Interpretable embeddings from molecular simulations using Gaussian mixture variational autoencoders

IOP Publishing Mach. Learn.: Sci. Technol. 1 (2020) 015012 https://doi.org/10.1088/2632-2153/ab80b7

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
19 December 2019

REVISED
4 March 2020

ACCEPTED FOR PUBLICATION
17 March 2020

PUBLISHED
27 April 2020

PAPER

Interpretable embeddings from molecular simulations using Gaussian mixture variational autoencoders

Yasemin Bozkurt Varolgüneş[1,2,3], **Tristan Bereau**[1] and **Joseph F Rudzinski**[1]

[1] Max Planck Institute for Polymer Research, Mainz 55128, Germany
[2] Department of Electrical & Electronics Engineering, Koc University, Sariyer, Istanbul 34450, Turkey
[3] Author to whom any correspondence should be addressed.

E-mail: bozkurty@mpip-mainz.mpg.de

Keywords: variational autoencoders, dimensionality reduction, clustering, Markov state models, molecular dynamics simulations

Supplementary material for this article is available online

Abstract

Extracting insight from the enormous quantity of data generated from molecular simulations requires the identification of a small number of collective variables whose corresponding low-dimensional free-energy landscape retains the essential features of the underlying system. Data-driven techniques provide a systematic route to constructing this landscape, without the need for extensive *a priori* intuition into the relevant driving forces. In particular, autoencoders are powerful tools for dimensionality reduction, as they naturally force an information bottleneck and, thereby, a low-dimensional embedding of the essential features. While variational autoencoders ensure continuity of the embedding by assuming a unimodal Gaussian prior, this is at odds with the multi-basin free-energy landscapes that typically arise from the identification of meaningful collective variables. In this work, we incorporate this physical intuition into the prior by employing a Gaussian mixture variational autoencoder (GMVAE), which encourages the separation of metastable states within the embedding. The GMVAE performs dimensionality reduction and clustering within a single unified framework, and is capable of identifying the inherent dimensionality of the input data, in terms of the number of Gaussians required to categorize the data. We illustrate our approach on two toy models, alanine dipeptide, and a challenging disordered peptide ensemble, demonstrating the enhanced clustering effect of the GMVAE prior compared to standard VAEs. The resulting embeddings appear to be promising representations for constructing Markov state models, highlighting the transferability of the dimensionality reduction from static equilibrium properties to dynamics.

1. Introduction

Particle-based computer simulations can provide unprecedented mechanistic insight into the driving forces of complex molecular systems, in contexts ranging from biochemistry to materials science [1–3]. These simulations rely on numerical integration of the relevant equations of motion as a means to navigate the system's conformational space. Due to the high dimensionality of this space, which prevents the exhaustive enumeration of all microstates, exploration is typically achieved through importance sampling [4]. Conformational sampling leads to an estimate of the potential energy landscape (PEL), which follows a Boltzmann distribution at equilibrium. Unfortunately, characterization of the PEL suffers from the so-called *curse of dimensionality* [5]—organization of the data in the high-dimensional space is challenging due to low population density. This problem is often remedied by projecting the PEL onto a lower-dimensional manifold, i.e. by performing a dimensionality reduction. By averaging over presumably unimportant degrees of freedom, the resulting low-dimensional surface represents a *free*-energy landscape (FEL). The ideal FEL distinguishes between microstates that are separated by large barriers on the PEL, yielding a partitioning of configuration space into collections of microstates, i.e. metastable basins. If all the largest barriers are

accounted for, intra-basin diffusion will occur much faster than inter-barrier crossing events, allowing an accurate, albeit coarse-grained, description of both the static and dynamical properties of the system.

The essential degrees of freedom that define the low-dimensional representation, commonly referred to as collective variables (CVs), are traditionally identified through expert physical/chemical intuition that is often rather specific for the particular system or process of interest [6–9]. Beyond the characterization of the FEL, these CVs can also be used for enhanced sampling [10], or for the construction of low-dimensional configuration-space discretizations, for instance when building Markov state models (MSMs) [11]. Although the manual selection of CVs can be extremely effective for practitioners with insight into the system, the approach is difficult to extend systematically and is susceptible to missing unanticipated or subtle features of the FEL that may nonetheless play an important role in the relevant phenomena. Data-driven techniques provide an alternative route by inferring the important features directly from the data. There is a long history of methods for finding an optimal low-dimensional representation from a given set of data, employing both linear (e.g. principal component analysis [12], time-lagged independent component analysis [13]) and non-linear (e.g. Isomap [14], diffusion map [15] and Sketchmap [16]) transformations.

In the last couple of years, there has been a growing interest in applying (deep) neural networks to automate the discovery of CVs [17–21]. One architecture that stands out as conceptually appealing is the autoencoder [22]. An autoencoder is a bow-tie-shaped network that forces an information compression in the bottleneck region. While the first half of the network (the encoder) reduces the input to a predefined lower dimension, the second half (the decoder) aims at transforming from the low-dimensional to the original representation. The weights of the neural network are tuned to minimize an objective or *loss* function, which typically penalizes deviations between input and output data. As such, the autoencoder aims at discovering a *latent space* (embedding) that faithfully describes the essential features of the high-dimensional input data. This makes autoencoders well suited for constructing low-dimensional FELs from molecular simulation data [17, 23, 24].

Traditional autoencoders lack continuity in the latent space, preventing interpolation between training points and, thus, its generative ability. Variational autoencoders (VAEs) remedy this limitation by modeling the input probability distribution using Bayesian inference [25]. VAEs enable sampling new data from the learned distribution (i.e. VAEs are generative models), and are also well suited to provide interpretable and disentangled data representations in the low-dimensional space [26]. Within the VAE framework, the latent distribution is forced to resemble a predefined probability distribution, called the *prior*. Although the VAE framework does not impose any particular prior distribution, it is often chosen as a normal distribution for computational convenience. This prior induces an 'anti-clustering' effect in the latent space, which can prohibit the identification of meaningful clusters and impede the construction of optimal FELs from molecular simulations. The autoencoder-based approaches were recently extended to explicitly incorporate the temporal nature of the data via a time lag in the network architecture [27, 28]. These time-lagged autoencoders aim to retain information about the slowest dynamical modes sampled in the underlying simulation trajectory and, as a consequence, may encourage metastable clustering in the latent space. However, they are also limited in terms of characterizing the hierarchy of long timescale processes [29] and only indirectly address the anti-clustering issue.

In this work, we propose to directly acknowledge the multi-basin structure of an ideal FEL by employing a Gaussian mixture model [30] as the prior distribution for the VAE latent space. The resulting *Gaussian mixture variational autoencoder* (GMVAE) retains the computational ease and reconstruction fidelity of traditional VAEs, while enforcing a more faithful description of the underlying physics: the resulting FEL clearly distinguishes between metastable basins separated by large free-energy barriers. We demonstrate the benefits of the GMVAE approach through explicit comparisons with the traditional VAE for two widely-studied toy models and for the standard benchmark system for conformational dynamics, alanine dipeptide, as well as a more challenging disordered peptide ensemble. To ensure the presence of distinct distributions in the latent space, the GMVAE introduces a categorical variable that (probabilistically) assigns each input configuration to the set of clusters. Thus, the GMVAE simultaneously performs dimensionality reduction and unsupervised clustering. Remarkably, the GMVAE clustering is capable of identifying the inherent dimensionality of the input data, in terms of the number of Gaussians required to categorize the data. In the case of hierarchical input data (i.e. data with distinct dimensionality depending on the level of resolution), we show that the GMVAE makes a reasonable prediction for the number of clusters, independent of the given hyperparameter, based on the dimensionality of the latent space and characteristics of the data. Beyond the representation of static equilibrium properties, by constructing MSMs from the GMVAE embedding, we show that our approach is also a promising avenue for accurately describing the long timescale *dynamical* properties of the data. In contrast to recent deep neural network approaches that aim to directly model the *propagator* of the system's dynamics [31, 32], the construction of MSMs from the learned

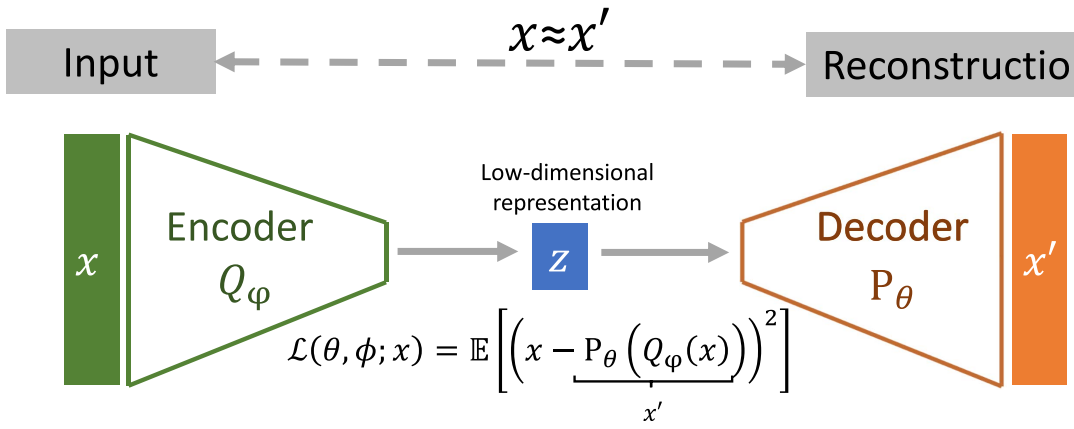


Figure 1. Schematic of an autoencoder architecture with mean-squared error reconstruction loss.

FEL offers a different strategy: explicitly testing to what extent a representation appropriate for the statics is directly amenable for the dynamics.

2. Theory and methods

2.1. Autoencoder

Autoencoders are special types of neural networks that are used for the task of representation learning in an unsupervised manner. They are composed of two connected parts: the encoder compresses the input signal to a low-dimensional representation, whereas, the decoder aims to reconstruct the input at full dimensionality from the reduced-space representation. The *reconstruction loss*, usually defined as either the mean-squared error or cross-entropy between the input, x, and the output, x', is minimized via backpropagation. Since the bottleneck dimension is typically much less than the original dimension, autoencoders learn the most compact representation of the input. Furthermore, because neural networks are universal function approximators, the learned data projections can generally preserve much more of the relevant information than with PCA or other basic linear projection techniques. Figure 1 shows the schematic structure of an autoencoder with mean-squared error loss. There are different types of autoencoders which are tailored for special tasks. For instance, sparse autoencoders impose sparsity constraints during optimization, whereas convolutional autoencoders utilize convolutional layers instead of fully-connected layers, in which case they learn the optimal filters. Variational autoencoders, which model the latent space probabilistically, are used for generative purposes, i.e. they can create new samples that look like the ones in the training dataset without simple data replication.

2.2. Variational autoencoder (VAE)

Variational autoencoders were introduced in [25]. In general, the theory of VAEs is approached from two different perspectives: variational inference and neural networks. This section starts with the former interpretation and then illustrates the connection between them. We mostly follow the notation and reasoning used in [33]. The input data and the latent variable are denoted by x and z, respectively.

The objective of the VAE is to find the posterior distribution $P(z|x)$, which can be written in terms of the likelihood $P(x|z)$, the prior $P(z)$, and the marginal probability density of x, $P(x)$, using Bayes law as

$$P(z|x) = \frac{P(x|z)P(z)}{P(x)} \,. \tag{1}$$

The denominator $P(x)$ is called the evidence and it could, in principle, be calculated using

$$P(x) = \int \mathrm{d}z P(x|z)P(z) \,, \tag{2}$$

once the prior is selected. However, the calculation is typically intractable, as it needs to be evaluated over all configurations of the latent variable z. Therefore, the posterior is approximated using *variational inference* with a chosen easy-to-evaluate family of distributions $Q_\phi(z|x)$, e.g. Gaussian functions, where ϕ is the variational parameter of the distribution. In particular, $P(z|x)$ is inferred using $Q_\phi(z|x)$ by reformulating the

problem within an optimization framework, such that the Kullback-Leibler divergence between $Q_\phi(z|x)$ and $P(z|x)$ is minimized. The KL divergence between Q and P is defined as

$$\begin{aligned} D_{\mathrm{KL}}[Q_\phi(z|x)||P(z|x)] &= \sum_z Q_\phi(z|x) \log \frac{Q_\phi(z|x)}{P(z|x)} \\ &= \mathbb{E}\left[\log \frac{Q_\phi(z|x)}{P(z|x)}\right] \\ &= \mathbb{E}[\log Q_\phi(z|x) - \log P(z|x)] \,. \end{aligned} \tag{3}$$

Equation (1) is then inserted into the posterior definition

$$\begin{aligned} D_{\mathrm{KL}}[Q_\phi(z|x)||P(z|x)] &= \mathbb{E}\left[\log Q_\phi(z|x) - \log \frac{P(x|z)P(z)}{P(x)}\right] \\ &= \mathbb{E}[\log Q_\phi(z|x) - \log P(x|z) - \log P(z) + \log P(x)] \,. \end{aligned} \tag{4}$$

Since the expectation is taken over z, $P(x)$ can be moved out of the expectation

$$D_{\mathrm{KL}}[Q_\phi(z|x)||P(z|x)] - \log P(x) = -\underbrace{\mathbb{E}[\log P(x,z) - \log Q_\phi(z|x)]}_{\mathrm{ELBO}(\phi)} \,. \tag{5}$$

The initial objective of minimizing the KL divergence between the exact and the approximate posterior is equivalent to maximizing the ELBO (Evidence Lower BOund), defined in equation (5).

Equation (5) can also be rewritten in terms of a different KL divergence:

$$D_{\mathrm{KL}}[Q_\phi(z|x)||P(z|x)] - \log P(x) = D_{\mathrm{KL}}[Q_\phi(z|x)||P(z)] - \mathbb{E}[\log P(x|z)] \,. \tag{6}$$

Here the neural network perspective comes into play, as depicted schematically in figure 2(a). $Q_\phi(z|x)$ acts like an encoder (*inference*), and transforms the data into the latent variable z. On the other hand, $P(z|x)$ (which can also be parametrized with the network parameter θ as $P_\theta(z|x)$[4]) generates the data from the latent representation, analogous to a decoder (*generator*). The parameters correspond to the weights and biases of the neural networks. Note that the initial aim is to minimize $D_{\mathrm{KL}}[Q_\phi(z|x)||P(z|x)]$, which is equivalent to minimizing the RHS of equation (6). The first term enforces the encoder to be similar to the chosen prior $P(z)$, which acts as a regularization, whereas the second term on the RHS deals with how well the reconstructions match the original input.

2.2.1. Standard selections for the family of inference distributions and for the prior distribution

In order to use equation (6) in an optimization procedure, both the family of distributions for inference, $Q_\phi(z|x)$, as well as the prior distribution, $P(z)$, must be specified. The most common assumption is that $Q_\phi(z|x)$ ($P(z)$) is a unimodal Gaussian distribution with mean $\mu(x)$ (0) and diagonal covariance $\Sigma(x)$ (1). Then, $D_{\mathrm{KL}}[Q_\phi(z|x)||P(z)]$ has a closed form solution:

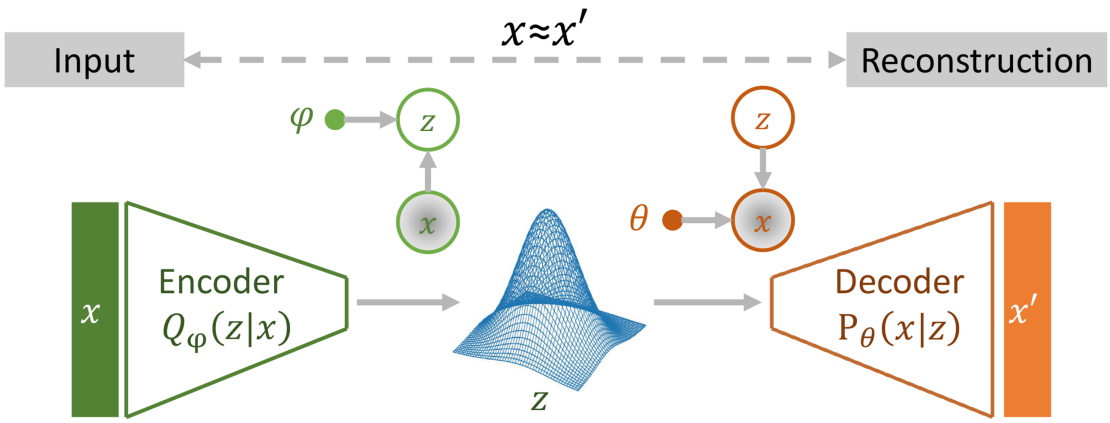
$$\begin{aligned} D_{\mathrm{KL}}[Q_\phi(z|x)||P(z)] &= D_{\mathrm{KL}}[\mathcal{N}(\mu(x), \Sigma(x))||\mathcal{N}(0,1)] \\ &= \frac{1}{2}\left(\mathrm{tr}(\Sigma(x)) + \mu(x)^T\mu(x) - d - \log\det(\Sigma(x))\right) \,, \end{aligned} \tag{7}$$

where d is the dimension of the Gaussian and tr denotes the trace. Although the unimodal Gaussian assumption simplifies the calculations, it also restricts the possible latent space representations, and may hinder the performance of the variational autoencoder by pushing the latent space to be described by highly-overlapping clusters.

2.3. Gaussian mixture variational autoencoder

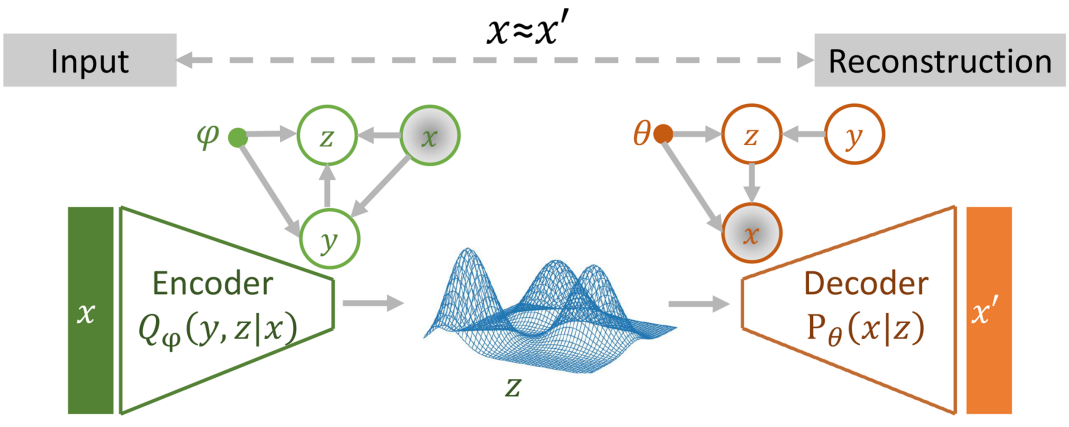
This section is largely distilled from the discussion and insights presented in [34]. The term Gaussian mixture variational autoencoder is open to misinterpretations. There exist several distinct architectures given this name, with variations in the choice of generative or inference models [30, 35–37]. In the present work, we take both the approximate posterior, (i.e. the family of distribution functions for inference), $Q_\phi(y, z|x)$, and the latent space distribution (i.e. the prior), $P(z)$, to be Gaussian mixtures. Note that we have introduced

[4]Both of the notations are used interchangeably.



$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{Q_\phi(x,z)}\left[\log P_\theta(x, z) - \log Q_\phi(z|x)\right]$$
$$\mathcal{L}(\theta, \phi; x) = \underbrace{\mathbb{E}_{Q_\phi(z|x)}[\log P_\theta(x|z)]}_{\text{Reconstruction}} - \underbrace{\mathrm{D}_{\mathrm{KL}}\left(Q_\phi(z|x)||P(z)\right)}_{\text{Regularization}}$$

(a) Schematic of a variational autoencoder with unimodal Gaussian prior.



$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{Q_\phi(y,z|x)}\left[\log P_\theta(x, y, z) - \log Q_\phi(y, z|x)\right]$$
$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{Q_\phi(y,z|x)}\left[\underbrace{\log \frac{P(y)}{Q_\phi(y|x)}}_{\text{Entropy}} + \underbrace{\log \frac{P_\theta(z|y)}{Q_\phi(z|x,y)}}_{\text{Regularization}} + \underbrace{\log P_\theta(x|z)}_{\text{Reconstruction}}\right]$$

(b) Schematic of a Gaussian mixture variational autoencoder.

Figure 2. (a) The VAE and (b) GMVAE architectures. In the probabilistic graph representation, circle nodes represent the random variables, and directed edges represent statistical dependencies between the variables in the two ends. Dot nodes are used to indicate the parameters of the model, while some of the nodes are intentionally filled to differentiate the observed random variables from the non-observed ones which are left empty.

a categorical variable, y, which identifies which Gaussian each particular data point belongs to. The inference model can be written as

$$Q_\phi(y, z|x) = Q_\phi(y|x)Q_\phi(z|x, y)\,. \tag{8}$$

The latent space is composed of k distinct Gaussians, i.e. $Q_\phi(z|x, y_i)$ is assumed to be Gaussian, where $i \in 0, 1, \ldots, k-1$. Thus, the approximate posterior becomes a Gaussian mixture.

Similar to equation (5), the ELBO can be written as

$$\mathrm{ELBO_m} = \mathbb{E}_{Q_\phi(y,z|x)}\left[\log P_\theta(x, y, z) - \log Q_\phi(y, z|x)\right]\,, \tag{9}$$

where the number of Gaussians, k, is a hyperparameter, and the subscript m is used to distinguish $\mathrm{ELBO_m}$ from the VAE ELBO. $P_\theta(x, y, z)$ can be written as $P_\theta(x, y, z) = P_\theta(x|y, z)P_\theta(z|y)P(y)$ using conditioning without any assumptions. Then, by assuming that x is conditionally independent of y, i.e.

Table 1. Distributions in the GMVAE model. Left (right) column corresponds to the distributions in the encoder (decoder) part.

$Q(z\|x,y)$	$=\mathcal{N}(\mu_z(x,y),\sigma_z^2(x,y))$	$P(y)$	$=\text{Uniform}(\frac{1}{k})$
$Q(y\|x)$	$=\text{Multinomial}(f(x))$	$P(z\|y)$	$=\mathcal{N}(\mu_z(y),\sigma_z^2(y))$
		$P(x\|z)$	$=\mathcal{N}(\mu_x(z),\sigma_x^2(z))$

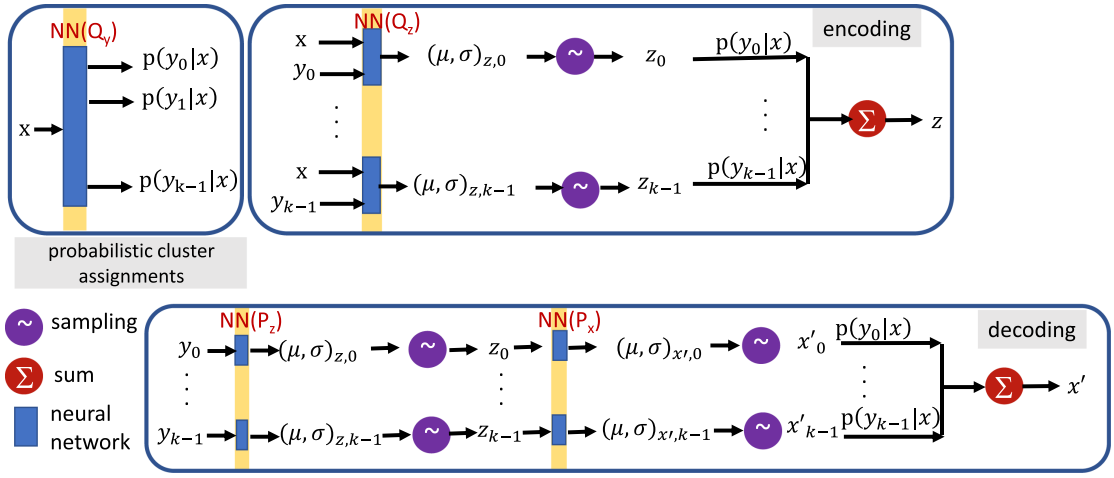


Figure 3. Schematic of the GMVAE workflow.

$P_\theta(x|y,z) = P_\theta(x|z)$ (see the graph representation in figure 2(b)), the joint probability can be expressed as

$$P_\theta(x,y,z) = P_\theta(x|z)P_\theta(z|y)P(y)\,. \tag{10}$$

By inserting equations (8) and (10) into equation (9), ELBO_m becomes

$$\begin{aligned}\text{ELBO}_\text{m} &= \mathbb{E}_{Q(y,z|x)}[\log P(y)P_\theta(z|y)P_\theta(x|z) - \log Q_\phi(y|x)Q_\phi(z|x,y)] \\ &= \mathbb{E}_{Q(y,z|x)}\left[\log P(y) - \log Q_\phi(y|x) + \log\frac{P_\theta(z|y)}{Q_\phi(z|x,y)} + \log P_\theta(x|z)\right].\end{aligned} \tag{11}$$

Similar to the VAE, the third and fourth terms represent regularization and reconstruction contributions to the loss, respectively. The initial prior on y is selected as a uniform multinomial distribution, while $\mathbb{E}_{Q(y,z|x)}[\log Q_\phi(y|x)]$ can be interpreted as a conditional entropy, reflecting how informative x is on y. To directly control the impact of the clustering relative to the other loss terms during training, we introduced a weighting factor, α, on the mutual information between x and y:

$$\text{ELBO}_\text{m} = \mathbb{E}_{Q(y,z|x)}\left[\log P(y) - \alpha\log Q_\phi(y|x) + \log\frac{P_\theta(z|y)}{Q_\phi(z|x,y)} + \log P_\theta(x|z)\right]. \tag{12}$$

Figure 3 presents a more detailed schematic of the GMVAE architecture, while table 1 presents a summary of the probability distributions utilized in the model. First, data points are probabilistically assigned to k clusters ($\text{NN}(\text{Q}_y)$). $Q(y|x)$ represents these cluster assignment probabilities, and has a multinomial distribution. Since each cluster is assumed to have Gaussian distribution in the latent space, the mean and variance of each of these Gaussians ($Q(z|x,y)$) are learned via the encoder part of the neural network ($\text{NN}(\text{Q}_z)$). The low-dimensional representation, z, is then obtained by first sampling and then taking the expected value of these samples, i.e. $z = \sum_{i=0}^{k-1} p(y_i|x)z_i$. As the first step in decoding, the moments of the corresponding low-dimensional representation z is learned by $\text{NN}(\text{P}_z)$ from each Gaussian-distributed individual cluster y_i, which is then followed by a sampling operation. $P(y)$ in the decoder is assumed to be uniformly distributed among the k clusters. Next, using the encodings, z_i's, the associated x reconstructions are obtained again by sampling from the x' by the $\text{NN}(\text{P}_x)$. Similar to the encoder, the decoder obtains a fixed reconstruction by taking the expected value of x'_i's.

2.3.1. Determination of cluster labels and thresholding scheme

The clustering within the GMVAE is probabilistic, i.e. each data point is assigned membership probabilities (between 0 and 1) to each of the clusters. Since most configurations are assigned predominantly to a single cluster, we perform a hard cluster assignment by assigning each data point to the cluster with highest

Table 2. Architecture specification and training hyperparameters.

	1D 4-well	Müller-Brown	Dipeptide	$AAQAA_3$ - I	$AAQAA_3$ - II
Number of clusters (k)	4	5	8	10	6
Input dimension (n)	1	2	25	60	126
Latent dimension (d)	1	1	2	2	2
Number of nodes ($NN(Q^y)$)	[16, 16]	[32]	[32]	[16, 16]	[128]
Number of nodes ($NN(Q^z)$)	[16, 16]	[16]	[16]	[16, 16]	[16]
Number of nodes ($NN(P^z)$)	[16, 16]	[16]	[16]	[16, 16]	[16]
Number of nodes ($NN(P^x)$)	[16, 16]	[128]	[128]	[16, 16]	[256]
α	0.5	0.05	0.05	0.3	0.95
Batch size	32000	5000	5000	10000	3000
Learning rate	0.00005	0.0001	0.00015	0.001	0.00005
Number of epochs	50	400	100	300	2000
Probability cut-off	None	None	None	0.95	0.98

membership probability. However, in cases where a configuration has similar membership probabilities for multiple clusters, this simple assignment may introduce errors when determining properties (e.g. transition probabilities) of the clusters. Thus, we also considered a different approach by enforcing a thresholding value for cluster assignment. More specifically, each configuration is only assigned to a cluster if the largest membership probability is above a chosen cut-off value. A naive coring scheme followed the thresholding operation such that the points that had been identified as noise were assigned back to their previous cluster index for all other dynamical analyses.

2.3.2. GMVAE architecture and training hyperparameters

The GMVAE algorithm was implemented in Tensorflow [38] and is available at (https://github.com/yabozkurt/gmvae). Training was performed in all cases with fully-connected layers, using the Adam optimization algorithm [39]. The Softmax activation function was used for probabilistic cluster assignments, while ReLu activation functions were employed in all hidden layers. The means were obtained without any activation, whereas Softplus activation was employed to obtain the variances. Table 2 shows the values of the hyperparameters for each example system. Default values were employed wherever the parameters are not specified. The $NN(\cdot)$'s correspond to the neural networks labeled in figure 3. $NN(Q_y)$ performs probabilistic cluster assignments, $NN(Q_z)$ is for learning the moments of each Gaussian distribution in the encoding, whereas $NN(P_z)$ and $NN(P_x)$ are for the decoding of the z and x, respectively. The lengths of the 'Number of nodes' entries correspond to the number of hidden layers. Hyperparameter optimization was carried out as follows. The number of nodes was initialized as [16, 16]. The number of nodes in the decoder ($NN(P_x)$) was then increased whenever a large and non-decreasing reconstruction loss was observed. Our overall observation for the considered examples is that the learning rate and batch size should be kept relatively low to promote the formation of distinct clusters. The VAE results (with unimodal Gaussian prior) that are provided as comparison are obtained using $k = 1$, while keeping the remaining parameters equal to the values in the corresponding GMVAE model.

2.4. Markov state models

Markov state models (MSMs) represent the dynamics generated by a molecular simulation trajectory as a series of memoryless jumps between a discrete set of states [40]. Given a configuration-space discretization, a transition probability matrix, $\mathbf{P}(\tau)$, is obtained by counting the transitions between pairs of states within a given lag time, τ, and then performing a maximum likelihood optimization [41]. The eigenvalues of $\mathbf{P}(\tau)$, $\{\lambda_i(\tau)\}$, are related to characteristic timescales of the system's dynamics:

$$t_i(\tau) = -\frac{\tau}{\ln|\lambda_i(\tau)|}\,, \tag{13}$$

where $t_i(\tau)$ is the timescale corresponding to the i^{th} eigenvalue, $\lambda_i(\tau)$. The time lag parameter τ is typically chosen by performing the 'implied timescale test', which assesses the Markovianity of $\mathbf{P}(\tau)$ through the convergence of its timescales with increasing τ. In other words, $\{t_i(\tau)\}$ is plotted as a function of τ, and τ is then chosen as small as possible such that the largest timescales are *sufficiently* converged. Once τ is chosen, the accuracy of $\mathbf{P}(\tau)$ is determined via the Chapman-Kolmogorov (CK) test, which compares the estimated and predicted probability decay out of a given state. The predicted values are obtained using the CK equation, i.e. using the Markovian property of the model:

$$p_{ij}(m\tau) = p_{ij}^m(\tau)\,, \tag{14}$$

where $p_{ij}(\tau)$ is the probability of transitioning from state i to state j within time τ, and m is a positive integer. The CK test is often performed on metastables states of the system—collections of quickly interconverting microstates.

Within the standard Markov state modeling workflow, microstates are typically defined on low-dimensional projections of the full-dimensional configuration space. Therefore, obtaining a relevant transformation of the molecular simulation data is the key. To this end, time-lagged independent component analysis (TICA) [13, 42] is one of the most commonly used dimensionality reduction methods, as its objective is to maximize the autocorrelation of the data at the given lag time, making it especially well suited for kinetic modeling purposes. Metastable states are typically obtained via a dynamical coarse-graining procedure, e.g. PCCA+ [43] whose objective is to retain an accurate description of the dominant eigenvectors of the transition probability matrix. The resulting metastable states are then used as representative collections of microstates for performing the CK test. In many cases, a coarse-grained MSM at the resolution of the metastable states is constructed, providing an easily interpretable, albeit often qualitative, picture of the long timescale processes. In this study, the GMVAE performs the dimensionality reduction and clustering simultaneously, yielding a coarse-grained description of configuration space directly, without the need for further dynamical clustering. The (coarse-grained) MSMs are constructed from the discretized trajectories obtained using the simple cluster assignment based on the GMVAE membership probabilities as described in section 2.3.1. MSM construction and analysis was performed using the PyEMMA package [44].

2.5. Peptide analysis

The helical propensity of the peptide was determined using the Lifson-Roig perspective, which assigns each residue to either a helical (h) or coil (c) state, according to the dihedral angles along the peptide backbone (i.e. the Ramachandran plot) [45, 46]. Therefore, the number of different conformations of the peptide is limited to 2^N, where N is the number of residues; $N = 15$ for $AAQAA_3$. The propensity of residue i to be part of a 'helical segment', $\langle h_i \rangle$, is then defined as the probability that residue i as well as its two neighboring residues are simultaneously found in a helical state. The average fraction of helical segments, $\langle f_h \rangle$, is obtained by averaging $\langle h_i \rangle$ over all residue positions: $\sum_{i=0}^{N-1} \frac{1}{N} \langle h_i \rangle$. To distinguish between partial helical structures occuring at the N- and C-terminus ends of the peptide backbone, we define $\langle h_N \rangle = \sum_{i=1}^{6} \frac{1}{6} \langle h_i \rangle$ and $\langle h_C \rangle = \sum_{i=8}^{13} \frac{1}{6} \langle h_i \rangle$. Note that the terminus residue from each end is not taken into consideration.

The dRMSD measures the average deviation of internal distances from the corresponding distances in a reference structure, and is calculated as

$$\mathrm{dRMSD}(\mathbf{X}(t), \mathbf{X}^r) = \sqrt{\sum_{i \neq j} (||\mathbf{X}_i(t) - \mathbf{X}_j(t)|| - ||\mathbf{X}_i^r - \mathbf{X}_j^r||)^2}\,, \tag{15}$$

where $\mathbf{X}(t)$ represents the conformation at time t, $\mathbf{X}^r$ is the conformation for the reference structure, and $||\cdot||$ denotes the Euclidean norm. Note that, unlike other RMSD metrics, no pre-alignment of structures is required. In this study, due to the large fluctuations of the end residues, two residues from each end of the peptide were excluded in the dRMSD calculations. dRMSD was calculated using the positions of the C_α atoms only. Helix, hairpin-like, and extended (coil) structures were separately considered as reference structures as illustrated in figure S12.

3. Results

Variational autoencoders (VAEs) have been previously applied for dimensionality reduction of molecular simulation data [18, 28, 47]. VAEs typically employ a normal distribution to represent both the prior distribution in the latent space and the family of distributions for variational inference. In this work, we extend traditional VAEs by representing these distributions with Gaussian mixture models. The resulting Gaussian mixture VAE (GMVAE) adopts the physics-based viewpoint that an optimal embedding of the simulation data should give rise to a free-energy landscape (FEL) with well-separated clusters of configurations, which correspond to metastable states that are separated by large barriers along the high-dimensional potential energy landscape. The GMVAE introduces a categorical variable, y, which represents the various underlying Gaussian distributions to which each configuration will be (probabilistically) assigned. As a consequence, the approach simultaneously performs a dimensionality reduction and clustering, while enabling direct control over the organization of configurations in the latent space. We demonstrate the properties of this architecture by considering two model systems and molecular simulations of alanine dipeptide as well as a more challenging disordered peptide ensemble. In the following, $\mathrm{X} \in \mathbb{R}^n$ represents the n dimensional input. The latent variable in the bottleneck is represented by $z \in \mathbb{R}^d, d \leq n$.

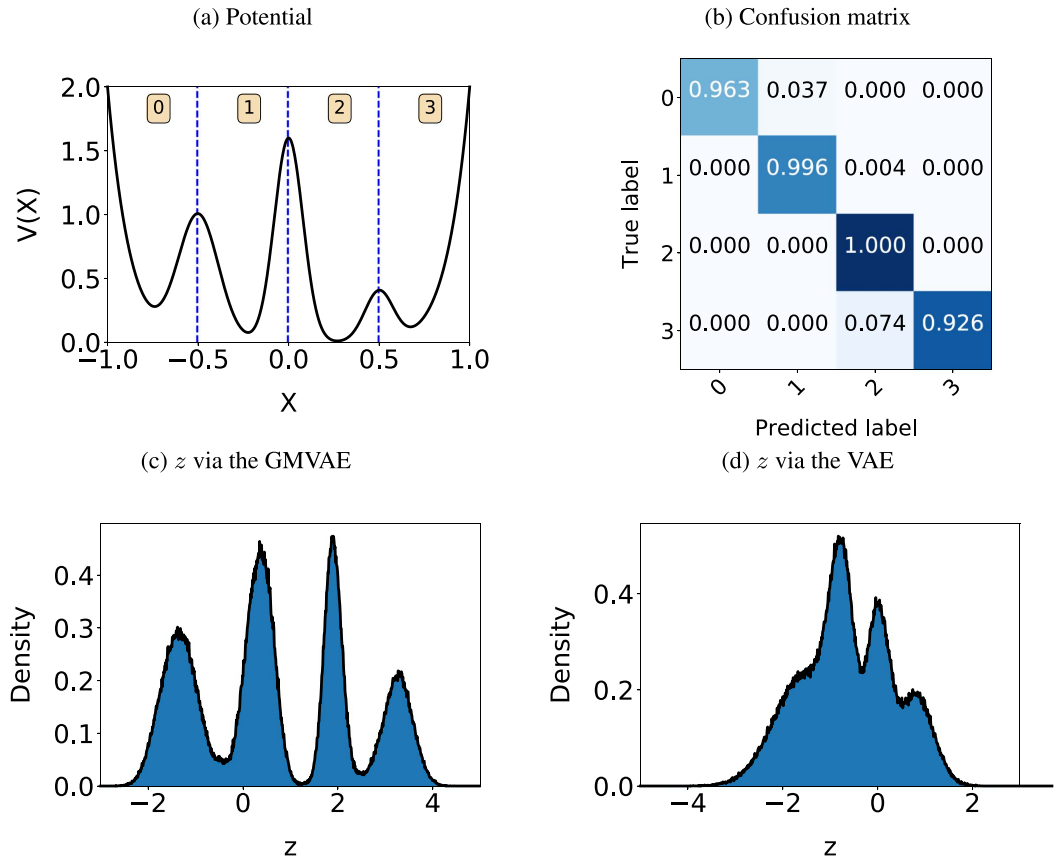


Figure 4. (a) 1D 4-well potential with the true labels. (b) Confusion matrix constructed with the true labels shown in (a) and the predicted labels obtained via the GMVAE. Population size increases from light to dark blue. Normalized histograms of the 1D latent variable via the (c) GMVAE and (d) VAE.

3.1. One-dimensional 4-well potential

We first consider a single particle in one dimension interacting with a 4-well external potential, which has been previously employed for testing methods associated with constructing MSMs [29, 48]. Figure 4(a) presents the potential, whose functional form and simulation details are given in section S.II. We employ a GMVAE with a latent space dimension of 1, which assesses the clustering performance of the architecture in the absence of any dimensionality reduction. The GMVAE was trained with $k = 4$ according to the parameters in table 2. Figure 4(b) presents the confusion matrix of the resulting model, which quantifies the probability that the model assigns a predicted label (x-axis) given the true label (y-axis). The true labels were determined using a coarse-grained representation of the system, where four metastable states are defined based on simple dividing surfaces, chosen as the maxima of the barriers between each potential well (dashed vertical lines in figure 4(a)). The GMVAE assigns the state labels with 97% overall accuracy.

Figure 4(c) shows a normalized histogram of z values. Without dimensionality reduction, the GMVAE largely retains the description of the input space within the latent dimension. As a consequence, the decoder is able to quite accurately reconstruct the input from the latent variable (See figure S1). This behavior is in stark contrast to traditional VAEs, which employ a Gaussian prior to represent the latent space distribution. As a result, anti-clustering effects can arise, leading to highly overlapping clusters of data in the reduced space. To demonstrate this effect, we constructed a traditional VAE for the present example. Figure 4(d) presents the corresponding normalized histogram of z values. In this case, even without a reduction in dimension, significant information is lost due to the constraint of the assumed prior distribution.

To further characterize the quality of the GMVAE clustering, we constructed an MSM from the trajectories of the predicted cluster IDs. Figure 5(a) presents the standard implied timescale test, which assesses the convergence of the characteristic timescales with increasing lag time parameter τ. Convergence indicates that the simulation dynamics, within the discrete-state representation, can be described within a Markovian approximation. The gray area indicates timescales that cannot be resolved by the model, since they are faster than the chosen lag time. From the test, the MSM with $\tau = 200$ was chosen for further analysis. The accuracy of this model was assessed with the Chapman-Kolmogorov test, which compares the simulated and predicted decay of probability from a chosen set of metastable states. Figure 5(b) demonstrates

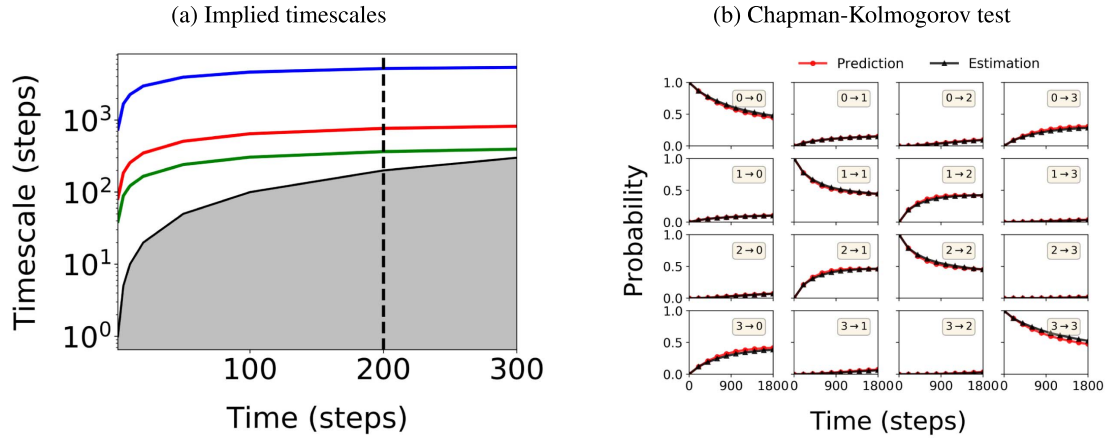


Figure 5. Markovianity check of the kinetic model built for 1D 4-well potential system. The MSM was constructed directly using the cluster labels obtained from the GMVAE. (a) Implied timescale test. (b) Chapman-Kolmogorov test (at lag = 200 steps).

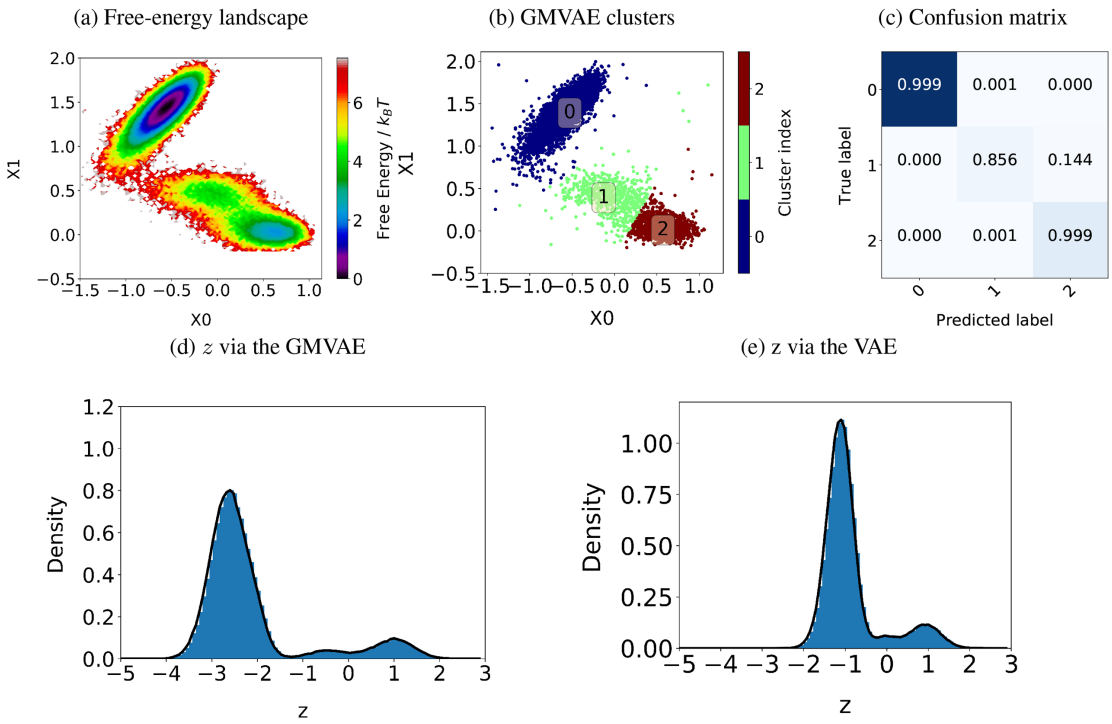


Figure 6. 2D Müller-Brown potential. (a) Free-energy landscape. (b) Clusters obtained from the GMVAE. (c) Confusion matrix with the true labels determined with linear dividing surfaces (figure S2(a)) and predicted labels obtained via the GMVAE. Population size increases from light to dark blue. Normalized histograms of the 1D latent variable via the (d) GMVAE (e) VAE.

that the predicted 'cluster dynamics' accurately represent the long timescale kinetic properties of the underlying simulation trajectory.

3.2. Müller-Brown potential

To assess both the dimensionality reduction and clustering performance of the GMVAE approach, we next consider a single Brownian particle in two dimensions interacting with an external Müller-Brown potential. The trajectory data was generated as the procedure suggested in [28] with the standard parameters [49] (see section S.III for more details). As depicted in figure 6(a), the resulting FEL contains two deep minima along with a less stable intermediate state. We employ a GMVAE that is trained with a latent space dimension of 1 and with $k = 5$, according to the parameters in table 2.

Despite employing $k = 5$, the resulting GMVAE model identified only 3 states with non-zero membership probabilities. Thus, somewhat remarkably, the GMVAE architecture was able to identify the inherent organization of the input data in the high-dimensional space, independent of the hyperparameter k.

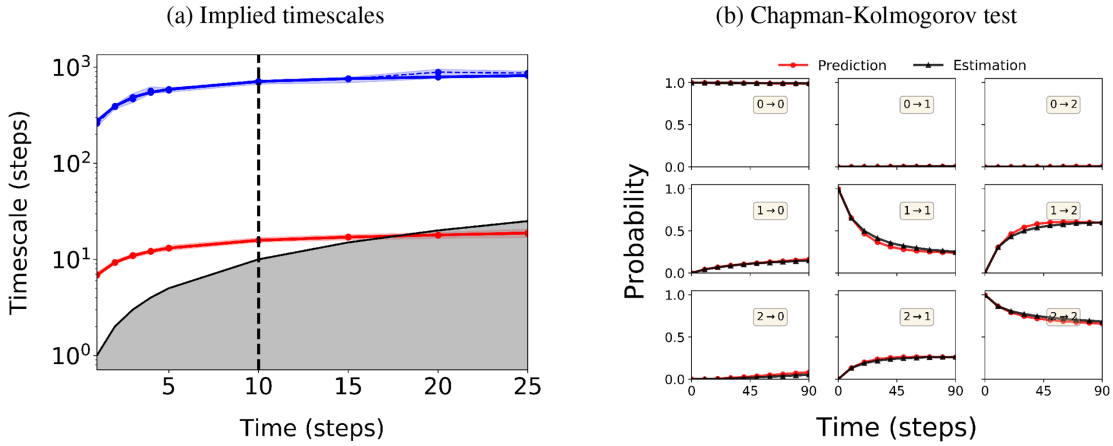


Figure 7. Markovianity check of the MSM built for 2D Müller-Brown potential via the GMVAE. (a) Implied timescales. (b) Chapman-Kolmogorov test (at lag = 10 steps).

Figure 6(b) shows the identified clusters. We define the true cluster labels in this case using linear dividing surfaces, as shown in figure S2(a). Figure 6(c) presents the confusion matrix from the GMVAE model with respect to these defined labels. Although it appears that there are errors in assigning state 1, this error is sensitively dependent on the precise definition of the true label dividing surfaces. Moreover, the overall classification accuracy is actually 99%, since state 1 corresponds to a very rarely sampled intermediate state. The model also demonstrates relatively high reconstruction accuracy (See figures S2(b) and S2(c)). Figures 6(d) and (e) present normalized histograms of z values obtained from the GMVAE model and a traditional VAE model trained on the same data, respectively. The low-dimensional representations obtained from the GMVAE clearly demonstrate a better separation of metastable states. Additionally, the ability of the GMVAE to learn a non-linear manifold is demonstrated in figure S3, with respect to the linear embedding obtained using time-lagged independent component analysis (TICA).

To further characterize the quality of the GMVAE clustering, we again constructed an MSM from the trajectories of the predicted cluster IDs. The implied timescale test (figure 7(a)) shows two dominant processes. The MSM with $\tau = 10$ was chosen for further analysis. Figure 7(b) presents the Chapman-Kolmogorov test, which further verifies the accuracy of the GMVAE embedding.

3.3. Alanine dipeptide

Alanine dipeptide is a representative model system for the characterization of conformational dynamics. Previous work [27, 31, 50–52] has shown that the (ϕ, ψ) backbone dihedral angles act as ideal collective variables for describing the metastable configurational basins and associated transition kinetics, making it an excellent system for testing the GMVAE framework within a more realistic molecular simulation context. Since in general the optimal set of input features is unknown *a priori*, we use this example to test the ability of the GMVAE to identify the proper collective variables from a larger set of input features. More specifically, we consider as input features both the normalized pairwise distances between heavy atoms as well as the (ϕ, ψ) dihedral angles (obtained from [53]). The pairwise distances were pre-processed using a kurtosis filter (with the threshold value of 0.03, see figure S4 for more detail), to reduce the input dimension by removing the low-variance features. The dihedral angles were pre-processed by applying sin and cos transformations in order to account for periodicity [54]. Figure 8(a) shows the FEL in the backbone dihedral angle space, with four labeled metastable basins corresponding to α_R, α_L, β, P_{II}, and γ conformations [55]. The gray lines are drawn for reference and do not represent any sort of optimal dividing surface.

Figure 9(a) presents the two-dimensional embedding found using the GMVAE, and figure 9(b) shows the simultaneously-obtained 6 clusters (indexed from 0 to 5) as a part of the GMVAE algorithm. The GMVAE again obtains a FEL that better separates clusters of conformations, relative to a standard VAE (figure S8). The distribution of these clusters on the Ramachandran plot (figure 8(b)) already strongly indicates their suitability for a kinetic analysis. The GMVAE clustering distinguishes all 5 of the metastable states, as well as a transition region between the α_R and β states (cluster 4). An MSM was again constructed from the coarse GMVAE cluster assignments. The implied timescale and Chapman-Kolmogorov tests are presented in figure 10, demonstrating the accuracy of this kinetic model.

We found in this example that, unlike the toy systems, the clustering obtained using the GMVAE did not appear to be completely robust. In particular, the precise clustering probabilities depend on the random

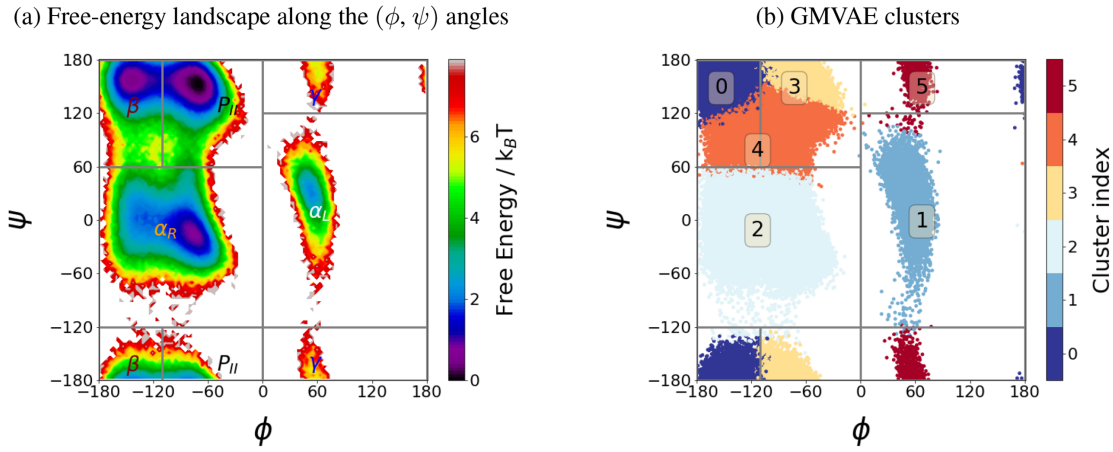


Figure 8. (a) Free energy landscape of alanine dipeptide. (b) GMVAE clusters on the Ramachandran plot.

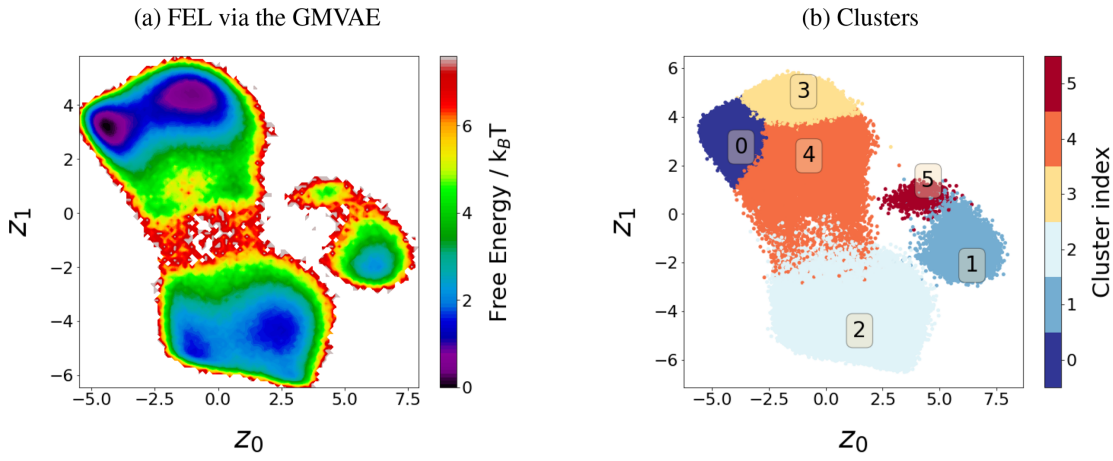


Figure 9. (a) FEL obtained for the alanine dipeptide by the GMVAE. The GMVAE clusters on the (b) GMVAE landscape.

effects of the training procedure (e.g. random weight initialization and the random shuffling of the input data). This issue was most pronounced for the lowest populated state, whose probability differs from the other states by two orders of magnitude (figure S5(b)). As a consequence, the γ state was not always sufficiently separated from the α_L state, resulting in a loss of one of the resolved kinetic processes (although the accuracy of the MSM remained intact, see figure S7). Despite this issue, the obtained FEL appeared rather robust with respect to changes in the random factors during training. We observed a much more robust clustering for all other applications considered.

3.4. $AAQAA_3$ peptide - I

As a more challenging test, we consider simulation trajectories of the capped helix forming peptide AC-$(AAQAA)_3$-NH2, which is a representative system for investigating helix-coil transitions. We employ a coarse-grained model [56], which describes the dominant attractive interactions, e.g. hydrogen bonding and effective hydrophobic interactions between side chains, with simple potentials between the C_α and C_β atoms. These interactions are the minimum required to sample the proper range of structures, (i.e. helix, coil, and hairpin-like). This model also represents excluded volume effects in near-atomic detail, which was demonstrated to be important for accurately characterizing the helix-coil kinetics. Here we employ a parametrization of the model that most accurately reproduces the experimental cooperativity of the helix-coil transition for $AAQAA_3$. As a result, hairpin-like structures appear to have relatively low metastability (similar to the intermediate state in the Müller-Brown example, and the γ state in alanine dipeptide), as we discuss further below. The model and simulation protocol are discussed further in the Supporting Information (stacks.iop.org/MLST/1/015012/mmedia), and also in [56, 57]. The considered simulation trajectories correspond to a disordered ensemble of peptide configurations, representing a stringent test for dimensionality and clustering methods [58].

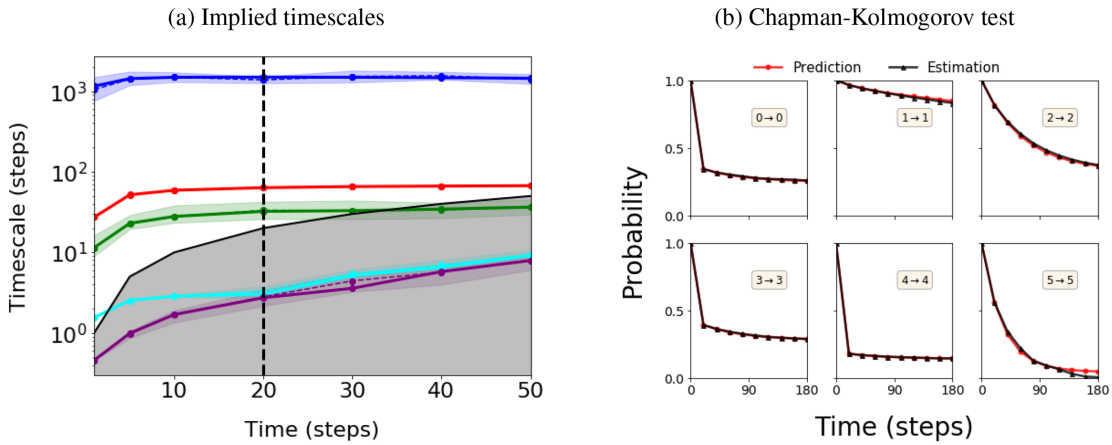


Figure 10. Markovianity check of the MSM built for alanine dipeptide via the GMVAE. (a) Implied timescales. (b) Chapman-Kolmogorov test (at lag = 20 steps).

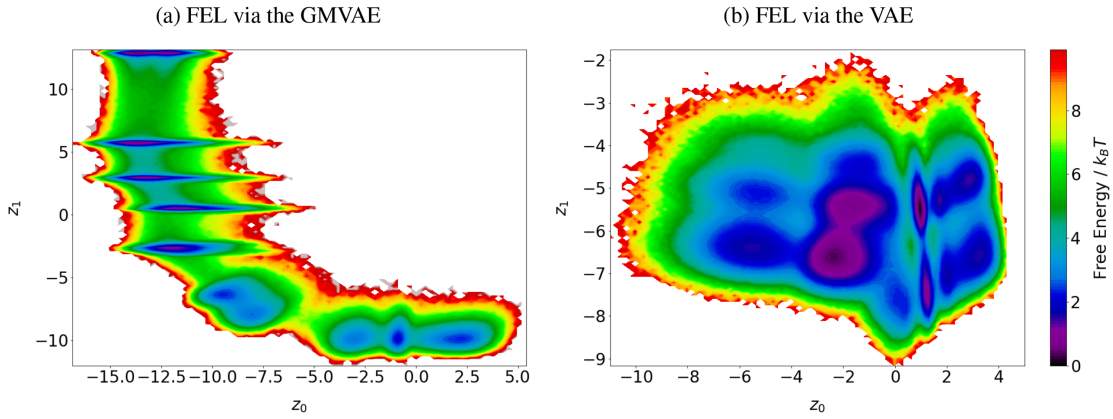


Figure 11. Free-energy landscapes of $AAQAA_3$ - I peptide obtained by (a) the GMVAE, and (b) the VAE.

Similar to alanine dipeptide, the set of sin and cos augmented (ϕ, ψ) dihedral angles along the peptide backbone were used as conformational descriptors. Thus, the input dimension is 60 for the 15-residue $AAQAA_3$ peptide. We chose to consider only a latent space dimension of 2, given that the ultimate goal of dimensionality reduction is often to reduce the high-dimensional description to something that is easily visualizable. Unlike the simple model systems above, the number of clusters, k, is completely unclear *a priori*. In fact, we expect that this ensemble to have a hierarchical structure, such that differing number of clusters may be appropriate depending on the chosen level of resolution. While we initially considered the GMVAE with varying number of clusters, we found that the number of 'non-zero clusters' (i.e. clusters with a significant probability of configuration assignment) was extremely insensitive to this choice, as discussed below. The GMVAE was trained according to the parameters in table 2. Also in contrast to the previous examples, there is no definitive reference kinetic model with corresponding known metastable states. Instead, the analysis below assesses the GMVAE embedding and clustering (in terms of both statics and kinetics) with respect to the landscapes obtained using a standard VAE and also following the standard MSM workflow (i.e. TICA [13, 42], see section 2.4 for more details).

Panels (a) and (b) of figure 11 show the FELs obtained using the GMVAE and the traditional VAE, respectively. As in the model systems, the GMVAE method results in a latent space description with highly separated clusters, while the traditional VAE yields more overlapping states. The two-dimensional TICA landscape (figure S19) also separates a number of clearly distinct states, although there are large diffuse regions with relatively low free-energy values. The clusters obtained via the GMVAE are shown in figure 12(a). Despite employing $k = 10$ and obtaining a landscape that appears to have approximately 10 distinct basins, only 7 states (labeled $0, 1, \ldots 6$) were assigned non-zero membership probabilities (see figure S9). Since standard metrics for analyzing peptide configurations do not yield a clear organization of the ensemble into a small number of metastable states, the distribution of these quantities are expected to be

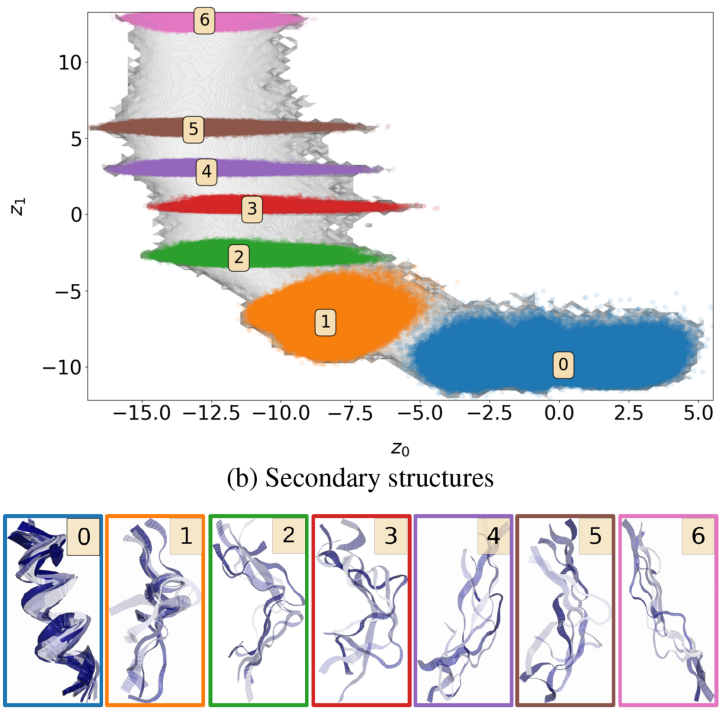


Figure 12. (a) The clusters obtained for the $AAQAA_3$ peptide - I by the GMVAE after thresholding. (b) The secondary structures closest to the cluster centers.

highly overlapping, even for a good clustering of the input data. Thus, to more easily visualize the characteristics of the GMVAE clusters, we applied a thresholding scheme, which removes configurations without a membership probability greater than 0.95 (see section 2.3.1 for details and figure S10 for cluster populations). Figure 12(b) shows 5 representative structures closest to the cluster centers. We stress that these images are intended to give the reader a rough idea of the types of structures contained in each cluster, but do not characterize the variance of structures within the clusters. This is a disordered ensemble and each cluster necessarily contains a diversity of structures. Nevertheless, figure 12(b) indicates that the GMVAE successfully distinguishes between distinct secondary structures within the simulation data.

To characterize the structural properties of the clusters quantitatively, we calculated the distribution of the average fraction of helical segments, $\langle f_h \rangle$. Figure 13(a) presents a heat map of $\langle f_h \rangle$ in the latent space. High $\langle f_h \rangle$ values (represented by blue) indicate the presence of helix and helix-like structures, whereas the lower values point to either hairpin- or coil-like secondary structures. There is an apparent trend of decreasing average helical content from the lower-right to upper-left regions of the latent space (i.e. from cluster 0 to 6). The VAE and TICA landscapes demonstrate similar trends (figures S23(b) and S19(b), respectively), although the VAE does not characterize partially-helical structures as clearly as the GMVAE. Figure S11 presents the intra-cluster distributions of $\langle f_h \rangle$, which can be used to assess the quality of the clustering (relative to an alternative clustering). We expect that an optimal clustering will result in tight, unimodal $\langle f_h \rangle$ distributions. The GMVAE clustering yields seemingly good distributions for the most and least helical clusters, while the partially-helical clusters appear broader and somewhat bimodal. For comparison, we consider three alternative clusterings obtained by performing a k-means clustering on a given landscape followed by the PCCA+ dynamical coarse-graining method [43] to define a set of metastable states (see section 2.4 for more details): (i) an alternative clustering of the GMVAE landscape (figure S16) , (ii) a clustering on the VAE landscape (figure S24) , and (iii) a clustering on the TICA landscape (figure S20). The alternative clustering scheme on the GMVAE landscape, (i), does not improve the intra-cluster distributions of $\langle f_h \rangle$, demonstrating that the GMVAE clustering is reasonable, given the GMVAE embedding. Similar results were obtained from the VAE clustering, with slightly broader distributions for the most and least helical states. The TICA clustering resulted in somewhat improved distributions, in the sense that they appear to be mostly unimodal, although some of the distributions appear to be slightly broader.

Figure 13(b) shows the $\text{dRMSD}_{\text{hel}}$ values of the projections, where the helicity increases as the $\text{dRMSD}_{\text{hel}}$ values decrease. These results are in agreement with the $\langle f_h \rangle$ analysis: as the cluster index increases from 0 to 6, the conformations tend to be more extended. The Supporting information (figures S12 and S13) contains

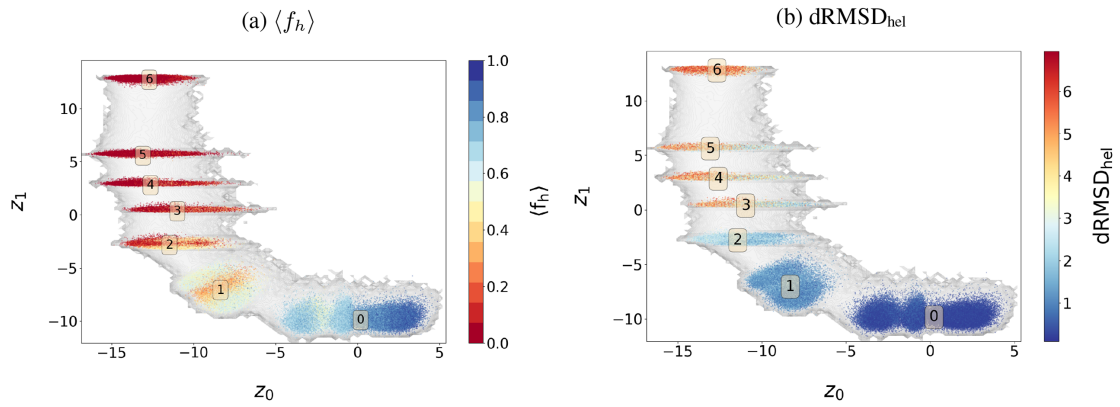


Figure 13. AAQAA$_3$ - I. (a) Average helical fraction, ⟨f_h⟩, analysis. Colors represent the ⟨f_h⟩ values of the corresponding projected data obtained from the GMVAE. (b) dRMSD$_{hel}$ analysis.

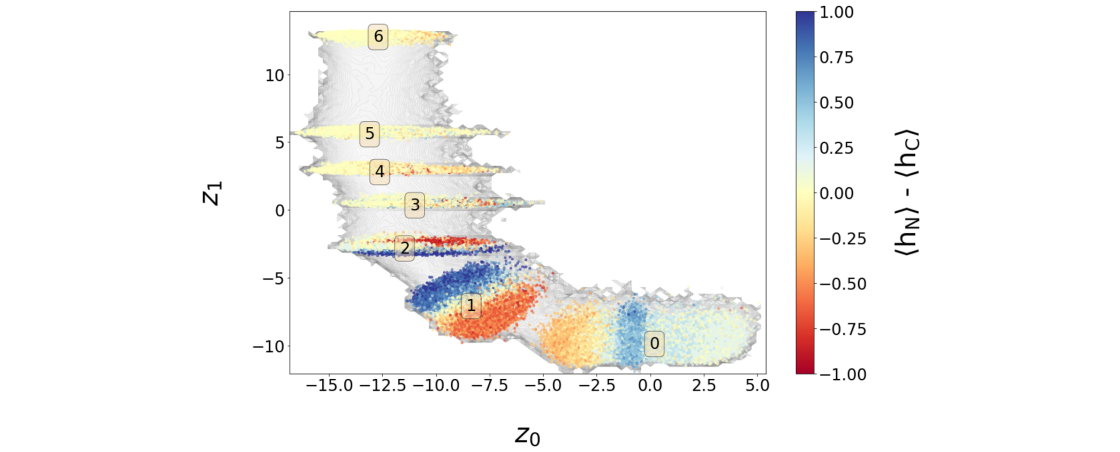


Figure 14. Analysis of partially-helical conformations for AAQAA$_3$ - I. Projections are colored according to ⟨h_N⟩ − ⟨h_C⟩ values.

additional characterization of the static properties of the clusters, which further validate the GMVAE embedding and clustering as a reasonable partitioning of the conformational landscape.

We also characterized the average fraction of helical segments on the N- and C-terminus sides of the peptide: ⟨h_N⟩ and ⟨h_C⟩, respectively (see section 2.5 for more details). Figure 14 presents the difference of these quantities, ⟨h_N⟩ − ⟨h_C⟩, plotted along the GMVAE embedding. Positive values (represented by blue) indicate conformations that contain helical structure on the N-terminus side of the peptide without helical structure on the C-terminus side. Conversely, negative values (represented by red) indicate conformations that contain helical structure on the C-terminus side of the peptide without helical structure on the N-terminus side. Values close to zero correspond to either fully helical or non-helical structures. Although the GMVAE embedding and clustering separate the most distinct structures in the ensemble (coils and full-helicies), some of the clusters (0, 1, 2) encompass partially-helical conformations on both sides of the peptide (see also figure S15). This is not ideal since kinetic barriers within a cluster will negatively impact the accuracy of a kinetic characterization at the cluster level. However, it appears that this issue may have more to do with the clustering than the embedding itself, since blue- and red-labeled structures appear to be reasonably separated on the landscape.

Similar to the other examples above, we also constructed an MSM directly from the discretized trajectories of GMVAE cluster indices. Although thresholding was applied in the results presented here (practically similar to coring methods for constructing kinetic models [59]), we found that this procedure had negligible effect on the accuracy of the resulting MSM. As shown in figure S14, the MSM constructed from the GMVAE clustering displayed significant errors in describing, e.g. the decay of probability out of the helix state. Perhaps this is not so surprising, since coarse-grained MSMs are often only used as a qualitative analysis tool, while higher-resolution kinetic models that characterize configuration space with many microstates are used for quantitative reproduction of simulation kinetics. Thus, to more carefully assess the GMVAE embedding and to more easily compare to the VAE and TICA results, we constructed a

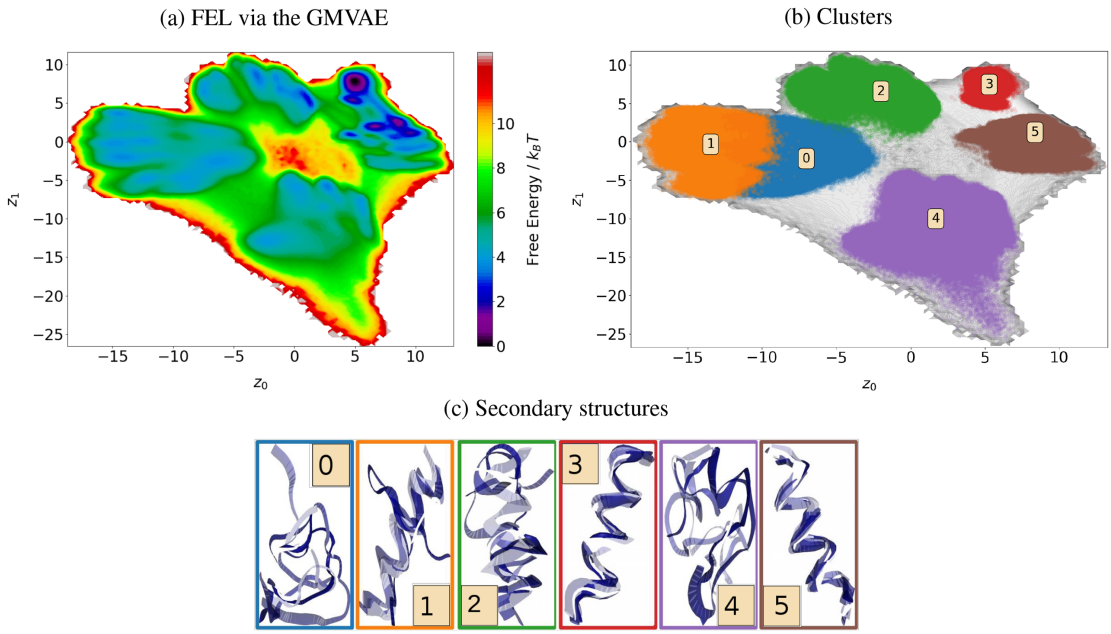


Figure 15. The GMVAE results for $AAQAA_3$ peptide - II. (a) Free-energy landscape. (b) The clusters obtained after thresholding. (c) The secondary structures closest to the cluster centers.

higher-resolution MSM by performing *k*-means to define microstates on the landscape (figure S16). Although the resulting model demonstrates improved accuracy according to the Chapman-Kolmogorov test, the probability decay out of the metastable states occurs on a fast timescale relative to the chosen lag time. This may be indicative of poorly defined dividing surfaces between metastable states. The kinetic models constructed from the VAE and TICA landscapes (figures S24 and S20, respectively) demonstrate similar quickly decaying probabilities. Although coring procedures could be applied to attempt to fix this problem, it indicates that there are fundamental limitations of all of these landscapes in terms of characterizing the long timescale simulation kinetics. There are several possible reasons for these difficulties, including (i) the limitation of our embeddings to two dimensions, (ii) the limitation of the chosen input features in characterizing kinetically-distinct structures, (iii) the presence of many low-lying barriers along the potential energy landscape of this disordered ensemble, and (iv) the poor sampling of relatively rare transitions to the full helix conformation. We partially address items (ii) and (iv) in the next section; however, a detailed investigation of these issues is beyond the scope of this initial study of the performance of the GMVAE, and is left for future work.

3.5. $AAQAA_3$ peptide - II

To investigate the impact of the low sampling of helical structures on the GMVAE embedding, as in the $AAQAA_3$ - I simulations presented above, we also considered a second set of simulations which primarily samples helical- and hairpin-like structures, while only rarely sampling fully-coiled structures. (Please see the Supporting Information for more details about the differences between the two sets of simulations). In addition to the dihedral angles, normalized pairwise distances between residues that are more than 3 residues apart were included as input features. Figure 15 presents the obtained GMVAE FEL (panel (a)), the corresponding clustering of 6 metastable states (panel (b)), and overlays of five structures that are closest to the cluster centers (panel (c)). The GMVAE embedding demonstrates significant separation of metastable states, relative to the landscape obtained with a standard VAE (figure S37(a)).

Similar to the previous ensemble ($AAQAA_3$ - I), figure 16 shows the separation of structures according to $\langle f_h \rangle$ (panel (a)), and $dRMSD_{hel}$ (panel (b)). The VAE and TICA landscapes demonstrate similar trends (figures S37 and S33, respectively). The intra-cluster $\langle f_h \rangle$ distributions are shown in figure S28. The majority of the fully-helical structures are in cluster 3 and 5, while clusters 0, 1, 2 and 4 contain hairpin-like structures as well as partial helicies. The coil structures are gathered in the bottom-most part of the landscape (in cluster 4), though not separated as a distinct cluster by the GMVAE. The distributions are broader and less unimodal than those determined from the previous set of simulations, although these can be somewhat improved with the alternative clustering scheme on the GMVAE landscape (figure S32). Similar results are also obtained from the VAE and TICA landscapes (figures S40 and S36, respectively).

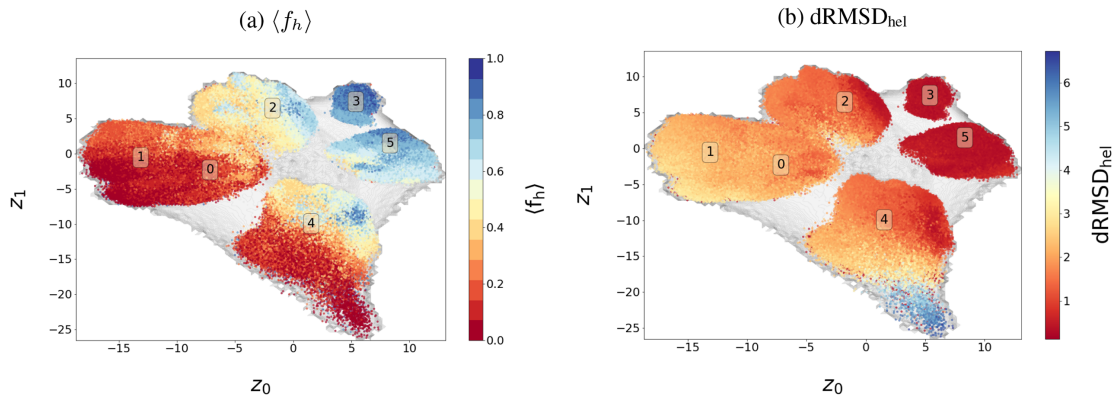


Figure 16. Projections for the AAQAA$_3$ peptide - II. (a) ⟨f_h⟩, (b) dRMSD$_{hel}$.

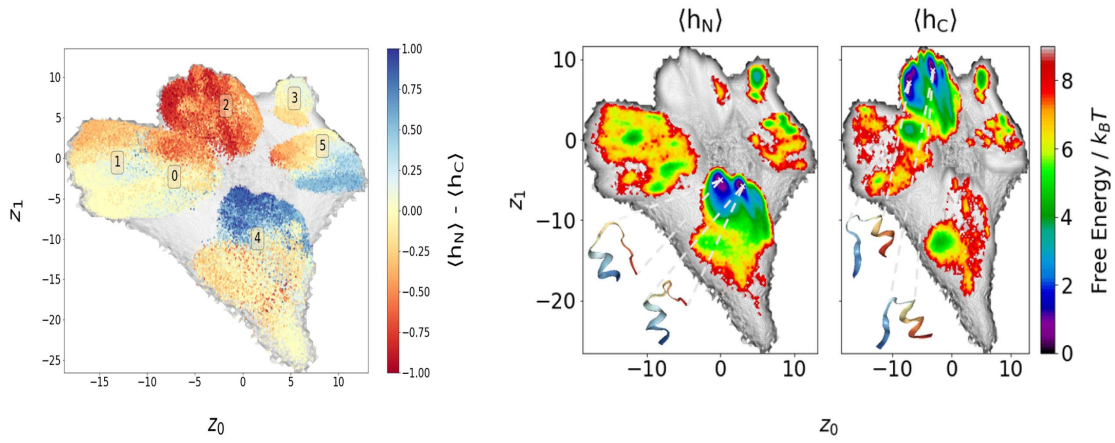


Figure 17. The N- and C-terminus end folding analysis for the AAQAA$_3$ peptide - II. (Left) The difference in the average values of the two-end foldings, $\langle h_N \rangle - \langle h_C \rangle$. (Right) Distribution of the N- (on the left, $\langle h_N \rangle \geq 0.8$) and C-end (on the right, $\langle h_N \rangle \leq -0.8$.).

Figure 17 presents the characterization of the N- and C-terminus, partially-helical conformations. In contrast to the AAQAA$_3$ - I embedding, the GMVAE embedding and clustering for AAQAA$_3$ - II more clearly separates the distinct types of structures. It appears that this difference may be due to the increased sampling of helical structures in AAQAA$_3$ - II, although the inclusion of pairwise distances as additional input features may also have played a role. N- and C-terminus partially-helical structures are mostly located in clusters 4 and 2, respectively, while both types of structures can be found to a lesser extent in cluster 5. Although the VAE and TICA landscapes also appear to largely distinguish between distinct partially-helical structures (figures S37 and S33, respectively), the GMVAE landscape provides a significantly better clustering of these two distinct sets of conformations.

Despite the improved description of partially-helical structures, the MSM constructed directly from the GMVAE clustering for AAQAA$_3$ - II displayed similar discrepancies to the model built for AAQAA$_3$ - I (figure S29). Moreover, the high-resolution MSMs constructed from the GMVAE, VAE, and TICA landscapes (figures S30, S38, and S34, respectively) displayed very fast decay of probability out of the identified metastable states, as in the AAQAA$_3$ - I example.

4. Discussion and conclusions

Variational autoencoders are quickly making an impact in the field of molecular simulations due to the inherent focus of the architecture on retaining the essential features of the system. Control over the *topology* of the latent space can increase the performance and interpretability of these methods by making a direct connection to the physics of the system through our physical intuition: an ideal free-energy landscape characterizes basins that are well-separated by the largest barriers along the higher-dimensional potential energy landscape. To explicitly enforce such features, we propose a Gaussian mixture model as the prior distribution in the latent space.

The performance of the Gaussian mixture variational autoencoder (GMVAE) was illustrated on two standard toy-model systems and on the standard benchmark alanine dipeptide, as well as on a challenging 15-residue-long disordered peptide. For each example, the GMVAE circumvents the aggregation of points in the latent space characteristic of traditional variational autoencoders. Instead, samples that are structurally distinct are clearly separated, leading to a latent space that displays apparent metastable basins and barriers. The GMVAE introduces a categorical variable that probabilistically assigns samples to a set of underlying clusters, each of which is Gaussian distributed. Thus, the approach combines the commonly distinct tasks of dimensionality reduction and clustering into a unified framework. In the absence of dimensionality reduction, the GMVAE retains the characteristics of the system within the latent space, while providing an accurate assignment between clusters. Remarkably, in the case of limited dimensionality reduction, the GMVAE identifies the inherent clustering of the input data, insensitive to the cluster-number hyperparameter.

Beyond statics, there have been several recent autoencoder architectures aiming at the characterization of molecular kinetics. Several of these methods directly incorporate kinetic information in the loss function, either by reconstructing time-lagged samples or by approximating the dynamical propagator [27–29, 31, 32, 60]. The interpretability of the latent space is becoming a feature of increasing interest: Hernández *et al* recently proposed an approach for identifying the most important input features for determining the one-dimensional latent-space representation within a time-lagged VAE framework [28], while Wang *et al* relied on a linear encoder to interpret the relevant coordinates of interest [60]. Here, we argue that incorporating physical constraints into the architecture helps to construct an interpretable model for the kinetics, even when kinetic information is not used for learning the representation. The GMVAE architecture attempts to better mimic the shape of an ideal free-energy landscape within the latent space. In particular, the presence of barriers that separate metastable clusters determines the relevant kinetic properties through the separation of timescales between intra- and inter-basin transitions.

Although incorporating time lag into learning the low-dimensional representation has been shown to help obtain better kinetic models, especially when geometric distance does not correspond to the kinetic distance, we report extremely encouraging results for constructing kinetic models from representations learned from static information alone. For the two toy models and for alanine dipeptide, the resulting Markov state models demonstrate excellent properties, as monitored by the implied timescale and Chapman-Kolmogorov (CK) tests. The disordered ensemble of the $AAQAA_3$ peptide proves more challenging: the CK test shows discrepancies for the decay of probability out of the longest-lived metastable states. Although higher-resolution MSMs constructed directly from the GMVAE landscape demonstrated an improved description of the simulation kinetics, the resulting model was unable to resolve all but the longest timescale processes. An MSM constructed from the TICA landscape demonstrated a slight improvement over this model, with respect to the CK test, but also exhibited a very fast decay of probabilities out of the identified metastable states, indicating a significant limitation in the time resolution of the model. These issues highlight the difficulty of characterizing such disordered ensembles, and motivate further investigation into the various possible causes. For example, comparisons of two distinct peptide ensembles clarified the role that sampling can play in distinguishing distinct partially-helical structures on the GMVAE landscape. It remains unclear to what extent the restriction of our embeddings to two dimensions or the choice of input features prevented the GMVAE (as well as the more standard methods considered) from better describing the simulation kinetics. Moreover, the presence of many low-lying barriers along the potential energy landscape of this disordered ensemble may cause fundamental challenges in obtaining a clear few-metastable-state characterization of the conformational landscape. Thus, we propose that, in conjunction with simpler test systems that clearly assess a method's performance, such examples are important for significant advancements in data-driven characterizations of molecular simulation trajectories.

While we defer a more detailed investigation of these issues for future work, we highlight the promising performance of the GMVAE demonstrated through our results. First, in the context of static equilibrium properties, the incorporation of the Gaussian mixture model as a prior distribution on the latent space closely links our physical intuition about ideal free-energy landscapes, resulting in an inherently more interpretable latent space. Secondly, our results show encouraging performance when constructing kinetic models from the learned representations—an aspect that is entirely absent in the loss function, representing an independent validation of the procedure.

Acknowledgments

The authors thank Kiran H Kanekal and Omar Valsson for critical reading of the manuscript. JFR is grateful to the BiGmax consortium and participants of the *BiGmax Big Data Summer School* for insightful

discussions. YBV acknowledges foreign collaborative research study support by The Scientific and Technological Research Council of Turkey, TÜBİTAK- BİDEB, under the 2214-A programme. TB acknowledges financial support by the Emmy Noether program of the Deutsche Forschungsgemeinschaft (DFG) and the long program Machine Learning for Physics and the Physics of Learning at the Institute for Pure and Applied Mathematics (IPAM).

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID iDs

Yasemin Bozkurt Varolgüneş https://orcid.org/0000-0002-9179-2458
Tristan Bereau https://orcid.org/0000-0001-9945-1271
Joseph F Rudzinski https://orcid.org/0000-0003-3403-640X

References

[1] Binder K 1995 *Monte Carlo and Molecular Dynamics Simulations in Polymer Science* (Oxford : Oxford University Press)
[2] Karplus M and Andrew McCammon J 2002 Molecular dynamics simulations of biomolecules *Nat. Struct. Molecular Biol.* **9** 646
[3] Bottaro S and Lindorff-Larsen K 2018 Biophysical experiments and biomolecular simulations: A perfect match? *Sci.* **361** 355–60
[4] Allen M P and Tildesley D P 1987 *Comput. Simulation of Liquids* (New York NY USA: Oxford Press)
[5] Bellman R E 2015 *Adaptive Control Processes: a Guided Tour Volume* (Princeton, NJ: Princeton University Press) 2045
[6] Kevrekidis I G, William Gear C, Hyman J M, Kevrekidid P G, Runborg O and Theodoropoulos C *et al* 2003 Equation-free, coarse-grained multiscale computation: Enabling mocroscopic simulators to perform system-level analysis *Commun. Math. Sci.* **1** 715–62
[7] Dobson C M 2003 Protein folding and misfolding *Nature* **426** 884
[8] Onuchic Je N and Wolynes P G 2004 Theory of protein folding *Curr. Opin. Struct. Biol.* **14** 70–5
[9] Weinan E, Engquist B, Xiantao Li, Ren W and Vanden-Eijnden E 2007 Heterogeneous multiscale methods: a review *Commun. Comput. Phys.* **2** 367–450
[10] Valsson O, Tiwary P and Parrinello M 2016 Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint *Annu. Rev. Phys. Chem.* **67** 159–84
[11] Husic B E and Pande V S 2018 Markov state models: From an art to a science *J. Am. Chem. Soc.* **140** 2386–96
[12] Pearson K 1901 On lines and planes of closest fit to systems of points in space *London Edinburgh Dublin Phil. Mag. J. Sci.* **2** 559–72
[13] Molgedey L and Schuster H G 1994 Separation of a mixture of independent signals using time delayed correlations *Phys. Rev. Lett.* **72** 3634
[14] Tenenbaum J B, De Silva V and Langford J C 2000 A global geometric framework for nonlinear dimensionality reduction *Science* **290** 2319–23
[15] Rohrdanz M A, Zheng W, Maggioni M and Clementi C 2011 Determination of reaction coordinates via locally scaled diffusion map *J. Chem. Phys.* **134** 03B624
[16] Ceriotti M, Tribello G A and Parrinello M 2011 Simplifying the representation of complex free-energy landscapes using sketch-map *Proc. Natl. Acad. Sci.* **108** 13023–8
[17] Chen W and Ferguson A L 2018 Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration *J. Comput. Chem.* **39** 2079–102
[18] Ribeiro J ao M L, Bravo P, Wang Y and Tiwary P 2018 Reweighted autoencoded variational Bayes for enhanced sampling (RAVE) *J. Chem. Phys.* **149** 072301
[19] Bonati L, Zhang Y-Y and Parrinello M 2019 Neural networks based variationally enhanced sampling *Proc. Natl Acad. Sci.* **116** 17641–7
[20] Ao M and Dinner A R 2005 Automatic method for identifying reaction coordinates in complex systems *J. Phys. Chem.* B **109** 6769–79
[21] Geiger P and Dellago C 2013 Neural networks for local structure detection in polymorphic systems *J. Chem. Phys.* **139** 164105
[22] Hinton G E and Salakhutdinov R R 2006 Reducing the dimensionality of data with neural networks *Science* **313** 504–7
[23] Doerr S, Ariz-Extreme I, Harvey M J and Gianni D F 2017 Dimensionality reduction methods for molecular simulations *arXiv preprint arXiv:1710.10629*
[24] Lemke T and Peter C 2019 Encodermap: Dimensionality reduction and generation of molecule conformations *J. Chem. Theory Comput.* **15** 1209–15
[25] Kingma D P and Welling M 2013 Auto-encoding variational Bayes *arXiv preprint arXiv:1710.10629*
[26] Adel T, Ghahramani Z and Weller A 2018 Discovering interpretable representations for both deep generative and discriminative models *Int. Conf. on Machine Learning* pp 50–9
[27] Wehmeyer C and Noé F 2018 Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics *J. Chem. Phys.* **148** 241703
[28] Hernández C X, Wayment-Steele H K, Sultan M M, Husic B E and Pande V S 2018 Variational encoding of complex dynamics *Phys. Rev.* E **97** 062412
[29] Chen W, Sidky H and Ferguson A L 2019 Capabilities and limitations of time-lagged autoencoders for slow mode discovery in dynamical systems *J. Chem. Phys.* **151** 064123
[30] Dilokthanakul N, Mediano P A M, Garnelo M, Lee M C H, Salimbeni H, Arulkumaran K and Shanahan M 2016 Deep unsupervised clustering with gaussian mixture variational autoencoders *arXiv preprint arXiv:1611.02648*
[31] Mardt A, Pasquali L, Hao W and Noé F 2018 VAMPnets for deep learning of molecular kinetics *Nat. Commun.* **9** 1–11

[32] Lusch B, Nathan Kutz J and Brunton S L 2018 Deep learning for universal linear embeddings of nonlinear dynamics *Nat. Commun.* **9** 4950
[33] Variational autoencoder: Intuition and implementation. (https://wiseodd.github.io/techblog/2016/12/10/variational-autoencoder/) Accessed: 2019 04–10
[34] Shu R 2016 Gaussian mixture VAE: Lessons in variational inference, generative models, and deep nets (http://ruishu.io/2016/12/25/gmvae/) Accessed: 2019 05–14
[35] Zhao Q, Honnorat N, Adeli E, Pfefferbaum A, Sullivan E V and Pohl K M 2019 Variational autoencoder with truncated mixture of gaussians for functional connectivity analysis *Int. Conf. on Information Processing in Medical Imaging* pp 867–879 Springer
[36] Nalisnick E Hertel L and Smyth P 2016 Approximate inference for deep latent gaussian mixtures *NIPS Workshop on Bayesian Deep Learning* vol 2
[37] Shi W, Zhou H, Miao N, Zhao S and Lei Li 2019 Fixing Gaussian mixture VAEs for interpretable text generation *arXiv preprint arXiv:1906.06719*
[38] Abadi M *et al* 2015 TensorFlow: Large-scale machine learning on heterogeneous systems (Software available from tensorflow.org)
[39] Kingma D P and Jimmy B 2014 Adam: A method for stochastic optimization *arXiv preprint arXiv:1412.6980*
[40] Bowman G R and Pande V S and Frank Noé 2014 *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* (Dordrecht, Netherlands: Springer Science and Business Media)
[41] Prinz J-H, Hao W, Sarich M, Keller B, Senne M, Held M, Chodera J D, Schütte C and Noé F 2011 Markov models of molecular kinetics: Generation and validation *J. Chem. Phys.* **134** 174105
[42] Pérez-Hernández G, Paul F, Giorgino T, De Fabritiis G and Noé F 2013 Identification of slow molecular order parameters for Markov model construction *J. Chem. Phys.* **139** 07B604_1
[43] Röblitz S and Weber M 2013 Fuzzy spectral clustering by PCCA+: Application to Markov state models and data classification *Adv. Data Anal. Classif.* **7** 147–79
[44] Scherer M K, Trendelkamp-Schroer B, Paul F, Perez-Hernandez G, Hoffmann M, Plattner N, Wehmeyer C, Prinz J-H and Noe F 2015 Pyemma 2: A software package for estimation, validation and analysis of markov models *J. Chem. Theory Comput.* **11** 5525–42
[45] Lifson S and Roig A 1961 On the theory of helix coil transition in polypeptides *J. Chem. Phys.* **34** 1963–74
[46] Doig A J 2008 The a-helix as the simplest protein model: Helix-coil theory, stability and design *Protein Folding, Misfolding and Aggregation* (Cambridge, Royal Society of Chemistry) pp 1–27
[47] Bhowmik D, Gao S, Young M T and Ramanathan A 2018 Deep clustering of protein folding simulations *BMC Bioinform.* **19** 484
[48] Schwantes C R and Pande V S 2015 Modeling molecular kinetics with tICA and the kernel trick *J. Chem. Theory Comput.* **11** 600–8
[49] Müller K and Brown L D 1979 Location of saddle points and minimum energy paths by a constrained simplex optimization procedure *Theor. Chim. Acta.* **53** 75–93
[50] Nüske F, Hao W, Prinz J-H, Wehmeyer C, Clementi C and Noé F 2017 Markov state models from short non-equilibrium simulations—analysis and correction of estimation bias *J. Chem. Phys.* **146** 094104
[51] Zhang J, Lei Y-K, Che X, Zhang Z, Yang Y I and Gao Y Q 2019 Deep representation learning for complex free-energy landscapes *J. Phys. Chem. Lett.* **10** 5571–6
[52] Chen W, Sidky H and Ferguson A L 2019 Nonlinear discovery of slow molecular modes using state-free reversible VAMPnets *J. Chem. Phys.* **150** 214114
[53] Markov Model MD Share 2019 (https://github.com/markovmodel/mdshare)
[54] Altis A, Otten M, Nguyen P H, Hegger R and Stock G 2008 Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis *J. Chem. Phys.* **128** 06B620
[55] Chodera J D, Singhal N, Pande V S, Dill K A and Swope W C 2007 Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics *J. Chem. Phys.* **126** 04B616
[56] Rudzinski J F and Bereau T 2018 Structural-kinetic-thermodynamic relationships identified from physics-based molecular simulation models *J. Chem. Phys.* **148** 204111
[57] Rudzinski J and 2018 Tristan Bereau The role of conformational entropy in the determination of structural-kinetic relationships for helix-coil transitions *Computation* **6** 21
[58] Kukharenko O, Sawade K, Steuer J and Peter C 2016 Using dimensionality reduction to systematically expand conformational sampling of intrinsically disordered peptides *J. Chem. Theory Comput.* **12** 4726–34
[59] Jain A and Stock G 2012 Identifying metastable states of folding proteins *J. Chem. Theory Comput.* **8** 3810–19
[60] Wang Y, Ribeiro J ao M L and Tiwary P 2019 Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics *Nat. Commun.* **10** 1–8