LEARNING TO PREDICT ENSEMBLES OF PROTEIN CONFORMATIONS FROM MOLECULAR DYNAMICS SIMULATION TRAJECTORIES

Bongjin Koo *
Simons Machine Learning Center
New York Structural Biology Center
New York, NY, USA
bkoo@nysbc.org

Patrick Jiang
School of Augmented Intelligence
Arizona State University
Tempe, AZ, USA
phjiang@asu.edu

Soumya Dutta
School of Molecular Sciences
Arizona State University
Tempe, AZ, USA
sdutta46@asu.edu

I. Can Kazan
Center for Biological Physics
Department of Physics
Arizona State University
Tempe, AZ, USA
John.Kazan@asu.edu

S. Banu Ozkan
Center for Biological Physics
Department of Physics
Arizona State University
Tempe, AZ, USA
Banu.Ozkan@asu.edu

Paul T Kim
Simons Machine Learning Center
New York Structural Biology Center
New York, NY, USA
pkim@nysbc.org

Abhishek Singharoy
School of Molecular Sciences
Arizona State University
Tempe, AZ, USA
asinghar@asu.edu

Tristan Bepler
Simons Machine Learning Center
New York Structural Biology Center
New York, NY, USA
tbepler@nysbc.org

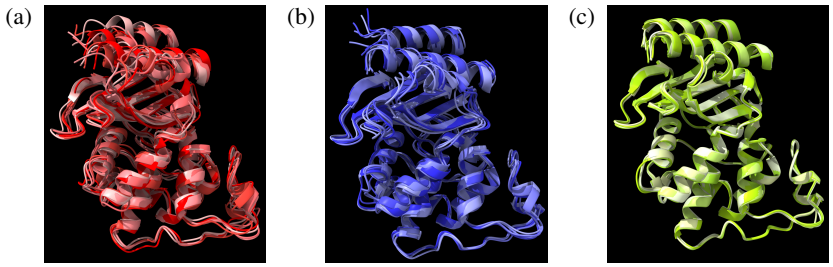

Figure 1: 3D structures of β-lactamase's one mutation from the test set experiment. Samples are (a) from T-REMD; (b) predicted using the finetuned AlphaFlow (`AlphaFlow-FT`); (c) predicted using the pretrained AlphaFlow (`AlphaFlow-PT`), given a sequence unseen during training. `AlphaFlow-FT` generates more diverse conformations. The Wasserstein distance between T-REMD and `AlphaFlow-FT` samples (1.67) is smaller than that between T-REMD and `AlphaFlow-PT` samples (2.15), which tells that `AlphaFlow-FT` samples follow the T-REMD ensemble distribution better.

Introduction. A group of heterogeneous conformations of a protein, also known as an *ensemble* of conformations, is a key to understanding protein functions. This is because many proteins are *mechanical machines* that perform tasks by changing their shapes. Nevertheless, the main focus of

*Corresponding author

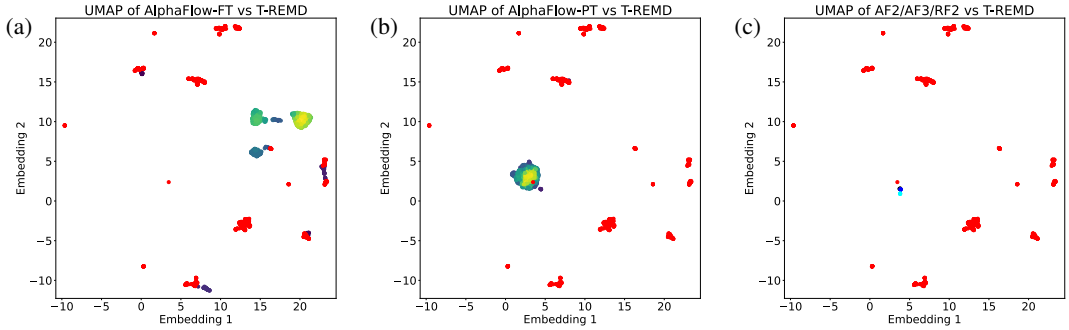

Figure 2: UMAP embedding of the samples of one mutation. Samples from T-REMD training set are embedded with samples from (a) `AlphaFlow-FT`; (b) `AlphaFlow-PT`; (c) AF2, AF3, and RF2. T-REMD samples are in red. Others in (a) and (b) are colored by their densities; more yellow means of higher density and more blue of less density. AF2 samples are in black, AF3 blue, and RF2 cyan in (c). Samples of `AlphaFlow-FT` cover better the distribution of T-REMD while those of other models cluster around a region.

protein structure prediction from a sequence thus far has been to accurately predict a single structure, e.g., AlphaFold (AF) [Abramson et al. (2024)] and ESMFold [Lin et al. (2023)]. Recently, works on predicting multiple conformations by subsampling MSAs (multiple sequence alignments) [del Alamo et al. (2022)] or by clustering MSAs [Wayment-Steele et al. (2024)] were introduced. While they can predict heterogeneous conformations, they are limited w.r.t. the diversity of predicted structures as well as the trainability on data other than Protein Data Bank (PDB) [Berman et al. (2000)] structures, such as on molecular dynamics (MD) simulation trajectories. AlphaFlow [Jing et al. (2024)] overcame this limitation by incorporating a Flow Matching (FM) [Lipman et al. (2023)] framework with AlphaFold as a denoising model. Since an FM model can generate diverse samples by transforming the initial samples from a prior distribution, AlphaFlow has a potential to generate ensembles of conformations. The authors showed that it can be trained on MD trajectories and generate physically feasible ensembles. In this paper, we look more closely into AlphaFlow's ability on learning MD ensembles that are generated using Temperature Replica Exchange Molecular Dynamics (T-REMD) [Qi et al. (2018)]. This is an exploratory study before improving its architecture for proposing our own model.

Experiments. Our dataset consists of 117 sets of trajectories; each set contains MD trajectories of a certain mutation (i.e., having a unique sequence) of β-lactamase. T-REMD was used to generate the trajectories. It is an enhanced sampling method that provides considerably better sampling of the conformational space of a protein compared to normal MD simulations. For a protein, 4 replicas were simulated in parallel at $310 - 340$ K. We finetuned AlphaFlow on these data (`AlphaFlow-FT`), starting from AlphaFlow pretrained on PDB structures (`AlphaFlow-PT`, provided by the original paper). First, we experimented on one mutation only (**`exp-train`**), i.e., trained the model on all trajectories of a mutation and compared the predicted samples to the training samples. Note that the input to the model for prediction is the sequence of the mutation. This experiment is to check if the model learns the distribution of the training data faithfully in the easiest setup. Next, out of all 117 sets of mutations, we trained on 85 sets with 10 sets for validation. The remaining 22 sets were held out as the test set (**`exp-test`**). During the test, the model predicts the ensembles of structures using the sequences of the mutations in the test set. The predicted structures per mutation are compared against the ones corresponding to each mutation to check if the former follows the distribution of the latter.

Results. To qualitatively check if the model learned well the distribution of the training data in **`exp-train`**, we embed the predicted and training samples together into low dimensional spaces using Uniform Manifold Approximation and Projection (UMAP) [McInnes et al. (2020)]. To quantitatively check it, we compute the Wasserstein distance between those samples, using RMSD (root-mean-square distance) as the cost. For **`exp-test`**, the checks are performed with the predicted and test samples. Figure 2 shows the UMAP embeddings of samples from **`exp-train`**. The predicted samples are (a) from `AlphaFlow-FT`, (b) from `AlphaFlow-PT`, and (c) from AF2/3

and RoseTTAFold2 (RF2) [Baek et al. (2023)]. The number of samples are 950, 1000, 1000, 5, 5, and 5 for the training data and the models, respectively. The samples from `AlphaFlow-FT` cover some of the clusters in the training samples (Figure 2(a)). On the contrary, the samples from `AlphaFlow-PT` cluster around one region (Figure 2(b)). Interestingly, this region corresponds to the one in which the samples from AF2/3 and RF2 fall. This shows that AlphaFlow that is not finetuned on MD trajectories tends to generate samples only near the PDB structures. The Wasserstein distance between samples of T-REMD and `AlphaFlow-FT` is 2.27, between T-REMD and `AlphaFlow-PT` 3.39, between T-REMD and AF2/AF3/RF2 3.43, 3.44, 3.52. This shows that `AlphaFlow-FT` samples follow that of the training data better than others. For **`exp-test`**, the mean/median (std.) of the Wasserstein distances, across the 22 mutations, between samples of T-REMD and `AlphaFlow-FT` is $2.49/2.45$ (0.68), and between T-REMD and `AlphaFlow-PT` $2.81/2.86$ (0.67). Figure 1 shows 3D structures of a mutation from this experiment. The structures from `AlphaFlow-FT` look more diverse than those from its counterpart. The Wasserstein distance between the structures from T-REMD and `AlphaFlow-FT` (1.67) is smaller than that between the structures from T-REMD and `AlphaFlow-PT` (2.15). This tells us that the samples from `AlphaFlow-FT` follow T-REMD test set, which is not used for training, more faithfully. The UMAP embeddings for all mutations from **`exp-test`** (not shown in the paper) confirm this trend.

Conclusions and Future Works. This paper investigates the viability of AlphaFlow for generating ensembles of conformations by training it on MD trajectories of various mutations of β-lactamase. The experiments show that AlphaFlow which is not finetuned on MD trajectories has difficulty in generating diverse conformations outside the narrow region around the static structures learned from PDB data. Our next step is to make the model generate more diverse conformations and more faithfully follow the ensembles generated by MD simulations. Potential directions include swapping the denoising model, i.e., AlphaFold, with another model, or using different input embeddings other than MSAs or Protein Language Models.

MEANINGFULNESS STATEMENT

This paper is a preliminary work towards developing a novel generative model for ensembles of protein conformations. Learning the representation of protein conformation ensembles is crucial for understanding protein functions and ultimately the mechanism of life. This paper explores the possibility of learning a reliable representation of protein conformation ensembles from MD trajectories using a state-of-the-art method. In doing so, we gained valuable insights into the workings of AlphaFlow and Flow Matching models, which help us develop our own improved generative model.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL `https://www.nature.com/articles/s41586-024-07487-w`. Publisher: Nature Publishing Group.

Minkyung Baek, Ivan Anishchenko, Ian R. Humphreys, Qian Cong, David Baker, and Frank DiMaio. Efficient and accurate prediction of protein structure using RoseTTAFold2, May 2023. URL `https://www.biorxiv.org/content/10.1101/2023.05.24.542179v1`. Pages: 2023.05.24.542179 Section: New Results.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–

242, January 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235. URL https://doi.org/10.1093/nar/28.1.235.

Diego del Alamo, Davide Sala, Hassane S Mchaourab, and Jens Meiler. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife*, 11:e75751, March 2022. ISSN 2050-084X. doi: 10.7554/eLife.75751. URL https://doi.org/10.7554/eLife.75751. Publisher: eLife Sciences Publications, Ltd.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. AlphaFold Meets Flow Matching for Generating Protein Ensembles, September 2024. URL http://arxiv.org/abs/2402.04845. arXiv:2402.04845 version: 2.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574. Publisher: American Association for the Advancement of Science.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling, February 2023. URL http://arxiv.org/abs/2210.02747. arXiv:2210.02747 [cs].

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. URL http://arxiv.org/abs/1802.03426. arXiv:1802.03426 [stat].

Ruxi Qi, Guanghong Wei, Buyong Ma, and Ruth Nussinov. Replica Exchange Molecular Dynamics: A Practical Application Protocol with Solutions to Common Problems and a Peptide Aggregation and Self-Assembly Example. *Methods in molecular biology (Clifton, N.J.)*, 1777:101–119, 2018. ISSN 1064-3745. doi: 10.1007/978-1-4939-7811-3_5. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6484850/.

Hannah K. Wayment-Steele, Adedolapo Ojoawo, Renee Otten, Julia M. Apitz, Warintra Pitsawong, Marc Hömberger, Sergey Ovchinnikov, Lucy Colwell, and Dorothee Kern. Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature*, 625(7996):832–839, January 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06832-9. URL https://www.nature.com/articles/s41586-023-06832-9. Publisher: Nature Publishing Group.