GRADUATION PROJECT for 2022/2023

The end of molecular dynamics era: Protein-Specific Conformation Generation Using Deep Learning

Undersupervision of:

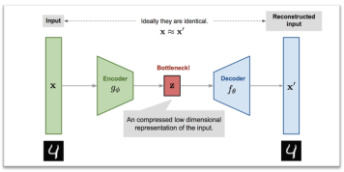
Dr. Tamer Mohamed Ibrahim

Abstract

Proteins are flexible molecules with dynamic states called conformations. Computational methods like molecular dynamics help define conformational space. This project uses deep learning to predict conformations, replacing MD simulations with generative neural networks. These networks minimize input-output differences, demonstrating their potential in studying molecular conformational space.

Methodology

Deep Learning Model: The methodology that has been used to carry out this project is based on building the deep learning models.



Dataset: The dataset we used to train the model was obtained from the MD Simulations for the four proteins: 2QKE,1Fq9,FGF2and 2P23

Validation

Cross validation: compare RMSD between the input and reconstructed output to generate TSV file for each pair and concatenate them into one TSV file.

Diversity : compare RMSD between the inputs to generate TSV file for each pair and concatenate them into one TSV files.The Same with reconstructed output files.

DSSP: We implemented a code with python that read the secondary structure assignments from two DSSP files, one for input and one for output, converted them to NumPy array, and finally calculated the Q3 score, which is a measure of the accuracy of the predicted secondary structure assignments.

Sampling: We used sampling interpolation function to sample from the latent space, Then decode the output vectors to generate those new conformations.

Conclusion

1. Goal: Improve sampling of molecular conformational space and replace molecular dynamics simulation.
2. Use generative neural networks to improve sampling of molecular conformational space.
3. Autoencoder trained on protein MD simulation-generated conformations.
4. Importance of Substantial Sampling: Accurate results required to justify experimental data or forecast outcomes before trials.
5. Use of Generative Neural Networks: Generate new, realistic protein conformations complementing existing ones
6. Application: Discover collective variables for extracting kinetic information or directing selection of underexplored regions

Results

MD simulation:

Trajectories Centring: To take frames every 100ps correctly, we needed to make control over atoms, so it won't go free, It is an important step, to take frames for the Deep learning training part.

Extracting Backbones: Extracting a new PDB file but whole frames or whole pdb file will not be used, Pdbs will be extracted that contains a new group containing only backbone (tertiary structure) **(CA, C, N) + CB.**

Taking frames every 100ps: skipping here is good and without it, this will extract billions of pdbs and total simulation will be thousands of ps and skip 100 ps it is nothing , Clustering or took representative it will affect our data and will not be valid for training our model "data will be decreased".

Autoencoder Model:

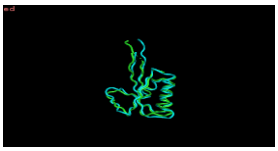
We needed to Convert timeframes coordinates to NumPy array format so that, model could deal with it. For the first step on this model, we extracted number of atoms and its conformations for each protein, the conformation of a protein refers to its 3D shape and arrangement of atoms.

Visualize all array: We needed to visualize all arrays to make sure that all trajectory files entered correctly, and it results each proteins' structure shape.

Dataset and Dataloader: A data loader in PyTorch is a utility that helps to load data in batches for efficient processing during training our model, the results of a data loader were batches of data that can be fed into a machine learning model.

Training Data: Validation

Reconstruction after training for example 2QKE:



Then we checked the following:

- **Loss function.**
- **Cross validation and concatenation.**
- **Diversity, Ramachandran plot, DSSP and Sampling.**

Team Members

Esraa Abdallah Maram Reda Elzahraa Saeed Rana Rizk Omar Mohammed