Improving dynamics-informed protein diffusion model

An MPhil Project Proposal

Vladimir Radenkovic, King's College
Supervisor: Mateja Jamnik
Co-supervisors: Julia Komorowska, Chaitanya Joshi

Abstract

The emergence of diffusion probabilistic models has provided a new pathway for protein design. These models have been successful in creating proteins with specific desired characteristics, such as certain symmetries or substructures. Previous work has shown that it's possible to condition the generative models on protein backbone dynamics, which has led to proteins with improved biochemical properties. This work aims to enhance dynamics-conditioned diffusion models by using Molecular Dynamics Simulations to obtain more detailed and accurate depiction of protein motions. By integrating modern AI tools with principles of protein physics, the goal of this project is to accelerate the design of proteins.

Introduction

Traditional Protein design methods are computationally demanding and require specialized knowledge of protein structures and interactions [1]. The advent of generative AI, particularly diffusion probabilistic models, has revolutionized this field [2] [3]. These models have shown great success in generating novel proteins conditioned on symmetries or substructures [4]. Recent work has shown that conditioning diffusion models on protein dynamics can lead to proteins with improved biochemical properties [5] [6]. In [5] authors leverage Normal Mode Analysis on coarse-grained representations of proteins to capture dynamical properties. Molecular Dynamics simulation (MDs) is widely used method to track atoms' movements over time, providing detailed and dynamic picture of protein's behaviour.

This project aims to enhance the first proposed dynamics-conditioned diffusion models with more accurate atomic motion depiction from MDs. The only previous work of this kind, [5], conditions the diffusion model on dynamics of protein backbone structure and uses less accurate estimation of the protein motion. We hypothesise that conditioning diffusion models on more accurate MDs data will enable reliable generation of rich protein structures with desired dynamic motion properties and more detailed representation of each residue. We will use atomic representation of the proteins, extract principal components of motion from the MDs [7] and condition the diffusion models directly on them with loss guided diffusion [8].We will evaluate our model by comparing MD trajectories of the generated structure with the target trajectories from the test set.

In the beginning we plan to evaluate this concept on three short proteins from [9]. We plan to demonstrate the concept for single protein and move to conditioning the model on motion information from all the three proteins. We plan to gradually move to harder tasks that will require strong generalisation by leveraging our model on datasets such as GPCRdb [10] and Misato [11], with multiple or longer proteins. One of the identified risks is that due to the lack of similarity in dynamic motion properties of proteins, the diffusion model may not be able to learn structure with desired dynamics and may fail to generalize on similar structures. To control this risk, we plan to implement our model on individual sequences from [9] and longer protein sequences. Once we prove our concept on such controlled data, we aim to tackle harder task of generalisation on different sequences. Additionally, we need to fin [5]d datasets with published MDs trajectories.

The goal of this project is to provide one of the first examples of usage of dynamic information for protein generation and pave the path for future research in dynamic-conditioned diffusion models for protein design.

We provide the success criterions for this project, starting from the minimum for success:

1. Provide proof of concept on individual proteins from [9].
2. Develop model which generalizes on all three proteins from [9].
3. Develop model which generalizes on longer sequences or hundreds of proteins from Misato or GPCRdb datasets.
4. Prove that our model generates structures with better dynamic properties than models in [5] and models without dynamic information [8].

Workplan

- 17th Nov – 24th Nov: Literature review. Develop better understanding of MDs and diffusion models for protein design.
- 24th Nov – 1st Dec: Continue studying the literature. Make appropriate choices of the unconditioned diffusion model and datasets. Detailed examination of the dataset and MD trajectories. Decide for small controllable data to start with.
- 1st Dec – 8th Dec: Setting up the workstation and computational resources. Implementation of model like Genie to develop better understanding of diffusion models for proteins.
- 8th Dec – 15th Dec: Implementation of model and data pipeline for experiments on three proteins from [9]. Implementation of diffusion model on single proteins of data: Conditioning diffusion model on eigenvectors from normal modes to gain experience with working with simpler model.
- 15th Dec – 1st Jan: Develop understanding of atomic representation of proteins. Creation of model and data pipeline for the experiments. Start with initial experiments on our model on single proteins.
- 1st Jan- 7th Jan: Christmas Holiday
- 7th Jan- 15th Jan: Continue with experiments. Identify bottlenecks, current limitations and potential improvements of the model.
- 15th Jan – 22nd January: Improve the model, continue with the experiments.
- 22nd Jan – 29th January: Analyse the results, run additional experiments to get meaningful insights and more information. Revaluate initial hypothesises from the observations and analysis of the results. Add additional hypotheses if necessary.
- 29th January – 5th February: Set up and start with experiments with all proteins from the [9]
- 5th February – 12th February: Identify bottlenecks, current limitations and potential improvements of the model, and continue the experiments.

- 12th February – 19th February: Analyse the results, run additional experiments or re-run previous experiments to get meaningful insights and more information. Revaluate initial hypothesises from the observations and analysis of the results. Add additional hypotheses if necessary.
- 19th February – 24th February: Project progress review: evaluate experiments and results to make a cross-section of previously done work to identify potential problems and unfinished tasks. Make conclusion about the work done so far and re-evaluate project goals and expectations.
- 24th February – 2nd March: Decide on dataset with longer or multiple proteins, acquire the dataset and implement the data and model pipeline.
- 2nd March – 9th March: Run initial experiments on subset of data to gain first insights about the method and the data.
- 9th March – 16th March: Make needed changes in the model, data and the experiments and rerun the experiments.
- 16th March – 23rd March: Analyse the results, run additional experiments or to get meaningful insights and more information. Revaluate initial hypothesises from the observations and analysis of the results. Add additional hypotheses if necessary.
- 23rd March – 30th March: Project progress review: evaluate all experiments and results to make a cross-section of previously done work to identify potential problems and unfinished tasks. Make conclusion about the work done so far and re-evaluate project goals and expectations.
- 30th March – 6th April: Start with writing of the first draft.
- 13th April – 4th May: Three weeks for contingencies.
- 4th May – 1st June: Four weeks for writing the dissertation.
- 1st June – 3rd June: Submission of the project dissertation.

References

[1] M. Suárez and A. Jaramillo, “Challenges in the computational design of proteins,” *Journal of the Royal Society Interface,* 2009.

[2] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan and S. Ganguli, “Deep Unsupervised Learning using Nonequilibrium Thermodynamics,” in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, 2015.

[3] J. Ho, A. Jain and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada., 2020.

[4] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky and N. Hanikel, “Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models,” Europe PMC, 2022.

[5] “Dynamics-Informed Protein Design with Structure Conditioning,” in *ICLR 2024 Conference Submission*.

[6] “DiffSim: Aligning Diffusion Model and Molecular Dynamics Simulation for Accurate Blind Docking,” in *ICLR 2024 Conference Submission*.

[7] RodrigoCossio-Peŕez, JulianaPalma and GustavoPierdominici-Sottile, “Consistent Principal Component Modes from Molecular Dynamics Simulations of Proteins,” *Journal of Chemical Information and Modeling,* 2017.

[8] J. Song, Q. Zhang, H. Yin, M. Mardani, M.-Y. Liu, J. Kautz, Y. Chen and A. Vahdat, “Loss-guided diffusion models for plug-and-play controllable generation,” in *Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pp. 32483–32498. PMLR*, 2023.

[9] M. Arts, V. G. Satorras, C.-W. Huang, D. Z.̈gner, M. Federici, C. Clementi, F. Noe, R. Pinsler and R. v. d. Berg, “Two for One: Diffusion Models and Force Fields for Coarse-Grained Molecular Dynamics,” *Journal of Chemical Theory and Computation,* no. 19, pp. 6151-6159, 2023.

[10] C. Munk, V. Isberg, S. Mordalski, K. Harpsøe, K. Rataj, A. S. Hauser, P. Kolb, A. J. Bojarski, G. Vriend and D. E. Gloriam, “GPCRdb: the G protein-coupled receptor database – an introduction,” *British Journal of Pharmacology,* vol. 173, no. 14, pp. 2195-2207, 2016.

[11] T. Siebenmorgen, F. Menezes, S. Benassou, E. Merdivan, S. Kesselheim, M. Piraud, F. J. Theis, M. Sattler and G. M. Popowicz, “MISATO - Machine learning dataset for structure-based drug discovery,” 2023. [Online].

[12] Y. Lin and M. AlQuraishi, “Generating Novel, Designable, and Diverse Protein Structures by Equivariantly Diffusing Oriented Residue Clouds,” in *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, Hawaii, USA, 2023.