Accelerating Drug Discovery through Active Learning-enhanced Virtual Screening

Data Science Institute
Columbia University

Yan Gong, Siyu Li, Shanzhao Qiao, Shuqing Shan, Zixiang Tang
Mentor: Pinyi Lu, Ph.D.

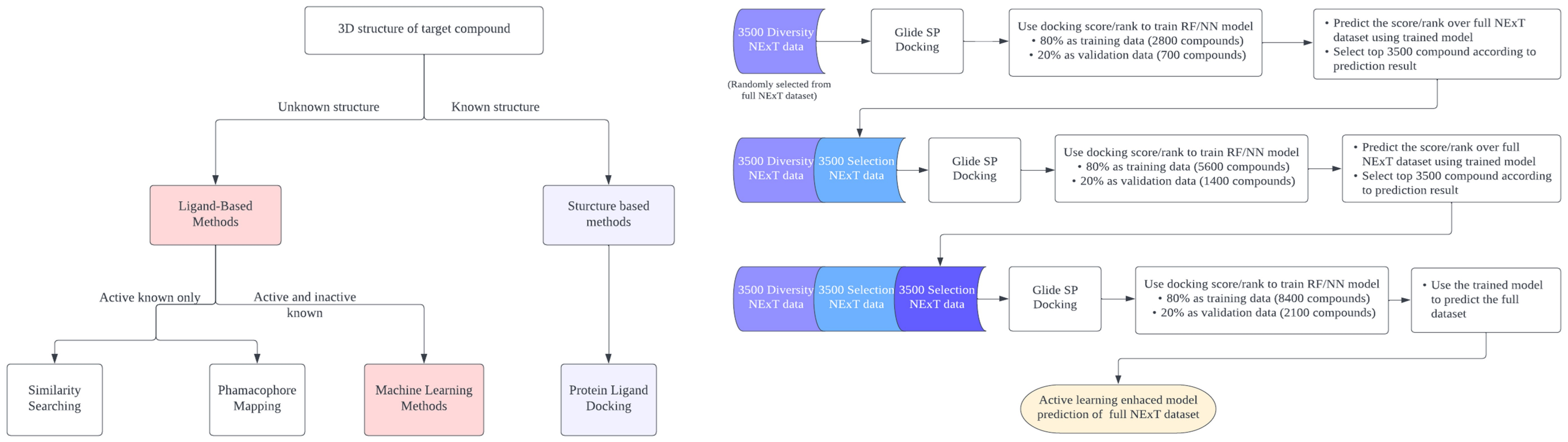
Data Science Capstone Project
with NIH

Frederick National Laboratory

NIH National Institutes of Health

Background: Drug Development & Computer-aid Drug Discovery

The drug development process is divided into four phases in a chronological order: drug discovery, preclinical studies, clinical trials, and FDA review. The first phase, drug discovery, is a time-consuming and costly process which could take approximately 6.5 years in tradition. Our project develops a machine learning pipeline in drug discovery process to reduce cost both timely and financially.



3D structure of target compound

Unknown structure

Known structure

Ligand-Based Methods

Sturcture based methods

Active known only

Active and inactive known

Similarity Searching

Phamacophore Mapping

Machine Learning Methods

Protein Ligand Docking

3500 Diversity NExT data

(Randomly selected from full NExT dataset)

Glide SP Docking

Use docking score/rank to train RF/NN model
- 80% as training data (2800 compounds)
- 20% as validation data (700 compounds)

- Predict the score/rank over full NExT dataset using trained model
- Select top 3500 compound according to prediction result

3500 Diversity NExT data

3500 Selection NExT data

Glide SP Docking

Use docking score/rank to train RF/NN model
- 80% as training data (5600 compounds)
- 20% as validation data (1400 compounds)

- Predict the score/rank over full NExT dataset using trained model
- Select top 3500 compound according to prediction result

3500 Diversity NExT data

3500 Selection NExT data

3500 Selection NExT data

Glide SP Docking

Use docking score/rank to train RF/NN model
- 80% as training data (8400 compounds)
- 20% as validation data (2100 compounds)

- Use the trained model to predict the full dataset

Active learning enhaced model prediction of full NExT dataset

Figure 1. Traditional computer-aid pipeline and active learning-enhanced pipeline

Methods: Active Learning-enhanced Virtual Screening

Virtual screening, a computational technique used in drug discovery to search libraries of small molecules, could identify those structures which are most likely to bind to a drug target. ATOM Modeling Pipeline (AMPL) is an end-to-end modular and extensible software pipeline for building and sharing machine learning models that predict key pharma-relevant parameters, with complete traceability and reproducibility of the model building and evaluation process. We combine the two: implement active learning of three rounds utilizing AMPL, each time docking with virtual screening to develop a better ensembled model. Google Colab and GCP were implemented for the project. The dataset used is the NExT Libraries, a general screening sets that were designed to identify small molecule lead compounds for HTS drug discovery projects.

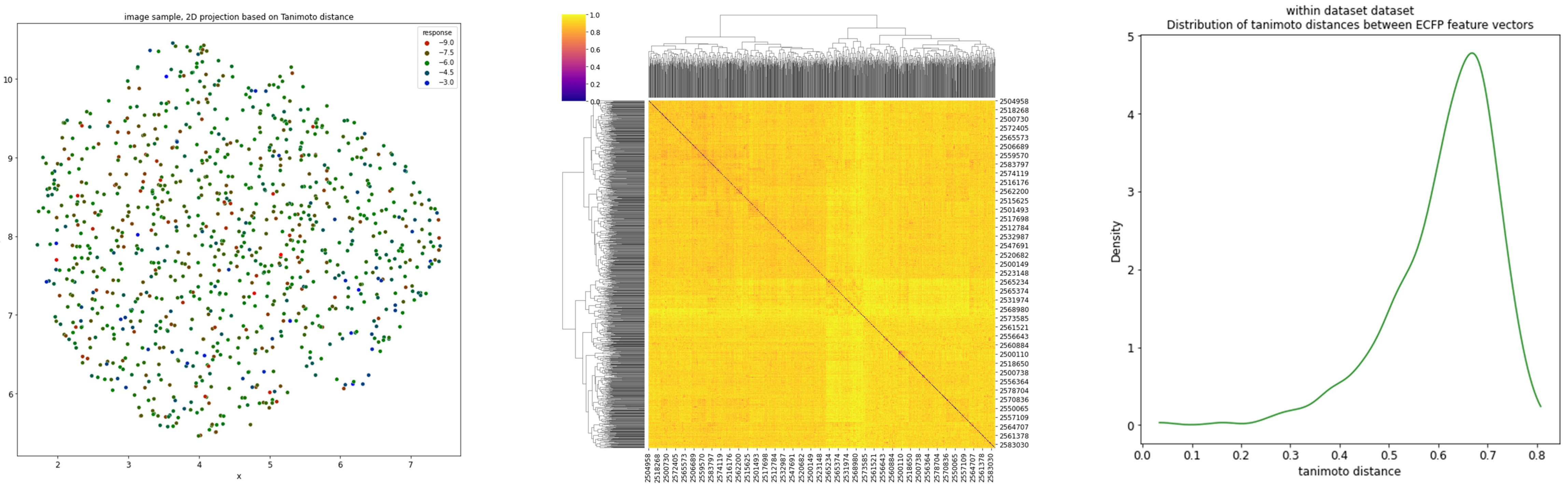


Figure 2. Diversity plot and self-similarity plot of NExT dataset from EDA.

Results

Random forest model and neural network were implemented, while two target variables including the docking score and ranking of compounds were predicted, so that four methods in total were performed and compared with each other. The difference between machine learning methods is not significant, while that between target variables is significant.

The best result is produced by predicting the docking score with random forest model, where among the top 20% of the compounds in NExT dataset, 45.9% of those compounds were predicted correctly by our model.

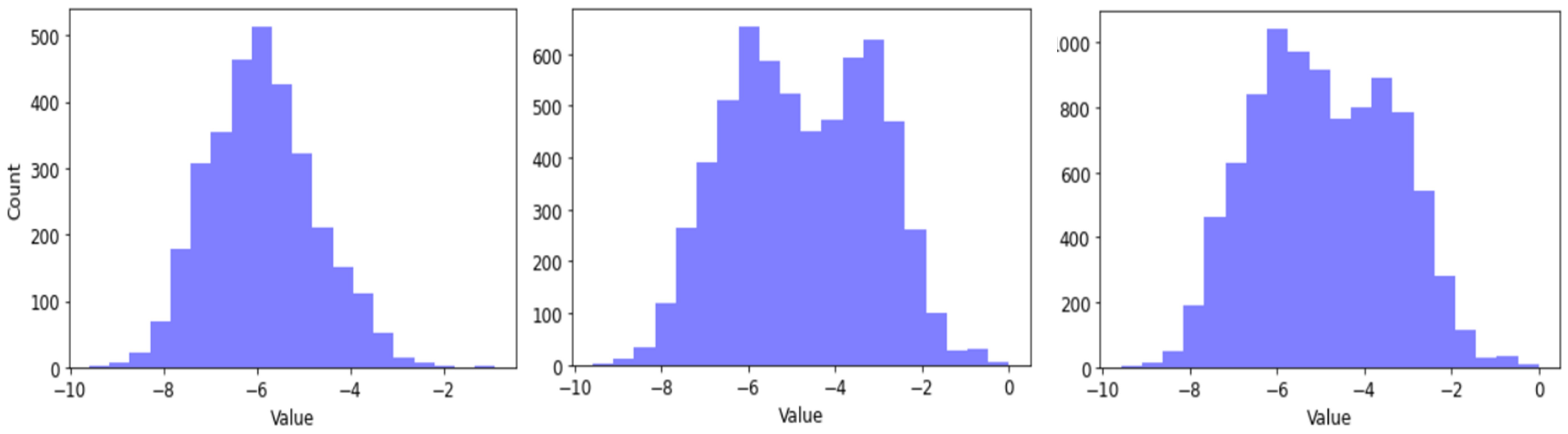


Figure 3. Distribution of predicted docking scores from three rounds of active learning. Round 1 to round 3 from the left to the right.

Conclusion

A direct prediction on docking score rendered a better result than predicting on ranking of compounds. Through the implementation of active learning with increasing number of rounds, the model performance became better with a decreasing speed.

Acknowledgments

This project was supported by The Accelerating Therapeutics for Opportunities in Medicine (ATOM) Consortium and Frederick National Laboratory for Cancer Research (FNLCR).

References

Minnich, Amanda J., et al. “AMPL: A Data-Driven Modeling Pipeline for Drug Discovery.” Journal of Chemical Information and Modeling, vol. 60, no. 4, American Chemical Society (ACS), Apr. 2020, pp. 1955–68. https://doi.org/10.1021/acs.jcim.9b01053.

Li, J., Fu, A. & Zhang, L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. Interdiscip Sci Comput Life Sci 11, 320–328 (2019). https://doi.org/10.1007/s12539-019-00327-w

Pagadala, N. S., Syed, K., & Tuszynski, J. (2017). Software for molecular docking: a review. Biophysical reviews, 9(2), 91–102. https://doi.org/10.1007/s12551-016-0247-1