Journal of Computational Biophysics and Chemistry

Research

Cite This: *J. Comput. Biophys. Chem.* **2023**, *22* (4), 489–501

Assessments of Variational Autoencoder in Protein Conformation Exploration

Sian Xiao, Zilin Song, Hao Tian and Peng Tao*

Department of Chemistry, Center for Research Computing, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas 75205, United States

*Corresponding author. E-mail: *ptao@smu.edu*

ABSTRACT: Molecular dynamics (MD) simulations have been extensively used to study protein dynamics and subsequently functions. However, MD simulations are often insufficient to explore adequate conformational space for protein functions within reachable timescales. Accordingly, many enhanced sampling methods, including variational autoencoder (VAE) based methods, have been developed to address this issue. The purpose of this study is to evaluate the feasibility of using VAE to assist in the exploration of protein conformational landscapes. Using three modeling systems, we showed that VAE could capture high-level hidden information which distinguishes protein conformations. These models could also be used to generate new physically plausible protein conformations for direct sampling in favorable conformational spaces. We also found that VAE worked better in interpolation than extrapolation and increasing latent space dimension could lead to a trade-off between performances and complexities.

KEYWORDS: Molecular dynamics; protein conformations; enhanced sampling; deep learning; variational autoencoder.

1. INTRODUCTION

All proteins are dynamic entities. The dynamics of proteins are bridges between the conformational ensembles and the corresponding functional states. In general, proteins carry out their biological functions by adopting a certain range of conformations. Therefore, sufficiently sampling the conformational space is critical for understanding how proteins fulfill their biological functions.[1] Many structural biology techniques, including X-ray crystallography,[2] cryo-electron microscopy,[3] nuclear magnetic resonance,[4] electron paramagnetic resonance,[5] and Förster resonance energy transfer[6] methods can be used to provide information about protein structures, dynamics, and conformational changes.[7]

Molecular dynamics (MD) simulations are computational tools to provide protein conformational changes by integrating the dynamical equations of motions starting from an initial structure and velocity of each particle in molecular models.[7,8] With the ever-increasing computational power, the accessible simulation timescale for protein systems has reached the microsecond range to provide great details and insights into protein dynamics and functions in typical computational studies.[9,10] However, challenges also persist that the high-dimensional conformational spaces can be rarely sufficiently sampled by brute force simulation.[11,12] Specifically, the transitions between distinct conformational states that are separated by high kinetical barriers may take a significantly longer time to occur than accessible timescales of MD simulations. The high-dimensional conformational landscapes of proteins could contain multiple energetically favored states which are separated by high potential barriers, which prevent the MD simulations from transitioning frequently among these states. The developments of methods for enhancing comprehensive and efficient sampling of protein conformational spaces are critical.[12–14]

Many enhanced sampling methods have been developed to address this issue, which in general fall into

Received: 1 December 2022
Accepted: 3 February 2023
Published: 27 March 2023

DOI: 10.1142/S2737416523500217
J. Comput. Biophys. Chem. **2023**, *22* (4), 489–501

two categories: collective variable based (such as metadynamics[14,15] and variationally enhanced sampling[16,17]) and collective variable free methods (such as replica exchange molecular dynamics[18,19] and integrated tempering sampling).[12,20–23] The deep-learning autoencoder models[24] present a powerful nonlinear dimensionality reduction technique to mine data-driven collective variables from MD trajectories.[25–32] This technique furnishes explicit and differentiable expressions for the highly abstract and differentiable collective variables, making it an ideal candidate for integration with enhanced sampling techniques to accelerate the exploration of the protein configurational space.[33]

In this study, we focused on the variational autoencoders (VAEs) for compressing and abstracting the hidden dynamical features. Three systems were used to assess the VAE model from multiple perspectives. The Calmodulin protein was used to evaluate whether the VAE could capture the substantial conformation differences in latent space. The Toho-1 protein was used to evaluate whether the VAE could correlate low-dimensional distances and high-dimensional conformational discrepancy and generate new physically plausible structures to start new simulations and complement the existing MD simulation trajectories. The ubiquitin protein was used to evaluate the influence of latent space dimensionality on VAE performance. We showed that the VAE-learned latent space trained with MD simulations as training data could retain key properties in high-dimensional conformational space and that it was possible to predict physically plausible conformations which are rarely accessible during the MD simulations. These reconstructed VAE-learned conformations can be used as seed conformations to initialize new simulations. We also observed that VAE worked better for interpolation than extrapolation, and there was a trade-off between performances when increasing the latent space dimension. We concluded that VAE could serve as a tool for exploring the protein conformations and demonstrated that the initial data preparation was important to construct a reliable VAE model.

2. METHODS

2.1. Molecular dynamics simulations

The initial structures of the modeled proteins were obtained from the Protein Data Bank (Calmodulin: 1CLL,[34] Ubiquitin: 1UBQ,[35] Toho-1: 5KMW[36]). The protonation states of titratable residues were determined under neutral pH. The protein systems were solvated in suitably sized water boxes, detailed in their respective results sections. In each system, sodium and chloride ions were added to maintain an ionic strength of 0.1 M. The classical CHARMM36m[37] force field and the CHARMM-modified TIP3P model[38] were used to simulate the protein and the solvent molecules, respectively. The preprocessing was conducted using CHARMM[39] (version c41b1).

Unless noted otherwise, the MD simulations were carried out as follows. After energy minimization, the systems were first heated from 110 to 310 K with a temperature increment of 20 K per 100 ps. Consequently, the systems were subjected to isothermal-isobaric (NPT) equilibrations at 310 K (equilibrium run), followed by canonical (NVT) simulations (product run) at 310 K. Snapshots of the product run were taken evenly between certain intervals and saved for further analysis. The time length for the MD simulation conducted for each system is noted in their respective results sections. In all simulations, the SHAKE[40] constraint was used to rigidify all covalent bonds in the solvent molecules and the proteins. The nonbonding interactions within 10 Å were treated explicitly. The Lennard–Jones interactions were smoothed out to zero at 12 Å. The long-range electrostatic interactions were accounted for using the particle mesh Ewald summation method.[41] The simulation was conducted using OpenMM[42] (version 7.6.0).

The trajectories were aligned by rigid rotation and translation using all heavy atoms as the reference. Cartesian coordinates of all heavy atoms were extracted and used as input features after a normalization procedure by min-max scaling. That is, the coordinate c_i^k ($c \in x, y, z$) for atom i in structure k is normalized as

$$c_i^k = \frac{c_i^k - \min(c_i)}{\max(c_i) - \min(c_i)}. \quad (1)$$

2.2. Modeling systems

Calmodulin (CaM) is a regulatory Ca^{2+}-binding protein involved in the regulation of many important biological processes.[43–45] EF-hand, the calcium-binding motif in CaM, is a helix-loop-helix structure comprising 12 residues. The binding of calcium ions causes the conformation transition of EF-hand loops and induces substantial rearrangement.[43,45] EF-loops I to IV of CaM comprise residue numbers 20–31, 56–67, 93–104, and 129–140, respectively (Fig. S1). Many experimental and computational studies have been conducted to determine the binding affinity and selectivity of four

EF-hand loops.[44–47] Ye *et al.* reported that the dissociation constants between EF-loops I–IV and Ca^{2+} were ordered as I > III ≈ II > IV.[46] Therefore, Ca^{2+} ions were removed sequentially with the reverse order of binding affinity (IV, II, III, I). A total of five systems were obtained and labeled as CAM0 through CAM4. The number represents the number of bound Ca^{2+} ions in the system. All five systems were solvated in 105 Å × 80 Å × 65 Å water boxes. Equilibrium runs for 21 nanoseconds (ns) and production runs for 200 ns were performed sequentially for each system. The trajectories of the production runs were saved every 40 picoseconds (ps), resulting in 5,000 frames for each system.

The β-lactamase enzymes are a family of hydrolytic enzymes expressed by infectious bacteria for hydrolyzing β-lactam based antibiotics and manifest clinically challenging antibiotic resistance. Toho-1 belongs to the extended-spectrum β-lactamase CTX-M enzymes, one important sub-family of class A serine-based β-lactamases (ASβLs),[36] and has efficient hydrolytic activity against penicillin and cephalosporin antibiotics.[48,49] The active site cavity of Toho-1, which is critical for its antibiotic activity, locates at the interface of two highly conserved domains (α/β and α). The three engineered mutations (Ala166/Asn274/Asn276) in the crystal structure were modified to Glu166/Arg274/Arg276 as in the wild-type enzyme. The system was solvated in cubic water boxes of 95 Å. Equilibrium runs for 20 ns and production runs for 400 ns were performed sequentially. The trajectory of the production run was saved every 20 ps, resulting in 20,000 frames.

Ubiquitin, a small protein with 76 residues, was used to conduct a relatively long simulation. The system was solvated in cubic water boxes of 75 Å. Equilibrium runs for 20 ns and production runs for 2 microseconds (μs) were performed sequentially. The trajectory of the production run was saved every 100 ps, resulting in a total number of 20,000 frames.

2.3. Variational autoencoders

Autoencoders are composed of two sequentially concatenated networks: an encoder network for data compression and a decoder network for reconstruction (Fig. 1). The encoder network receives a d-dimensional input feature vector associated with example $x \in R^d$ and encodes it into a p-dimensional vector, z with $z \in R^p$. In other words, the role of the encoder is to learn how to model the function $z = f(x)$. The encoded vector, z, is also known as the latent vector or the latent feature representation. Typically, for an undercomplete autoencoder, the dimensionality of the latent vector is less than that of the input examples (hourglass-shaped architecture, $p < d$). Then, the decoder decompresses $\hat{x}$ from the low-dimensional latent vector, z, as a function $\hat{x} = g(z)$.[50] The decoder module of an autoencoder shares some conceptual similarities to a generative model. They both receive a latent vector z as the input and return an output $\hat{x}$ in the same space as x. However, the major difference between the two is that we do not know the distribution of z in the autoencoder, while in a generative model, the distribution of z is fully characterizable. It is possible to generalize an autoencoder into a generative model. VAE is such an approach.

In VAE, the encoder network is modified with additional optimization constraints on the latent vector distribution: the mean (μ) and variance (σ^2). During the training of a VAE, the model is forced to match

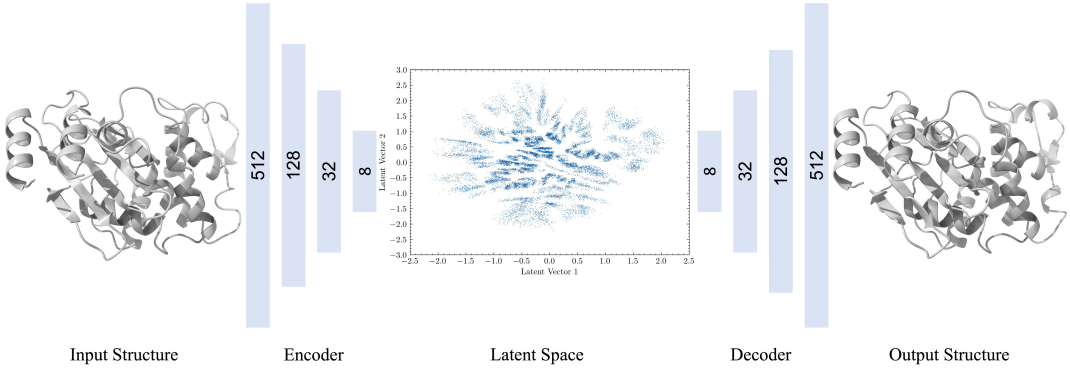

Fig. 1. (Color online) VAE architecture. The Cartesian coordinates of all heavy atoms are extracted as inputs. The encoder network with decreasing numbers of neurons in hidden layers encodes high-dimensional inputs to a low-dimensional latent space. The decoder network with increasing numbers of neurons in hidden layers projects latent space back to protein structures.

DOI: 10.1142/S2737416523500217
J. Comput. Biophys. Chem. **2023**, *22* (4), 489–501

these moments with a certain distribution (commonly the standard normal distribution). The latent vector distribution can therefore be regularized during the training such that the VAE latent space could reasonably correlate with the intrinsic distributions of the sample inputs. After the VAE model is built, the decoder module could be used to generate new examples, $\tilde{x}$, by feeding arbitrary values of z vectors.

We implemented the VAEs with the Keras[51] frontend in TensorFlow[52] v2.6.2. A detailed implementation is provided in the Supporting Information.

2.4. Principal component analysis (PCA)

PCA is a commonly used linear dimensionality reduction method that projects each data point onto fewer principal components (PCs) than original dimensions while preserving as much of the data's variability (i.e., statistical information) as possible.[53–55]

When a set of p original variables is replaced by an optimal set of PCs, the performance is incremental with the number of PCs used for projection. It can be measured by the variability associated with the set of retained PCs. The sum of variances of the p original variables (as well as all p PCs) is the trace of the covariance matrix S and the quality of a given PC j is the proportion of total variance that it accounts for, shown as follows:

$$\pi_j = \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} = \frac{\lambda_j}{\mathrm{trace}(S)}. \tag{2}$$

As a result of PCs' incremental nature, more total variance can be explained if more PCs are retained. A predetermined percentage of the total variance explained can be used to decide how many PCs should be retained. All PCAs in this study were performed using scikit-learn[56] v1.0.1.

2.5. Performance criteria

Pearson correlation coefficient (PCC) is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables $\mathrm{cov}(X, Y)$ and the product of their standard deviations σ. It is a normalized measurement of the covariance, such that the result always has a value between -1 and 1. Given two sets of data, X and Y, with n samples respectively, PCC is calculated as

$$\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})}}, \tag{3}$$

where x and y represent individual samples in X and Y, and $\bar{x}$ and $\bar{y}$ represent the averaged values.

Spearman correlation coefficient (SCC) is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function. The Spearman correlation coefficient is calculated as

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}, \tag{4}$$

where $d_i = R(x_i) - R(y_i)$ is the difference between the ranks R of individual sample x_i and y_i in datasets X and Y, respectively.

Root-mean-square deviation (RMSD) is a measure of the differences between a molecular structure and a reference. Given a molecular structure with a total of N atoms and a reference structure r^0, the RMSD of structure r is calculated as

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i^0 - Ur_i)^2}, \tag{5}$$

where r_i and r_i^0 are the coordinates of atom i in structures r and r^0, respectively. U is the translational and rotational operator that minimizes the RMSD distances by rigid fitting.

Root-mean-square fluctuation (RMSF) is a parameter to evaluate the flexibility of individual residues. RMSF_i measures how much an individual residue i fluctuates around its average position $\bar{r}_i$ in a simulation with T frames. RMSF_i is calculated as

$$\mathrm{RMSF}_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_i(t) - \bar{r}_i)^2}, \tag{6}$$

where $r_i(t)$ is the coordinate for residue i in frame t.

Discrete Optimized Protein Energy (DOPE)[57] is a potential function used to evaluate the quality of predicted protein structure. The lower the DOPE score, the better the model. DOPE scores were calculated using the modeling package MODELLER[58] (version 10.2).

3. RESULTS

3.1. Evaluation of encoder with latent space

The construction of VAE enforces the latent vectors transformed from the input samples to evenly distribute in the latent space by imposing an additional optimization constraint. This constraint utilizes the

Kullback–Leibler (KL) divergence[59] between the batched latent vectors and a standard normal distribution, in addition to the reconstruction error, as the loss function. Due to the intrinsic property of the KL divergence, distances among data points in low-dimensional space are not necessarily guaranteed to always preserve when the distances among these data points are large in the high-dimensional space.[60] In this study, we investigated how data points were distributed and spaced out in the latent space, which reflects how well and reasonably the encoder module projects the high-dimensional data onto low-dimensional latent space. We first assessed the performance of VAE models to distinguish different groups of conformations with significant differences. Then we evaluated how the shift of their corresponding low-dimensional representations reflects the changes in high-dimensional conformations.

3.1.1. *Calcium binding states of calmodulin*

All 25,000 frames of the five CAM systems were aligned to the first frame of CAM0. The VAE model was trained using all 25,000 frames and the latent space was provided in Supporting Information (Fig. S2). To investigate whether VAE could capture the critical conformational change of CaM upon binding with Ca^{2+} ions, the last 160 ns (last 4,000 frames) of each system were projected onto the latent space (Fig. 2). The first 40 ns was considered as the equilibrium stage and the rest was considered to have relatively stable overall conformational evolutions.

The simulations of five different systems CAM0 through CAM4 were separated into five regions in the latent space (Fig. 2(a)). Five points were randomly picked from five different regions, respectively. The original structures and decoded structures were aligned and illustrated (Figs. 2(b)–2(f)). The stacked original structures and the RMSD between each other are provided in Supporting Information (Fig. S3 and Table S1).

In this system, VAE retained the knowledge learned from the high-dimensional conformational space in the latent space. The fidelity of the decoded structures with reference to the original structures demonstrates that VAE has a good capability to compress and reconstruct data. Also, conformations with significant differences were well separated in the latent space, while the projections of similar conformations aggregate in the latent space.

3.1.2. *Structure correspondence of Toho-1*

An averaged configuration for Toho-1 was calculated after aligning all 20,000 conformations from all frames by rigid rotation and translation. This averaged configuration was used as the reference to measure the temporal drifting of each snapshot. A VAE model was trained using the aligned structures and the projection of simulation data was presented in Supporting Information (Fig. S4). For validation purposes, we replaced the coordinates of individual residue in every simulation frame with its coordinates in the average structure. The reason for doing so is to measure the sensitivity of

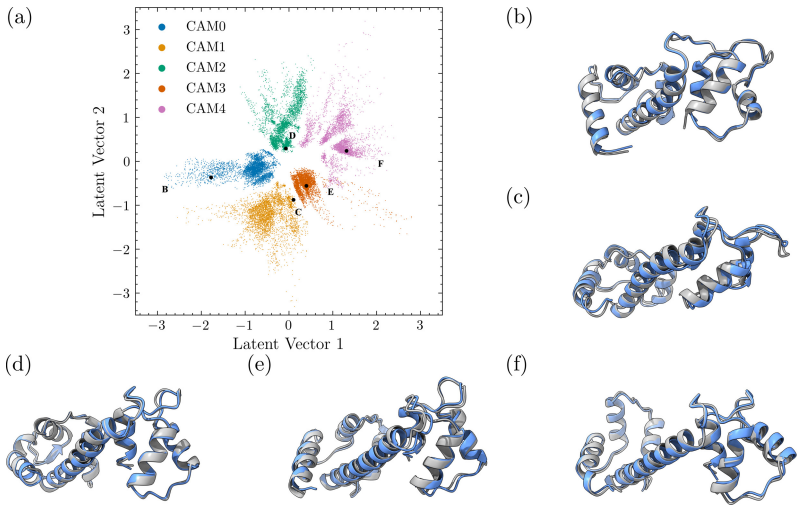

Fig. 2. (Color online) (a) Projection of the last 160 ns in the simulation of each system onto latent space; Encoded (dark gray) and decoded structures (cornflower blue) for five systems (b) CAM0, (c) CAM1, (d) CAM2, (e) CAM3 and (f) CAM4.

DOI: 10.1142/S2737416523500217
J. Comput. Biophys. Chem. **2023**, *22* (4), 489–501

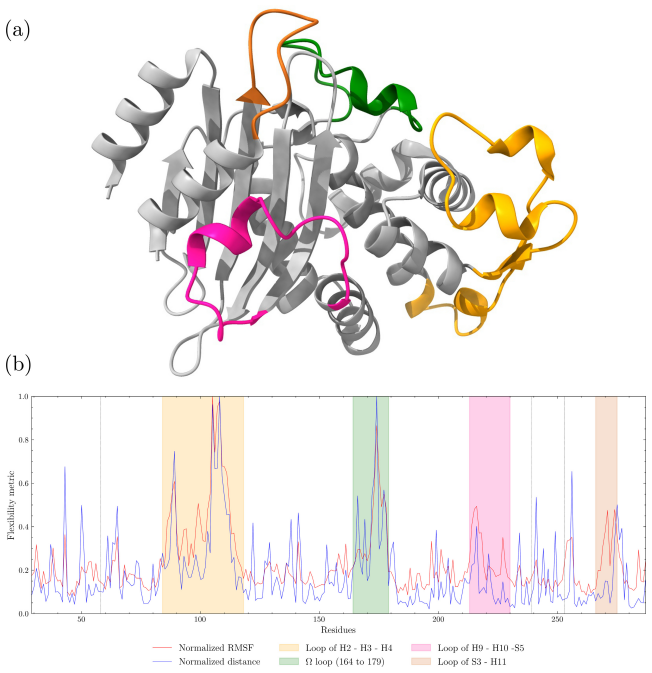

Fig. 3. (Color online) Evaluation of Toho-1 VAE model. (a) Most flexible secondary structures for Toho-1; (b) Normalized RMSF (red) and normalized displacement (blue). The position of three missing residues (58, 239, 253) in 5KMW was marked by dashed lines.

VAE latent space to structural fluctuations. The calculations were carried out separately for each individual residue to obtain respective average displacement. The original coordinates and modified coordinates were projected onto the latent space as two different latent points using the VAE encoder. The shift between these points was calculated for all 20,000 frames to obtain an average shift value. The 1-norm distance between the two points for each pair of original coordinates and modified coordinates was calculated. The 1-norm distance between two points (x_1, y_1) and (x_2, y_2) was calculated as

$$d = |x_1 - x_2| + |y_1 - y_2|. \quad (7)$$

Both RMSF and 1-norm distance were calculated and normalized by dividing individual values by the maximum values to obtain the measurement of relative flexibility for each residue in Toho-1 (Fig. 3(b)). Key secondary structures with large RMSF values in three domains are highlighted (Fig. 3(a)). The orange structure represents the loop connecting H2, H3, and H4. The green structure represents the ω loop of Toho-1. Residues 213 through 230, connecting H9, H10, and S5, are illustrated in brown. Residues 266 through 275, connecting S3 and H11, were illustrated in magenta.

The structural fluctuations calculated by RMSFs, and low-dimensional VAE latent space displacements have an adequate correlation with PCC around 0.78 (Fig. 3(b)). This positive correlation demonstrated that sufficient biological information has been preserved in the latent space and the distance in latent space could be used to evaluate the conformational deviations.

3.2. Structural decoding capability

Latent space could be used to explore protein conformational landscapes and even predict novel conformations based on existing MD simulations. It is proposed that points sampled in the latent space not belonging to the projected trajectories could lead to new physically plausible conformations. These conformations could complement pre-existing samplings and serve as new seeds for additional simulations. Therefore, we selected several points in the latent space not covered by existing trajectories of Toho-1 for evaluation purposes. First, these points were decoded to the three-dimensional Cartesian coordinates. CHARMM program was used to minimize the system energy to ensure the local structural integrity of generated structures, using the previously mentioned CHARMM36m force field. The optimized structures were subjected to MD simulations using OpenMM. Similar to other OpenMM simulation procedures, the systems were optimized and then heated (the temperature increment is 10 K per picosecond). After 200 ps NPT equilibrations at 310 K, these systems were subjected to 20 ns NVT simulations, in which snapshots were taken every 20 ps. Aligned structures before and after minimizations and the RMSDs between them were provided in Supporting Information (Fig. S6 and Table S2).

In Fig. 4, the chosen points were represented by larger blue points. The decoded structures from these chosen points were projected back to the latent space using the encoder and represented as purple star points. The structures after minimization were also projected back to the latent space using the encoder and represented as red diamond points. The snapshots of NVT simulations were then encoded and represented by red points.

The points selected outside the region of existing trajectories shifted more than those picked inside after energy minimizations. This is not surprising because encoding distribution is regularized during the training so the data on its latent space is more likely to distribute inside than outside. For the standard normal distribution, 68% of the observations lie within 1

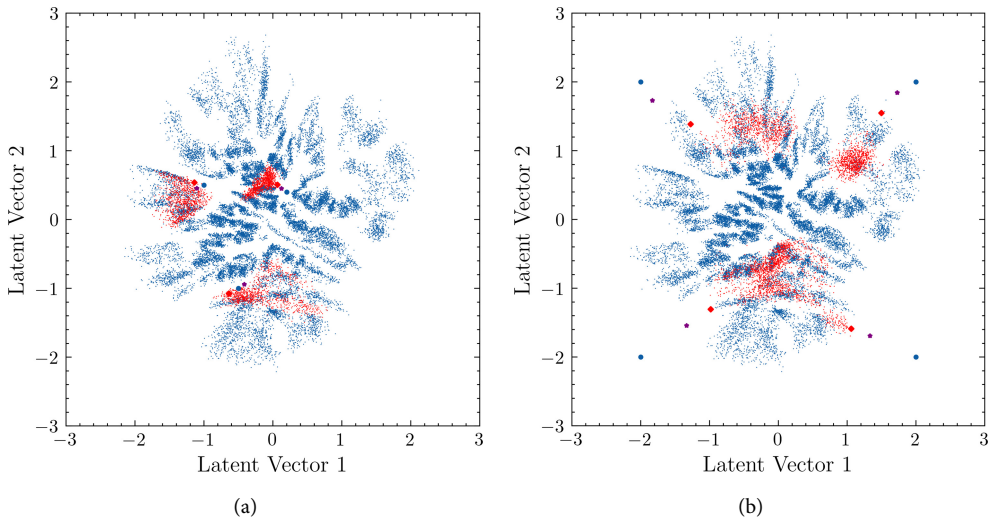

Fig. 4. (Color online) New points encoded by NVT simulations from decoded structures of given points. (a) Inside area covered by existing trajectories (points (0.2, 0.4), (−0.5, −1), (−1, 0.5)). (b) Outside area covered by existing trajectories (points (−2, −2), (−2, 2), (2, −2), (2, 2)).

standard deviation of the mean; 95% lie within two standard deviations of the mean; and 99.9% lie within 3 standard deviations of the mean. In other words, VAE works better for interpolation than extrapolation.

Starting from the seed structures decoded by the randomly picked points in latent space, these relatively short simulations filled the empty space and complemented the currently sampled region. This demonstrated the capability and effectiveness of VAE in exploring protein conformational spaces and complementing the existing MD simulation trajectories.

In conclusion, the VAE encoder could produce a low-dimensional conformational landscape from existing MD simulation trajectories. The decoder could generate new confirmations from unsampled areas in latent space through projection to the high-dimensional conformational space. To a certain extent, VAE has good abilities of interpolations and fair abilities of extrapolations to generate conformations with good structural quality. The generated structures can be used as seed structures for further simulations to sample new conformations.

3.3. Latent space dimensionality

It is anticipated that VAE could perform better as the latent space dimension increases, similar to PCA as more total variance can be explained by retaining more PCs. The latent space dimensionality is usually set to two for convenience in visualizations and manual seed-picking. With the help of computations, the latent dimension could certainly be set to a higher number and the sampling can be carried out in higher dimensions without human intervention. Therefore, we evaluated the impact of latent space dimensionality on the reconstruction quality of the VAEs. The 20,000 frames of ubiquitin simulations as training data were fed into a 4-layer VAE model with different numbers of latent dimensions. Different numbers of neurons were assigned to the hidden layers in each case but the ratio between adjacent layers is set to four. The results of different performance metrics versus latent space dimensions were plotted for comparison (Fig. 5).

Two correlation-based metrics were used to measure how well the information is preserved in the latent space. Generally, both PCC and SCC values decrease with increasing latent space dimensions. For the decoder module, RMSD and absolute percent error of the DOPE score were used to compare the discrepancies between the training and decoded structures. The absolute percent error of the DOPE score, omitting the one-dimension result, had an upward trend whereas the average RMSD values had a declining trend (Fig. 5).

Another observation is that the model training processes failed more with the increasing dimensions. When the model training fails, the encoder projects every frame to the same point in latent space. We further investigated how latent space dimensions influenced the latent vectors. Three VAE models were trained with latent space dimensions of two, four, and five, respectively. The reconstructed performance of 4D VAE is similar to 2D VAEs, while the performance of

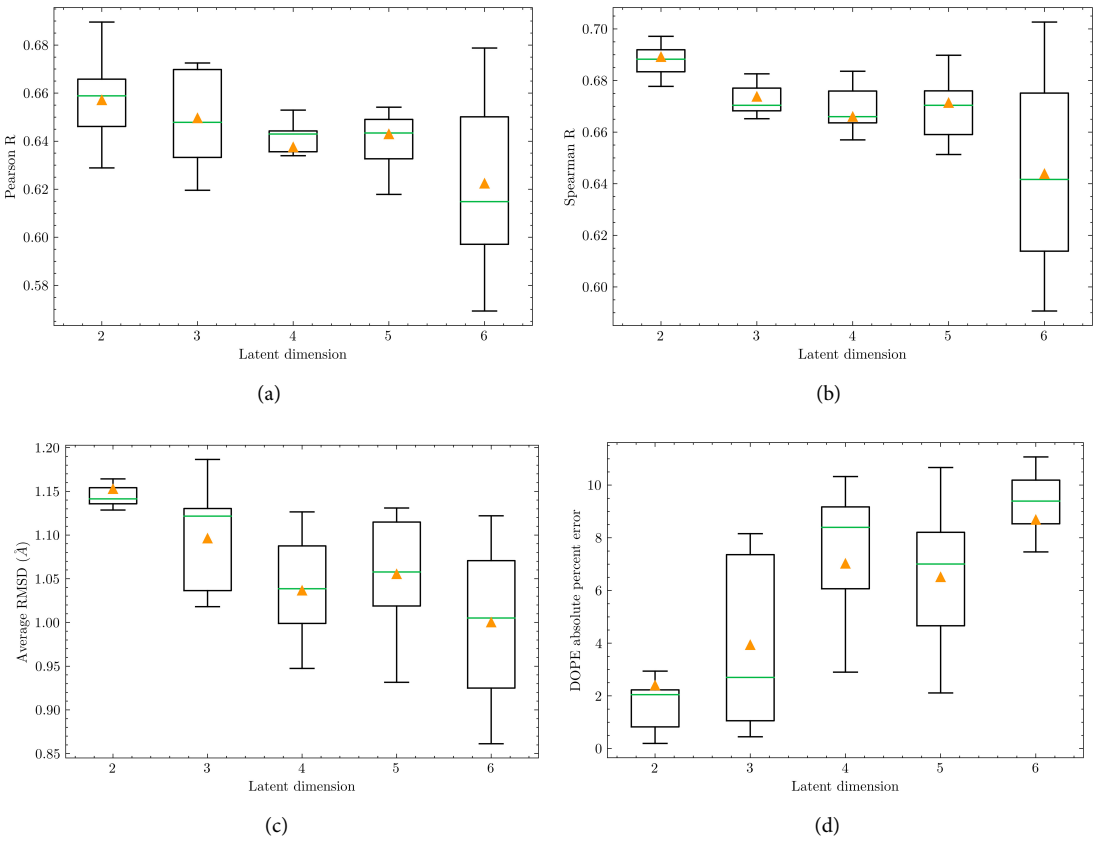

Fig. 5. (Color online) Performance assessment results for ubiquitin simulations using VAE of various dimensions of latent spaces. (a) Pearson correlation coefficient, (b) Spearman correlation coefficient, (c) average RMSD for all simulation structures, and (d) absolute percent error of DOPE score for all simulation structures. Four hidden layers' variational autoencoder was implemented for each model with different latent space dimensions. For each model, the results were obtained from 10 parallel model buildings.

5D VAE is significantly better than the 2D and 4D VAEs (Fig. 6).

The distributions of latent vector values projected from simulation frames are plotted in Supporting Information (Fig. S7) and the variances are listed in Table 1.

Two latent vectors of 4D have extremely small variances, indicating that the encoder projected points to almost the same position in these two dimensions (Table 1). In this case, these two dimensions do not contain significant information from the high-dimensional space and are trivial in the latent space. With only two predominant latent vectors taking effect, the 4D model performance is similar to the 2D model on RMSD (Fig. 6). The 5D model has four

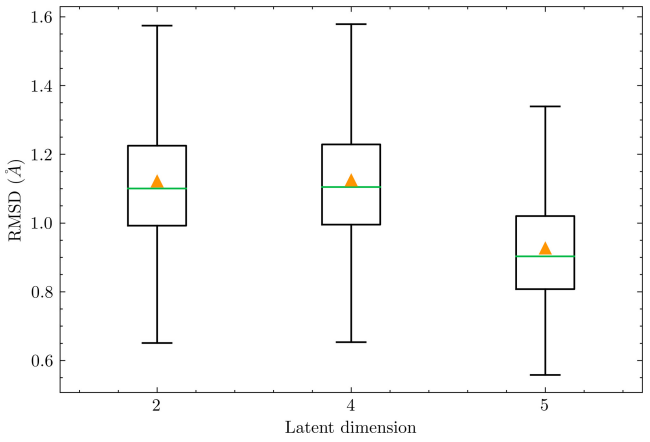

Fig. 6. (Color online) RMSD values with different latent space dimensions.

Table 1. Variances of latent vectors in different VAE models.

	2D	4D	5D
Latent Vector 1	0.49694	3.8911×10^{-5}	0.51871
Latent Vector 2	0.51171	4.3879×10^{-6}	0.46156
Latent Vector 3	N/A	0.52630	0.52309
Latent Vector 4	N/A	0.52567	0.50389
Latent Vector 5	N/A	N/A	4.1286×10^{-6}

DOI: 10.1142/S2737416523500217

predominant latent vectors carrying information from the high-dimensional space and had improved performance.

4. DISCUSSION

Variational autoencoders (VAEs) are deep latent embedding models consisting of two modules in an autoencoder structure. The encoder network, an inference model, learns to map data points to low-dimensional latent vectors, while the decoder network, a generative model, learns to reconstruct high-dimensional data points from the latent encodings.[61] The ability of VAEs trained on structures from MD simulations as a tool for enriching the molecular conformational space sampling was assessed in this paper from several aspects.

First, a good latent representation space should convey sufficient information on protein conformations. The remarkable ability to differentiate conformations with substantial differences was shown in the latent space of VAEs. Calmodulin had a large conformational change upon binding with calcium ions. Those systems with different amounts of Ca^{2+} bound with EF-loops had large conformational differences, which were captured by VAE. Different groups were located at different regions in latent space with certain gaps. Meanwhile, the fluctuations during MD simulations of residues correspond to the shift in latent space. The magnitude of the shift in latent space could reflect the local structure flexibility, indicating that the minor conformational differences could also be captured by VAE.

Second, VAE was expected to serve as a tool for enhanced sampling methods. Enhanced sampling could be accomplished by selecting points in latent space as needed. For instance, points could be chosen at the "frontiers" of covered regions,[22] or by prior knowledge about a specific protein.[32] In this study, new points were randomly picked from the unsampled regions, and the decoded structures were employed as seed structures for MD simulations. Energy minimizations were conducted to guarantee local structure integrity before simulations. The structure deviations of points inside the simulation-sampled region were smaller than those of points outside the sampled region. This indicates that VAE could be used to generate physically plausible structures and VAE works better in interpolations than extrapolations to generate new conformations. This was consistent with the fact that VAE distributes data in latent space to approximate normal distribution so that the outlying regions have lower probabilities to be visited (lower probabilities to form thermodynamically accessible structures).

Increasing the latent space dimensions appears to be a solution for high-resolution reconstructions of protein conformations with VAE as larger latent dimensions allow more information in the original sample space to be retained. However, the difficulty of robustly training the VAE models increases rapidly with the latent dimensions, given the same number of training samples. The optimization constraint of VAE requires that the latent space distribution of each batch of input approximates a normal distribution. The sparse distributions of the high dimensional latent vectors during the training would likely fail to approximate the normal distribution in the same dimension. This explains why the model building with higher latent space dimensions failed more frequently than those with lower dimensions. In the comparative study, not all dimensions successfully obtained enough variances to convey information, which explains the volatility of model performance when increasing the latent space dimensions. The sampling benefits from better model performance but there are trade-offs between visualization, performance, and obstacles with increasing latent space dimensions. Currently, the sampling is based on two-dimensional latent space. The sampling with higher dimensions could be promising with better computing resources.

VAE exhibits enormous potential to be a tool for exploring protein conformations and can be further developed for enhanced sampling. Based on our investigations, we have several further perspectives for VAE applications. In this study, the input features are normalized Cartesian coordinates of all heavy atoms so the decoder module could easily rebuild the decoded protein structures. Many other descriptors, such as pairwise alpha carbon ($C\alpha$) distances of the backbone, and torsional angles, could also be used as input features. Different descriptors might lead to varying effects in the representation performance and generated structure rationality. It is observed that more flexible and unstructured regions of the protein are less accurately reconstructed than the secondary structures. Introducing additional representation variables for specific areas could possibly increase desired structural information embedded into the model and enhance the encoding. However, this might cause a problem in decoding part to reconstruct structures from low dimensional space. This is because those local structural variables may be redundant and dependent on each other. Input data quality is another key factor since

DOI: 10.1142/S2737416523500217
J. Comput. Biophys. Chem. 2023, 22 (4), 489–501

building a better VAE model demands more high-quality data. One suggestion is that the input data should cover simulations starting from as many known conformations as possible to expand the knowledge of the VAE model. Concentrating around a certain starting structure leads to miss of key information about other conformations with substantial conformation differences. In this case, the VAE model can be hardly used to explore those completely different conformations. Moreover, normal MD simulations may contain excessive snapshots surrounding structures at low free energy states. Using these excessive snapshots in training may lead to undesired bias in the model. Filtration of the excessive simulation frames might be a practical technique to address this. Alternatively, training the autoencoder model using data from metadynamics simulation could be another suitable way to address the issue of excessive low-energy snapshots and high dimensional vectors in sparse latent space.

In addition to the applications presented in this study, the presented autoencoder models have many potential applications in protein simulations and analysis. For example, the latent space and associated decoder could be used to enhance the sampling efficiency of protein simulations. The latent space could be used for clustering analysis and Markov State Model development to investigate the kinetics of protein conformation transitions.

5. CONCLUSION

Variational autoencoders are unsupervised learning models designed to encode an input to a low-dimensional latent space and decode it for reconstruction. The encoded latent vectors are therefore expected to capture the key representational information of the input space. In this work, we have evaluated the viability of using variational autoencoders to assist protein conformational landscape exploration. VAEs are demonstrated to be capable of retaining high-dimensional information to distinguish protein conformations and generate yet-to-be-accessed protein conformations for initializing further simulations. It is also noteworthy that VAE works better for interpolation than extrapolation and increasing latent space dimension can lead to a trade-off between performances and obstacles. VAE could serve as a tool to explore the protein conformations with future studies.

ACKNOWLEDGMENTS

S.X. is grateful to Mayar Tarek Ibrahim and Dr. Francesco Trozzi for fruitful discussions. Computational time was generously provided by the Southern Methodist University Center for Research Computing.

FUNDING INFORMATION

The research reported in this article was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award No. R15GM122013.

DATA AVAILABILITY STATEMENT

The processing codes and processed data presented in this study can be found in the GitHub repository at *https://github.com/smu-tao-group/VAE_protein_assessment*.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Degiacomi, M. T. Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space. *Structure* **2019**, *27*, 1034–1040.e3.
2. Garman, E. F. Developments in X-ray Crystallographic Structure Determination of Biological Macromolecules. *Science* **2014**, *343*, 1102–1108, Publisher: American Association for the Advancement of Science.
3. Wrapp, D.; Wang, N.; Corbett, K. S.; Goldsmith, J. A.; Hsieh, C.-L.; Abiona, O.; Graham, B. S.; McLellan, J. S. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **2020**, *367*, 1260–1263, Publisher: American Association for the Advancement of Science.
4. Wüthrich, K. Protein Structure Determination in Solution by Nuclear Magnetic Resonance Spectroscopy. *Science* **1989**, *243*, 45–50, Publisher: American Association for the Advancement of Science.
5. Sahu, I. D.; McCarrick, R. M.; Lorigan, G. A. Use of Electron Paramagnetic Resonance To Solve Biochemical Problems. *Biochemistry* **2013**, *52*, 5967–5984, Publisher: American Chemical Society.

6. Sahoo, H. Förster resonance energy transfer — A spectroscopic nanoruler: Principle and applications. *Journal of Photochemistry and Photobiology C: Photochemistry Reviews* **2011**, *12*, 20–30.
7. Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99*, 1129–1143.
8. Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nature Structural Biology* **2002**, *9*, 646–652, Number: 9 Publisher: Nature Publishing Group.
9. Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation* **2013**, *9*, 3878–3888, Publisher: American Chemical Society.
10. Stone, J. E.; Hallock, M. J.; Phillips, J. C.; Peterson, J. R.; Luthey-Schulten, Z.; Schulten, K. Evaluation of Emerging Energy-Efficient Heterogeneous Computing Platforms for Biomolecular and Cellular Simulation Workloads. *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. **2016**; pp 89–100.
11. Nemec, M.; Hoffmann, D. Quantitative Assessment of Molecular Dynamics Sampling for Flexible Systems. *Journal of Chemical Theory and Computation* **2017**, *13*, 400–414, Publisher: American Chemical Society.
12. Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q. Enhanced sampling in molecular dynamics. *The Journal of Chemical Physics* **2019**, *151*, 070902, Publisher: American Institute of Physics.
13. Henzler-Wildman, K.; Kern, D. Dynamic personalities of proteins. *Nature* **2007**, *450*, 964–972, Number: 7172 Publisher: Nature Publishing Group.
14. Laio, A.; Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566, Publisher: Proceedings of the National Academy of Sciences.
15. Valsson, O.; Tiwary, P.; Parrinello, M. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annual Review of Physical Chemistry* **2016**, *67*, 159–184.
16. Valsson, O.; Parrinello, M. Variational Approach to Enhanced Sampling and Free Energy Calculations. *Physical Review Letters* **2014**, *113*, 090601, Publisher: American Physical Society.
17. Bonati, L.; Zhang, Y.-Y.; Parrinello, M. Neural networks-based variationally enhanced sampling. *Proceedings of the National Academy of Sciences* **2019**, *116*, 17641–17647, Publisher: Proceedings of the National Academy of Sciences.
18. Swendsen, R. H.; Wang, J.-S. Replica Monte Carlo Simulation of Spin-Glasses. *Physical Review Letters* **1986**, *57*, 2607–2609, Publisher: American Physical Society.
19. Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **1999**, *314*, 141–151.
20. Yang, L.; Liu, C.-W.; Shao, Q.; Zhang, J.; Gao, Y. Q. From Thermodynamics to Kinetics: Enhanced Sampling of Rare Events. *Accounts of Chemical Research* **2015**, *48*, 947–955, Publisher: American Chemical Society.
21. Gao, Y. Q. An integrate-over-temperature approach for enhanced sampling. *The Journal of Chemical Physics* **2008**, *128*, 064105, Publisher: American Institute of Physics.
22. Zhang, J.; Gong, H. Frontier Expansion Sampling: A Method to Accelerate Conformational Search by Identifying Novel Seed Structures for Restart. *Journal of Chemical Theory and Computation* **2020**, *16*, 4813–4821, Publisher: American Chemical Society.
23. Wu, X.; Xu, L.-Y.; Li, E.-M.; Dong, G. Application of molecular dynamics simulation in biomedicine. *Chemical Biology & Drug Design* **2022**, *99*, 789–800, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cbdd.14038.
24. Hinton, G. E.; Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507, Publisher: American Association for the Advancement of Science.
25. Jin, Y.; Johannissen, L. O.; Hay, S. Predicting new protein conformations from molecular dynamics simulation conformational landscapes and machine learning. *Proteins: Structure, Function, and Bioinformatics* **2021**, *89*, 915–921, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.26068.
26. Sultan, M. M.; Wayment-Steele, H. K.; Pande, V. S. Transferable Neural Networks for Enhanced Sampling of Protein Dynamics. *Journal of Chemical Theory and Computation* **2018**, *14*, 1887–1894, Publisher: American Chemical Society.
27. Tian, H.; Jiang, X.; Trozzi, F.; Xiao, S.; Larson, E. C.; Tao, P. Explore Protein Conformational Space With Variational Autoencoder. *Frontiers in Molecular Biosciences* **2021**, *8*, 781635.
28. Ramil, M.; Boudier, C.; Goryaeva, A. M.; Marinica, M.-C.; Maillet, J.-B. On Sampling Minimum Energy Path. *Journal of Chemical Theory and Computation* **2022**, *18*, 5864–5875, Publisher: American Chemical Society.
29. Belkacemi, Z.; Gkeka, P.; Lelièvre, T.; Stoltz, G. Chasing Collective Variables Using Autoencoders and Biased Trajectories. *Journal of Chemical Theory and Computation* **2022**, *18*, 59–78, Publisher: American Chemical Society.
30. Wehmeyer, C.; Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of Chemical Physics* **2018**, *148*, 241703, Publisher: American Institute of Physics.
31. Chen, W.; Ferguson, A. L. Molecular enhanced sampling with autoencoders: On-the-fly collective variable

discovery and accelerated free energy landscape exploration. *Journal of Computational Chemistry* **2018**, *39*, 2079–2102, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.25520.

32. Tian, H.; Jiang, X.; Xiao, S.; La Force, H.; Larson, E. C.; Tao, P. LAST: Latent Space Assisted Adaptive Sampling for Protein Trajectories. 2022; http://arxiv.org/abs/2204.13040 (accessed 2022-10-21), arXiv:2204.13040 [q-bio].
33. Chen, W.; Tan, A. R.; Ferguson, A. L. Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design. *The Journal of Chemical Physics* **2018**, *149*, 072312, Publisher: American Institute of Physics.
34. Chattopadhyaya, R.; Meador, W. E.; Means, A. R.; Quiocho, F. A. Calmodulin structure refined at 1.7 Å resolution. *Journal of Molecular Biology* **1992**, *228*, 1177–1192.
35. Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. Structure of ubiquitin refined at 1.8Å resolution. *Journal of Molecular Biology* **1987**, *194*, 531–544.
36. Langan, P. S.; Vandavasi, V. G.; Weiss, K. L.; Cooper, J. B.; Ginell, S. L.; Coates, L. The structure of Toho1 -lactamase in complex with penicillin reveals the role of Tyr105 in substrate recognition. *FEBS Open Bio* **2016**, *6*, 1170–1177, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2211-5463.12132.
37. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nature Methods* **2017**, *14*, 71–73, Number: 1 Publisher: Nature Publishing Group.
38. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79*, 926–935, Publisher: American Institute of Physics.
39. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry* **2009**, *30*, 1545–1614, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21287.
40. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Computational Physics* **1977**, *23*, 327–341.
41. Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993**, *98*, 10089–10092, Publisher: American Institute of Physics.
42. Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology* **2017**, *13*, e1005659, Publisher: Public Library of Science.
43. Linse, S.; Helmersson, A.; Forsén, S. Calcium binding to calmodulin and its globular domains. *Journal of Biological Chemistry* **1991**, *266*, 8050–8054, Publisher: Elsevier.
44. Yang, J. J.; Gawthrop, A.; Ye, Y. Obtaining Site-Specific Calcium-Binding Affinities of Calmodulin. *Protein and Peptide Letters* **2003**, *10*, 331–345.
45. Tan, Q.; Ding, Y.; Qiu, Z.; Huang, J. Binding Energy and Free Energy of Calcium Ion to Calmodulin EF-Hands with the Drude Polarizable Force Field. *ACS Physical Chemistry Au* **2022**, *2*, 143–155, Publisher: American Chemical Society.
46. Ye, Y.; Lee, H.-W.; Yang, W.; Shealy, S.; Yang, J. J. Probing Site-Specific Calmodulin Calcium and Lanthanide Affinity by Grafting. *Journal of the American Chemical Society* **2005**, *127*, 3743–3750, Publisher: American Chemical Society.
47. Kohagen, M.; Lepšík, M.; Jungwirth, P. Calcium Binding to Calmodulin by Molecular Dynamics with Effective Polarization. *The Journal of Physical Chemistry Letters* **2014**, *5*, 3964–3969, Publisher: American Chemical Society.
48. Ambler, R. P.; Coulson, A. F. W.; Frère, J. M.; Ghuysen, J. M.; Joris, B.; Forsman, M.; Levesque, R. C.; Tiraby, G.; Waley, S. G. A standard numbering scheme for the class A -lactamases. *Biochemical Journal* **1991**, *276*, 269–270.
49. Nitanai, Y.; Shimamura, T.; Uchiyama, T.; Ishii, Y.; Takehira, M.; Yutani, K.; Matsuzawa, H.; Miyano, M. The catalytic efficiency (kcat/Km) of the class A -lactamase Toho-1 correlates with the thermal stability of its catalytic intermediate analog. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2010**, *1804*, 684–691.
50. Raschka, S.; Liu, Y. H.; Mirjalili, V.; Dzhulgakov, D. *Machine Learning with PyTorch and Scikit-Learn: Develop Machine Learning and Deep Learning Models with Python*; Packt Publishing Ltd, **2022**; Google-Books-ID: SVxaEAAAQBAJ.
51. Chollet, F. Keras: The Python deep learning API. 2015; htttps://keras.io.
52. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. {TensorFlow}: A system for {Large-Scale} machine learning. 12th USENIX symposium on operating systems design and implementation (OSDI 16). **2016**; pp 265–283.
53. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1901**, *2*, 559–572, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14786440109462720.
54. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, *2*, 37–52.

55. Jolliffe, I. T.; Cadima, J. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2016**, *374*, 20150202, Publisher: Royal Society.
56. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
57. Shen, M.-Y.; Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Science* **2006**, *15*, 2507–2524, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1110/ps.062416606.
58. Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics* **2016**, *54*, 5.6.1–5.6.37, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpbi.3.
59. Kullback, S.; Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics* **1951**, *22*, 79–86, Publisher: Institute of Mathematical Statistics.
60. Trozzi, F.; Wang, X.; Tao, P. UMAP as a Dimensionality Reduction Tool for Molecular Dynamics Simulations of Biomacromolecules: A Comparison Study. *The Journal of Physical Chemistry B* **2021**, *125*, 5022–5034, Publisher: American Chemical Society.
61. Hawkins-Hooker, A.; Depardieu, F.; Baur, S.; Couairon, G.; Chen, A.; Bikard, D. Generating functional protein variants with variational autoencoders. *PLOS Computational Biology* **2021**, *17*, e1008736, Publisher: Public Library of Science.

DOI: 10.1142/S2737416523500217
J. Comput. Biophys. Chem. **2023**, *22* (4), 489–501