampling in these variables, the simulation will be driven by spontaneous thermal fluctuations to leak into those coordinates corresponding to its slow motions, and the diffusion map embedding will reveal the emergence of pathways associated with these slow modes. Driving the system orthogonal to the slow attractive manifold will lead to large increases in the free energy, and consequently, the FES parametrized by $\vec{\psi}^{\,0}$ will have high free energy everywhere outside of a narrow region corresponding to the slow manifold. This signature would suggest that the order parameters in $\vec{\psi}^{\,0}$ are poor descriptors of the slow motions and should be discarded in favor of those parametrizing the slow manifold in the next round of umbrella sampling.

By discarding poor variables in $\vec{\psi}^{\,0}$ and adding emergent order parameters suggested by the diffusion map, we generate a new set of order parameters in which to conduct a new round of umbrella sampling, which we denote $\vec{\psi}^{\,1}$. Alternating between umbrella sampling and diffusion mapping in this fashion, we repeat this process until no further order parameters emerge. At this point, the final set of order parameters, $\vec{\psi}^{\,\mathrm{FINAL}}$, may be considered a good set of global variables with which to parametrize the FES and characterize the fundamental dynamical motions of the system.

The calculation of the Boltzmann weights of the snapshots in the biased ensemble in the diffusion map adaptation described in Sec. II B 2, assumes that no additional slow motions exist in coordinates orthogonal to the FES used to perform the reweighting. The convergence of the iterative procedure toward the FES parametrized by $\vec{\psi}^{\,\mathrm{FINAL}}$ substantiates this assumption, and suggests that the diffusion map embeddings of the umbrella sampling trajectories conducted in $\vec{\psi}^{\,\mathrm{FINAL}}$ represent good reconstructions of the intrinsic manifold of the system. The final umbrella sampling trajectories provide a set of snapshots drawn from a distribution with good sampling of the accessible phase space, from which converged thermodynamic averages may be extracted.

As described in Sec. II B 1, the umbrella sampling restraining potentials may only be constructed in collective variables which are known functions of the atomic coordinates of the system. As is typical of nonlinear dimensionality reduction techniques, the functional dependences for the order parameters identified by the diffusion map are not explicitly furnished by the approach. Consequently, umbrella sampling may not be performed directly in these variables and must instead be performed in proxy physical variables (e.g., torsional angles, interatomic distances, and local densities) whose dependence of the atomic coordinates of the system is known. A crucial step, therefore, in the bootstrapped diffusion map approach (Fig. 2) is the successful correlation of the variables identified by the diffusion map with physical variables, and in the event that physical variables exhibiting good correlation with the diffusion map variables cannot be found, the approach fails. As we have previously noted,[13] systematic approaches to screen candidate physical variables, such as those developed by Peters, Beckham, and Trout[35,36] and Ma and Dinner,[2] may be of use in ameliorating this shortcoming.

At this juncture, we note that in contrast to nonlinear techniques, the functional dependence upon the atomic coordinates of parametrizations resulting from linear dimensionality reduction—the top principal components in PCA, for example—is explicitly known, thereby permitting umbrella sampling to be performed directly in the low-dimensional order parameters. Within the framework of our iterative approach, this feature would provide a significant advantage to a reformulation of PCA to operate on biased data over the adapted diffusion map approach. However, since highly nonlinear intrinsic manifolds are poorly parametrized by hyperplane approximations, the inherent linearity of PCA typically results in low-dimensional embeddings requiring more dimensions than the corresponding embeddings synthesized by nonlinear techniques.[7,13] Since the number of independent umbrella sampling simulations over a grid of a specified resolution increases exponentially with grid dimensionality, the capacity of the diffusion map approach to generate economical low-dimensional parametrizations of a system arguably outweighs the absence of an explicit functional dependence of the diffusion map variables upon the atomic coordinates of the system.

C. Molecular simulations of alanine dipeptide

All simulations were conducted using the GROMACS 4.0.2 molecular dynamics simulation suite[37,38] in which alanine dipeptide was modeled using the all-atom OPLS-AA/L force field[39,40] residing in a cubic box of side length 3 nm containing 868 TIP3P water molecules[41] at a density of 0.979 g/cm^3. Periodic boundary conditions were implemented in all dimensions. The Dundee PRODRG2 server was used to construct the initial coordinates of the peptide.[42] Peptide bond lengths were fixed to permit a larger integration timestep. The Lennard-Jones interactions were smoothly switched to zero at a cutoff of 1.3 nm. The particle-mesh Ewald method[43] with a real space cutoff of 1.4 nm and a reciprocal space grid spacing of 0.12 nm was used to evaluate the electrostatic interactions. The box side length was sufficiently large to preclude nonbonded interactions with multiple periodic images. Simulations were conducted in the NPT ensemble at 298 K and 1 bar, using a Nosé-Hoover thermostat[44,45] and Parrinello-Rahman barostat.[46] The equations of motion were integrated using the leap-frog algorithm with a time step of 2 fs. High energy overlaps in the initial configuration were removed by steepest descent energy minimization, which was followed by 10 ps of simulation in which the peptide was held fixed to permit solvent relaxation. The peptide was then released and the entire system was allowed to relax over the course of a 50 ps equilibration run. The final configuration of this run was used as the initial state for the unbiased molecular dynamics and biased umbrella sampling simulations described in Sec. III, during which snapshots were saved every 2 ps. Umbrella sampling is natively supported by the GROMACS simulation engine, as is also the case for many other molecular simulation packages, including AMBER[47] and CHARMM.[48]

Each 1 ns umbrella sampling simulation required approximately 2 h on a 2.34 GHz Intel Core 2 Duo processor. For the final four-dimensional round of umbrella sampling (Sec. III D), solution of the WHAM equations over the 1296 umbrella runs required 20 min on one core of a 2.66 GHz Dual-Core Intel Xeon. Construction of the 36,786-by-36,786 **M** matrix was dominated by the computation of the pairwise RMSD distances, requiring a total of 27 h on the same processor. Computation of the eigenvectors of **M** was conducted in 2 h on three 2.77 GHz Intel Core 2 Quad processors (12 cores) using the implicitly restarted Arnoldi method provided in the PARPACK library.[49]

III. RESULTS AND DISCUSSION

We now demonstrate the umbrella bootstrapped diffusion map approach by applying it to simulations of alanine dipeptide (*N*-acetyl- L-alanine-*N*′-methylamide, AcAlaNHMe) in explicit solvent (Fig. 3). This has become the *de facto* test system for new biophysical molecular simulation techniques[2,7,50–59] and is known to exhibit moderately high ($\sim$10 $k_B T$) free energy barriers.[21,60] The approach furnishes a low-dimensional diffusion map embedding depicting the dynamic connectivity between metastable states, providing a globally meaningful "skeleton" of the topology of the intrinsic manifold. Furthermore, this embedding is of particular use in guiding the subsequent analysis of individual state-to-state transitions, a demonstration of which we provide in the supplementary material[61] by the application of local PCA to a selected isomerization pathway.

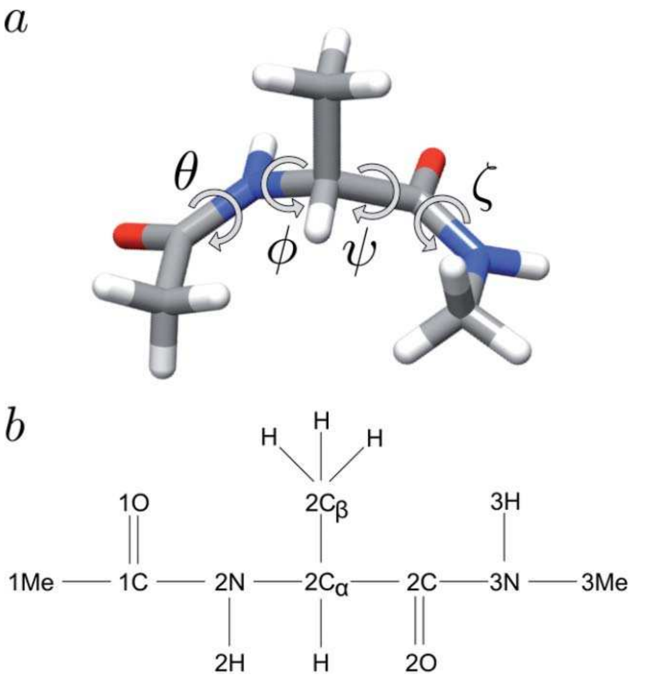

FIG. 3. Alanine dipeptide. (a) Molecular representation illustrating the four dihedral angles θ (1O-1C-2N-2C$_\alpha$), ϕ (1C-2N-2C$_\alpha$-2C), ψ (2N-2C$_\alpha$-2C-3N), and ζ (2C$_\alpha$-2C-3N-3H). The dihedral angle is defined as the right-handed rotation around the vector connecting the second atom to the third; 0° corresponds to the *cis* conformation. (b) Schematic representation with atomic labels using a naming scheme similar to that employed by Ma and Dinner (Ref. 2).

It is well known the backbone ϕ and ψ dihedral angles of alanine dipeptide (Fig. 3) are good variables with which to parametrize the solvated phase FES, but they have been shown to be inadequate to characterize the transition state ensemble of at least one isomerization transition.[2,56,57] In the terminology of Du *et al.* (Sec. II A), they provide a good parametrization of the *transition coordinates*, but a poor characterization of the *reaction coordinates*.[15] While we anticipated the recovery of the ϕ and ψ angles in our final set of order parameters, $\vec{\psi}^{\text{ FINAL}}$, our results suggest that these two dihedrals should be supplemented by the θ and ζ backbone dihedrals to provide an essentially complete parametrization of the important dynamical motions.

A. Determination of initial umbrella variables $\vec{\psi}^{\,0}$

In the first step of our procedure, we conducted a 50 ns unbiased molecular dynamics simulation at 298 K and 1 bar as described in Sec. II C to determine an initial set of order parameters in which to conduct the first round of umbrella sampling. Recording snapshots every 2 ps, we applied the diffusion map in its original formulation to the 25000 snapshot trajectory, employing a value of $\epsilon = 3.35 \times 10^{-4}$. The appropriate range of ϵ values for each system was calculated as described in Sec. II A, and in all cases we elected to employ ϵ values toward the lower end of this range. Regarding ϵ as the bandwidth of the Gaussian kernel with which the pairwise RMSD distances between snapshots are combined [Eq. (1)], RMSD$_{ij}$ values larger than ϵ are effectively discarded by the application of the kernel. Accordingly, a small ϵ value reduces the size of the neighborhood "seen" by each snapshot, and leads to low-dimensional embeddings with enhanced separation of the metastable states and the transition paths between them. As will become apparent in the figures below, this choice results in diffusion map embeddings in which the metastable states typically appear as vertices linked by essentially one-dimensional, linear pathways.

A gap in the eigenvalue spectrum [Fig. S1 (Ref. 61)] suggested that embeddings be constructed in the top two nontrivial eigenvectors *evec2* and *evec3*. Two-dimensional embeddings of this simulation trajectory are presented in Fig. 4, where the points in the embedding are colored according to the ϕ and ψ backbone angles. As illustrated in Fig. 4(a), *evec3* and ϕ are negatively correlated, possessing a correlation coefficient of −0.58. Conversely, Fig. 4(b) illustrates the positive correlation between *evec2* and ψ, which possess a correlation coefficient of 0.78. A Ramachandran plot of the trajectory (Fig. 5) shows that the simulations were restricted to the upper left corner of the plot due to the presence of high free energy barriers which prevent good sampling of phase space by unbiased molecular dynamics simulations.

We now wish to employ biased simulations to extend our sampling of phase space into the surrounding region, with the correlation of the ϕ and ψ dihedrals with the diffusion map

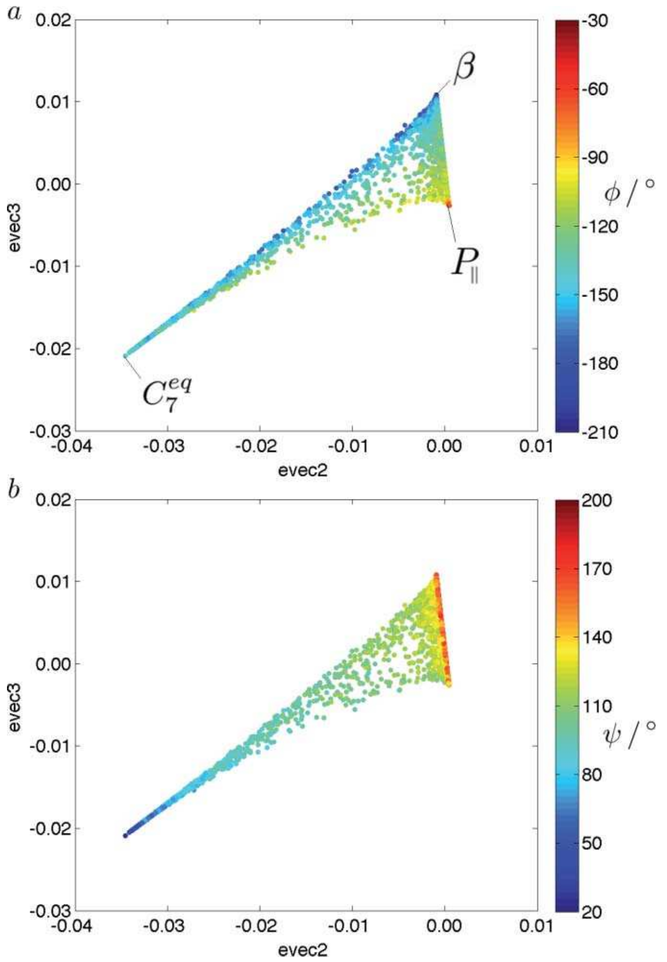

FIG. 4. Diffusion map embedding of the initial 50 ns unbiased molecular dynamics trajectory into the top two eigenvectors *evec2* and *evec3*. Data points are colored according to the (a) ϕ and (b) ψ dihedral angles. The labels in (a), which also apply to panel (b), correspond to the metastable conformational basins depicted in the ϕ/ψ FES in Fig. 6 with the θ and ζ dihedrals occupying the $-180/180^{\circ}$ position. *Evec2* and *evec3* show good correlation with the ψ and ϕ dihedral angles, respectively. As shown in Fig. 5, the simulation sampled ϕ=[[-180°, -34°],[151°, 180°]] and ψ=[[-180°, -164°],[24°, 180°]]. The 360° periodicity of the dihedral angles facilitated a better demonstration of the *evec3*-ϕ correlation in (a) by subtracting 360° from those ϕ angles in the range [151°, 180°] shifting them to [-209°, -180°], and permitting the explored ϕ angles to be represented as a continuous range spanning [-209°, -34°]. Similarly, to better illustrate the *evec2*-ψ correlation in (b), 360° was added to those ψ angles in the range [-180°, -164°] shifting them to [180°, 196°] and permitting the explored ψ angles to be represented as a continuous range spanning [24°, 196°].

variables suggesting that the first round of umbrella sampling be conducted in $\vec{\psi}^{\,0} = [\phi, \psi]$. The question remains, however, of how far out we should extend the umbrella sampling. In this case, since both umbrella variables are angles spanning [-180°, 180°) with 360° periodicity, we elect simply to perform the umbrella sampling over the entire domain.

B. Round 0

The initial umbrella sampling runs were performed in $\vec{\psi}^{\,0} = [\phi, \psi]$ by partitioning the two-dimensional phase space with 360° periodicity in each dimension into a 20°-by-20° square grid. An 1 ns biased simulation was run on each of the 324 vertices, employing a two-dimensional harmonic restraining potential in ϕ and ψ around each vertex, with a force constant of 100 kJ/mol rad^2 in each dimension. The grid spacing was specified to provide adequate coverage of the phase space in a computationally manageable number of simulations, and the restraining potential force constant tuned to permit sufficient overlap between the trajectories at neighboring grid points to allow solution of the WHAM equations[31] (Sec. II B 1). Solving these equations for the set of 324 umbrella simulations resulted in the FES in ϕ/ψ presented in Fig. 6, in which seven metastable basins are apparent, showing some consistency with prior simulations of alanine dipeptide using the OPLS-AA force field with an implicit solvent model.[21,62] We then subsampled the umbrella sampling data set by retaining only every tenth snapshot to generate a reasonably sized ensemble of 32 432 snapshots which spanned the ϕ/ψ FES. Employing Eqs. (10) and (11), we chose to generate a

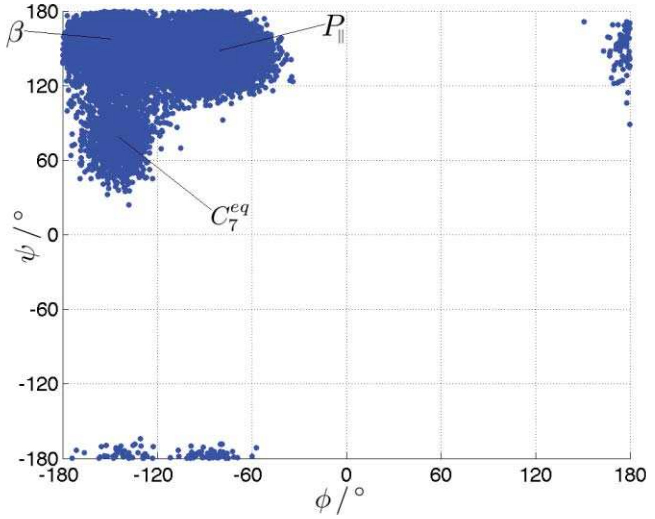

FIG. 5. A Ramachandran plot of the 25 000 snapshots constituting the 50 ns unbiased molecular dynamics trajectory. The presence of the high free energy barriers apparent in Fig. 6 restricted the simulation to explore only the upper left corner of the plot. The labels correspond to the metastable conformational basins depicted in the ϕ/ψ FES in Fig. 6 with the θ and ζ dihedrals occupying the $-180/180^{\circ}$ position.

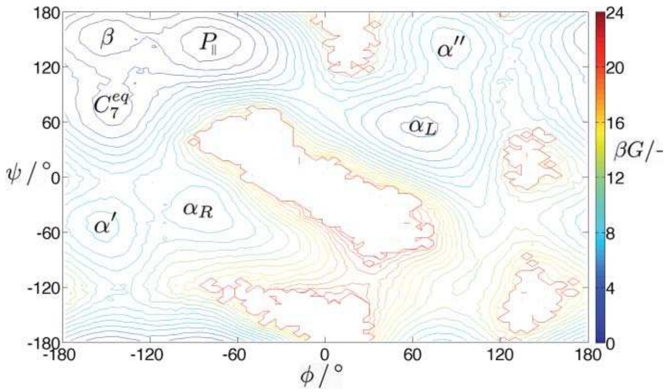

FIG. 6. The ϕ/ψ FES resulting from the Round 0 umbrella sampling simulations in $\vec{\psi}^{\,0} = [\phi, \psi]$. Contours in βG, where $\beta = 1/k_B T$ and G is the Gibbs free energy, are plotted in unit steps from $\beta G = 0$ to 22. Basin labels are assigned on the basis of the ϕ and ψ dihedral angles, and indicate the metastable conformational states.

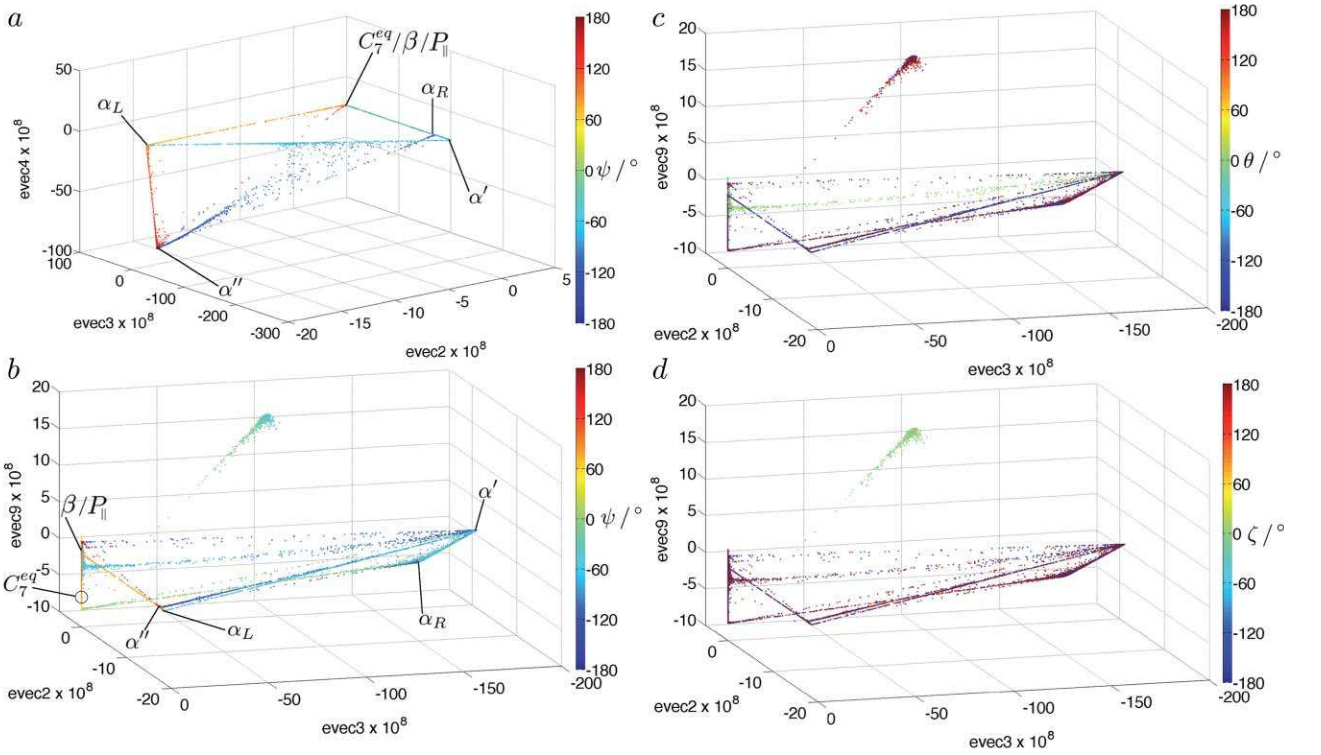

FIG. 7. Embeddings of the Round 0 umbrella sampling simulations in $\vec{\psi}^{\,0}=[\phi,\psi]$. (a) An embedding in [*evec2*, *evec3*, *evec4*] where the data points are colored according to the ψ dihedral angle. (b,c,d) Embeddings in [*evec2*, *evec3*, *evec9*] where the data points are colored according to the (b) ψ, (c) θ and (d) ζ dihedral angles. The labels in (a) and (b) correspond to the metastable conformational basins depicted in the ϕ/ψ FES in Fig. 6 with the θ and ζ dihedrals occupying the $-180/180°$ position. Since the C_7^{eq} state in panel (b) does not lie at a vertex, the appropriate region on the line segment is indicated by a lollipop rather than a line. The labeling in (b) also applies to panels (c) and (d). Observe that different combinations of eigenvectors provide good parametrizations for different dynamical transitions, and that the simulations have "leaked" into the θ and ζ coordinates.

sample of size $s=1\times10^{16}$ over the discrete state space, in which zero multiplicity, $\hat{c}_i=0$, was assigned to only the 101 highest free energy snapshots. The adapted diffusion map approach was applied to the remaining 32 331 snapshots with $\epsilon=3.35\times10^{-4}$.

The resulting eigenvalue spectrum was found to exhibit a gap after the tenth eigenvalue [Fig. S2 (Ref. 61)], suggesting that embeddings be constructed in the top nine nontrivial eigenvectors. While this result implies that a nine-dimensional parametrization is required to provide an adequate global embedding of the trajectory, an analysis of three-dimensional projections of the embedding in triplets of eigenvectors reveals that different eigenvectors serve as good order parameters for different dynamic transitions between the metastable states illustrated in Fig. 6. For example, Fig. 7(a) presents a three-dimensional projection of the embedding in [*evec2*, *evec3*, *evec4*] in which the data points are colored according to the ψ angle, and which provides a good parametrization for the $\alpha_L\leftrightarrow\alpha''$ transition, but not the transitions involving the C_7^{eq}, β and $P_{||}$ basins, which are collapsed together in this projection. Conversely, Fig. 7(b) presents an embedding in [*evec2*, *evec3*, *evec9*], which is a better parametrization for the $\alpha_R\leftrightarrow C_7^{\text{eq}}$ transition, but results in poorer separation of the α_L and α'' states.

The jump from a two-dimensional intrinsic manifold in the determination of good initial umbrella sampling variables in Sec. III A, to a nine-dimensional manifold in the present case, apparently arises from driving the system to explore many additional metastable states, the transitions between which are parametrized by additional dynamical variables. If we had instead extended out the sampling only slightly beyond the region explored by the system (Fig. 5), rather than performing umbrella sampling over the entire periodic ϕ/ψ domain, we anticipate that exploration of only a few additional metastable states would have led to the emergence of fewer additional variables, and the commensurate increase in measured system dimensionality in a more gradual manner.

We now search for "leakage" of the $\vec{\psi}^{\,0}=[\phi,\psi]$ umbrella simulations into variables other than those in which the sampling was conducted. Specifically, we look for pathways in the diffusion map embedding which are not well-characterized by either of these two dihedral angles, where the existence of such pathways would indicate that the system has evolved along additional slow coordinates. We would like to correlate such motions with physical order parameters, to permit umbrella sampling in these directions in a subsequent round of simulations. In embeddings constructed in [*evec2*, *evec3*, *evec9*], we observe excursions of the system along pathways characterized by rotations around the θ and ζ backbone dihedrals (Fig. 3), as illustrated in Figs. 7(c) and 7(d) in which the data points are colored according to the θ and ζ angles, respectively. Since θ and ζ appear to be good descriptors of additional slow dynamical motions, we include them in the next round of umbrella sampling.

C. Round I

We elected to incorporate the θ and ζ angles independently in this round of umbrella sampling, first sampling in $\vec{\psi}^{\,1a}=[\phi,\psi,\theta]$. The three-dimensional phase space was partitioned into a 40°-by-40°-by-40° cubic grid, and 1 ns simulations run on each of the 729 vertices, restrained to the locale of the grid point by a harmonic biasing potential with a force constant of 50 kJ/mol rad^2 in each dimension. Compared to Round 0, the increased dimensionality of the phase space demanded the use of a coarser grid, and correspondingly looser restraining potentials, to permit coverage of the space with a practicable number of simulations. Retaining every twentieth snapshot and choosing to generate a sample of size $s=1\times 10^{16}$ over the discrete state space, the 1214 highest free energy snapshots were assigned multiplicities of zero, leaving an ensemble of 35 272 snapshots to which we applied the adapted diffusion map approach with $\epsilon = 5.53\times 10^{-4}$. A spectral gap after the twelfth eigenvalue suggested embeddings be constructed in the top eleven non-trivial eigenvectors [Fig. S3 (Ref. 61)]. (Note that, in principle, the ensemble of snapshots generated by this round of umbrella sampling could have been supplemented with that from the previous round, but to avoid the complexities of dealing with data sampled over differently spaced grids of different dimensionalities, we elected not to mix data from different rounds of umbrella sampling in this demonstration of the approach.)

Compared to the Round 0 sampling in $\vec{\psi}^{\,0}=[\phi,\psi]$, in the present round we additionally drive the system in the θ dihedral angle. This results in the splitting of each of the metastable basins identified in Fig. 6 into states with $\theta=-180/180°$ (recalling that dihedral angles have 360° periodicity) and $\theta=0°$ [Fig. S4c (Ref. 61)]. The sparsity of intermediate values is attributable to the planar character of the peptide bond described by this dihedral, which exhibits a bimodal distribution of angles centered around the energetically favorable $\theta=-180/180°$ and $\theta=0°$ planar orientations.

Coloring the points in the embedding according to their free energies on the $\phi/\psi/\theta$ FES computed by solution of the WHAM equations [Fig. S4e (Ref. 61)] suggests that the $C_7^{\mathrm{eq}}\leftrightarrow\alpha_R$ transition with the θ dihedral locked into the $-180/180°$ position is the low free energy pathway for this isomerization, although other routes involving the $\theta=0°$ dihedral position do exist. The paucity of snapshots linking the $\theta=-180/180°$ and $\theta=0°$ isomers of α_R indicate that rotations around the θ dihedral with the ϕ and ψ dihedrals in a configuration corresponding to the α_R state ([($\phi\approx-90°$, $\psi\approx-30°$), see Fig. 6] is highly unfavorable, likely due to the sterics of this conformational transition.

This umbrella sampling run did not show any leakage into the ζ coordinate, with all snapshots from the ensemble exhibiting $\zeta=-180/180°$ dihedral values [Fig. S4d (Ref. 61)]. This suggests that the leakage into the ζ coordinate in the $\vec{\psi}^{\,0}=[\phi,\psi]$ umbrella run occurred at a high free energy region in ϕ/ψ space, unsampled in the $\vec{\psi}^{\,1a}=[\phi,\psi,\theta]$ runs, which implemented looser restraining potentials and a coarser grid over the higher dimensional phase space. In this regard, the $\vec{\psi}^{\,1a}=[\phi,\psi,\theta]$ umbrella sampling simulations and associated diffusion map embedding are self-consistent, and represent a good parametrization of the system. Nevertheless, due to the ζ leakage observed in Round 0, we proceed to conduct umbrella sampling incorporating the ζ dihedral.

We conducted a second round of three-dimensional umbrella sampling in an identical manner to before, but now in $\vec{\psi}^{\,1b}=[\phi,\psi,\zeta]$. As demonstrated in Fig. S5 (Ref. 61) by the reparametrization of the WHAM FES into $\phi/\psi/\theta$ using Eq. (9), although most snapshots in the ensemble possess $\theta=-180/180°$ values, the fact that some do possess dihedrals in the $\theta=0°$ position indicates the existence of excursions in this coordinate. Since equally tight restraining potentials were employed in the umbrella sampling conducted in $\vec{\psi}^{\,1a}=[\phi,\psi,\theta]$ and $\vec{\psi}^{\,1b}=[\phi,\psi,\zeta]$, the absence of leakage in the ζ coordinate in the former runs, combined with the observed leakage in θ in the latter, suggests that rotations around the θ dihedral angle are associated with a lower free energy barrier than those around the ζ angle.

In both of the umbrella runs in this section, different combinations of eigenvectors were found to provide good embeddings for different dynamical transitions. Analysis of each individual transition demonstrated that each could be well-described by motions in one or more of the four dihedral angles ϕ, ψ, θ and ζ, [see caption of Fig. S4 (Ref. 61) for a specific example] suggesting that while relatively many order parameters are required for a complete global parametrization of the phase space—eleven in the case of the $\vec{\psi}^{\,1a}=[\phi,\psi,\theta]$ umbrella simulations—the underlying dynamical transitions may be locally characterized by motions in these four dihedral angles. It may initially seem counter-intuitive that an eleven-dimensional space is required to describe a system in which the state-to-state transitions are adequately described by four parameters, but we must recognize that the precise dynamical motions along the transition pathways parametrized by the eleven diffusion map variables are unknown, possibly highly nonlinear, functions of the atomic coordinates of the system. We simply observe that these transition pathways are all reasonably well-parametrized by changes in the four dihedral angles ϕ, ψ, θ, and ζ, whereas the actual structural transitions along these pathways are almost certainly far more complicated motions, possibly involving couplings between dihedral and planar angles, interatomic distances, sidechain rotameric states, and collective solvent coordinates.[2]

D. Round II

We next performed four-dimensional umbrella sampling in $\vec{\psi}^{\,2}=[\phi,\psi,\theta,\zeta]$. The phase space was partitioned into a 60°-by-60°-by-60°-by-60° grid comprising 1296 vertices. A 1 ns simulation was conducted at each vertex, employing a harmonic biasing potential centered on the grid point with a force constant of 34 kJ/mol rad^2 in each of the four dimensions. In response to the increased phase space dimensionality with respect to the previous round of umbrella sampling, the grid was once again coarsened and the restraining potentials loosened. We retained every thirtieth snapshot of the subsequent trajectories and chose to generate a sample of size $s=1\times 10^{18}$ over the discrete state space. The 6457 highest free energy snapshots were assigned zero multiplicity, resulting in an ensemble of 36,786 snapshots to which the adapted diffusion map approach was applied with $\epsilon = 3.35 \times 10^{-4}$. A spectral gap after the tenth eigenvalue [Fig. S6 (Ref. 61)] suggested that embeddings be constructed in the top nine non-trivial eigenvectors.

As in Round I, all individual transitions could be parametrized by motions in one or more of the four dihedral angles. This suggests that we terminate the umbrella bootstrapped diffusion map approach at this round, since we have generated a self-consistent description of the underlying dynamics: four-dimensional umbrella sampling in ϕ, ψ, θ, and ζ resulted in a diffusion map embedding in which each state-to-state transition may be parametrized by this same set of physical variables (Fig. 2). To illustrate that state-to-state transitions within a selected subset of metastable states may be characterized by motions in one or more dihedral angle, in Fig. 8 we present embeddings of the data in [*evec2*, *evec5*, *evec6*] colored according to the ϕ and ψ dihedral angles. For clarity of viewing, only those snapshots with $[90^\circ \leq \theta, \zeta \leq 180^\circ]$ or $[-180^\circ \leq \theta, \zeta < -90^\circ]$ are plotted, where – recalling the bimodal nature of the distribution of these dihedral angles around $-180/180^\circ$ and 0° – we effectively visualize only those snapshots with θ and ζ in the $-180/180^\circ$ position. Additional embeddings visualizing the remainder of the data set are presented in Fig. S7.[61] We recognize that partitioning the data according to the θ and ζ dihedrals to facilitate visualization necessarily obscures state-to-state transitions in these angles, which would be better illustrated in a different representation of the data.

Given the bimodal distribution of the θ and ζ dihedrals around $-180/180^\circ$ and 0°, we wished to determine which of the two states was more stable, finding that the molecular conformations residing within the global free energy minimum possessed both θ and ζ in the $-180/180^\circ$ position. In order to determine whether restricting the θ or ζ dihedral to occupy the less favorable 0° position causes a larger increase in the free energy of the system, we first partitioned the umbrella sampling data into two groups according to the θ dihedral. One group contained those snapshots centered on the 0° position with $[-90^\circ \leq \theta < 90^\circ]$, while the other group contained the remaining conformations with θ values centered around the $-180/180^\circ$ position. By examining the free energies computed by solution of the WHAM equations over the *complete* data set, we found the global free energy minimum of the $\theta = 0^\circ$ group to be $\sim 11\ k_B T$ higher than that of the $\theta = -180/180^\circ$. A similar calculation for the ζ dihedral showed the $\zeta = 0^\circ$ group to be $\sim 14\ k_B T$ higher than that of the $\zeta = -180/180^\circ$ group. This calculation suggests that restricting the θ dihedral to occupy the less favorable 0° position causes a marginally smaller increase in the free energy of the system compared to restricting the ζ dihedral to occupy this position.

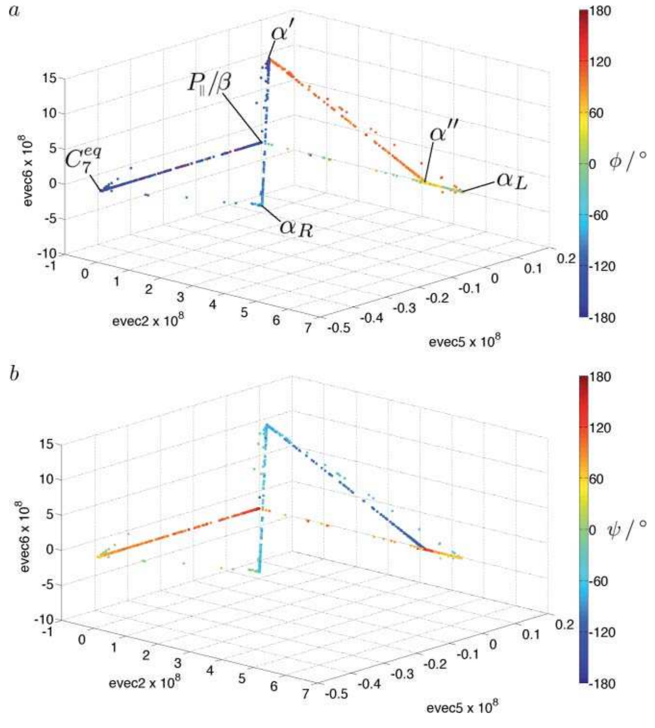

FIG. 8. Embeddings of the Round II umbrella sampling simulations in $\vec{\psi}^2 = [\phi, \psi, \theta, \zeta]$ into [*evec2*, *evec5*, *evec6*]. For clarity of viewing, only those data points with $\theta, \zeta = -180/180^\circ$ dihedral angles are visualized, and are colored according to the (a) ϕ and (b) ψ dihedral angles. The labels in (a) correspond to the metastable conformational basins depicted in the ϕ/ψ FES in Fig. 6 with the θ and ζ dihedrals occupying the $-180/180^\circ$ position. The labeling in (a) also applies to panel (b). Observe that all state-to-state transitions are associated with motions in the ϕ and/or ψ dihedral angles. Additional embeddings considering the data points pertaining to other θ and ζ positions are available in Fig. S7 (Ref. 61).

E. Discussion

The results of Round I suggest that the free energy barrier for the θ dihedral angle to access the less favorable 0° position from the more favorable $-180/180^\circ$ conformation is lower than the corresponding barrier in ζ. Furthermore, the results of Round II indicate that once such excursions occur, peptide conformations locked into the $\theta = 0^\circ$ position are slightly lower in free energy than those locked into the $\zeta = 0^\circ$ state. Our findings are in agreement with the established result the ϕ and ψ backbone dihedrals provide a good two-dimensional parametrization of the dynamical motions of alanine dipeptide. A better three-dimensional description is provided by preferentially incorporating the θ, as opposed to the ζ, dihedral angle, whereas a four-dimensional characterization in ϕ, ψ, θ, and ζ provides an essentially complete parametrization of the fundamental dynamical motions. We observe that umbrella sampling in the two well-known order parameters ϕ and ψ without pursuing the iterative procedure developed in this work would not have revealed the importance of the remaining two dihedrals in characterizing the dynamical motions of the peptide.

Our results demonstrate that the "leakage" of the simulation along order parameters beyond those in which the umbrella sampling was conducted can depend on the coarseness of the umbrella sampling grid and the tightness of the restraining potentials employed. A finer grid packs more simulations more densely over the phase space, and permits the use of tighter restraining potentials which better constrain each simulation around its grid point. Compared to a coarser grid with looser potentials, this arrangement leads to better sampling of high free energy regions, potentially enabling the simulation to "leak" from such high free energy locales into regions of phase space which would be otherwise inaccessible. Specifically, the relatively fine grid and tight restraining potentials employed in the Round 0 umbrella simulations in $\vec{\psi}^{\,0} = [\phi, \psi]$ permitted the simulation to leak into both the θ and ζ coordinates, whereas the Round I umbrella simulations in $\vec{\psi}^{\,1a} = [\phi, \psi, \theta]$, employing a coarser grid and looser potentials, did not exhibit any ζ leakage. Although this observation would suggest that an exhaustive application of the umbrella bootstrapped diffusion map approach would require infinitesimally spaced simulations each employing infinitely tight biasing potentials, only the relatively low lying regions of the FES are typically of interest since spontaneous thermal fluctuations permit the molecular system to surmount only those barriers a few $k_B T$ in height, leaving higher free energy regions effectively inaccessible. Although such barriers are sufficiently high to trap molecular simulation trajectories and frustrate unbiased sampling of phase space, they are low enough to permit adequate sampling of the accessible phase space using a manageable number of relatively well-spaced umbrella sampling simulations. In this work, the final $\vec{\psi}^{\,2} = [\phi, \psi, \theta, \zeta]$ umbrella sampling FES spans a free energy range of $\sim 38\ k_B T$ [Fig. S7a (Ref. 61)], suggesting that the umbrella sampling simulations from which this landscape was constructed adequately explored the thermally accessible phase space available to the system.

The terminal Round II four-dimensional embedding provides a low-dimensional representation of the topology of the phase space accessible to alanine dipeptide, describing the relative locations and connectivity of its stable and metastable states. In the supplementary material,[61] we demonstrate the utility of this description in guiding the subsequent principal components analysis of a selected $C_7^{\mathrm{eq}} \leftrightarrow \alpha_R$ isomerization pathway.

IV. CONCLUSIONS

The diffusion map[16,19,20] is a nonlinear dimensionality reduction technique which may be applied to molecular simulation trajectories to systematically extract global order parameters with which to parametrize the fundamental dynamical motions of the system. For molecular systems possessing high free energy barriers, unbiased simulation trajectories can become trapped in local free energy minima and exhibit poor sampling of barrier regions. The application of dimensionality reduction techniques to such simulation data may result in locally good low-dimensional parametrizations within well-sampled free energy wells, but typically precludes the construction of a globally valid description.

Umbrella sampling[14] is a well established simulation technique which improves sampling of phase space by applying biasing potentials to drive the system along specific order parameters. This technique results in a biased ensemble of simulation data to which the diffusion map in its original formulation may not immediately be applied. In this work, we develop an adaptation of the diffusion map approach to permit its application to biased data sets. Furthermore, we propose an iterative methodology to facilitate the systematic determination of order parameters for systems exhibiting high free energy barriers by interleaving successive rounds of umbrella sampling and the adapted diffusion map approach. Commencing with an initial round of umbrella sampling in a set of putative variables thought to be reasonable descriptors of the important motions of the system, the adapted diffusion map approach is applied to the biased simulation data to detect the evolution of the system in any additional important dynamical modes. Order parameters characterizing these additional motions are then fed back into a subsequent round of umbrella sampling, and the process is repeated until no new dynamical motions emerge.

Nonlinear dimensionality reduction techniques typically do not provide the functional dependence of the variables parametrizing the low-dimensional embedding upon the atomic coordinates of the system. Since the biasing potentials in umbrella sampling may only be constructed in variables which are known functions of the atomic coordinates, a key component of this proposed methodology is the translation of the order parameters identified by the diffusion map into physical variables in which to perform umbrella sampling. No systematic means yet exists to perform this mapping, and in the present work candidate physical variables must be manually selected on the strength of their correlation with the diffusion map order parameters. This crucial step in the approach may be made more robust by employing systematic techniques to screen combinations of candidate variables such as those proposed by Peters, Beckham, and Trout[35,36] and Ma and Dinner.[2]

We note that linear dimensionality reduction techniques, such as PCA, explicitly furnish the functional dependence of the low-dimensional parametrizations upon the atomic coordinates. This would seem to suggest that in the iterative framework, we should replace the adapted diffusion map approach with an adaptation of PCA formulated for biased data sets, thereby permitting umbrella sampling to be conducted directly in the low-dimensional variables. However, it is a well-known deficiency of linear dimensionality reduction techniques that the low-dimensional embeddings constructed for a given data set are typically of higher dimensionality than the embeddings resulting from the application of nonlinear approaches.[7] Since the computational expense of umbrella sampling increases exponentially with the number of variables in which sampling is conducted, the more parsimonious low-dimensional descriptions provided by nonlinear techniques, such as the diffusion map, can offer a significant advantage.

We have demonstrated the iterative methodology, which we term the umbrella bootstrapped diffusion map approach, in an application to alanine dipeptide in explicit solvent, which is known to exhibit free energy barriers on the order of tens of $k_B T$. We systematically determined that the ϕ, ψ, θ, and ζ backbone dihedral angles provide a good four-dimensional parametrization of the important dynamical motions, and in the supplementary material[61] we illustrate the utility of the global diffusion map embedding in guiding subsequent analysis of particular dynamical transitions by applying PCA to one of several identified $C_7^{\mathrm{eq}} \leftrightarrow \alpha_R$ isomerization pathways.

Overall, our results demonstrate the utility of our proposed methodology in analyzing biased simulation data and in providing an efficient means to systematically develop low-dimensional parametrizations for molecular systems exhibiting high free energy barriers. We hope that the procedures developed herein will prove useful in the systematic determination of low-dimensional descriptions for more complex biophysical systems.

ACKNOWLEDGMENTS

A.Z.P. acknowledges the (U.S.) Department of Energy (DOE) (Grant No. DE-SC-0002128) and the Princeton Center for Complex Materials (MRSEC Grant No. DMR-0819860). P.G.D. acknowledges the support of the National Science Foundation (Collaborative Research in Chemistry Grant No. CHE-0908265). I.G.K. acknowledges partial support by the Department of Energy (Grant No. DE-SC-0002097).

[1] S. S. Cho, Y. Levy, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **103**, 586 (2006).
[2] A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769 (2005).
[3] A. E. García, Phys. Rev. Lett. **68**, 2696 (1992).
[4] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins: Struct., Funct., Genet. **17**, 412 (1993).
[5] R. Hegger, A. Altis, P. H. Nguyen, and G. Stock, Phys. Rev. Lett. **98**, 028102 (2007).
[6] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, Proc. Natl. Acad. Sci. U.S.A. **103**, 9885 (2006).
[7] H. Stamati, C. Clementi, and L. E. Kavraki, Proteins: Struct. Funct. Bioinf. **78**, 223 (2009).
[8] P. I. Zhuravlev, C. K. Materese, and G. A. Papoian, J. Phys. Chem. B **113**, 8800 (2009).
[9] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, Proc. Natl. Acad. Sci. U.S.A. **107**, 13597 (2010).
[10] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, New York, 2001).
[11] G. Hummer and I. G. Kevrekidis, J. Chem. Phys. **118**, 10762 (2003).
[12] B. E. Sonday, M. Haataja, and I. G. Kevrekidis, Phys. Rev. E **80**, 031102 (2009).
[13] A. L. Ferguson, S. Zhang, I. Dikiy, A. Z. Panagiotopoulos, P. G. Debenedetti, and A. J. Link, Biophys. J. **99**, 3056 (2010).
[14] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[15] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[16] M. Belkin and P. Niyogi, Neural Comput. **15**, 1373 (2003).
[17] J. B. Tenenbaum, V. de Silva, and J. C. Langford, Science **290**, 2319 (2000).
[18] S. T. Roweis and L. K. Saul, Science **290**, 2323 (2000).
[19] 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).
[20] R. R. Coifman and S. Lafon, Appl. Comput. Harmon. Anal. **21**, 5 (2006).
[21] D. S. Chekmarev, T. Ishida, and R. M. Levy, J. Phys. Chem. B **108**, 19487 (2004).
[22] A. Singer, R. Erban, I. G. Kevrekidis, and R. R. Coifman, Proc. Natl. Acad. Sci. U.S.A. **106**, 16090 (2009).
[23] R. Erban, T. A. Frewen, X. Wang, T. C. Elston, R. Coifman, B. Nadler, and I. G. Kevrekidis, J. Chem. Phys. **126**, 155103 (2007).
[24] E. Plaku, H. Stamati, C. Clementi, and L. E. Kavraki, Proteins: Struct., Funct., Bioinf., **67**, 897 (2007).
[25] A. Kentsis, T. Gindin, M. Mezei, and R. Osman, PLoS ONE, **2**, e446 (2007).
[26] R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, IEEE Trans. Image Process. **17**, 1891 (2008).
[27] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, in *Advances in Neural Information Processing Systems*, edited by Y. Weiss, B. Schölkopf, and J. Platt, Neural Information Processing Systems (NIPS) (MIT Press, Cambridge, MA, 2006), Vol. 18, pp. 955–962.
[28] K. Y. Sanbonmatsu and A. E. García, Proteins: Struct., Funct., Genet., **46**, 225 (2002).
[29] A. L. Ferguson, P. G. Debenedetti, and A. Z. Panagiotopoulos, J. Phys. Chem. B **113**, 6405 (2009).
[30] T. F. Miller, E. Vanden-Eijnden, and D. Chandler, Proc. Natl. Acad. Sci. U.S.A. **104**, 14559 (2007).
[31] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic, San Diego, 2002).
[32] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).
[33] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, J. Comput. Chem. **13**, 1011 (1992).
[34] B. Roux, Comput. Phys. Commun. **91**, 275 (1995).
[35] B. Peters and B. L. Trout, J. Chem. Phys. **125**, 054108 (2006).
[36] B. Peters, G. T. Beckham, and B. L. Trout, J. Chem. Phys. **127**, 034109 (2007).
[37] E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Modeling. **7**, 306 (2001).
[38] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, J. Comput. Chem. **26**, 1701 (2005).
[39] W. L. Jorgensen and J. Tirado-Rives, J. Am. Chem. Soc. **110**, 1657 (1988).
[40] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen, J. Phys. Chem. B **105**, 6474 (2001).
[41] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[42] A. W. Schüttelkopf and D. M.F. van Aalten, Acta Crystallogr. **60**, 1355 (2004).
[43] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).
[44] S. Nosé, J. Chem. Phys. **81**, 511 (1984).
[45] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).
[46] M. Parinello and A. Rahman, J. Appl. Phys. **52**, 7182 (1981).
[47] D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, J. Comput. Chem. **26**, 1668 (2005).
[48] B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, J. Comput. Chem. **30**, 1545 (2009).
[49] K. J. Maschhoff and D. C. Sorensen, in *Third International Workshop on Applied Parallel Computing, Industrial Computation and Optimization*, Lecture Notes in Computer Science, Vol. 1184, edited by J. Wasniewski, J. Dongarra, K. Madsen, and D. Olesen (Springer, New York, 1996), pp. 478–486.
[50] J. Apostolakis, P. Ferrara, and A. Caflisch, J. Chem. Phys. **110**, 2099 (1999).
[51] T. A. Frewen, G. Hummer, and I. G. Kevrekidis, J. Chem. Phys. **131**, 134104 (2009).
[52] M. B. Kubitzki and B. L. De Groot, Biophys. J. **92**, 4262 (2007).
[53] C. A. F. de Oliveira, D. Hamelberg, and J. A. McCammon, J. Chem. Phys. **127**, 175105 (2007).
[54] E. Vanden-Eijnden and M. Venturoli, J. Chem. Phys. **130**, 194103 (2009).
[55] M. Bonomi, A. Barducci, and M. Parrinello, J. Comput. Chem. **30**, 1615 (2009).
[56] C. Velez-Vega, E. E. Borrero, and F. A. Escobedo, J. Chem. Phys. **130**, 225101 (2009).
[57] P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Natl. Acad. Sci. U.S.A. **97**, 5877 (2000).
[58] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, J. Chem. Phys. **125**, 024106 (2006).
[59] C. Bartels and M. Karplus, J. Comput. Chem. **18**, 1450 (1997).
[60] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, J. Chem. Phys. **126**, 155101 (2007).
[61] See supplementary material at http://dx.doi.org/10.1063/1.3574394 for eigenvalue spectra and additional diffusion map embeddings pertaining to the application of the umbrella bootstrapped diffusion map approach to alanine dipeptide, and a local principal component analysis of a selected isomerization pathway identified by the approach.
[62] P. Liu, Q. Shi, E. Lyman, and G. A. Voth, J. Chem. Phys. **129**, 114103 (2008).