Exploring the conformational ensembles of protein-protein complex with transformer-based generative model

Jianmin Wang[a], Xun Wang[f, h], Yanyi Chu[d], Chunyan Li[e], Xue Li[f], Xiangyu Meng[f], Yitian Fang[g], Kyoung Tai No[a,*], Jiashun Mao[c,*], Xiangxiang Zeng[b,*]

[a] The Interdisciplinary Graduate Program in Integrative Biotechnology, Yonsei University, Incheon 21983, Korea

[b] College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, 410082, China

[c] School of Medical Information and Engineering, Southwest Medical University, Luzhou 646000, Sichuan, China

[d] Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA

[e] School of Informatics, Yunnan Normal University, Kunming, China

[f] School of Computer Science and Technology, China University of Petroleum, Qingdao, 266580, Shandong, China

[g] School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200030, P.R. China

[h] High Performance Computer Research Center, University of Chinese Academy of Sciences, Beijing, 100190, China

ABSTRACT

Protein-protein interactions are the basis of many protein functions, and understanding the contact and conformational changes of protein-protein interactions is crucial for linking protein structure to biological function. Although difficult to detect experimentally, molecular dynamics (MD) simulations are widely used to study the conformational ensembles and dynamics of protein-protein complexes, but there are significant limitations in sampling efficiency and computational costs. In this study, a generative neural network was trained on protein-protein complex conformations obtained from molecular simulations to directly generate novel conformations with physical realism. We demonstrated the use of a deep learning model based on the transformer architecture to explore the conformational ensembles of protein-protein complexes through MD simulations. The results showed that the learned latent space can be used to generate unsampled conformations of protein-protein complexes for obtaining new conformations complementing pre-existing ones, which can be used as an exploratory tool for the analysis and enhancement of molecular simulations of protein-protein complexes.

Keywords: protein-protein complex, molecular dynamics, Transformer, deep generative model, conformational ensembles

INTRODUCTION

Proteins are the basic components of life , and almost all proteins perform their functions through non-covalent interactions with proteins, peptides, nucleotides, and small molecules, and so on[1,2]. Protein-protein interactions (PPIs) play a key role in many biological processes. Understanding the 3D structure of protein-protein complexes can help to understand many biological processes and molecular mechanisms, providing useful insights for drug and protein design[2–7]. The number of protein-protein complex structures that were determined by experimental techniques such as X-ray crystallography and cryo-electron microscopy (cryo-EM) is increasing. However, obtaining these experimental structures is still quite time-consuming and costly, and high-resolution structures determination and dynamic mechanism measurements are still experimentally challenging. Computational modeling of protein-protein complexes based on limited experimental data is a helpful complement and alternative to limited experimental approaches. Traditional approaches consist of protein-protein docking and homology modeling[8–10]. Deep learning (DL) has recently made unprecedented advances in the field of computational biology[11], in particular the various applications of AlphaFold2 [12],AlphaFold-Multimer[6] and RoseTTAFold[13] for the prediction of protein structures and protein-protein complexes. Unfortunately, experimental structures usually only capture static conformations in the low-energy state for protein-protein binding. Although molecular modeling or DL have been applied to predict protein-protein complex structures, their sampling of stable functional conformations, protein binding kinetics, and conformational space remains challenging[2–4,14–16].

Molecular dynamics (MD) simulations have wide applications for the conformations and dynamics of proteins and protein-protein complexes[2,3,14,17–19]. MD samples from potential conformations of biomolecular systems to determine the most energetically favorable subspace in the conformational space. These simulations have greatly benefited from hardware and software developments such as graphics processing unit (GPU), AMBER[20,21], GROMACS[22], and OpenMM[23]. Although these advances have been made, MD simulations to produce dynamic conformational ensembles remain challenging in terms of computational cost and efficient sampling of the conformational ensembles. Protein-protein binding affinity is often stronger , and the process of binding and unbinding often requires longer time scales[3]. Although many MD

simulations and enhanced sampling methods have been developed for PPIs modeling, the conformational ensembles of protein-protein complex remains largely underexplored.

Deep learning and generative artificial intelligence have recently made unprecedented advances in computational biology[11,24,25]. Generative deep learning has been successful for the sampling of protein conformations[25], such as long short-term memory (LSTM)[26], autoencoders (AEs)[27,28], variational autoencoders (VAEs)[29–33], generative adversarial networks (GANs)[31,34], score-based models[35], energy-based models[36], Transformer[31,37,38], and active learning[33]. Training of deep generative models relies on conformations extracted from molecular dynamics simulations. Aims to learn the data distribution of molecular dynamics simulations is used to generate novel realistic ensembles of protein conformations[2,4,14].

In this study, we proposed a Transformer-based generative framework to accelerate the sampling and exploration of the conformational ensembles of protein-protein complexes. First, the crystal structure of the barnase-barstar complex was performed in 100 ns MD production simulations. The iterations were repeated to obtain six trajectories meeting specific criteria. Then the trajectories data were split into 300 ensembles (1000 frame/ensemble) as a training set. We use the Transformer-based generative neural network to learn the 3D conformational distribution of protein-protein complexes in the training set. Our model is named AlphaPPImd, which directly generates novel conformations of protein-protein complex and generates conformations beyond the MD time scale. The results of scoring the generated conformational models of protein-protein complexes show that most of the generated conformational models were acceptable models. Conformational ensembles analysis shows that the learned latent space can be used to generate unsampled conformations of protein-protein complexes. We visualize the key amino acid residues and residue pairs of the protein-protein complexes with higher weights captured by the attention mechanism of the model. The results show that the deep generative model learns key residues from multiple MD trajectories that influence the folding and dynamical mechanisms of protein-protein complexes, provides mechanistic insights into protein-protein binding. Generative deep learning can be utilized as one of the optional tools to enhance the exploration of the conformational ensembles of protein-protein complexes and the analysis of key amino acid residues.

MATERIALS AND METHODS

Molecular Dynamics Simulations

The crystal structure of the barnase-barstar complex (PDB ID: 1BRS)[39] was used to build simulation system and has been widely used in previous research[3,14]. We downloaded the crystal structure from the Protein Data Bank (PDB)[40,41], and extracted chain A and chain D as the initial complex structure by removing the ligand and crystal water. The missing hydrogen atoms were added by tleap module in AmberTools[42]. The system was neutralized by adding Na^+ and Cl^- ions , and solvated in a 12 Å periodic bounding box of TIP3P water molecules[43]. The topology and coordinate files of the system were prepared using the tleap module in AmberTools and the AMBER ff14SB force field[44].

The simulated system was first energy minimized by performing a 500 steps typical ensemble (NVT) simulation with the Langevin integrator, and then further equilibrated by performing a 10,000 steps NPT simulation at 300 K. Particle Mesh Ewald (PME) algorithm was applied to the calculation of long-range electrostatic interactions[45]. A cutoff 1 nm for direct space interactions was used and the SHAKE algorithm was set to constrain the lengths of all bonds involving hydrogen atoms[46]. The simulation time step was set to 2 fs , and then six independent 100 ns MD production were performed. All simulations were conducted by OpenMM 7.7[23].

Model Architecture of AlphaPPImd

In this section, we introduce the Transformer-based AlphaPPImd framework. The approach is inspired by a recent research in exploring the protein conformational ensembles[31], which utilized deep generative models to capture the physical states of proteins that are difficult to analyze by traditional molecular dynamics.

According to the important role of the attention mechanism in the Transformer[47], Fernández proposed a frame-to-frame Transformer based on the extension of the basin-encoded trajectories[48]. In this study, we are concerned with converting the basin-encoded frame n (the source) of protein-protein complex to the basin-encoded frame n+1 (the target) of protein-protein complex.

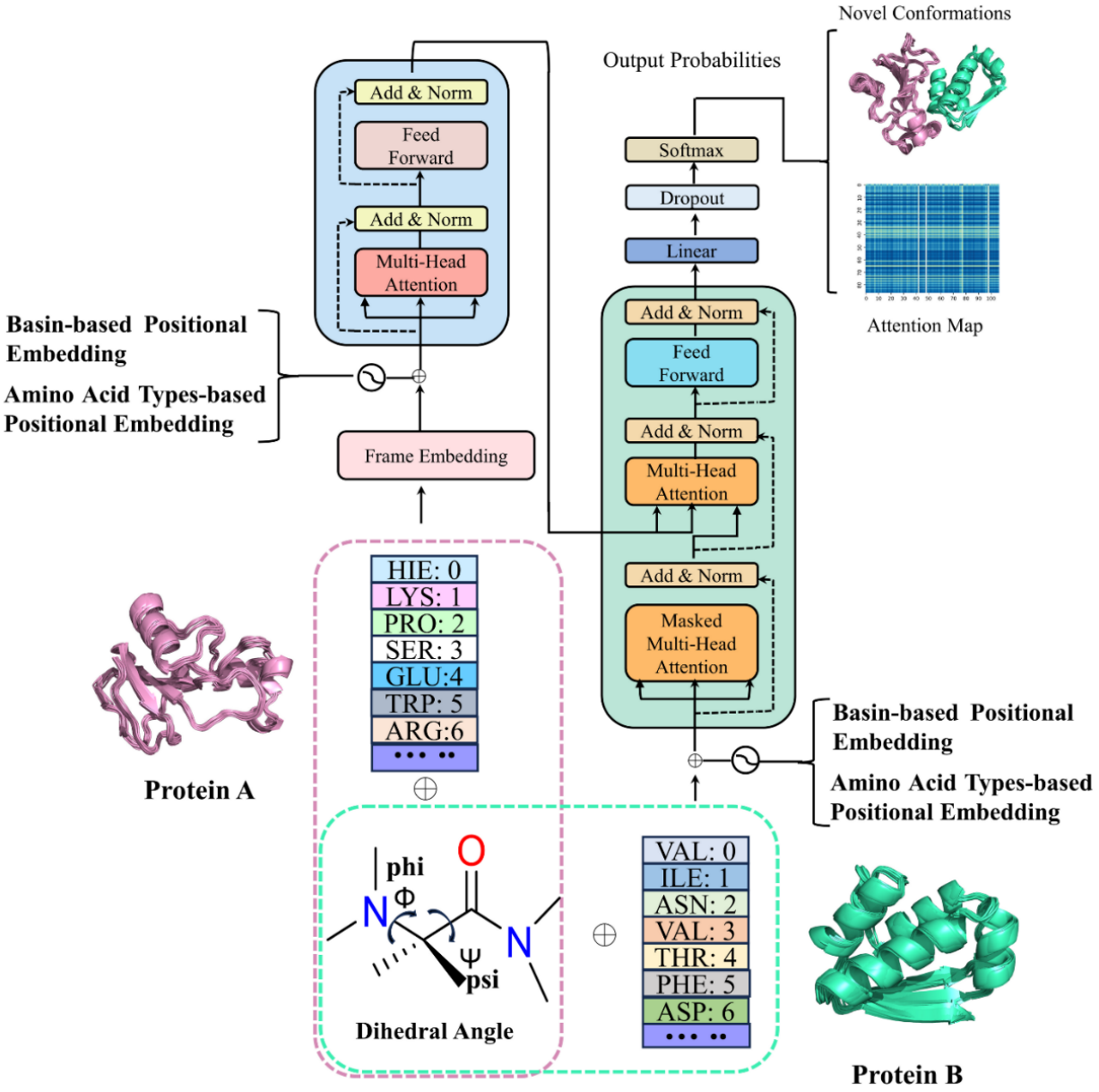

Figure 1. Overview of the AlphaPPImd architecture. The Transformer input is the basin-encoded frames of protein-protein complex. The torsion ϕ , ψ angles of the all residues of the protein-protein complex are required to be encoded, and then the folding state of the protein-protein complex is represented as

cooperative contribution to the protein-complex folding indicated by ϕ , ψ . Multiple short MD trajectories in atomic detail are utilized to train Transformer to propagate protein-protein complex dynamics beyond the time scales accessible to MD. The Transformer outputs novel states of protein complex indicated by ϕ , ψ that are decoded as conformational model via Modeller[49,50].

The core of the AlphaPPImd framework is self-attention mechanism , which can capture key amino acid residue pairs that affect the folding of protein-complex from MD trajectories. AlphaPPImd consists of the following modules (**Figure 1**). First, the MD trajectories of the protein-protein complex are preprocessed to get sequence length, sequence composition and amino acid residue types for both chains. The torsional ϕ , ψ angles of selected residues in the trajectories are calculated to represent the folding state as a string of occupancies in the four basins indicated by the ϕ , ψ plot[31,48]. Second, the embedding module adds two positional embeddings to the amino acid sequence embedding to generate the frame embedding. Two positional embeddings, correlating the basins of basin-encoded frame with their relative positions in the protein sequence, and correlating the basins of basin-encoded frame with the types of amino acids at each position. Then, each frame embedding of protein-protein complex MD trajectories via the embedding module was input to encoder module of AlphaPPImd, which contains the masked multi-head self-attention mechanism, attention score, and feature optimization module. To represent the folding information of the protein-protein complex, we used two encoder modules to capture the folding state of each of the two chains. Next, The decoder module of AlphaPPImd is designed to learn and capture the cooperative contribution of different types and positions of residues of protein-protein complex to protein-protein folding. The decoder module mainly consists of multi-head self-attention layers, layer normalization layers and feedforward layers, and then apply dropout technique to enhance the robustness. Finally, the prediction module iteratively generates the basins for next frame, and then the Modeller(v10.2) reconstructed protein-protein complex conformational model derived from the extended basin-encoded trajectory[49,50].

Attention Mechanism of AlphaPPImd

Attention mechanism is widely used to regulate the weights of feature vectors. The multi-head self-attention layer of the AlphaPPImd decoder module learns the interactions between specific residue pairs. The attention function can be thought of as a mapping between a query (Q) and the output of a key-value pairs (K-V). AlphaPPImd takes the residue embedding of protein-protein complex as Q , regards the global protein-protein complex features as K and V , and calculates the attention weight by using Q and K . The calculation formula is as follows:

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \qquad (1)$$

The attention layer updates the values of the statistical features based on the usefulness of critical residue pairs for the folding of protein-protein complex.

Model Training and Hyperparameters

The six independent 100 ns MD trajectories of the barnase-barstar complex were split into 300 ensembles, each ensemble consisting of 1000 frames. MD trajectories are preprocessed to keep only protein atoms. Each MD runs contributes a limited set of physical snapshots of protein-protein complex[48]. Each frame in the trajectory was represented as a basins encoded by Φ, Ψ indication. Thus, the torsional state of the protein-protein complex is reduced to a textual representation that retains the dominant subordinate features of the dynamics. All the trajectory ensembles are used as training sets, the validation set is always the trajectory with index 0 and the test set is always the trajectory with index 8.

Model training was conducted on Ubuntu Linux 20.04.6 LTS system. The CPU is an Intel(R) Xeon(R) Gold 6226R CPU@ 2.90 GHz. The memory is 376 GB. The model was trained on the single NVIDIA RTX A5000 GPU, the code language is Python 3.7.12, and the model was constructed by TensorFlow 2.8.0 and Keras 2.8.0. The training consists of 10 epochs and each epoch contains 300 subepochs.

To optimize the model, we use the RMSprop optimizer (learning_rate=0.001) , and the loss function is sparse_categorical_crossentropy , and the metrics is accuracy. The number of

trainable parameters for each component of the Transformer model in this study can be found in **Supporting Table S1**.

Evaluation Metrics

The following metrics were used in this study to assess model performance: accuracy, root mean square deviation (RMSD), and a quality measure for protein- protein complex models (DockQ)[51].

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \qquad (2)$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Root-mean-square deviation (RMSD). RMSD was utilized to quantify conformational differences between the trained and decoded conformations for protein-protein complexes.

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (r_i - \hat{r}_i)^2} \qquad (3)$$

We evaluate the accuracy of the generated conformations by DockQ score[51], a quality score of protein-protein models widely used in the computational structural biology community.

$$\text{DockQ} = \frac{1}{3}\left[F_{nat} + \frac{1}{1+\left(\frac{LRMS}{8.5}\right)^2} + \frac{1}{1+\left(\frac{iRMS}{1.5}\right)^2}\right] \qquad (4)$$

DockQ is a score between 0 and 1 , presented based on all the evaluation criteria used in the Critical Assessment of Prediction Interaction (CAPRI). DockQ score is classified as follows: “incorrect”($0.00 \leq \text{DockQ} < 0.23$), “acceptable quality” ($0.23 \leq \text{DockQ} < 0.49$), “medium quality” ($0.49 \leq \text{DockQ} < 0.80$) and “high quality” ($\text{DockQ} \geq 0.80$).

RESULTS AND DISCUSSION

MD Trajectories Representation of PPI complex

The RMSD distributions of protein-protein complex for six independent 100 ns MD simulations are shown in **Supporting Figure S1**. The barnase-barstar complex in the training ensembles of the MD trajectories consists of two different chains and 197 residues (barnase chain: 108 residues, barstar chain: 89 residues), and thus only 195 pairs (barnase chain: 107 pairs, barstar chain: 88 pairs) of ϕ , ψ dihedral angles can be estimated for every time step of the trajectories. As shown in **Figure 2a and 2b**, the Ramachandran plot of the torsional information of the barnase and barstar chain from an ensemble of MD trajectories used in this study, which contains 1000 frames. With each point in the plot representing the backbone torsional state of a residue in one of the frames in the trajectory.

The points of the plot were classified into 4 clusters (basins) labeled 0, 1, 2, and 3 by the KMeans algorithm (**Figure 2c and 2d**). The centroids of each cluster are recorded and stored for reconstruction of the full atomistic model of the barnase-barstar complex from the basin-encoded torsional states. Then, each frame of trajectories was converted into a character vector of an alphabet consisting of 4 symbols corresponding to 4 clusters. A similar representation procedure was performed for all 300 ensembles in the MD trajectories dataset for the barnase-barstar complex.

In summary, the barnase-barstar complex is a heterodimer. There are distinctly different residue basins encoded in the two chains. Thus, implying that there are significant differences in generating novel basin-encoded frames and reconstructing conformational models with individual proteins.

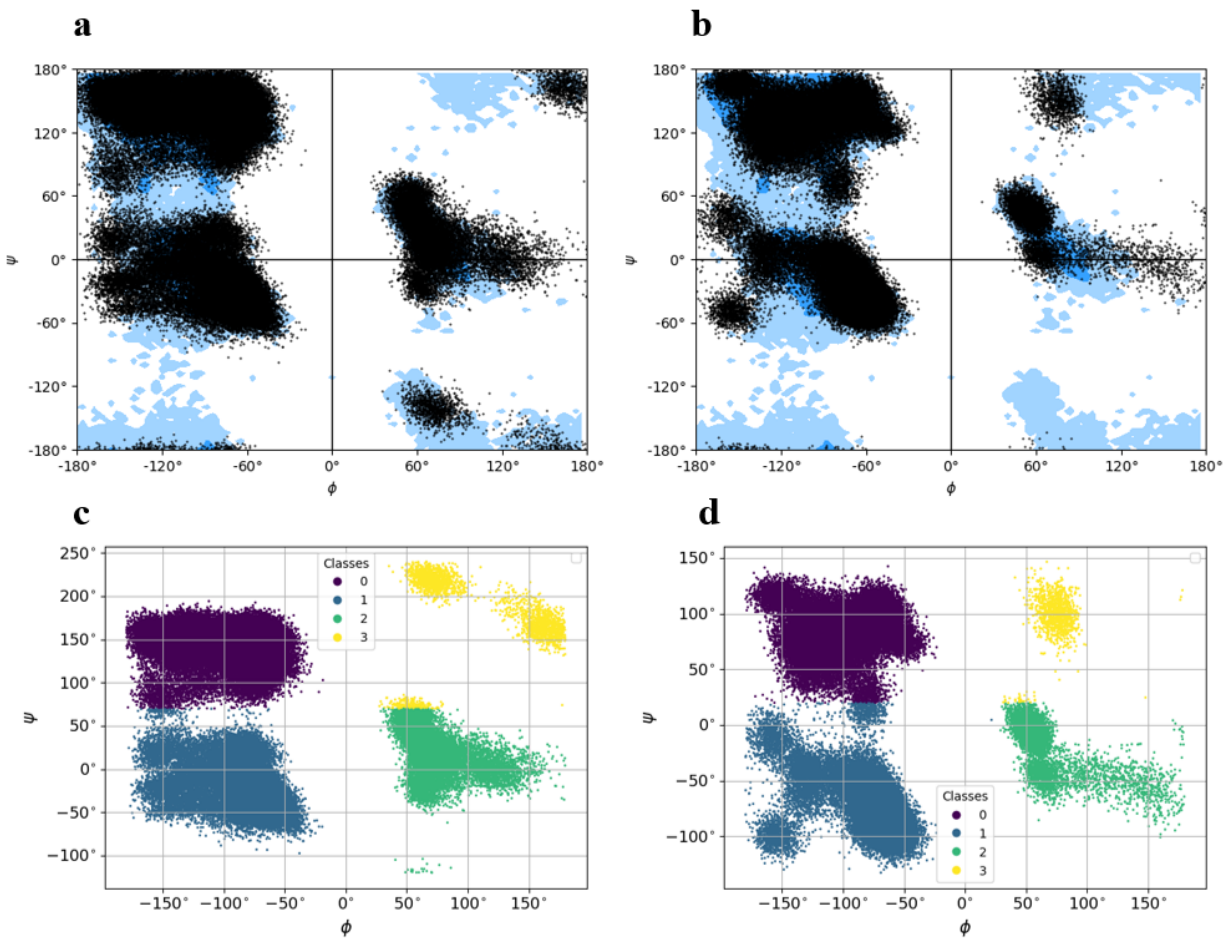

Figure 2. a, Ramachandran plot (allowed : dark blue, marginally allowed : lighter blue) of the first trajectory in the barnase chain set. Each point in the plot is a ϕ , ψ pair of one residue in one frame of the trajectory. **b**, Ramachandran plot (allowed : dark blue, marginally allowed : lighter blue) of the first trajectory in the barstar chain set. Each point in the plot is a ϕ , ψ pair of one residue in one frame of the trajectory. **c**, Classification of the points in the Ramachandran plot as belonging to 1 of 4 basins in the barnase chain. **d**, Classification of the points in the Ramachandran plot as belonging to 1 of 4 basins in the barstar chain.

Model Performance and Conformational Reconstruction

The trajectory ensemble datasets of barnase-barstar complex were selected as the datasets. The model was trained with 300 ensembles of the barnase-barstar complex, each ensemble contains

1000 frames. The accuracy and loss of the Transformer model of the barnase-barstar complex obtained by the training and validation sets of the basin-encoded trajectories are shown in **Figure 3a and 3b**. The average training accuracy of the model is 0.995 and the average validation accuracy is 0.999. Although the AlphaPPImd soon achieved stable performance, to further refine the Transformer model and enrich the MD conformational distribution learned by the model, we used multiple MD trajectories as a dataset. Rapid convergence of the sequence that remains the same in subsequent expansion cycles can be avoided by introducing entropy in the frame prediction module.

A random frame from trajectory of the test set was taken as input , and 100 basin-encoded frames were generated by the trained AlphaPPImd framework. The barnase-barstar complex conformation of the input frame is shown as in **Figure 3c**. The generated extended basin-encoded frames were reconstructed to obtain the conformational model by Modeller[49,50]. The overlap models of the generated conformation and the input conformation are shown in **Figure 3d**. The results show the ability of the model in successfully sampling and expanding the conformations. The Ramachandran plot of the reference structure of the barnase-barstar complex is shown in **Figure 3e**. The Ramachandran plot of the generated conformational model with Φ, Ψ constraints derived by using the center point is shown in **Figure 3f**. The comparison of Figure 3e and 3f shows that the overall conformation of the reference structure is preserved in the corresponding generated conformation. The results show that Modeller correctly performs the dihedral restraints.

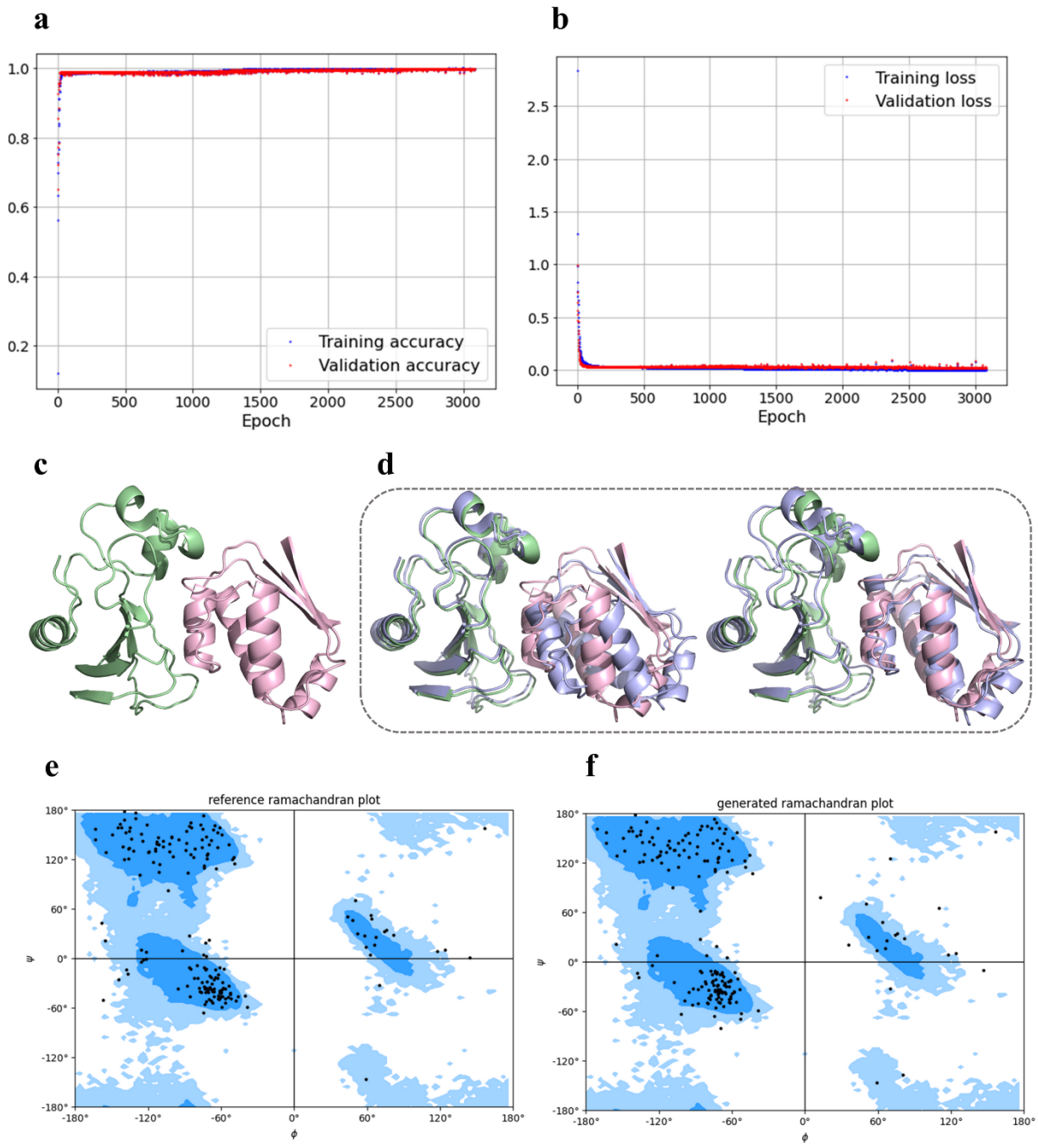

Figure 3. a and b, Training and validation set accuracy and loss for the Transformer neural networks. **c**, The conformation of the barnase-barstar complex in one frame of the test sets. **d**, Overlap of the conformation of the generated barnase-barstar complex(color: lightblue) with the reference conformation. **e**, Ramachandran plot of reference conformation. **f**, Ramachandran plot of generated conformation.

Evaluation of Generated Conformational Ensembles

The MD trajectories of the barnase-barstar complex were randomly selected 1000 frames input into the model to generate 1000 novel basin-encoded frames, which were then reconstructed as novel models of the barnase-barstar complex. The RMSD distribution of 1000 frames of conformations randomly selected from the MD trajectories of the barnase-barstar complex is shown in **Figure 4a**, and the RMSD distribution of 1000 conformations of the barnase-barstar complex that are generated by the AlphaPPImd is illustrated in **Figure 4b**. The RMSD of the conformations in the MD ensembles of the barnase-barstar complex are all less than 2 Å. The RMSD of the conformations in the generated ensembles of barnase-barstar complexes are mostly less than 2Å. The result reveals that the AlphaPPImd framework has high performance in generating novel conformational models of protein-protein complex.

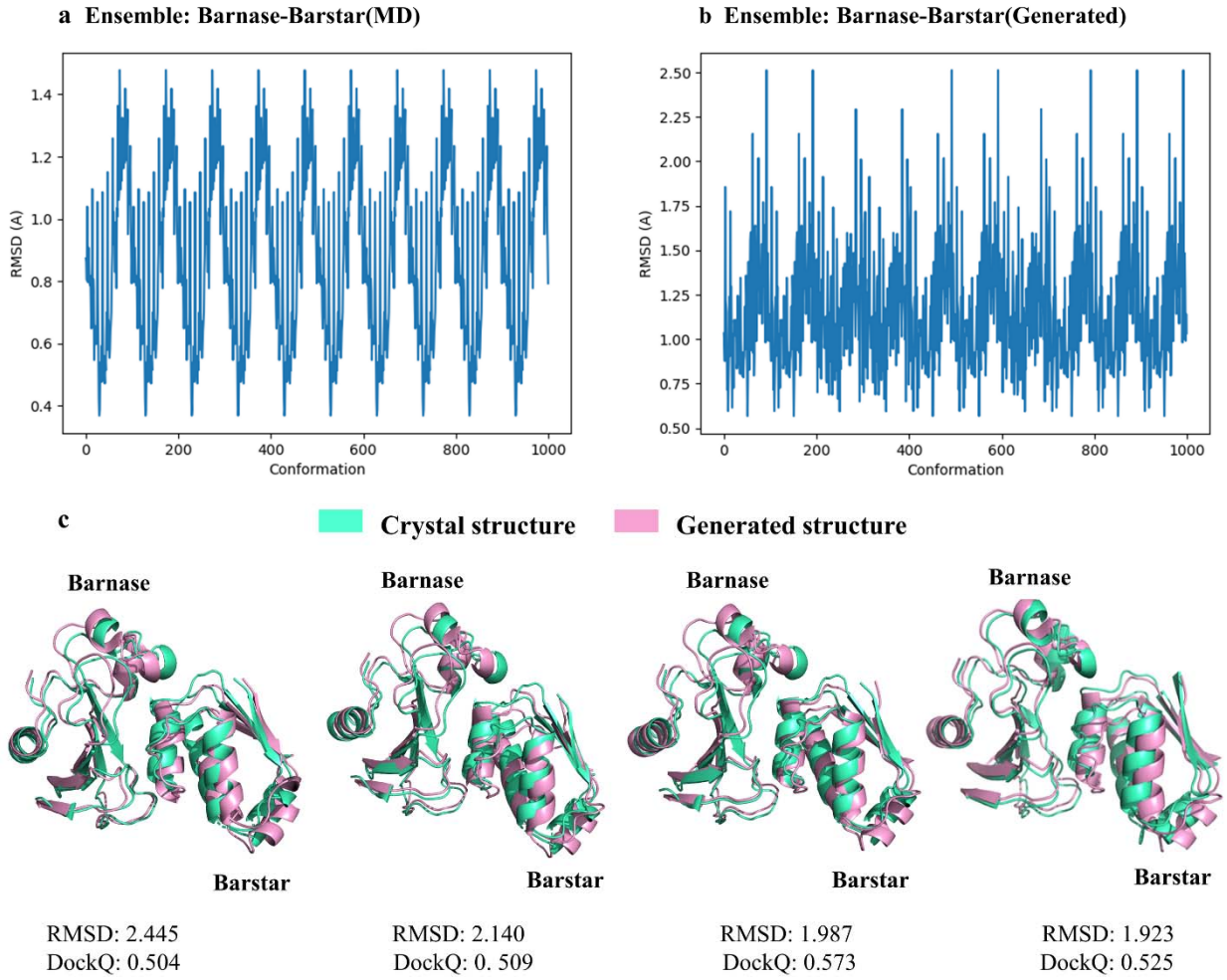

Figure 4. The RMSD distribution of protein-protein complex conformations. **a**, the RMSD distribution of 1000 frames of the barnase-barstar complex MD trajectories. **b**, the RMSD distribution of 1000 conformations of the barnase-barstar complex generated by AlphaPPImd. **c**, Overlap of the generated conformational models of the barnase-barstar complex (color: greencyan) with the crystal structure of the barnase-barstar complex (color: pink), show the RMSD and DockQ score of the generated conformational models.

We selected four representative conformations with RMSD close to 2 Å from the 1000 conformational models of barnase-barstar complexes generated by the AlphaPPImd model. Then, they were overlapped with the crystal structure , and the RMSD and DockQ score were calculated as shown in **Figure 4c**. As the RMSD becomes smaller, the DockQ score of the generated conformations of the barnase-barstar complex is becoming larger, meaning that the novel protein-protein complex conformational models generated by AlphaPPImd are closer to the experimental results, with higher accuracy and acceptability. Although the deep generative model generated many acceptable conformational models of protein-protein complex, their quality was low compared to the physical conformation of protein-protein complex MD. A possible reason is that the basin-encoded frame is a simplified textual representation.

Conformational Space Analysis

The principal component analysis (PCA) is widely used for understanding the dynamics of biological systems[52]. For generating PCA data for ensembles, each frame of the conformational models in the ensembles are overlapped on the initial crystal structure. Then features of each conformational models are calculated for the projection in two-dimensional space. The conformational structures of 1000 randomly selected frames from the MD trajectories of the barnase-barstar complex and 1000 generated structures by the AlphaPPImd model were performed by PCA via ProDy[53]. As shown in **Figure 5**, the MD systems and the sampling ensembles share the same conformational space. The sampling ensembles of the deep generative model covers the MD trajectory ensembles , and samples novel conformations beyond the MD time scale. Although there are a few incorrect conformational models in the conformations

generated by the deep generative model, it generates more diverse conformations. The MD simulations of protein-protein complex can be complemented with their conformational generation models to enhance the understanding of the system dynamics of protein-protein complex.

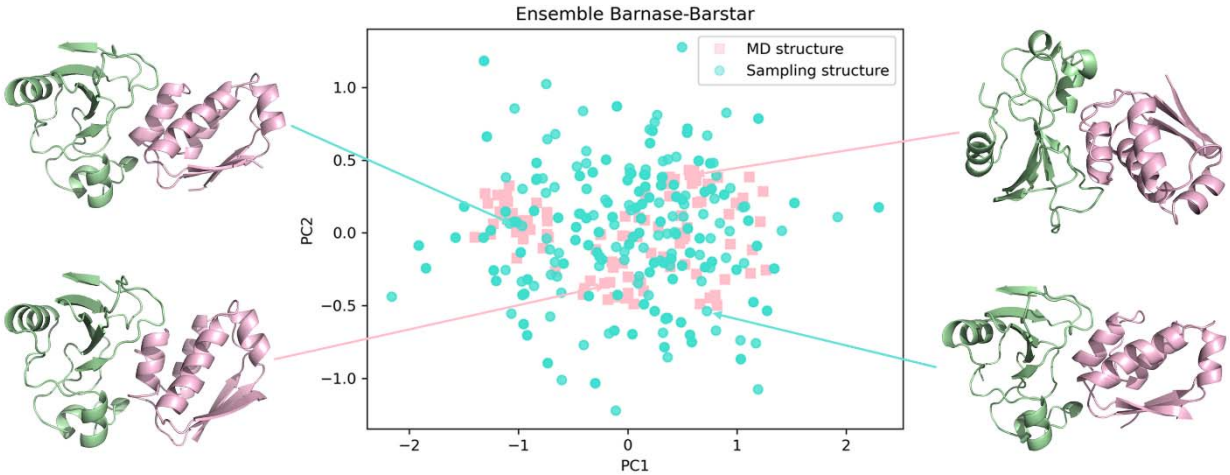

Figure 5. The PCA visualization of the conformational ensembles distributions of protein-protein complex: MD ensembles (color: greencyan) and sampling ensembles (color: pink).

Explainability Analysis

The knowledge of amino acid residues and residue pairs from different partners that are critical to the folding, mechanical and dynamical mechanisms of protein-protein complexes is the basis for understanding many protein functions. The attention mechanism of AlphaPPImd captures the weights between key residues, and provides mechanistic insights into protein-protein binding[54–56]. We explored the key residues and residue pairs between the two chains that have an impact on the dynamics and folding of the barnase-barstar complex by attention score. As shown in **Figure 6a**, different colors are applied with attention weights ranging from 0 to 1. The higher the intensity of red color has the higher attention weight for the particular residue pairs, indicating that its inherent amino acids contribute the most to the dynamics and folding of the barnase-barstar complex, whereas the higher the intensity of blue color has the lower attention weight for the particular residue pairs whose inherent amino acids contribute the least to the dynamics and folding of the barnase-barstar complex. We showed the residues with higher attention weights on

the crystal structure of the barnase-barstar complex, as shown in **Figure 6b** (drawn by Pymol[57]). The results show that the key residues captured by the AlphaPPImd model are mainly located at the interface of protein-protein interactions, loop and helix. It means that deep generative model captures key residues from the MD trajectories of the barnase-barstar complex that affect its dynamics and folding, which can be used to complement the MD results.

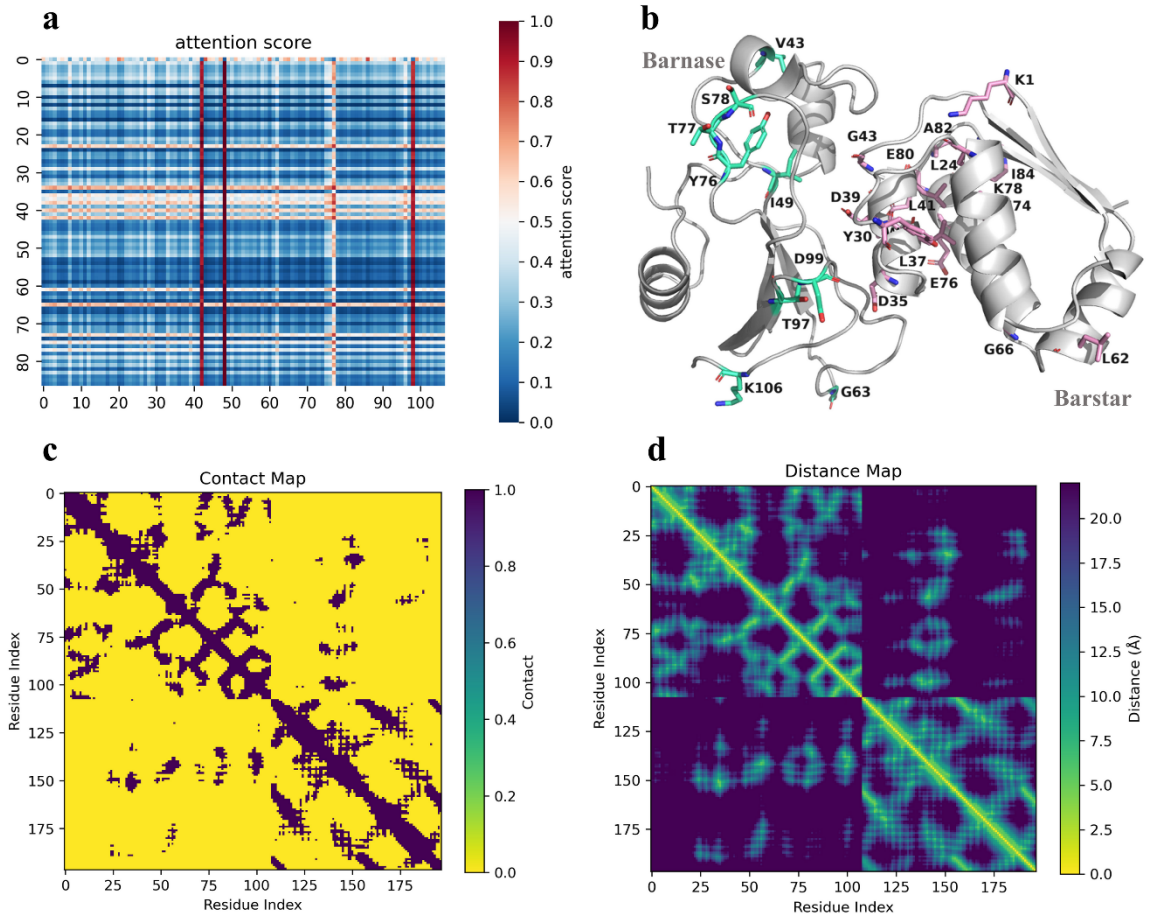

Figure 6. The explainability analysis. **a**, The attention score map of amino acid residue pairs between the two chains of the barnase-barstar complex. **b**, The mapping of attentional residues in the 3D structure of barnase-barstar complex. The key residues of barnase and barstar are highlighted in greencyan and pink stick, respectively. **c**, The contact map of the crystal structure of the barnase-barstar complex. **d**, The distance map of the crystal structure of the barnase-barstar complex.

The contact map is a binary two dimensional matrix representation of residue interactions in the protein structure. The distance map is a two dimensional matrix representation of the distances between all possible residue pairs in the protein structure, where residue-residue distances are shown as a color gradations. We calculated the contact map (cutoffs =13 Å) and distance map of the barnase-barstar complex , as shown in **Figure 6c and 6d**. We were able to observe that most of the key residue pairs between the two chains revealed by the contact/distance map, which appeared in residue pairs with higher attention weights. The continuous contact/distance maps of the MD trajectories of protein-protein complexes are taken as representations to train the deep generative model to generate novel conformational models, potentially producing higher resolution models of protein-protein complex.

CONCLUSION

MD simulations are widely utilized to investigate the conformational ensembles and dynamics of protein-protein complexes, but there are significant limitations in sampling efficiency and computational cost. Deep generative models have achieved unprecedented advancements in the field of computational biology. Generative deep learning has been broadly used for protein conformational sampling. There is still a scarcity of applications of deep generative models to explore the conformational ensembles of protein-protein complexes.

In this study, we present a strategy for accelerated conformational sampling of protein-protein complex by combining Transformer-based generative model with MD simulation trajectories, which can extend protein-protein physical states that are difficult to discover by traditional MD simulation. AlphaPPImd can generate conformations beyond the MD time scale. The results of scoring the generated conformational models of the protein-protein complex showed that most of the generated conformational models were acceptable. The conformational ensembles analysis showed that the learned latent space could be used to generate unsampled conformations of protein-protein complex. The higher weights of residues and residue pairs captured by the attention mechanism were visualized. Deep generative mold learns critical residues from multiple MD trajectories that influence the folding and dynamical mechanisms of protein-protein complexes, and provides mechanistic insights in protein-protein binding. Although there are a

few incorrect conformational models in the generated conformations, it generates more diverse conformations. The deep generative model generated many acceptable conformational models of protein-protein complex, but their quality was low compared to the physical conformation of MD. Training the deep generative model with continuous contact/distance maps of protein-protein complex MD trajectories as inputs or conditions may generate higher resolution conformational models of protein-protein complexes.

In the future, we can combine the rich conformational information of protein-protein complexes, the prediction of protein-protein interactions/binding affinity[58], PPI binding site prediction[59], the prediction of PPI-modulator interaction[60], and the generative design of PPI modulators for accelerating the study of PPI targets and the design/screening of modulators[61,62].

ASSOCIATEDCONTENT

Data Availability

The datasets and source code are publicly available at https://github.com/AspirinCode/AlphaPPImd.

Supporting Information

Supporting Materials, Supporting Tables, and Supporting Figures.

AUTHOR CONTRIBUTIONS

J. Wang collected data, developed the model, analyzed the data, and wrote the manuscript; J. Mao and C. Li developed the model and analyzed the data; X. Wang, Y. Chu, X. Li, X. Meng and Y. Fan helped to refine the research through constructive discussions and revised the manuscript; K. No, J. Mao, X. Zeng supported and supervised the project, interpreted the results, and wrote revisions to the manuscript.

Notes

Any additional relevant notes should be placed here.

ACKNOWLEDGMENTS

This research was supported by the Yonsei University graduate school “Integrative Biotechnology”.

REFERENCES

(1) Nooren, I. M. A.; Thornton, J. M. Diversity of Protein–Protein Interactions. *EMBO. J.* **2003**, *22* (14), 3486–3492.

(2) Dasgupta, B.; Tiwari, S. P. Explicit versus Implicit Consideration of Binding Partners in Protein–Protein Complex to Elucidate Intrinsic Dynamics. *Biophys. Rev.* **2022**, *14* (6), 1379–1392.

(3) Wang, J.; Miao, Y. Protein–Protein Interaction-Gaussian Accelerated Molecular Dynamics (PPI-GaMD): Characterization of Protein Binding Thermodynamics and Kinetics. *J. Chem. Theory. Comput.* **2022**, *18* (3), 1275–1285.

(4) Siebenmorgen, T.; Zacharias, M. Evaluation of Predicted Protein–Protein Complexes by Binding Free Energy Simulations. *J. Chem. Theory. Comput.* **2019**, *15* (3), 2071–2086.

(5) *Correction to Identifying and Overcoming the Sampling Challenges in Relative Binding Free Energy Calculations of a Model Protein:Protein Complex . J. Chem. Theory. Comput*.

(6) Evans, R.; O’Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J.; Ronneberger, O.; Bodenstein, S.; Zielinski, M.; Bridgland, A.; Potapenko, A.; Cowie, A.; Tunyasuvunakool, K.; Jain, R.; Clancy, E.; Kohli, P.; Jumper, J.; Hassabis, D. Protein Complex Prediction with AlphaFold-Multimer. *bioRxiv* **2022**, 2021.10.04.463034.

(7) Bryant, P.; Pozzati, G.; Zhu, W.; Shenoy, A.; Kundrotas, P.; Elofsson, A. Predicting the Structure of Large Protein Complexes Using AlphaFold and Monte Carlo Tree Search. *Nat. Commun. 2022 13:1* **2022**, *13* (1), 1–14.

(8) Vajda, S.; Kozakov, D. Convergence and Combination of Methods in Protein–Protein Docking. *Curr. Opin. Struct. Biol.* **2009**, *19* (2), 164–170.

(9) Siebenmorgen, T.; Zacharias, M. Computational Prediction of Protein–Protein Binding Affinities. *Wiley. Interdiscip. Rev. Comput. Mol. Sci.* **2020**, *10* (3), e1448.

(10) Kundrotas, P. J.; Zhu, Z.; Janin, J.; Vakser, I. A. Templates Are Available to Model Nearly All Complexes of Structurally Characterized Proteins. *Proc. Natl. Acad. Sci. U S A* **2012**, *109* (24), 9438–9441.

(11) Sapoval, N.; Aghazadeh, A.; Nute, M. G.; Antunes, D. A.; Balaji, A.; Baraniuk, R.; Barberan, C. J.; Dannenfelser, R.; Dun, C.; Edrisi, M. Current Progress and Open Challenges for Applying Deep Learning across the Biosciences. *Nat. Commun.* **2022**, *13* (1), 1728.

(12) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589.

(13) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Dustin Schaeffer, R.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; Van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Christopher Garcia, K.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science (1979)* **2021**, *373* (6557), 871–876.

(14) Martin, J.; Frezza, E. A Dynamical View of Protein-Protein Complexes: Studies by Molecular Dynamics Simulations. *Front. Mol. Biosci.* **2022**, *9*, 970109.

(15) Basu, S.; Wallner, B. Finding Correct Protein–Protein Docking Models Using ProQDock. *Bioinformatics* **2016**, *32* (12), i262–i270.

(16) Harmalkar, A.; Mahajan, S. P.; Gray, J. J. Induced Fit with Replica Exchange Improves Protein Complex Structure Prediction. *PLoS Comput. Biol.* **2022**, *18* (6), e1010124.

(17) Allison, J. R. Computational Methods for Exploring Protein Conformations. *Biochem. Soc. Trans.* **2020**, *48* (4), 1707–1724.

(18) 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* (8), 915–921.

(19) Hunkler, S.; Diederichs, K.; Kukharenko, O.; Peter, C. Fast Conformational Clustering of Extensive Molecular Dynamics Simulation Data. *J. Chem. Phys.* **2023**, *158* (14).

(20) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *Wiley. Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3* (2), 198–210.

(21) Case, D. A.; Cheatham III, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688.

(22) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1*, 19–25.

(23) 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. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS. Comput. Biol.***2017**, *13* (7), e1005659.

(24) Agajanian, S.; Alshahrani, M.; Bai, F.; Tao, P.; Verkhivker, G. M. Exploring and Learning the Universe of Protein Allostery Using Artificial Intelligence Augmented Biophysical and Computational Approaches. *J. Chem. Inf. Model.* **2023**, *63* (5), 1413–1428.

(25) Zheng, L.-E.; Barethiya, S.; Nordquist, E.; Chen, J. Machine Learning Generation of Dynamic Protein Conformational Ensembles. *Molecules* **2023**, *28* (10), 4047.

(26) Tsai, S. T.; Kuo, E. J.; Tiwary, P. Learning Molecular Dynamics with Simple Language Model Built upon Long Short-Term Memory Neural Network. *Nat. Commun. 2020 11:1* **2020**, *11* (1), 1–11.

(27) Gupta, A.; Dey, S.; Hicks, A.; Zhou, H.-X. Artificial Intelligence Guided Conformational Mining of Intrinsically Disordered Proteins. *Commun. Biol.* **2022**, *5* (1), 610.

(28) Tatro, N. J.; Das, P.; Chen, P.-Y.; Chenthamarakshan, V.; Lai, R. ProGAE: A Geometric Autoencoder-Based Generative Model for Disentangling Protein Conformational Space. **2020**.

(29) Tian, H.; Jiang, X.; Trozzi, F.; Xiao, S.; Larson, E. C.; Tao, P. Explore Protein Conformational Space with Variational Autoencoder. *Front. Mol. Biosci.* **2021**, *8*, 781635.

(30) Tian, H.; Jiang, X.; Xiao, S.; La Force, H.; Larson, E. C.; Tao, P. LAST: Latent Space-Assisted Adaptive Sampling for Protein Trajectories. *J. Chem. Inf. Model.* **2022**, *63* (1), 67–75.

(31) Schwing, G.; Palese, L. L.; Fernández, A.; Schwiebert, L.; Gatti, D. L. Molecular Dynamics without Molecules: Searching the Conformational Space of Proteins with Generative Neural Networks. *arXiv preprint arXiv:2206.04683* **2022**.

(32) Mansoor, S.; Baek, M.; Park, H.; Lee, G. R.; Baker, D. Protein Ensemble Generation through Variational Autoencoder Latent Space Sampling. *22ioRxiv* **2023**, 2023.08. 01.551540.

(33) Kleiman, D. E.; Shukla, D. Active Learning of the Conformational Ensemble of Proteins Using Maximum Entropy VAMPNets. *J. Chem. Theory. Comput.* **2023**, *19* (14), 4377–4388.

(34) Janson, G.; Valdes-Garcia, G.; Heo, L.; Feig, M. Direct Generation of Protein Conformational Ensembles via Machine Learning. *Nat. Commun.* **2023**, *14* (1), 774.

(35) Lu, J.; Zhong, B.; Tang, J. Score-Based Enhanced Sampling for Protein Molecular Dynamics. *arXiv preprint arXiv:2306.03117* **2023**.

(36) Du, Y.; Meier, J.; Ma, J.; Fergus, R.; Rives, A. Energy-Based Models for Atomic-Resolution Protein Conformations. In *International Conference on Learning Representations*.

(37) Zeng, W.; Cao, S.; Huang, X.; Yao, Y. A Note on Learning Rare Events in Molecular Dynamics Using Lstm and Transformer. *arXiv preprint arXiv:2107.06573* **2021**.

(38) Chennakesavalu, S.; Rotskoff, G. M. Data-Efficient Generation of Protein Conformational Ensembles with Backbone-to-Side-Chain Transformers. *J. Phys. Chem. B* **2024**.

(39) Buckle, A. M.; Schreiber, G.; Fersht, A. R. Protein-Protein Recognition: Crystal Structural Analysis of a Barnase-Barstar Complex at 2.0-. ANG. Resolution. *Biochemistry* **1994**, *33* (30), 8878–8889.

(40) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic. Acids. Res.* **2000**, *28* (1), 235–242.

(41) Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chao, H.; Chen, L.; Craig, P. A.; Crichlow, G. V; Dalenberg, K.; Duarte, J. M. RCSB Protein Data Bank (RCSB. Org): Delivery of Experimentally-

Determined PDB Structures alongside One Million Computed Structure Models of Proteins from Artificial Intelligence/Machine Learning. *Nucleic. Acids. Res.* **2023**, *51* (D1), D488–D508.

(42) Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I.; Brozell, S. R.; Cerutti, D. S.; Cheatham III, T. E.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E. *Amber 2021*; University of California, San Francisco, 2021.

(43) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935.

(44) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory. Comput.* **2015**, *11* (8), 3696–3713.

(45) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593.

(46) 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. *J. Comput. Phys.* **1977**, *23* (3), 327–341.

(47) Vaswani, A.; Brain, G.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural. Inf. Process. Syst.* **2017**, *30*.

(48) Fernández, A. Artificial Intelligence Deconstructs Drug Targeting in Vivo by Leveraging a Transformer Platform. *ACS Med. Chem. Lett.* **2021**, *12* (7), 1052–1055.

(49) Šali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234* (3), 779–815.

(50) Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinformatics.* **2016**, *54* (1), 5.6.1-5.6.37.

(51) Basu, S.; Wallner, B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLoS. One.* **2016**, *11* (8), e0161879.

(52) Yang, L. W.; Eyal, E.; Bahar, I.; Kitao, A. Principal Component Analysis of Native Ensembles of Biomolecular Structures (PCA_NEST): Insights into Functional Dynamics. *Bioinformatics* **2009**, *25* (5), 606–614.

(53) Bakan, A.; Meireles, L. M.; Bahar, I. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **2011**, *27* (11), 1575–1577.

(54) Chu, Y.; Zhang, Y.; Wang, Q.; Zhang, L.; Wang, X.; Wang, Y.; Salahub, D. R.; Xu, Q.; Wang, J.; Jiang, X.; Xiong, Y.; Wei, D. Q. A Transformer-Based Model to Predict Peptide–HLA Class I Binding and Optimize Mutated Peptides for Vaccine Design. *Nature Machine Intelligence 2022 4:3* **2022**, *4* (3), 300–311

(55) Mao, J.; Wang, J.; Zeb, A.; Cho, K. H.; Jin, H.; Kim, J.; Lee, O.; Wang, Y.; No, K. T. Transformer-Based Molecular Generative Model for Antiviral Drug Design. *J. Chem. Inf. Model.* **2023**.

(56) Wang, J.; Mao, J.; Wang, M.; Le, X.; Wang, Y. Explore Drug-like Space with Deep Generative Models. *Methods* **2023**, *210*, 52–59

(57) Schrödinger, L. L. C. The PyMOL Molecular Graphics System, Version 1.8. November 2015.

(58) Li, X.; Han, P.; Chen, W.; Gao, C.; Wang, S.; Song, T.; Niu, M.; Rodriguez-Patón, A. MARPPI: Boosting Prediction of Protein–Protein Interactions with Multi-Scale Architecture Residual Network. *Brief. Bioinform.* **2023**, *24* (1), 1–10.

(59) Wang, S.; Chen, W.; Han, P.; Li, X.; Song, T. RGN: Residue-Based Graph Attention and Convolutional Network for Protein-Protein Interaction Site Prediction. *J. Chem. Inf. Model.* **2022**, *62* (23), 5961–5974.

(60) Sun, H.; Wang, J.; Wu, H.; Lin, S.; Chen, J.; Wei, J.; Lv, S.; Xiong, Y.; Wei, D. Q. A Multimodal Deep Learning Framework for Predicting PPI-Modulator Interactions. *J. Chem. Inf. Model.* **2023**, *63* (23), 7363–7372.

(61) Wang, J.; Chu, Y.; Mao, J.; Jeon, H. N.; Jin, H.; Zeb, A.; Jang, Y.; Cho, K. H.; Song, T.; No, K. T. De Novo Molecular Design with Deep Molecular Generative Models for PPI Inhibitors. *Brief. Bioinform.* **2022**, *23* (4), 1–10.

(62) Wang, J.; Mao, J.; Li, C.; Xiang, H.; Wang, X.; Wang, S.; Wang, Z.; Chen, Y.; Li, Y.; Sun, H.; No, K. T.; Song, T.; Zeng, X. Interface-Aware Molecular Generative Framework for Protein-Protein Interaction Modulators. *bioRxiv* **2024**, 2023.10.10.557742.