

Quantifying Unbiased Conformational Ensembles from Biased Simulations Using ShapeGMM

Published as part of *Journal of Chemical Theory and Computation* virtual special issue “Machine Learning and Statistical Mechanics: Shared Synergies for Next Generation of Chemical Theory and Computation”.

Subarna Sasmal, Triasha Pal, Glen M. Hocky,* and Martin McCullagh*



Cite This: *J. Chem. Theory Comput.* 2024, 20, 3492–3502



Read Online

ACCESS |



Metrics & More



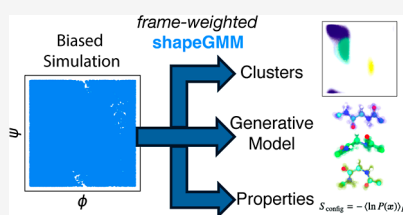
Article Recommendations



Supporting Information

ABSTRACT: Quantifying the conformational ensembles of biomolecules is fundamental to describing mechanisms of processes such as protein folding, interconversion between folded states, ligand binding, and allosteric regulation. Accurate quantification of these ensembles remains a challenge for conventional molecular simulations of all but the simplest molecules due to insufficient sampling. Enhanced sampling approaches, such as metadynamics, were designed to overcome this challenge; however, the nonuniform frame weights that result from many of these approaches present an additional challenge to ensemble quantification techniques such as Markov State Modeling or structural clustering.

Here, we present rigorous inclusion of nonuniform frame weights into a structural clustering method entitled shapeGMM. The result of frame-weighted shapeGMM is a high dimensional probability density and generative model for the unbiased system from which we can compute important thermodynamic properties such as relative free energies and configurational entropy. The accuracy of this approach is demonstrated by the quantitative agreement between GMMs computed by Hamiltonian reweighting and direct simulation of a coarse-grained helix model system. Furthermore, the relative free energy computed from a shapeGMM probability density of alanine dipeptide reweighted from a metadynamics simulation quantitatively reproduces the underlying free energy in the basins. Finally, the method identifies hidden structures along the actin globular to filamentous-like structural transition from a metadynamics simulation on a linear discriminant analysis coordinate trained on GMM states, illustrating how structural clustering of biased data can lead to biophysical insight. Combined, these results demonstrate that frame-weighted shapeGMM is a powerful approach to quantifying biomolecular ensembles from biased simulations.



Conformational ensembles of molecules dictate many of their thermodynamic properties. Conventional molecular dynamics (MD) simulations allow us to sample models of these ensembles but suffer from the so-called rare event problem. A variety of enhanced sampling techniques, such as Metadynamics (MetaD),^{1,2} Adaptive Biasing Force,³ Gaussian accelerated MD,⁴ and Temperature Accelerated MD/Driven Adiabatic Free Energy Dynamics,^{5,6} have been developed to promote faster sampling by effectively heating some degrees of freedom.⁷ Unfortunately, due to the biased sampling of many of these approaches, it is not obvious how to use the biased configurations in methods such as Markov State Models (MSMs)^{8,9} and/or structural clustering approaches that quantify the conformational ensemble. Here, we adapt shapeGMM,¹⁰ a probabilistic structural clustering method, to rigorously quantify the unbiased conformational ensembles generated from biased simulations. The result is a high dimensional Gaussian mixture model (GMM) characterizing the unbiased landscape that can be used to extract important thermodynamic quantities and to give additional insight beyond the low dimensional projections often used to represent free energy landscapes.

Meaningful quantification of conformational ensembles from large molecular simulations requires the grouping of similar frames by using a clustering algorithm. Clustering algorithms for molecular simulation can be grouped into two categories: temporal and structural. Temporal clustering, such as spectral clustering of the transition matrix,^{11,12} has been successfully applied to MD trajectories to achieve kinetically stable clusters for use in objects like MSMs.^{13–15} Enhanced sampling techniques, however, can distort the underlying kinetics of the system, making temporal clustering difficult to apply properly in these circumstances. While there have been efforts to build MSMs from enhanced sampling data^{16,17} it still remains a challenge.¹⁸ Additionally, building MSMs relies on an initial structural clustering step, making it critical to perform this step accurately, even in the context of enhanced sampling.

Received: February 22, 2024

Revised: April 5, 2024

Accepted: April 5, 2024

Published: April 25, 2024



Structural clustering involves partitioning either frames or feature space into a finite number of elements. This can be achieved from enhanced sampling data, but care must be taken to properly account for the nonuniform weights of the frames.

Previous efforts to use structural clustering algorithms on enhanced sampling simulations have focused on partitional, as opposed to model-based, algorithms. The main results of partitional algorithms are cluster populations that can be reweighted based on enhanced sampling frame weights to estimate the unbiased populations.^{16,19} Model-based clustering algorithms offer many advantages over partitional algorithms, the most relevant being that the resulting probability density can be used to predict clusterings on new data and estimate thermodynamic properties of the underlying ensemble. Reweighting the cluster populations of model-based algorithms a posteriori is, however, not satisfactory for methods such as GMMs, as the frame weights will affect the determination of additional model parameters. It is possible to use multiple copies of frames to approximately account for the frame weights, but this can yield intractably large trajectories and inaccuracies due to rounding.

In this work, we present an adaptation to shapeGMM,¹⁰ a probabilistic structural clustering method on particle positions, to directly account for nonuniform frame weights. As opposed to introducing copies of input data and maintaining uniform weights, the current method directly accounts for nonuniform frame weights and is thus more efficient and scalable than the alternative. In the next section, we briefly introduce the shapeGMM method and the adaptations necessary to account for nonuniform frame weights. This is followed by a demonstration of the method on three examples of increasing difficulty, specifically demonstrating that our proposed choices of frame weights from MetaD simulations result in a reliable clustering procedure. We show in benchmark cases how this method can yield thermodynamic quantities directly and use the complex case of actin flattening to show how a weighted shapeGMM can give physical insight into the conformations sampled, in a case where unbiased simulation would not be a practical option. In addition, frame-weighted shapeGMM is implemented in an easy-to-use python package (`pip install shapeGMMTorch`).

2. THEORY AND METHODS

2.1. Overview of ShapeGMM. In shapeGMM, a particular configuration of a macromolecule is represented by a particle position matrix, \mathbf{x}_i , of order $N \times 3$, where N is the number of particles being considered for clustering. To account for translational and rotational invariance, the proper feature for clustering purposes is an equivalence class

$$[\mathbf{x}_i] = \{\mathbf{x}_i \mathbf{R}_i + \mathbf{1}_N \vec{\xi}_i^T : \vec{\xi}_i \in \mathbb{R}^3, \mathbf{R}_i \in \text{SO}(3)\} \quad (1)$$

where $\vec{\xi}_i$ is a translation in \mathbb{R}^3 , \mathbf{R}_i is a rotation $\mathbb{R}^3 \rightarrow \mathbb{R}^3$, and $\mathbf{1}_N$ is the $N \times 1$ vector of ones. $[\mathbf{x}_i]$ is thus the set of all rigid body transformations, or orbit, of \mathbf{x}_i .

The shapeGMM probability density is a Gaussian mixture given by

$$P(\mathbf{x}_i) = \sum_{j=1}^K \phi_j N(\mathbf{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad (2)$$

where the sum is over the K Gaussian mixture components, ϕ_j is the weight of component j , and $N(\mathbf{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ is a normalized multivariate Gaussian given by

$$N(\mathbf{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp\left[-\frac{1}{2}(\mathbf{g}_i^{-1} \mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{g}_i^{-1} \mathbf{x}_i - \boldsymbol{\mu})\right]}{\sqrt{(2\pi)^{(3N)} \det \boldsymbol{\Sigma}}} \quad (3)$$

where $\boldsymbol{\mu}$ is the mean structure, $\boldsymbol{\Sigma}$ is the covariance, and $\mathbf{g}_i^{-1} \mathbf{x}_i$ is the element of the equivalence class, $[\mathbf{x}_i]$, that minimizes the squared Mahalanobis distance in the argument of the exponent. Determining the proper transformation, \mathbf{g}_i , is achieved by translating all frames to the origin and then determining the optimal rotation matrix. Cartesian and quaternion-based algorithms for determining optimal rotation matrices are known for two forms of the covariance were considered $\boldsymbol{\Sigma} \propto \mathbf{I}_{3N}$ ^{20,21} or $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_N \otimes \mathbf{I}_3$,^{22,23} where $\boldsymbol{\Sigma}_N$ is the $N \times N$ covariance matrix and \otimes denotes a Kronecker product. In this paper, we employ only the more general Kronecker product covariance.

2.2. Incorporating Nonuniform Frame Weights in shapeGMM. Previously, each frame in shapeGMM was considered to be equally weighted. Approximate weighting of frames could be taken into account by including frames multiple times in the training data to give them more importance; however, this introduces the imprecision of rounding to the nearest integer and can be extremely computationally expensive due to the large increase in amount of training data. Here, we take nonuniform frame weights into account by performing weighted averages in the Expectation Maximization estimate of model parameters $\{\hat{\phi}_j, \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j\}$, consistent with other fixed-weight GMM procedures.²⁴ Considering a normalized set of frame weights, $\{w_i\}$ where $\sum_{i=1}^M w_i = 1$ for M frames, their contribution to the probability can be accounted for by weighting the estimate of the posterior distribution of latent variables

$$\gamma_{Z_i}(j) = w_i \frac{\hat{\phi}_j N(\mathbf{x}_i | \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)}{\sum_{j=1}^K \hat{\phi}_j N(\mathbf{x}_i | \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)} \quad (4)$$

The frame weight will propagate to the estimate of component weights, means, and covariances in the Maximization step through $\gamma_{Z_i}(j)$

$$\hat{\phi}_j = \sum_{i=1}^M \gamma_{Z_i}(j) \quad (5)$$

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{i=1}^M \gamma_{Z_i}(j) \mathbf{g}_{i,j}^{-1} \mathbf{x}_i}{\sum_{i=1}^M \gamma_{Z_i}(j)} \quad (6)$$

$$\hat{\boldsymbol{\Sigma}}_j = \frac{\sum_{i=1}^M \gamma_{Z_i}(j) \langle \hat{\boldsymbol{\Sigma}}_N \rangle_i}{\sum_{i=1}^M \gamma_{Z_i}(j)} \otimes \mathbf{I}_3 \quad (7)$$

Additionally, the log likelihood per frame is computed as a weighted average

$$\ln(L) = \sum_{i=1}^M w_i \ln\left(\sum_{j=1}^K \hat{\phi}_j N(\mathbf{x}_i | \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)\right) \quad (8)$$

2.3. Choosing Number of Clusters. Performing shapeGMM requires the user to choose a number of clusters,

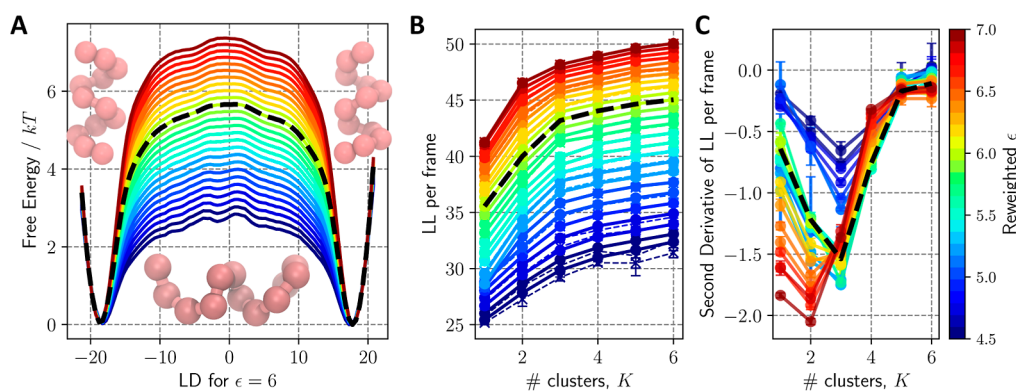


Figure 1. Beaded helix ϵ reweighting. Trajectory data for a 12 bead polymer having $i, i + 4$ interactions with strength $\epsilon = 6$ was reweighted to predict the ensemble for ϵ values ranging from 4.5 to 7 in increments of 0.1. (A) The corresponding free energies as a function of the linear discriminant (LD) between the two helices are plotted with ϵ values denoted by the color bar on the right-hand. The weights per frame were fed in to shapeGMM to perform a cluster scan. (B) The resulting log likelihood per frame as a function of number of clusters from the cluster scan. (C) Second derivative of the curves from (B). Error bars in (B,C) are estimated as the standard deviation from three different training sets. The true curve for $\epsilon = 6$ is given in black dashed lines in all three panels.

K . The “optimal” choice will be system and problem specific and potentially has no correct answer. The choice is no different if you consider uniformly or nonuniformly weighted frames. We used a cluster scan with a combination of the elbow method and cross validation to assess if our choice of K is reasonable. A good choice of clusters based on this approach is to find the number of clusters where the increase in log-likelihood with K is decreasing fastest, which we can evaluate by choosing the minimum of the second derivative of $\ln(L)$ with respect to the number of clusters. In practice, this works well for simple systems, but it may be hard to pick a “best” choice for more complex systems, so we may seek a choice that is physically interpretable.

2.4. Assigning Frames to Clusters. After the model parameters have been fit using fuzzy assignments, individual frames are assigned to the cluster in which that frame has the largest likelihood [largest value, $\gamma_{z_i}(j)$]. This is the standard procedure for clustering from a GMM and is no different for the frame-weighted version.

2.5. Implementation. We have completely rewritten shapeGMM in PyTorch for computational efficiency and the ability to use GPUs. The current implementation takes an array of frame weights as an optional argument to both the fit and predict functions (the code defaults to uniform weights). The PyTorch implementation is significantly faster than the original version and is available both on github (<https://github.com/mccullaghlab/shapeGMMTorch>) and PyPI (`pip install shapeGMMTorch`). Examples are also provided on that github, and all examples from this paper are provided in a second github page discussed below.

2.6. Choosing Training Sets. For nonuniformly weighted frames, the choice of training set may be important. We have attempted a variety of training set sampling schemes and have found that, at least for the frame weight distributions that we have encountered, uniform sampling of the training data is at least as good as any importance sampling scheme. We discuss this further and show results for three different training set selection schemes for the beaded helix system in Section S1.

2.7. Biasing and Weighting Frames. If configuration \mathbf{x} is generated from an MD simulation at constant T and V then $P(\mathbf{x}) \propto \exp[-H(\mathbf{x})/k_B T]$ where H is the system’s Hamiltonian.²⁵ If \mathbf{x} is generated from an MD simulation at a different

state point (e.g., different T) or with a different Hamiltonian, it is sampled from a different distribution $Q(\mathbf{x})$. Samples from Q can be reweighted to P with weights²⁵

$$w(\mathbf{x}) \propto \frac{P(\mathbf{x})}{Q(\mathbf{x})} \quad (9)$$

from which averages over P can be estimated. This approach is effective only if Q and P are finite over the same domain. Nonetheless, eq 9 underlies many enhanced sampling approaches, for example, it is the basis of the original formulation of umbrella sampling.²⁶ By including weights in shapeGMM, we can predict the importance of clusters at nearby state-points or for similar systems.

2.8. Thermodynamic Quantities from ShapeGMM. Many Thermodynamic quantities can be computed from fit shapeGMM probability densities. One such quantity is the configurational entropy

$$S_{\text{config}} = - \int P(\mathbf{x}) \ln P(\mathbf{x}) \, d\mathbf{x} = - \langle \ln P(\mathbf{x}) \rangle_P \quad (10)$$

The configurational entropy has an analytic solution for a single multivariate Gaussian but for the general mixture of multivariate Gaussians we use sampling and Monte Carlo integration to approximate the integral.

To do so accurately requires that we generate points from the shapeGMM objects and not just use the trajectory on which the object was fit. We introduced a `generate` function as an attribute to a fit shapeGMM object that produces configurations sampled from the underlying trained distribution.

The second Thermodynamic quantity we consider is the free energy cost to move from one distribution to another. This is also known as the relative entropy or Kullback–Leibler divergence and the cost to go from distribution Q to distribution P is given by

$$D_{\text{KL}}(P||Q) = \int P(\mathbf{x}) \ln \frac{P(\mathbf{x})}{Q(\mathbf{x})} \, d\mathbf{x} = \langle \ln \frac{P(\mathbf{x})}{Q(\mathbf{x})} \rangle_P \quad (11)$$

Here, again, we generate points from distribution P and average the difference in log likelihoods of these points in P and Q to assess this value. It should be noted that this is a nonequilibrium free energy and is thus not necessarily

symmetric.^{27,28} The quantity can prove useful in applications, for example measuring the free energy cost to shift a distribution from an apo to a ligand-bound state, for example.^{29,30}

A symmetric metric is useful when comparing the similarity of two distributions. Here we opt for the Jensen–Shannon divergence (JSD)³¹ given by

$$\text{JSD}(P\|Q) = \frac{1}{2 \ln 2} (D_{\text{KL}}(P\|M) + D_{\text{KL}}(Q\|M)) \quad (12)$$

where $M = \frac{1}{2}(P + Q)$ is the midpoint distribution between P and Q . JSD is restricted to between 0 and 1.

All three of these measures were implemented in the `similarities` library of the `shapeGMM` code. They use point generation and Monte Carlo sampling to assess the integrals and thus return both the mean value and the standard error.

3. RESULTS AND DISCUSSION

3.1. Proof of Concept: Reweighting the Beaded Helix.

To demonstrate the accuracy of the frame-weighted `shapeGMM` process we perform Hamiltonian reweighting of a nonharmonic beaded helix previously studied in refs 10 and 32. The system is composed of 12 beads connected in a sequential fashion by stiff harmonic bonds. Every fifth pairwise interaction is given by an attractive Lennard-Jones potential with well depth ϵ . The value of ϵ relative to kT dictates the stability of an α -helix-like structure as compared to a completely disordered state. Additionally, because of the symmetry of the model, both the left- and right-handed helices have equal probability no matter the value of ϵ . A value of $\epsilon = 6$ in reduced units forms stable helices while allowing transitions between the two folded states; here, we performed a long unbiased trajectory to sample both left and right states, as well as possibly intermediates (see Sec. A1 for details).

`ShapeGMM` suggests that three clusters is a good choice for a simulation of the beaded helix with $\epsilon = 6$. Shown in blacked dash line in Figure 1A is the unbiased free energy for this system computed as $F(s) = -\ln P(s)$ for a linear discriminant (LD) reaction coordinate.³³ By performing a scan over the number of clusters on 100k frames from an unbiased trajectory, we identify three clusters as the optimal number by observing a definite kink in the curves in Figure 1B and the presence of a minimum in the second derivative in Figure 1C. These clusters correspond to the left- and right- helical states as well as a partially unfolded intermediate cluster, examples shown in Figure 1A.

Reweighted clustering of the beaded helix system predicts that the prevalence of the partially unfolded intermediate will disappear at $\epsilon = 6.5$. To demonstrate this, we performed frame-weighted `shapeGMM` cluster scans of our trajectory at $\epsilon = 6$ with weights corresponding to ϵ values ranging from 4.5 to 7.0 in increments of 0.1. Given that the samples come from a Boltzmann distribution, the weights for each frame given by eq 9 are $w_i(\epsilon) = e^{(U(x_i|\epsilon=6) - U(x_i|\epsilon))/(k_B T)}$. The log likelihood of the `shapeGMM` fits as a function of the number of clusters is shown in Figure 1B,C with ϵ values indicated in the color bar on the right. We see that as ϵ increases from 6, the minimum in the second derivative moves from 3 clusters to 2 cluster. The transition occurs between $\epsilon = 6.4$ and $\epsilon = 6.5$. This suggests that a simulation run at ϵ values of greater than 6.4 (in reduced units) will not exhibit the partially unfolded third cluster.

These results are consistent with the increasing free energy barrier height as a function of ϵ depicted in Figure 1A.

The reweighting of ϵ for the beaded helix example also predicts that only one cluster will be present for a small ϵ . In Figure 1B, the elbow at 3 clusters is evident for ϵ values as low as $\epsilon = 5$ and becomes less pronounced below this threshold. While a minimum at 3 clusters is still observed in the second derivative plot for $\epsilon = 4.5$, the trend is clear that as ϵ becomes small the choice of anything other than 1 cluster is less well supported by the elbow heuristic. This is an expected result, and consistent with the reduced free-energy barriers observed for small ϵ in Figure 1A, as ϵ approaches thermal energy, the prevalence of anything other than an unfolded state is entropically unfavorable.

`ShapeGMM` reweighted clustering also produces quantitatively accurate probability densities for the beaded helix. To demonstrate this, we compute a reweighted `shapeGMM` object ($\epsilon = 6 \rightarrow 8$) to a `shapeGMM` object trained on an unbiased trajectory at $\epsilon = 8$, which we refer to as ground truth (GT). Because, as predicted, transitions at $\epsilon = 8$ are very unlikely, this object is trained on simulations, each with 100k frames, initiated from left and right helices, and concatenated. Two controls are included that are fit to the $\epsilon = 6$ trajectory without reweighting: the predicted 3 cluster object and that same object with only the cluster populations reweighted to $\epsilon = 8$. To quantitatively compare between two probability densities we use two similarity metrics, both described above in more detail and introduced as eqs 10 and 11: Jensen–Shannon divergence (JSD) and change in configurational entropy S_{config} . These similarity metrics between the GT and the three different `shapeGMM` objects are tabulated in Table 1. JSD is a

Table 1. Similarity Measures between Three Beaded Helix Probability Densities Fit from a Simulation with $\epsilon = 6$, Q , and the “Ground-Truth” (GT) Probability Density Fit to a Simulation at $\epsilon = 8$ ^a

K	Q	JSD(GT Q)	$\Delta S_{\text{config}}/R$
	ϵ_R		
3	6.0	0.401(2)	7.22(3)
3	6.0/8.0 ^b	0.357(2)	4.30(2)
2	8.0	0.0071(3)	0.00(2)

^aThe reweighted probability densities are denoted by the number of clusters, K , and the value of ϵ used in reweighting, ϵ_R . The three Q s are $K = 3$ clusters and weighted to $\epsilon_R = 6.0$, $K = 3$ clusters from $\epsilon_R = 6.0$ with only the cluster populations reweighted to $\epsilon = 8$, and $K = 2$ clusters completely reweighted to $\epsilon_R = 8$. The similarity measures are the Jensen–Shannon divergence (JSD) and the difference in configurational entropy $\Delta S_{\text{config}} = S_{\text{config}}^Q - S_{\text{config}}^{\text{GT}}$. Error in the last digit is included in parentheses and is estimated as Monte Carlo sampling errors in estimating the integrals. ^bOnly the cluster populations are reweighted to $\epsilon = 8$ in this probability density.

symmetric metric bounded between 0 and 1 where 0 indicates no divergence and 1 indicates complete divergence between the two probabilities. The reweighted `shapeGMM` object demonstrates a very small JSD (0.0071 ± 0.0003) to the GT as compared to either of the $\epsilon = 6$ objects (0.357 ± 0.002 or 0.401 ± 0.002). This trend holds true when comparing relative S_{config} 's with the difference in S_{config} between the reweighted and GT $\epsilon = 8$ `shapeGMM` probabilities being within error of 0. These results indicate that the $\epsilon = 8$ reweighted `shapeGMM` probability density is nearly identical with the GT.

3.2. Conformational States of Alanine Dipeptide from Metadynamics Simulations. Alanine Dipeptide (ADP) in a vacuum is a common benchmark system for methods designed to sample and quantify conformational ensembles. In this work, we demonstrate that ADP MetaD simulations can be used directly to achieve equilibrium clustering by using various estimates of the frame weights. In Well-tempered MetaD (WT-MetaD), a history dependent bias is generated by adding Gaussian hills to a grid at the current position in collective variable (CV) space^{2,34} such that the bias at time t for CV value position s_i is given by

$$V(s_i, t) = \sum_{\tau < t} h e^{-V(s_i, \tau)/\Delta T} e^{-Q(x(\tau) - s_i)^2/2\sigma^2} \quad (13)$$

where h is Gaussian height, and σ is the width, and $T + \Delta T$ is an effective sampling temperature for the CVs. Rather than setting ΔT , one typically chooses the bias factor $\gamma = (T + \Delta T)/T$, which sets the smoothness of the sampled distribution.^{2,34} Asymptotically, a free energy surface (FES) can be estimated from the applied bias by $F(s) = -\frac{\gamma}{\gamma-1} V(s, t \rightarrow \infty)$ ^{34,35} or using a reweighting scheme.^{34,36} In MetaD, frames are generated from a time dependent Hamiltonian, so the choice of frame weights for clustering is not obvious. Reweighting of MetaD trajectories to compute free energy surfaces was accomplished through several different schemes.

For a static bias V added to the initial Hamiltonian, the weight of a frame given by eq 9 would be $w_i = e^{V(s_i)/k_B T}$. Our first choice of frame weights (termed “bias”) corresponds to using this formula even though the bias is time-dependent. A second choice that removes some of the time-dependence is to use $w_i = e^{(V(s_i) - c(t))/k_B T}$, where $c(t) = -k_B T \ln(e^{-V(s(t))/k_B T})$ is the bias averaged over the CV grid at a fixed time. The quantity $V(s_i) - c(t)$ is called the “reweighting bias” and can be computed automatically in PLUMED,³⁷ hence we term clustering using this scheme “rbias”. Finally, we evaluate another commonly used approach to compute Boltzmann weights of each frame postfacto,³⁸ which in the case of WT-MetaD would correspond to $w_i = e^{-E_{\text{final}}(s(x_i))/k_B T} = e^{\gamma/(\gamma-1)V_{\text{final}}(s(x_i))/k_B T}$; we label these weights “fbias”. Other more sophisticated reweighting schemes have also been proposed, e.g. in refs 38 and 39, but we did not test these here because, as will be seen, the bias, rbias, and fbias approaches all worked well for our test system. However, shapeGMM, as implemented, is capable of using any choice of frame weights. We include “uniform” weights as a control.

For assessing the best choice of weights, we performed a 100 ns WT-MetaD simulation on ADP biasing backbone dihedral angles ϕ and ψ using bias factor 10, saving every 1 ps to generate 100,000 frames (see Section A1 for full details). The five atoms involved in the ϕ and ψ dihedral angles were chosen for shapeGMM clustering. The coordinates of these atoms and the frame weights from the four different schemes were fed into shapeGMM. The log likelihood per frame of the resulting fits as a function of number of clusters is shown in Figure 2A. In general, the three nonuniformly weighted clustering objects result in significantly higher log likelihoods than the uniform weights for equivalent numbers of clusters $K > 2$, indicating a better fit to the underlying data. The significant kink in the cluster scans at 2 clusters indicates that at least 2 clusters are necessary for a

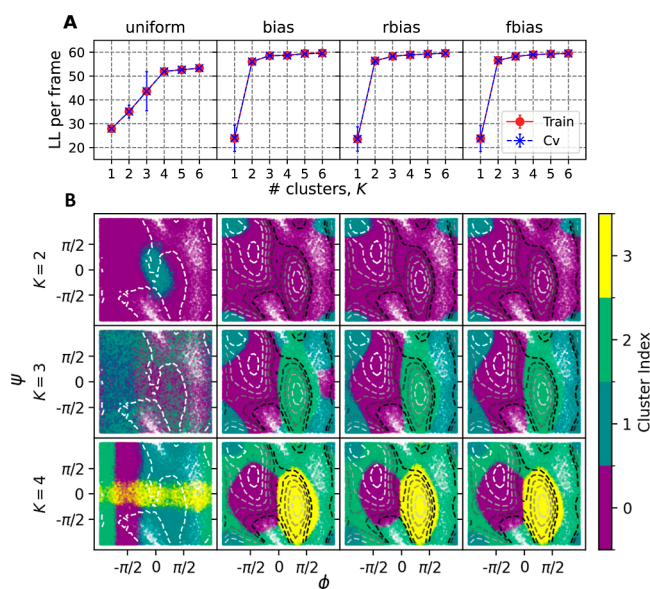


Figure 2. WT-MetaD simulation for ADP with BF 10. Each column represents a particular choice of weights been used in frame-weighted SGMM. (A) Cluster scans for each choice of frame weights using 50k frames, 4 training sets and 10 attempts for each case. (B) Clusterings performed for $K = 2, 3, 4$ are shown by coloring each of 100k sampled points by their cluster assignment. Contour lines indicate the underlying free energy surface as computed from the WT-MetaD simulation via reweighting with the different choice of weights. Contours indicate free energy levels above the minimum from 1 to 11 kcal/mol with a spacing of 2 kcal/mol.

good fit to the data; there is still substantial increase going from 2 to 3 clusters, however, indicating that there may be additional insight gained at $K = 3$ and above, as we shall see.

Nonuniform frame-weighted shapeGMM produces physically relevant clusterings. Figure 2B indicates how sampled points in ϕ and ψ space are assigned to two, three, or four clusters when using each of the choices of frame weights, with the underlying free energy landscape computed from a weighted histogram with the same choices of weights as used for the clustering indicated by contour lines. Clustering with uniform weights has little correlation with the underlying free energy landscape, whereas performance is much better when using any of the nonuniform weighting schemes. Weighted clustering with $K = 2$ tends to split the landscape into one cluster covering the most extended upper-left “C5” basin near $(-2, 2)$, while using a second cluster to cover the rest of the landscape (see ref 40 for a naming convention). However, a higher number of clusters allows for separating the upper left basin into its two constituent states, C5 and “C7eq” at $(-2, 1)$, while also revealing the presence of the minor “C7ax” state at $(1, -1)$. Slight differences in contour FES correspond with slight differences in the weighted cluster assignments; for example, in the $K = 3$ case the upper left and bottom left parts of the axial basin are disconnected at $\Psi = 0$ for bias weights but connected for rbias and fbias weights.

Nonuniform frame-weighted shapeGMM also works for standard (untempered) MetaD^{1,34} with $\Delta T \rightarrow \infty$. For untempered MetaD, we favor rbias weights because the final bias is not static and the instantaneous bias diverges, meaning that initial frames receive no weight. In Figure S2, we show that shapeGMM clustering with rbias weights performs much better than equally weighted frames, and results are

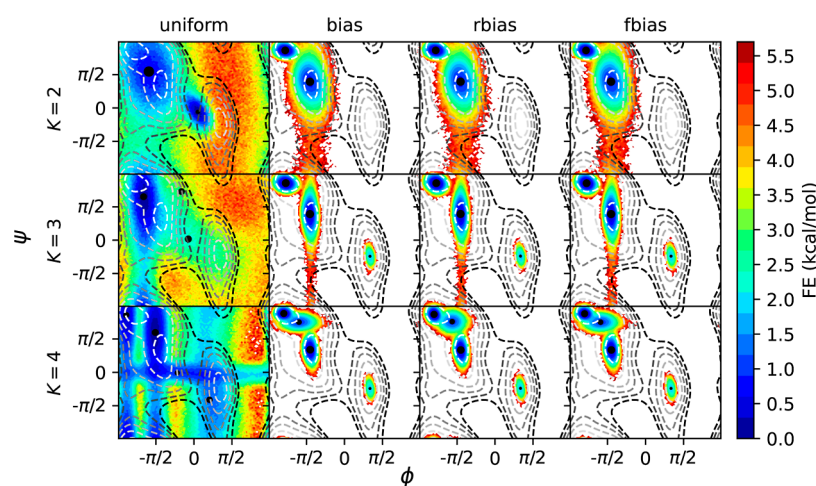


Figure 3. FE profiles obtained from GMM objects trained on BF = 10 MetaD data. Each column corresponds to a different choice of bias and each row corresponds to a different number of clusters used. These are computed as unweighted histograms from 1M samples obtained from each GMM object. Black circles placed on the FEs are the centers calculated from the reference structures corresponding to different clusters, with the size indicating their relative population. Contour lines indicate the underlying free energy surface as computed from the WT-MetaD simulation, positioned at 1.0 to 11.0 kcal/mol with a spacing of 2 kcal/mol above the global minimum.

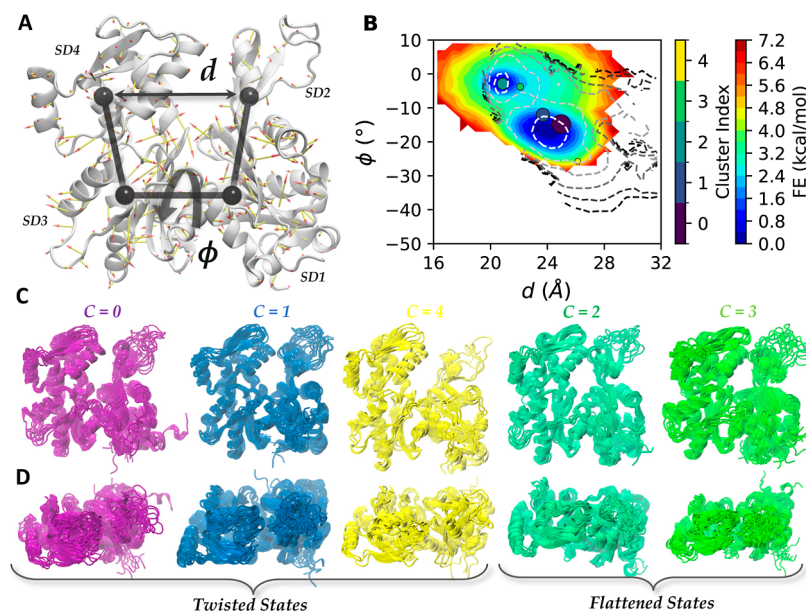


Figure 4. (A) Cartoon representation of Actin monomer. The arrows representing the magnitude and directions of the LD vector acting on 375 C_α atoms. SD1 to SD4 are four subdomains defined for the monomer.⁴¹ d is the distance between center of masses (COMs) of subdomains SD2 and SD4. ϕ is the dihedral angle defined using COMs of SD2-SD1-SD3-SD4 respectively. (B) FES calculated by performing an unweighted histogram of ~ 1 M samples generated from GMM. Contour lines represent the reweighted FE obtained from restarted OPES-MetaD trajectory using fbias frame weights. Contours are positioned at 1 to 11 kcal/mol with a spacing of 2 kcal/mol above the global minimum. Colored circles are the locations for different cluster centers weighted by relative population. (C) Snapshots of frames belonging to different clusters (front view). (D) Top view for the same.

comparable to our study with WT-MetaD, indicating that frame-weighted shapeGMM can be extended to this method as well.

Nonuniform frame-weighted shapeGMM probability densities quantitatively capture the correct free energy basins. Because we know that the free energy in dihedral space is a good proxy for the configuration space of ADP, we here quantify the accuracy of our GMM fits (which are 15-dimensional objects) by predicting this FE landscape directly from the GMMs. To do so, we generate 1M samples in Cartesian space from each GMM object and compute the FES from an unweighted histogram of the backbone dihedral

angles. **Figure 3** shows a comparison of these predicted FES with the reference FES computed directly from the WT-MetaD bias, as described above. Here we see that uniform weights produces FES that span all of dihedral space but whose minima are not centered on the true minima.

In contrast, the FESs generated from the nonuniform weighting schemes demonstrates that the clustering above captures the nature of the underlying FES as well as could be expected given a limited number of clusters. FES for $K = 2$ captures the primary C7 equatorial global minimum and C5 metastable state, while going to three or more clusters also allows resolution of the minor C7 axial basin. As should be

expected, the GMM objects only resolve the configurational landscape of our system around the minima, and cannot resolve (nonconvex) high free energy regions. Importantly, we note that the results reflect an intrinsic error due to the fact that we are fitting an anharmonic landscape to a locally harmonic model, resulting in an overestimate of the FES away from the minima. We can also compute a FES that covers the entire energy landscape using a Monte Carlo procedure described in Section S3, resulting in FES shown in Figure S3 that are qualitatively correct but which also reflect the inherent overestimation of the Gaussian model.

The comparison of FESs can be further quantified by difference metrics, which also provide an alternative metric to choose the best method or best number of clusters. In Figure S4 we show both the root-mean-squared error (RMSE) for the sampled region and the JSD as compared to the reference FES. While the uniform weights perform poorly, we see that all other weights do comparably well for 3 or more clusters. Using RMSE as a metric, rbias weights are the most accurate by a small margin, and a five state clustering is the best within the range of $K = 2$ to $K = 6$. Additionally, we compute the change in configurational entropy between all shapeGMM objects and the MetaD ground truth (ΔS_{config} in Table S1). The trend is similar to the other metrics in that the weighted objects all have a smaller magnitude ΔS_{config} compared to the uniform weights. We also include a modified uniform weight shapeGMM object (uniform_{modf} in Table S1) in which we reweight only the cluster populations (ϕ_i) after the shapeGMM fit using final bias weights. ΔS_{config} values for these objects are almost identical to the unmodified uniform object, indicating that simply reweighting cluster populations is unsatisfactory for shapeGMM.

3.3. Elucidating Conformational States of the Actin Monomer. Up to this point, we have established that we can accurately train a GMM with data weighted from MetaD or Hamiltonian reweighting for small systems. In this section, we demonstrate that this approach can provide insight into the data for a complex biochemical problem. The actin cytoskeleton, composed of filaments of actin, plays major roles in a wide range of active biological processes, including cell motility and division.^{42–44} Actin filaments are noncovalent polymers that form from head-to-tail assembly of globular actin (G-actin), which is a 375-amino acid protein consisting of four primary subdomains (Figure 4A). Each actin monomer contains a bound nucleotide that is in the form of ATP in G-actin and is eventually hydrolyzed to ADP as filaments “age”.^{43,45} The polymerization from G-actin to filamentous actin (F-actin) results in a flattening of the protein which is characterized by a reduction of the ϕ dihedral angle shown in Figure 4A.⁴³ An open question in the field is whether the flat state is metastable in solution, or whether it is only stabilized when contacting the end of a filament.⁴⁶ Additionally, structural intermediates along the flattening pathway remain elusive.

Previous efforts to directly sample the flattening of G-actin have proven to be difficult. These efforts employed umbrella sampling or MetaD on two experimentally defined coordinates ϕ and d and demonstrate the difficulty in sampling the conformational landscape of actin, either because restraining those coordinates traps you in the starting state, or because a MetaD bias can quickly push you into unphysical regions of configuration space.^{41,47} Other related efforts have investigated the role of flattening on ATP hydrolysis catalyzed by actin, and

analogous transitions in the homologous proteins Arp2 and Arp3.^{45,47–51} None of these previous studies have identified intermediate structures that might occur during flattening.

Here, we report for the first time biased MD simulations that sample reversibly the flat to twisted transition of actin by using our method to produce a position linear discriminant analysis (posLDA)³³ coordinate separating the two states. To determine the LDA reaction coordinate, we performed two short MD simulations starting from each of these states and used 10 ns from the twisted and 5 ns from the flat state (shorter because it eventually flattens;⁴⁸ see Section A1 for full details). We then performed iterative alignment of all frames in both states (using positions of all 375 C_α atoms) to the global mean and covariance as described in ref 33. LDA on the resulting aligned trajectory yielded a single posLDA coordinate that separates the twisted and flat states. The coefficients for the posLDA coordinate separating the two states is illustrated using a porcupine plot in Figure 4A. We then performed the OPES variant of WT-MetaD^{52,53} along this reaction coordinate as described in Section A1.

Frame-weighted shapeGMM trained on an OPES MetaD trajectory indicates that five distinct structural states can be occupied during a twisted to flat transition of actin. The trajectory generated contains two full round trip trajectories between flat and twisted states as measured by changes in ϕ (Figure S6), which provides sufficient sampling to investigate the observed conformations and approximate relative free energies. The FES estimated from this approach is shown in Figure S6. To increase the number of samples available for clustering purposes, we initiated new simulations using a fixed bias taken from the end of the simulation as described in Section A1. A cluster scan using these additional frames (see Figure S7) shows small kinks at $K = 3$ and $K = 5$, and in Figure 4B,C we show results for $K = 5$ in more detail. Reasonable agreement between the training set and the cross validation set in Figure S7 demonstrates a lack of overfitting on this data set.

The FES computed from the shapeGMM probability density ($K = 5$) agrees well with the MetaD free energy. Figure 4B shows the FESs computed from the shapeGMM probability density (in the colormap) and the MetaD (in the contours). The FESs are shown in the space of the ϕ and d coordinates illustrated in Figure 4A which have been used to describe the G- to F-actin transition, for better comparison with earlier MD studies.^{41,47} The MetaD simulation was performed in ϕ and the LD coordinate so was reweighted into these coordinates using the same weights fed into shapeGMM. There is impressively good quantitative agreement between the surfaces up to 3 kcal/mol ($\sim 5k_B T$) considering the very high dimensionality of the GMM. The agreement around the energy minima in this space indicates that the shapeGMM probability density is a good representation of the MetaD simulation results for these regions.

The five state shapeGMM model is in contrast to the two states that would be predicted just by looking at a 2D free energy projection. Overlain on the FES depicted in Figure 4B are circles indicating the average ϕ and d for the structures assigned to each cluster, with the size indicating their relative population. The five state clustering detected two clusters in the flat F-actin like basin ($\phi \sim -3$) and three states in or around the twisted basin ($\phi < -10$). The 2D FESs either in d and ϕ (Figure 4B) or in the sampled ϕ and LD (Figure S6) space have two basins. Clustering in this space would thus likely yield two states. The five-state shapeGMM probability

density, however, quantitatively matches the 2D FES thus demonstrating the potential oversimplification achieved in lower dimensional clusterings.

Figure 4C,D shows representative snapshots from the frames assigned to each cluster in two different orientations. To give some interpretation to these three different states, we have computed the average root-mean-square deviation (RMSD) to several published crystal or CryoEM structures of actin alone (twisted), in a filament (flattened), or in complex with an actin binding protein for the C_α atoms available in all crystal structures (numbers 7–38, 53–365 out of a total of 375). The twisted states ($C = 0, 1, 4$) all have lower RMSD to twisted than flat actin subunits, while the converse is true for the flat states ($C = 2, 3$). State $C = 4$, which is the most twisted, has the lowest rmsd to the starting structure 1NWK⁵⁴ (1.67 Å) and ADP-bound actin 1J6Z⁵⁵ (1.73 Å) than do clusters 0 and 1 (2.59 Å, 2.48 Å). It is expected based on earlier work that our simulations would produce a less twisted equilibrium state for ATP-bound actin than what is seen in the crystal structure (which was solved with a nonhydrolyzable ATP analog⁵⁴). What is interesting is that the clustering algorithm still picks up on this more twisted state as a possible structure despite the fact that early frames in the trajectory have relatively low weight (since they have little bias applied at that point).

Interestingly, states $C = 0$ and $C = 1$ have equally low rmsd to actin structures in complex with another protein as to the twisted structures considered, for example 2.59 and 2.48 Å rmsd to the twisted starting structure 1NWK, but 2.28 and 2.09 Å to the structure of actin complexed with the protein profilin (3UBS⁵⁶), which is how a large fraction of actin monomers are found in cells. This suggests that our weighted GMM models may be able to point us toward biologically relevant configurations within a conformational ensemble.

Within the flat states, the most noteworthy difference appears to be in the disordered D-loop (upper right), with cluster 3 having a significantly higher variance than cluster 2. This difference is also evident if we look at the root-mean-squared-fluctuations (RMSF) of the D-Loop residues shown in Figure S8. This lower RMSF state ($C = 2$) could correspond to one of the intermediates previously probed through MetaD simulations along a disordered-folded pathway for the D-loop, which were metastable for the ATP-bound actin used in our study, but would be expected to become more stabilized after conversion to ADP.⁵⁷ Meanwhile, on close inspection ($C = 3$) it seems to contain some more disordered structures and some partially folded structures, meaning that the higher variance could be a result of combining two subpopulations into one single state. As it stands, both flattened states have higher RMSF than all twisted states, suggesting a coupling between D-loop structure and twisting that was previously ascribed to nucleotide state (ATP vs ADP), as opposed to the conformational transition which results in ATP hydrolysis, and this would be an interesting question to consider in the future.

4. CONCLUSIONS

In this work, we present a probabilistic structural clustering protocol that can rigorously account for nonuniform frame weights. This ability allows shapeGMM to be applied, directly, to reweighted or enhanced sampling simulation data to achieve a clustering of the underlying Hamiltonian of interest. Additionally, we demonstrate that the resulting shapeGMM probability density is a good approximation to the underlying unbiased probability and can thus be used to calculate

important thermodynamic quantities, such as relative free energies and configurational entropies. To do so, we took advantage of our ability to generate biomolecular configurations from the trained clustering model; this is a unique and powerful advantage of using a probabilistic clustering model that operates directly in position space, which has not been previously exploited to our knowledge.

By applying our method to the flattening of G-actin, we have shown that this approach is capable of picking out physically meaningful structural clusters even for highly complex systems and illustrates how structural clustering on biased data can provide additional insights that would be difficult to obtain only by looking at the free-energy projected into low dimensional coordinates.

In summary, our work represents a significant advance in our ability to quantify biomolecular ensembles. In the future, we envision this approach to be useful in quantifying important biophysical processes, such as ligand binding and allosteric regulation.

APPENDIX A

A1. Simulation Details

Input files, shapeGMM objects, and analysis codes used to generate all figures are available from a github repository for this article: <https://github.com/hocky-research-group/weighted-SGMM-paper>. The simulation input files and plumed parameter files are also included in a PLUMED-NEST repository under the name plumID:24.009.

Beaded Helix. A 12-bead model designed to have two energetic ground states as left- and right-handed helices³² was simulated in LAMMPS.⁵⁸ 11 harmonic bonds between beads having rest length 1.0 and spring constant 100 form a polymer backbone. Lennard-Jones (LJ) interactions between every $i, i + 4$ pair of beads with $\sigma = 1.5$ and a cutoff length of 3.0 give rise to the helical shape. The ϵ value of this interaction dictates the stability of the helices and was the focus of our reweighting. Simulations were performed with $\epsilon = 6$ as the baseline and with $\epsilon = 8$ and $\epsilon = 4.5$ to assess the accuracy of the reweighting scheme. All nonbonded $i, i + 2$ and farther also have a repulsive WCA interaction with $\epsilon = 3.0$ and $\sigma = 3.0$ added to prevent overlap, with the ϵ for $i, i + 2$ reduced by 50%. Simulations at temperature 1.0 were performed using “fix nvt” using a simulation time step of 0.005 and a thermostat time step of 0.5. A folding/unfolding trajectory of length 50,000,000 steps was generated and analyzed as above. Here, all parameters are in reduced (LJ) units.

Alanine Dipeptide in Vacuum. Alanine dipeptide simulations were performed using GROMACS 2019.6 with PLUMED 2.9.0-dev. GROMACS mdp parameter and topology files are obtained from previous PLUMED Tutorials (Belfast-7: Replica Exchange I). AMBER99SB-ILDN force field is used with a time step of 2 fs. *NPT* ensemble is sampled using a velocity rescaling thermostat and Berendsen barostat with a temperature of 300 K and pressure 1 bar. For MetaD simulations we used PACE = 500, SIGMA = 0.3 Å (for both ϕ and ψ) and HEIGHT = 1.2 kcal/mol. PLUMED input files are available in our paper's github repository for complete details.

Actin Monomer. Actin simulations were also performed using GROMACS 2019.6 with PLUMED 2.9.0-dev. G-actin with a bound ATP was built and equilibrated at 310 K as described previously.⁴⁸ The structure of the twisted, ATP-bound actin is derived from the crystal structure with PDB ID

1NWK,⁵⁴ while that in the flat state is taken from PDB ID 2ZWH,⁵⁹ with the nucleotide, magnesium ion, and surrounding water replaced with ATP as described previously. MD simulation for ~5 ns was performed to relax the starting structure. *NPT* simulation was performed with a 2 fs time step. Parrinello–Rahman barostat is used along with a velocity rescaling thermostat with a temperature of 310 K and pressure 1 bar. For OPES we used PACE = 500, BIASFACTOR = 12, BARRIER = 15.0 kcal/mol and a multiple time step stride of 2. Two UPPER_WALLS were employed ~ -1° and 31 Å for ϕ and d respectively. We also used one UPPER_WALLS at +40.0 and one LOWER_WALLS at -40.0 for the posLDA coordinate. All the walls used were quadratic with a spring constant of KAPPA = 500 kcal/mol/nm². PLUMED input files are available in our paper's github repository.

We performed ~1 s of sampling along this LD coordinate and dihedral angle ϕ using the On the Fly Probability Enhanced Sampling variant of MetaD (OPES-MetaD).^{52,53} This method uses a kernel density estimate of the probability distribution over the whole space for biasing rather than building this bias through the sum of Gaussians. The bias at time t for CV value s_i is given by the expression

$$V(s_i) = k_B T \left(\frac{\gamma - 1}{\gamma} \right) \log \left(\frac{P_t(s_i)}{Z_t} + \varepsilon \right) \quad (14)$$

Here, $P_t(s)$ is the current estimate of the probability distribution, Z_t is a normalization factor. Finally, $\varepsilon = \exp\left(\frac{\Delta E}{k_B T \gamma - 1}\right)$ is a regularization constant that ensures the maximum bias that can be applied is ΔE . OPES-MetaD data can be reweighted similarly to standard WT-MetaD, using the exponential of the bias (which is similar to *rbias* for MetaD) or using the estimated free energy of each frame from the final bias.⁵²

We chose the OPES variant of MetaD because (a) literature precedent suggests that it converges more quickly than standard WT-MetaD, and (b) it allows us to set an free-energy cutoff above which bias is not applied (in this case 15 kcal/mol) which limits the amount of unphysical exploration, in a similar manner to Metabasin-MetaD that we previously showed was desirable for this problem.⁴⁷ Even with this energy cutoff, we needed to include upper and lower walls to prevent overflattening or overtwisting observed here and in prior attempts by us.⁴⁸

A cluster scan on our OPES trajectory (Figure S7) showed a large difference between training and cross-validation curves. Hence we decided to generate additional training frames. We did this by taking the bias accumulated after 900 ns of OPES simulation and started four 1 ns simulations with random velocities from each of 191 initial configurations from the initial trajectory (separated by 5 ns each), saving every 5 ps; this resulted in ~153k frames available for clustering. The resulting training and cross-validation curves are in much better agreement as discussed in the main text; hence, these data were used for clustering and analysis.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jctc.4c00223>.

Choosing training data; clustering untempered metadynamics; ADP FES computed by evaluating GMM on

WT-MetaD samples; error analysis for GMM Free energies; FEs from GMM for cluster size 5 and 6; OPES-MetaD simulation of Actin (~1 μ s); cluster scans; variance of D-loop in actin; configurational entropies from GMMs (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Glen M. Hocky – Department of Chemistry, New York University, New York, New York 10003, United States; Simons Center for Computational Physical Chemistry, New York University, New York, New York 10003, United States; orcid.org/0000-0002-5637-0698; Email: hockyg@nyu.edu

Martin McCullagh – Department of Chemistry, Oklahoma State University, Stillwater, Oklahoma 74078, United States; orcid.org/0000-0002-8603-4388; Email: martin.mccullagh@okstate.edu

Authors

Subarna Sasmal – Department of Chemistry, New York University, New York, New York 10003, United States; orcid.org/0000-0003-1599-7539

Triasha Pal – Department of Chemistry, New York University, New York, New York 10003, United States

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.jctc.4c00223>

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

MM would like to acknowledge funding from the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number R01AI166050 and the National Science Foundation under award 2238706. S.S., T.P., and G.M.H. were supported by the National Institutes of Health through the award R35GM138312. S.S. was also partially supported by a graduate fellowship from the Simons Center for Computational Physical Chemistry (SCCPC) at NYU (SF grant no. 839534). This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise, and simulations were partially executed on resources supported by the SCCPC at NYU.

■ REFERENCES

- (1) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.
- (2) Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **2008**, *100*, 020603.
- (3) Darve, E.; Pohorille, A. Calculating free energies using average force. *J. Chem. Phys.* **2001**, *115*, 9169–9183.
- (4) Miao, Y.; Feher, V. A.; McCammon, J. A. Gaussian accelerated molecular dynamics: Unconstrained enhanced sampling and free energy calculation. *J. Chem. Theory Comput.* **2015**, *11*, 3584–3595.
- (5) Maragliano, L.; Vanden-Eijnden, E. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett.* **2006**, *426*, 168–175.
- (6) Abrams, J. B.; Tuckerman, M. E. Efficient and direct generation of multidimensional free energy surfaces via adiabatic dynamics without coordinate transformations. *J. Phys. Chem. B* **2008**, *112*, 15742–15757.

- (7) Hénin, J.; Lelièvre, T.; Shirts, M. R.; Valsson, O.; Delemotte, L. Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0]. *Living J. Comput. Mol. Sci.* **2022**, *4*, 1583.
- (8) Singhal, N.; Snow, C. D.; Pande, V. S. Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.* **2004**, *121*, 415–425.
- (9) Kasson, P.; Kelley, N. W.; Singhal, N.; Vrljic, M.; Brunger, A. T.; Pande, V. S. Ensemble molecular dynamics yields submillisecond kinetics and intermediates of membrane fusion. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11916–11921.
- (10) Klem, H.; Hocky, G. M.; McCullagh, M. Size-and-shape space gaussian mixture models for structural clustering of molecular dynamics trajectories. *J. Chem. Theory Comput.* **2022**, *18*, 3218–3230.
- (11) Deuffhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Lin. Algebra Appl.* **2000**, *315*, 39–59.
- (12) Deuffhard, P.; Weber, M. Robust Perron cluster analysis in conformation dynamics. *Lin. Algebra Appl.* **2005**, *398*, 161–184.
- (13) Keller, B.; Daura, X.; Van Gunsteren, W. F. Comparing geometric and kinetic cluster algorithms for molecular simulation data. *J. Chem. Phys.* **2010**, *132*, 074110.
- (14) Peng, J. H.; Wang, W.; Yu, Y. Q.; Gu, H. L.; Huang, X. Clustering algorithms to analyze molecular dynamics simulation trajectories for complex chemical and biological systems. *Chin. J. Chem. Phys.* **2018**, *31*, 404–420.
- (15) Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.* **2021**, *121*, 9722–9758.
- (16) Marinelli, F.; Pietrucci, F.; Laio, A.; Piana, S. A kinetic model of Trp-cage folding from multiple biased molecular dynamics simulations. *PLoS Comput. Biol.* **2009**, *5*, No. e1000452.
- (17) Tiwary, P.; Parrinello, M. From metadynamics to dynamics. *Phys. Rev. Lett.* **2013**, *111*, 230602.
- (18) Ray, D.; Parrinello, M. Kinetics from Metadynamics: Principles, Applications, and Outlook. *J. Chem. Theory Comput.* **2023**, *19*, 5649–5670.
- (19) Bonomi, M.; Barducci, A.; Parrinello, M. Reconstructing the Equilibrium Boltzmann Distribution from Well-Tempered Metadynamics. *J. Comput. Chem.* **2009**, *30*, 1615–1621.
- (20) Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A* **1978**, *34*, 827–828.
- (21) Horn, B. K. P. Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Am. A* **1987**, *4*, 629.
- (22) Goodall, C. Procrustes Methods in the Statistical Analysis of Shape. *J. Roy. Stat. Soc. B* **1991**, *53*, 285–321.
- (23) Theobald, D. L.; Wuttke, D. S. Empirical bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian procrustes problem. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18521–18527.
- (24) Gebru, I. D.; Alameda-Pineda, X.; Forbes, F.; Horaud, R. EM Algorithms for Weighted-Data Clustering with Application to Audio-Visual Scene Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2402–2415.
- (25) Tuckerman, M. E. *Statistical Mechanics: Theory and Molecular Simulation*; Oxford University Press, 2023.
- (26) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.
- (27) Qian, H. Relative entropy: Free energy associated with equilibrium fluctuations and nonequilibrium deviations. *Phys. Rev. E* **2001**, *63*, 042103.
- (28) Lindorff-Larsen, K.; Ferkinghoff-Borg, J. Similarity Measures for Protein Ensembles. *PLoS One* **2009**, *4*, No. e4203.
- (29) Ming, D.; Wall, M. E. Quantifying allosteric effects in proteins. *Proteins* **2005**, *59*, 697–707.
- (30) Wall, M. E. Ligand binding, protein fluctuations, and allosteric free energy. *AIP Conf. Proc.* **2006**, *851*, 16–33.
- (31) Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theor.* **1991**, *37*, 145–151.
- (32) Hartmann, M. J.; Singh, Y.; Vanden-Eijnden, E.; Hocky, G. M. Infinite switch simulated tempering in force (FISST). *J. Chem. Phys.* **2020**, *152*, 244120.
- (33) Sasmal, S.; McCullagh, M.; Hocky, G. M. Reaction Coordinates for Conformational Transitions Using Linear Discriminant Analysis on Positions. *J. Chem. Theory Comput.* **2023**, *19*, 4427–4435.
- (34) Bussi, G.; Laio, A. Using metadynamics to explore complex free-energy landscapes. *Nat. Rev. Phys.* **2020**, *2*, 200–212.
- (35) Dama, J. F.; Parrinello, M.; Voth, G. A. Well-tempered metadynamics converges asymptotically. *Phys. Rev. Lett.* **2014**, *112*, 240602.
- (36) Tiwary, P.; Parrinello, M. A time-independent free energy estimator for metadynamics. *J. Phys. Chem. B* **2015**, *119*, 736–742.
- (37) The PLUMED consortium. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **2019**, *16*, 670–673.
- (38) Schäfer, T. M.; Settanni, G. Data reweighting in metadynamics simulations. *J. Chem. Theory Comput.* **2020**, *16*, 2042–2052.
- (39) Giberti, F.; Cheng, B.; Tribello, G. A.; Ceriotti, M. Iterative unbiasing of quasi-equilibrium sampling. *J. Chem. Theory Comput.* **2020**, *16*, 100–107.
- (40) Tobias, D. J.; Brooks, C. L. Conformational equilibrium in the alanine dipeptide in the gas phase and aqueous solution: A comparison of theoretical results. *J. Phys. Chem.* **1992**, *96*, 3864–3870.
- (41) Saunders, M. G.; Tempkin, J.; Weare, J.; Dinner, A. R.; Roux, B.; Voth, G. A. Nucleotide regulation of the structure and dynamics of G-actin. *Biophys. J.* **2014**, *106*, 1710–1720.
- (42) Pollard, T. D.; Cooper, J. A. Actin, a Central Player in Cell Shape and Movement. *Science* **2009**, *326*, 1208–1212.
- (43) Dominguez, R.; Holmes, K. C. Actin Structure and Function. *Annu. Rev. Biophys.* **2011**, *40*, 169–186.
- (44) Pollard, T. D. Actin and Actin-Binding Proteins. *Cold Spring Harbor Perspect. Biol.* **2016**, *8*, a018226.
- (45) McCullagh, M.; Saunders, M. G.; Voth, G. A. Unraveling the mystery of ATP hydrolysis in actin filaments. *J. Am. Chem. Soc.* **2014**, *136*, 13053–13058.
- (46) Zsolnay, V.; Katkar, H. H.; Chou, S. Z.; Pollard, T. D.; Voth, G. A. Structural basis for polarized elongation of actin filaments. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 30458–30464.
- (47) Dama, J. F.; Hocky, G. M.; Sun, R.; Voth, G. A. Exploring valleys without climbing every peak: more efficient and forgiving metabasin metadynamics via robust on-the-fly bias domain restriction. *J. Chem. Theory Comput.* **2015**, *11*, 5638–5650.
- (48) Hocky, G. M.; Dannenhoffer-Lafage, T.; Voth, G. A. Coarse-grained directed simulation. *J. Chem. Theory Comput.* **2017**, *13*, 4593–4603.
- (49) Hocky, G. M.; Sindelar, C. V.; Cao, W.; Voth, G. A.; De La Cruz, E. M. Structural basis of fast-and slow-severing actin–cofilactin boundaries. *J. Biol. Chem.* **2021**, *296*, 100337.
- (50) Singh, Y.; Hocky, G. M.; Nolen, B. J. Molecular dynamics simulations support a multistep pathway for activation of branched actin filament nucleation by Arp2/3 complex. *J. Biol. Chem.* **2023**, *299*, 105169.
- (51) Mukadam, F.; Peña Ccoa, W. J.; Hocky, G. M. Molecular simulation approaches to probing the effects of mechanical forces in the actin cytoskeleton. *Cytoskeleton* **2024**.
- (52) Invernizzi, M.; Parrinello, M. Rethinking metadynamics: from bias potentials to probability distributions. *J. Phys. Chem. Lett.* **2020**, *11*, 2731–2736.
- (53) Invernizzi, M.; Piaggi, P. M.; Parrinello, M. Unified approach to enhanced sampling. *Phys. Rev. X* **2020**, *10*, 041034.
- (54) Graceffa, P.; Dominguez, R. Crystal structure of monomeric actin in the ATP state: structural basis of nucleotide-dependent actin dynamics. *J. Biol. Chem.* **2003**, *278*, 34172–34180.

(55) Otterbein, L. R.; Graceffa, P.; Dominguez, R. The crystal structure of uncomplexed actin in the ADP state. *Science* **2001**, *293*, 708–711.

(56) Porta, J. C.; Borgstahl, G. E. Structural basis for profilin-mediated actin nucleotide exchange. *J. Mol. Biol.* **2012**, *418*, 103–116.

(57) Pfaendtner, J.; Branduardi, D.; Parrinello, M.; Pollard, T. D.; Voth, G. A. Nucleotide-dependent conformational states of actin. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 12723–12728.

(58) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; et al. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **2022**, *271*, 108171.

(59) Oda, T.; Iwasa, M.; Aihara, T.; Maéda, Y.; Narita, A. The nature of the globular-to fibrous-actin transition. *Nature* **2009**, *457*, 441–445.