LAST: Latent Space Assisted Adaptive Sampling for Protein Trajectories

Hao Tian [1] **Xi Jiang** [2] **Sian Xiao** [1] **Hunter La Force** [1] **Eric Larson** [3] **Peng Tao** [1]

Abstract

Molecular dynamics (MD) simulation is widely used to study protein conformations and dynamics. However, conventional simulation suffers from being trapped in some local energy minima that are hard to escape. Thus, most computational time is spent sampling in the already visited regions. This leads to an inefficient sampling process and further hinders the exploration of protein movements in affordable simulation time. The advancement of deep learning provides new opportunities for protein sampling. Variational autoencoders are a class of deep learning models to learn a low-dimensional representation (referred to as the latent space) that can capture the key features of the input data. Based on this characteristic, we proposed a new adaptive sampling method, latent space assisted adaptive sampling for protein trajectories (LAST), to accelerate the exploration of protein conformational space. This method comprises cycles of (i) variational autoencoders training, (ii) seed structure selection on the latent space and (iii) conformational sampling through additional MD simulations. The proposed approach is validated through the sampling of four structures of two protein systems: two metastable states of *E. Coli* adenosine kinase (ADK) and two native states of Vivid (VVD). In all four conformations, seed structures were shown to lie on the boundary of conformation distributions. Moreover, large conformational changes were observed in a shorter simulation time when compared with conventional MD (cMD) simulations in both systems. In metastable ADK simulations, LAST explored two transition paths toward two stable states while cMD became trapped in an energy basin. In VVD light state simulations, LAST was three times faster than cMD simulation with a similar conformational space.

[1]Department of Chemistry, Center for Research Computing, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas 75206, United States [2]Department of Statistical Science, Southern Methodist University, Dallas, Texas 75206, United States [3]Department of Computer Science, Southern Methodist University, Dallas, Texas 75206, United States. Correspondence to: Peng Tao <ptao@smu.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

1. Introduction

Molecular dynamics (MD) simulation has wide application on the study of protein conformations and dynamics. Recent developments in bio-computing, such as Anton (Shaw et al., 2009), AMBER (Salomon-Ferrer et al., 2013) and OpenMM (Eastman et al., 2017), have enabled the simulation time scale to milliseconds, which promotes the research in sampling protein motions and structure landscapes (Prinz et al., 2011; Lindert et al., 2013). However, the time scales of many protein functions exceed the time scales achievable through traditional MD simulations. Moreover, protein sampling suffers from being trapped within local energy minima, proving difficult to escape. (Krivov, 2011; Brotzakis et al., 2018) As a result, most of the computational time is typically spent in sampling previously visited regions, which hinders efficient exploration of protein conformational space.

Many enhanced sampling methods have been developed to address this issue. These methods can be classified into two types. In the first type, biasing potentials are introduced to expand the sampling space, such as metadynamics (Barducci et al., 2011; Raiteri et al., 2006) and Gaussian-accelerated MD (Hamelberg et al., 2004). In the second type, seed structures are selected as restarts for iterative MD simulations. This is referred to as adaptive sampling and numerous methods have been proposed that differ in seed selection methods. Markov state models have been applied to cluster conformations into microstates (Bowman et al., 2010); parallel cascade selection MD (PaCS-MD) (Harada & Kitao, 2013) and nontargeted PaCS-MD (Harada & Kitao, 2015) calculates the root-mean-square deviation (RMSD) to select top snapshots; frontier expansion sampling (Zhang & Gong, 2020) conducts dimensionality reduction with principal component analysis and Gaussian mixture models to select frontier structures.

Recent innovations in deep learning have provided new insights into sampling protein conformational space. (Chen & Ferguson, 2018; Hawkins-Hooker et al., 2021) Autoen-

coders (AEs) and variational autoencoders (VAEs) are a class of deep learning models that learn a representation (encoding) which can capture the key features of input data. Several studies have demonstrated the success of AEs and VAEs in their applications to protein conformations and functions. (Ramaswamy et al., 2021; Jin et al., 2021; Bandyopadhyay & Mondal, 2021; Guo et al., 2021) In a previous work (Tian et al., 2021), the authors showed that VAEs are capable of learning a low-dimensional representation (referred to as the latent space) of protein systems. Through a quantitative study, the learned latent space is shown to be able to represent conformational characteristics. This property indicates that the larger differences two protein conformations have, the farther their corresponding latent points are from each other.

In this study, we proposed a new adaptive sampling method, latent space assisted adaptive sampling for protein trajectories (LAST), to accelerate the exploration of protein conformational space. This method iterates through three steps. First, a VAE is trained using previous MD simulations. Then, seed structures are selected on the learned latent space. Finally, additional simulations are conducted with these selected seeds. To quantify the performance, we applied LAST on four conformations in two protein systems: two metastable states of *E. Coli* adenosine kinase (ADK) and two native states of Vivid (VVD). ADK conformations are projected onto two intrinsic angles while VVD conformations to RMSDs of two native structures. Our results showed that seed structures were consistently located on the boundary of sampled conformational distributions in all four conformations regardless of protein projection methods on reduced coordinates. We further compared the sampling efficiency between LAST and conventional MD (cMD). In both systems, large conformational changes were observed in a shorter time in LAST simulations. To be specific, LAST explored two transition paths toward two stable states while cMD being trapped in an energy basin in the metastable ADK simulations. In VVD simulations, LAST only took one third of cMD simulation time to discover a similar conformational space.

2. Methods

2.1. Variational Autoencoder

An autoencoder is a type of deep learning models that aims to encode a high-dimensional input to a low-dimensional latent space through an encoder module and decode it back to the original dimensions through a decoder module. By minimizing the differences between inputs and outputs, known as reconstruction loss, the latent space is expected to learn a low-dimensional representation of the input space. However, the latent space in an AE is not well-constrained and leads to unsatisfying results when sampling in the latent space. (Wetzel, 2017) To overcome this issue, variational autoencoders add an optimization constraint on the latent space to follow a certain distribution.

The encoder module $q_\phi(z|x)$ is an inference model that transforms data x into output latent variable z, being parametrized with ϕ. In reverse, the decoder module $p_\theta(x|z)$ is a generative model that transforms latent variable z into output data $\hat{x}$, being parametrized with θ. Both models are trained simultaneously with a joint distribution as $p(x,z) = p_\theta(x|z)p(z)$. $p(z)$ is the constraint distribution for latent space and typically is chosen as a normal distribution. (Doersch, 2016) The tractable variational Bayes approach is used to approximate the intractable posterior $p_\theta(z|x) = p_\theta(x|z)p(z)/(\int p_\theta(x|z)p(z)dz)$ by maximizing the Evidence Lower Bound (ELBO):

$$\begin{aligned}\mathcal{L}(\phi,\theta;x) =& \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \\ &- KL(q_\phi(z|x)||p(z)) \leq \log p_\theta(x)\end{aligned} \tag{1}$$

where KL is the Kullback-Leibler divergence.

In our implementation, the VAE model is developed using Keras package (Chollet et al., 2015) with Tensorflow backend (Abadi et al., 2016).

2.2. Molecular Dynamics Simulations

The initial structures of four conformations in two protein systems: two metastable states (PDB ID 1DVR and 2AK3) of *E. Coli* adenosine kinase (ADK) and two native states (PDB ID 2PD7 and 3RH8) of Vivid (VVD) were taken from the Protein Data Bank (PDB) (Berman et al., 2000) . In each conformation, ligands and crystal waters were removed and chain A was extracted as the starting structure. The system was further solvated in a box of TIP3P water molecules and neutralized using Na^+ and Cl^-. Simulation files were generated using tleap (Case et al., 2005) with the AMBER ff14SB force filed (Maier et al., 2015). 100 ps NVT Langevin MD simulations were carried out, followed by 200 ps NPT simulations at 1 atm and 300 K. In each round of LAST method, one 100 ps MD simulation was conducted for each seed structure. Particle Mesh Ewald (PME) algorithm was used to calculate long-range electrostatic interactions. The simulation time step was set as 2 fs. All simulations were conducted with OpenMM 7 (Eastman et al., 2017).

2.3. Latent Space Assisted Adaptive Sampling for Protein Trajectories

LAST method includes three steps, and its workflow is shown in Figure 1. First, a variational autoencoder is trained using all previous simulations. Second, outliers are selected on the latent space and their corresponding protein structures

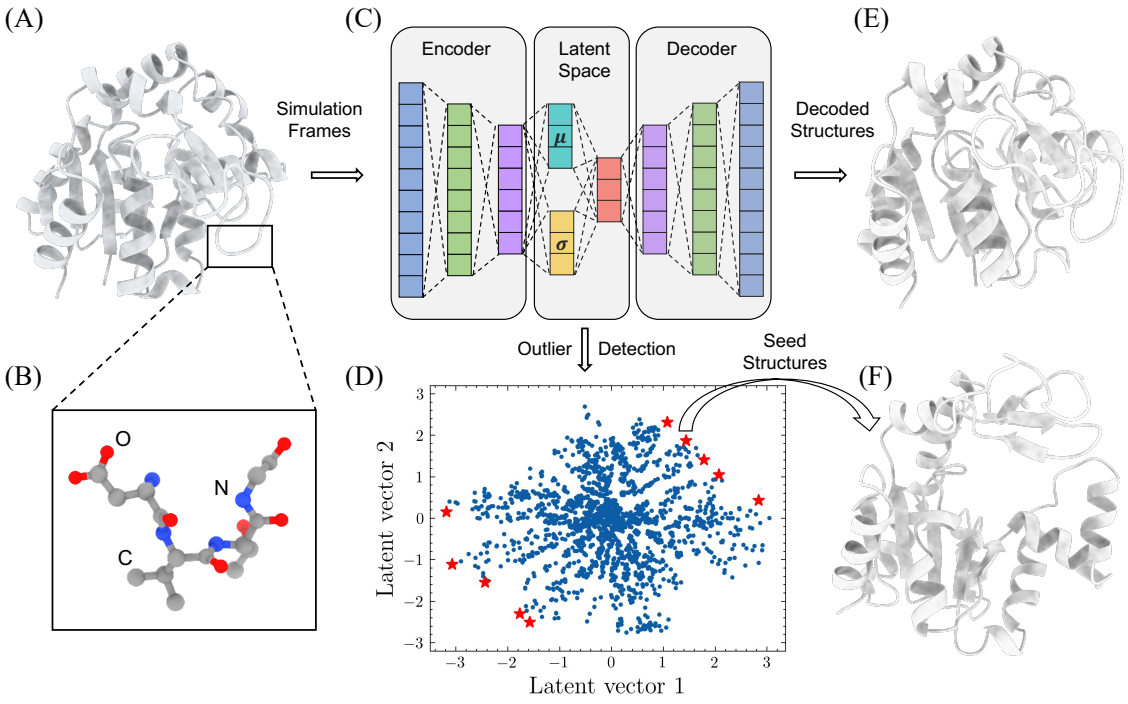

Figure 1. Workflow of LAST method. Heavy atoms in protein simulations are extracted as features to train a variational autoencoder model. In the latent space, outliers are selected based on the joint distribution of multivariate points in the latent space. Protein structures corresponding to these outliers are treated as seed structures to initiate additional simulations. LAST iterative conducts these steps until convergence.

are treated as seeds. Third, additional MD simulations are conducted from seed structures.

VAE Training

In each iteration, some preprocessing procedures are needed. The simulation trajectories are firstly aligned to the first frame and heavy atoms are extracted with Cartesian coordinates being expanded as a feature vector (Figure 1A-B). Then, each feature is transformed to a range of 0 to 1 using min-max linear scaling, which is used to construct a dataset for VAE training.

The architecture of VAE model is shown in Figure 1C. In the current study, we design the encoder model being composed of three hidden layers with size of 512, 128 and 32 and decoder model with size of 32, 128 and 512. The number and size of hidden layers can be adjusted based on the size of proteins. The dimension of latent space is set as two for the simplicity and ease of visualization.

Seeds Selection

Appropriate seed selection method is needed to expedite the sampling of protein conformational space. In LAST, seeds are selected on the two-dimensional learned latent space of VAE, which has two important characteristics to enable an efficient seeds selection. First, as demonstrated in our previous work, the distance between two data points on the latent space is meaningful. Two structurally similar proteins have a shorter distance between their corresponding latent vectors. Second, the sampling distribution of latent space in the VAE is similar to a normal distribution due to the KL divergence term in the loss function. As for the distribution of the VAE latent space of protein conformations, the most common protein structures are encoded in the center of the latent space while structurally different proteins are encoded on the boundary. In a data distribution, outliers refer to those points that differ significantly from other data. Based on the above two points, it is reasonable to treat outliers on the latent space as seeds to accelerate conformational space exploration, as their conformations deviate the majority of the sampled ones.

To implement this method, we propose three improvements

to make LAST computationally efficient.

1. The latent space of VAE does not strictly follow a normal distribution. This is mainly because of relatively strong emphasis on reconstruction loss and lesser emphasis on KL divergence during VAE training. The reconstruction loss term controls the quality of latent space data reconstruction (how well the VAE can reconstruct a protein structure) and KL divergence term constrains the distribution of the latent space (to what degree the latent space needs to follow a normal distribution). Therefore, in order to have a well-constructed and normal regularized latent space, appropriate weights are needed to be set for both terms. This is a challenging task with finetuning by hand, as the sample size keeps growing linearly with additional MD simulations in each round. Therefore, instead of trying to find weights to balance the reconstruction loss and KL divergence, we allow the latent space not strictly follow a normal distribution and use a non-parametric multivariate kernel density estimator to fit the latent space. The estimator is developed in Python statsmodels library (Seabold & Perktold, 2010).

2. The distribution of the latent space might be skewed or kurtotic. In such cases, one side of probability density function (PDF) will have a long tail with low values. This could lead to the situation that all selected seed structures lie in the long tail side, and the corresponding protein structures of these seeds might be similar to each other. Seeds gathering on one side of latent space distribution decreases the chance to explore more structurally different conformations and thus leads to a less efficient protein sampling process. To partially overcome this issue, we used the cumulative distribution function (CDF) to select outliers: data points on the two sides of the CDF are evenly selected. This improvement prevents sampling similar seeds on the boundary of protein conformational spaces.

3. Still, seed structures might be similar to each other. Nontargeted PaCS-MD proposed a nonredundant selection rule which calculates pairwise RMSDs between the current simulation cycle and seeds selected in all the past cycles. (Harada et al., 2019) Protein configurations with large RMSD are then selected as new seeds in the current cycle. We took reference from this idea to select seeds. Outliers from two ends of the estimated CDF are picked sequentially while the pairwise RMSDs to previously selected seeds are calculated. We required the RMSD threshold should be greater than 1 $\mathring{A}$. If not, LAST discards this outlier and moves to the next. Moreover, LAST is a memory method: the selected seed structures are stored for RMSD calculation in future iterations, which avoids repeated sampling in the same conformational region and further improves the sampling efficiency.

One example of seeds selection result is shown in Figure 1D, where seeds are highlighted in red stars in the latent space visualization.

Additional MD Simulations

Short MD simulations are conducted in each round. In the current study, 10 seeds are selected in one round and 100ps simulation is done starting from each seed. Thus, the total simulation time in each round is 1ns. The detail of these simulations is described in section 2.2.

The above three steps are iteratively done until convergence. Here, we design the convergence criterion by calculating the mean RMSD of Cα atoms with regard to the starting protein structure. The iterative sampling process is terminated once the mean RMSD stops to increase for successive five rounds or reaches the maximum round number.

Algorithm 1 Latent space assisted adaptive sampling for protein trajectories

```
Prepare simulation files.
Conduct 100 ps NVT and 200 ps NPT simulations.
while iteration is not reaching the maximum round do
    Align trajectories and extract Cartesian coordinates.
    Train a VAE model.
    Fit latent space with a non-parametric multivariate kernel density estimator.
    Select 10 outliers based on CDF and get seed structures.
    Conduct 100 ps simulation for each seed.
    if mean RMSD is converged then
        Stop iteration.
    end if
end while
```

The LAST algorithm is summarized in Algorithm 1.

3. Results

Four structures of two protein systems (ADK and VVD) were prepared for MD simulations as described in section 2.2. For each protein structure, 100 ps NVT and 200 ps NPT simulations were conducted. During the iterative process, all previous simulations were aligned to the first frame with Cartesian coordinates of heavy atoms being extracted as a feature vector to represent protein conformation. Afterwards, a variational autoencoder model was trained. 10 seed structures were selected with additional 100 ps simulation starting from each of them. Therefore, each iteration takes 1 ns simulation time.

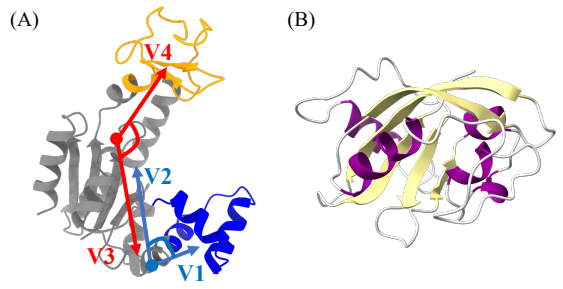

Figure 2. Structures of (A) ADK and (B) VVD. ADK is composed of a CORE domain (grey), a LID domain (orange) and a NMP domain (blue). LID-CORE and NMP-CORE angles are calculated by four vectors to represent protein conformations. VVD protein is colored at secondary structure level using VMD.

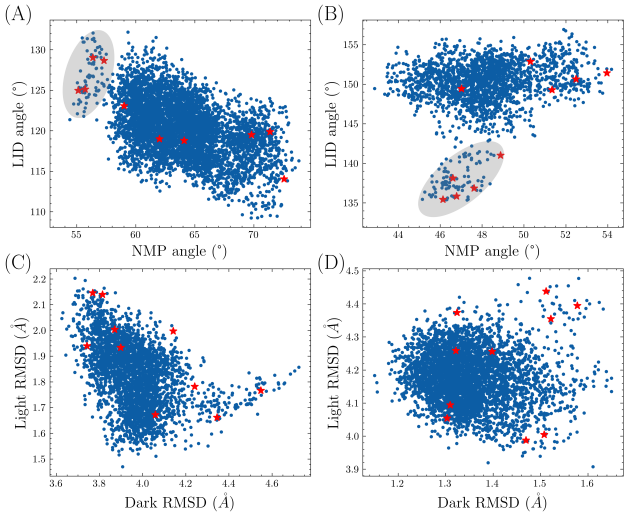

Figure 3. Seed structure distribution on low-dimensional protein representations. (A-B) ADK protein conformations are represented in LID-CORE and NMP-CORE angle vectors. (C-D) VVD protein conformation are represented in RMSDs with regard to the native dark and native light states. Seed structures are represented in red stars. Two less sampled regions are shown in grey circles.

ADK protein is composed of a rigid CORE domain, a lid-shaped ATP-binding domain (LID) and an AMP-binding domain (NMP). Many computational studies have shown ADK to carry out large conformational transitions between the closed state to the open state during the ATP-ADP catalyzation process. (Unan et al., 2015; Matsunaga et al., 2012) Four vectors that form NMP-CORE and LID-CORE angles have been widely used to characterize ADK protein conformation. VVD is a light-oxygen-voltage domain that undergoes global conformational changes upon perturbation. VVD is shown to be flexible in the native light state and relatively stable in the native dark state. (Matsunaga et al., 2012) ADK and VVD structures are illustrated using ChimeraX (Goddard et al., 2018) (Figure 2).

Proper low-dimensional protein representations are needed to evaluate the quality of seed selection. In the current study, ADK protein structure is projected to LID-CORE and NMP-CORE 2D angle plot. We followed the same reside selection rule to calculate vectors and angles. (Tian et al., 2021) For VVD structure, 2D root-mean-square deviation (RMSD) with reference to the native dark and light structures was used to show the sampled protein conformational space.

Both the angle plot in ADK and RMSD plot in VVD were used to display the protein conformation of seed structures (Figure 3). In each subplot, seed structures are highlighted as red stars. In two metastable ADK conformations (Figure 3A-B), seed structures mainly locate in the less sampled regions with small or large LID/NMP angles. It should be noted that there are two significantly less sampled regions in both plots (grey regions in Figure 3A and B).

This indicates that the variational autoencoder can capture the structural differences of protein conformations within the learned latent space. In the native dark and native light VVD conformations (Figure 3C-D), seed structures are also shown to be evenly distributed in the boundary of protein conformational space defined by RMSD to two native VVD structures.

To compare the effectiveness of LAST to conventional molecular dynamics simulations, the sampled protein conformational space in each round of LAST method is displayed together with cMD sampled conformations under the same simulation time. Figure 4 shows the protein conformations in 1 ns, 5 ns, 10 ns and 15 ns for both LAST and cMD. Specifically, cMD was conducted twice independently. It is shown that under the same simulation time, LAST can explore more protein conformations than cMD. Moreover, the trained variational autoencoder can consistently learn a low-dimensional protein representation in the latent space, regardless of the growing number of simulations and changing shape of conformational space, and guide MD simulations to explore less sampled regions. In contrast, there are limited new conformations being explored in cMD simulations from 10 ns to 15 ns, indicating that it might be trapped in a local energy minimum.

We continued the LAST simulation of ADK until convergence. For comparison, two independent 500 ns cMD simulations were conducted. Both protein conformations with free energy being calculated are shown in Figure 5. The LAST sampling method took 22 iterations (22 ns simulation time) and explored two paths from the metastable state to the two native states. This aligns with the computational finding

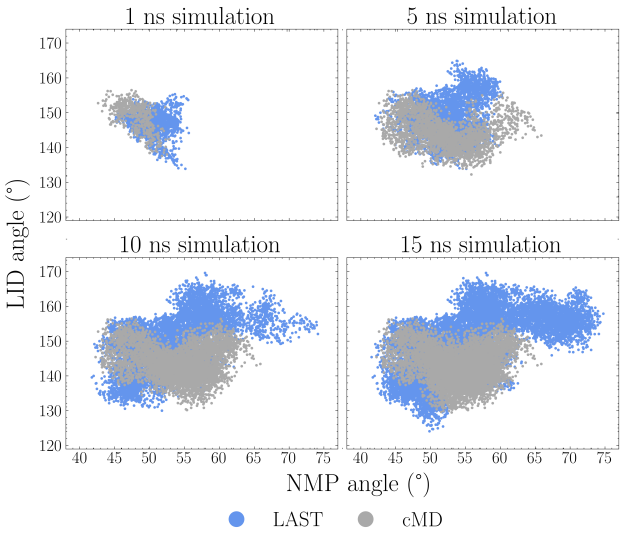

Figure 4. Comparison of ADK conformational spaces of LAST and cMD. Protein conformations are shown in blue at iteration 1, 5, 10 and 15 in LAST method. Protein conformations produced by cMD are shown in grey with the same simulation time. In each round LAST explored a larger conformational space compared with cMD.

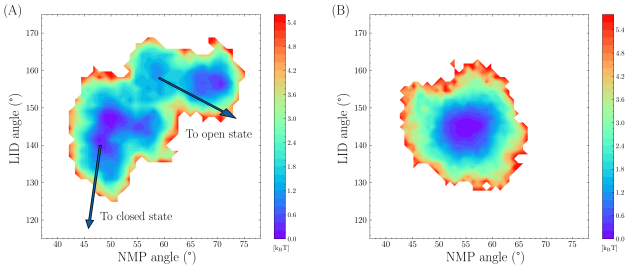

Figure 5. Free energy spaces of (A) LAST and (B) cMD produced protein conformations. LAST took 22 iterations to complete. Two independent 500 ns cMD were conducted for comparison. LAST explored two paths to the open and closed states while cMD being trapped in a local free energy minimum.

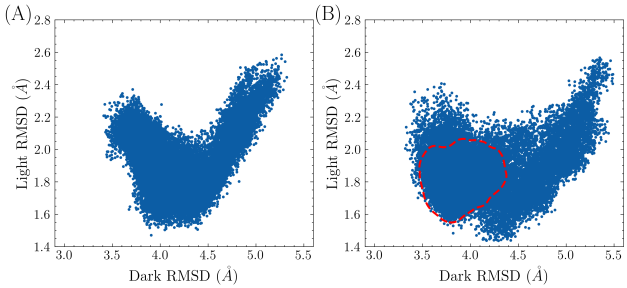

Figure 6. Comparison of VVD conformational spaces of (A) LAST and (B) cMD. LAST took 30 iterations to convergence. The conformational space in cMD with the same simulation time is circled on (B) with red dotted lines. cMD simulation was continued to 100 ns to have a similar space.

that ADK protein undergoes conformational transitions between the open and the closed states. (Formoso et al., 2015) Moreover, the ADK free energy landscape is similar to a previous study (Unan et al., 2015), which the 2AK3 structure (NMP angle of 50° and LID angle of 155°) is shown to lie in a less preferred state and quickly moves to other stable states. The two energy basins also agree with the free energy calculation results that the closed states are more stable than the open state. However, cMD simulation was being trapped in an energy basin, which is also indicated in the free energy plot of LAST. This demonstrates that LAST has a better chance to jump out of energy minimum without biasing potential energies.

For the VVD system, LAST simulation took 30 iterations (30 ns simulation time) to converge. The conformational space is illustrated in Figure 6A. Two cMD simulations were independently conducted with the same simulation time. Their conformational space is highlighted as the red dotted circle in Figure 6B. The region sampled by cMD overlaps greatly with the LAST sampled region but with small coverage of the overall conformational space. To compare the efficiency of two methods, these two cMD simulations were continued while this 2D RMSD map being monitored. It took 100 ns simulation time for cMD simulations to have a similar space shape with LAST.

It is worth noting that the convergence criterion used in this study does not represent the "true" convergence of protein systems. The notion of "true" convergence, as discussed in previous studies, (Romo & Grossfield, 2011; Sawle & Ghosh, 2016; Knapp et al., 2011) is elusive in simulations. More appropriate criteria are needed for the convergence signal in adaptive sampling, through either numerical indicators or self-consistency checks.

The mean RMSDs with regard to the starting protein structure in each iteration were calculated for both systems and are shown in Figure 7. Mean RMSDs are presented with black lines and standard deviation shown in red lines for each round. The maximum and minimum RMSD values are shown as the upper and lower bound in the colored regions. Currently, we set the patience as 5: the iteration loop stops if the maximum mean RMSD does not increase in 5 successive rounds. For the simulation in ADK system, RMSD starts with 2$\AA$, gradually increases to 3.5$\AA$, and stops at iteration 22. In contrast, the RMSDs in VVD system are smaller and the total simulation lasts longer with a total of 30 iterations.

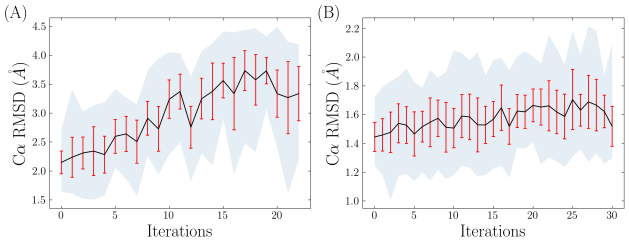

Figure 7. Mean RMSDs in (A) ADK and (B) VVD systems. Mean RMSD values are connected with black lines. Standard deviation in each iteration is plotted as vertical red lines. The gap between minimum and maximum RMSD values are colored as light grey background.

4. Conclusion

In this study, we present an adaptive sampling method, latent space assisted adaptive sampling for protein trajectories, to accelerate the exploration of protein conformational spaces. LAST iterates through variational autoencoder training, seed selection and additional short MD simulations. LAST differs with previous methods in seed selection where the outliers in the learned latent space are selected and treated as seed structures. We tested LAST method in ADK and VVD protein systems, each with different low-dimensional protein representations. In both systems, LAST can capture the key protein characteristics and select seeds that lie in the boundary of conformational space. For ADK simulations, LAST explored two transition paths that are in agreement with previous findings. For VVD simulations, LAST is three times faster than conventional MD for exploring the same conformational regions.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. {TensorFlow}: A system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pp. 265–283, 2016.

Bandyopadhyay, S. and Mondal, J. A deep autoencoder framework for discovery of metastable ensembles in biomacromolecules. *The Journal of Chemical Physics*, 155(11):114106, 2021.

Barducci, A., Bonomi, M., and Parrinello, M. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5):826–843, 2011.

Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., and Westbrook, J. The protein data bank and the challenge of structural genomics. *Nature structural biology*, 7(11):957–959, 2000.

Bowman, G. R., Ensign, D. L., and Pande, V. S. Enhanced modeling via network theory: Adaptive sampling of markov state models. *Journal of Chemical Theory and Computation*, 6(3):787–794, 2010.

Brotzakis, Z. F., Limongelli, V., and Parrinello, M. Accelerating the calculation of protein–ligand binding free energy and residence times using dynamically optimized collective variables. *Journal of Chemical Theory and Computation*, 15(1):743–750, 2018.

Case, D. A., Cheatham III, T. E., Darden, T., Gohlke, H., Luo, R., Merz Jr, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. The amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, 2005.

Chen, W. and Ferguson, A. L. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *Journal of Computational Chemistry*, 39(25):2079–2102, 2018.

Chollet, F. et al. Keras. https://keras.io, 2015.

Doersch, C. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

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., et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7): e1005659, 2017.

Formoso, E., Limongelli, V., and Parrinello, M. Energetics and structural characterization of the large-scale functional motion of adenylate kinase. *Scientific reports*, 5(1): 1–8, 2015.

Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H., and Ferrin, T. E. Ucsf chimerax: Meeting modern challenges in visualization and analysis. *Protein Science*, 27(1):14–25, 2018.

Guo, X., Du, Y., Tadepalli, S., Zhao, L., and Shehu, A. Generating tertiary protein structures via interpretable graph variational autoencoders. *Bioinformatics Advances*, 1(1):vbab036, 2021.

Hamelberg, D., Mongan, J., and McCammon, J. A. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *The Journal of Chemical Physics*, 120(24):11919–11929, 2004.

Harada, R. and Kitao, A. Parallel cascade selection molecular dynamics (pacs-md) to generate conformational transition pathway. *The Journal of Chemical Physics*, 139(3): 07B611_1, 2013.

Harada, R. and Kitao, A. Nontargeted parallel cascade selection molecular dynamics for enhancing the conformational sampling of proteins. *Journal of Chemical Theory and Computation*, 11(11):5493–5502, 2015.

Harada, R., Sladek, V., and Shigeta, Y. Nontargeted parallel cascade selection molecular dynamics based on a nonredundant selection rule for initial structures enhances conformational sampling of proteins. *Journal of Chemical Information and Modeling*, 59(12):5198–5206, 2019.

Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A., and Bikard, D. Generating functional protein variants with variational autoencoders. *PLoS Computational Biology*, 17(2):e1008736, 2021.

Jin, Y., Johannissen, L. O., and Hay, S. Predicting new protein conformations from molecular dynamics simulation conformational landscapes and machine learning. *Proteins: Structure, Function, and Bioinformatics*, 89(8): 915–921, 2021.

Knapp, B., Frantal, S., Cibena, M., Schreiner, W., and Bauer, P. Is an intuitive convergence definition of molecular dynamics simulations solely based on the root mean square deviation possible? *Journal of Computational Biology*, 18(8):997–1005, 2011.

Krivov, S. V. The free energy landscape analysis of protein (fip35) folding dynamics. *The Journal of Physical Chemistry B*, 115(42):12315–12324, 2011.

Lindert, S., Bucher, D., Eastman, P., Pande, V., and McCammon, J. A. Accelerated molecular dynamics simulations with the amoeba polarizable force field on graphics processing units. *Journal of Chemical Theory and Computation*, 9(11):4684–4691, 2013.

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015.

Matsunaga, Y., Fujisaki, H., Terada, T., Furuta, T., Moritsugu, K., and Kidera, A. Minimum free energy path of ligand-induced transition in adenylate kinase. *PLoS computational biology*, 8(6):e1002555, 2012.

Prinz, J.-H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., Chodera, J. D., Schütte, C., and Noé, F. Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics*, 134(17):174105, 2011.

Raiteri, P., Laio, A., Gervasio, F. L., Micheletti, C., and Parrinello, M. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *The Journal of Physical Chemistry B*, 110(8):3533–3539, 2006.

Ramaswamy, V. K., Musson, S. C., Willcocks, C. G., and Degiacomi, M. T. Deep learning protein conformational space with convolutions and latent interpolations. *Physical Review X*, 11(1):011052, 2021.

Romo, T. D. and Grossfield, A. Block covariance overlap method and convergence in molecular dynamics simulation. *Journal of Chemical Theory and Computation*, 7(8): 2464–2472, 2011.

Salomon-Ferrer, R., Gotz, A. W., Poole, D., Le Grand, S., and Walker, R. C. Routine microsecond molecular dynamics simulations with amber on gpus. 2. explicit solvent particle mesh ewald. *Journal of Chemical Theory and Computation*, 9(9):3878–3888, 2013.

Sawle, L. and Ghosh, K. Convergence of molecular dynamics simulation of protein native states: Feasibility vs self-consistency dilemma. *Journal of Chemical Theory and Computation*, 12(2):861–869, 2016.

Seabold, S. and Perktold, J. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, pp. 61. Austin, TX, 2010.

Shaw, D. E., Dror, R. O., Salmon, J. K., Grossman, J., Mackenzie, K. M., Bank, J. A., Young, C., Deneroff, M. M., Batson, B., Bowers, K. J., et al. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the conference on high performance computing networking, storage and analysis*, pp. 1–11, 2009.

Tian, H., Jiang, X., Trozzi, F., Xiao, S., Larson, E. C., and Tao, P. Explore protein conformational space with variational autoencoder. *Frontiers in molecular biosciences*, 8, 2021.

Unan, H., Yildirim, A., and Tekpinar, M. Opening mechanism of adenylate kinase can vary according to selected molecular dynamics force field. *Journal of computer aided molecular design*, 29(7):655–665, 2015.

Wetzel, S. J. Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. *Physical Review E*, 96(2):022140, 2017.

Zhang, J. and Gong, H. Frontier expansion sampling: a method to accelerate conformational search by identifying novel seed structures for restart. *Journal of Chemical Theory and Computation*, 16(8):4813–4821, 2020.