ML & DL Techniques for CPI & Binding Affinity

Roziena Badree and Mengriu Mao

Hunter College at the City University of New York

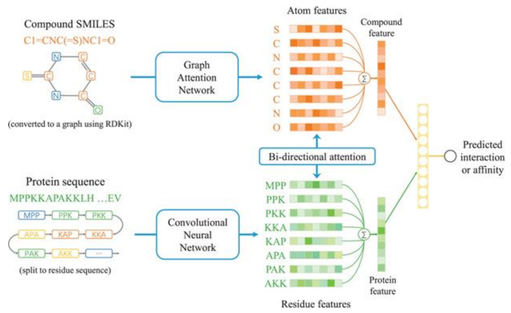
Introduction

The identification of compound–protein interactions (CPIs) is a crucial step in drug discovery. Since the experimental determination of CPIs is expensive and time-consuming, the computational model provides a promising and efficient alternative.

CPI prediction is a binary classification task, which determines if there is a molecular interaction between a compound-protein pair while binding affinity prediction is a regression task, which focuses on predicting a continuous value that represents the binding strength between a compound-protein pair.

We explored two state-of-the-art techniques for predicting CPI and binding affinity: a bi-directional attention neural network and a language model named ESM-2 developed by Meta.

BACPI Model



The BACPI portion of our project uses features obtained by a GAT and fingerprinting representation of compounds and a CNN for representations of proteins. For the compounds, we use RDKit to convert each SMILES format into a graph representation and employ the GAT to extract information (such as atom types, aromaticity and chemical bond types) from the graph. For the proteins, our CNN takes each protein's AA sequence and learns the feature representation of the protein. Finally, an attention-based bi-directional neural network is used to integrate the representations of compounds and proteins and predict the interaction and binding affinity of the input compound–protein pair.

ESM-2

An ESM-2 model was created for each protein in a compound-protein pair in our datasets. They were trained − at scales of 8 million to 15 billion parameters − to predict the identity of randomly selected amino acids in a protein sequence by observing their context in the rest of the sequence causing the model to learn the dependencies between the amino acids. In other words, the model randomly removes amino acids from the input and the result is fed into a BERT-style encoder only transformer and via self-supervised learning, the model learned how to predict the missing amino acids. The output is the three-dimensional coordinates and confidences of the binding sites of the protein.

For clarity, the proteins are pre-trained with an ESM-2 model and the compounds are fingerprinted using RDKit. Then, these results are passed into various supervised and unsupervised machine learning models to predict the CPI or binding affinity for a compound-protein pair.

Results

C. elegans		Human	
Model Name	RMSE	Model Name	RMSE
Extra Trees	0.189474	Extra Trees	0.228321
Random Forest	0.211261	Random Forest	0.242715
Gradient Boosted	0.273941	SVM	0.291737
SVM	0.297722	Gradient Boosted	0.307586
Logistic Regression	0.322886	Logistic Regression	0.339053
KNN	0.360259953	Linear Regression	0.363745
ElasticNet Regression	0.372044	ElasticNet Regression	0.372946
Multi-Layer Perceptron	0.407379	KNN	0.385137992
GaussianNB	0.409982	Perceptron	0.408501
Local Outlier Factor	0.50318407	Multi-Layer Perceptron	0.408501
MultinomialNB	0.518796	MultinomialNB	0.455701
Perceptron	0.705601	Local Outlier Factor	0.496150446
Kmeans	0.777395731	GaussianNB	0.777464
Linear Regression	1778409083	Kmeans	0.834972822

Model Name	IC50	EC50	Ki	Kd
BACPI	0.74	0.78	0.8	1.08
Random Forest	0.792437	-	0.989753	1.184066
MLP Regressor	0.917291	-	1.083637	1.259279
Ridge Linear Regression	0.949984	-	3.034943	2.285426
Lasso Linear Regression	1.013284	-	1.422459	1.367706
Support Vector Machine (Gaussian RBF)	1.186531	-	1.198874	1.421241

We were unable to gather results for the CPI task for the BACPI model and therefore cannot compare the BACPI and ESM-2 models for the CPI task. For the CPI prediction − of the ESM-2 models, the Extra Trees model performed the best for the C. elegans and human datasets. For the binding affinity prediction, BACPI performed the best across the datasets for which we have data.

Discussion

For CPI prediction, the Extra Trees model be the best model (so far, until we have BACPI results) because the algorithm creates many unpruned decision trees from the dataset. In the case of classification, predictions are made using majority voting from the decision trees.

For the binding affinity prediction, BACPI performed the best across the datasets for which we have data. The only other model that came close is the ESM Random Forest model. Studies have shown that decision trees like the Random Forest model works well for tabular and structured data (vision and audition) with small sample sizes while neural networks perform better on structured data with larger sample sizes.

References

1. Ferruz, Noelia & Hocker, Birte. (2022). Towards Controllable Protein design with Conditional Transformers.

2. Karimi, M., Wu, D., Wang, Z., & Shen, Y. (2019). DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. Bioinformatics (Oxford, England), 35(18), 3329–3338.

3. Min Li, Zhangli Lu, Yifan Wu, and YaoHang Li. BACPI: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction. Bioinformatics, 38(7), March 2022, pp. 1995–2002.

4. Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, et al. "Evolutionary-Scale Prediction of Atomic Level Protein Structure with a Language Model." bioRxiv, 2022. .

5. Toutain, P. L.; Bousquet-Melou, A. (2002-12-14). Free Drug Fraction vs. Free Drug Concentration: A Matter of Frequent Confusion. Journal of Veterinary Pharmacology and Therapeutics. Wiley inc. 25 (6): 460–463.

6. Whitford, David. 2013. Proteins: Structure and Function. J. Wiley & Sons.