DeepLPI: A Novel Drug Repurposing Model based on Ligand-Protein Interaction Using Deep Learning

Bomin Wei, Princeton International School of Mathematics and Science, Princeton, NJ, 08540

Background

COVID-19 drug development faces long R&D time and low success rate.

- **Drug repurposing** finds effective cures from existing drugs to lower drug R&D time and cost.
- **Protein-Ligand Interaction (PLI) or drug-target interaction (DTI)** is essential for drug repurposing. It indicates whether a candidate drug can bind to a target protein and thus inhibit its function to cure the disease.
- **Computational-based methods** have been developed to predict DTI and to reduce the size of drug pool and speed up drug discovery.

Problems with current methods

- Some models have low accuracy because they select features based on expert knowledge of the target protein, which loses key information.
- Others have limited data because they reply on 3D structure input that are hard to obtain.
- No detailed analysis on the generalization ability to unseen drugs/targets.

Objective

This project builds a deep neural network-based model to predict PLI and verify it on COVID-19 application.

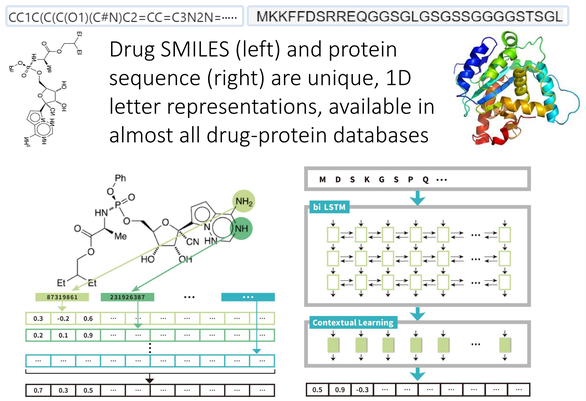
Highlights

- Using **NLP-inspired** embedding methods to treat 1-D drug molecular and protein sequence input for higher accuracy.
- Using a new model architecture that combined CNN and LSTM to capture local and global information together. The LSTM module gives a better connection between the molecular and protein sequences.
- To test the model's **zero-shot** performance **on realistic test dataset**, where the drug or target protein from a new disease are likely not appeared in the training dataset.
- **Repurpose drugs for COVID-19** by deploying the trained model on COVID-19 feature proteins.

Method

A new architecture is proposed, including two major innovations:

1. Use simple, widely available 1D drug/protein sequence as input



Drug SMILES (left) and protein sequence (right) are unique, 1D letter representations, available in almost all drug-protein databases

Fig 1. Principle of NLP-inspired deep learning-based methods to embed drug SMILES strings (Mol2vec) and protein sequences (ProSE).

2. Use a composite architecture employing CNN and LSTM to extract features locally and globally together.

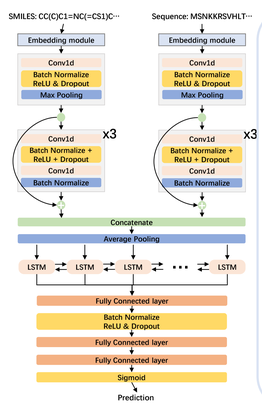


Fig 2. Architecture of DeepLPI model. The model uses raw strings of molecular SMILES and protein sequences as inputs. The embedded vectors for the drug SMILES and the protein sequences are then fed into the respective head module and ResNet-based CNN module to extract features, which were concatenated, pooled (max-pooling operation), encoded by a bi-LSTM layer, and finally fed into an MLP module. The final output is passed through a sigmoid function for binary classification to predict binding/non-binding labels.

Data

BindingDB: high quality data after cleaning the consensus database.

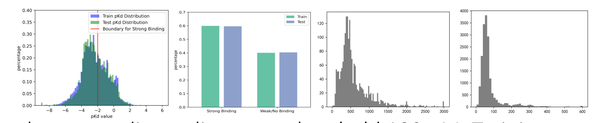


Figure 2. Kd-labeled data were used and converted to binary values according to literature threshold 100 nM. Training data was in balance. **BindingDB** is a commonly-used benchmark database for drug-protein studies with **about 2.4 million** entries of drug-protein interaction data in total.

Result: Validation

Our model reaches high accuracies on realistic, unseen drugs & proteins.

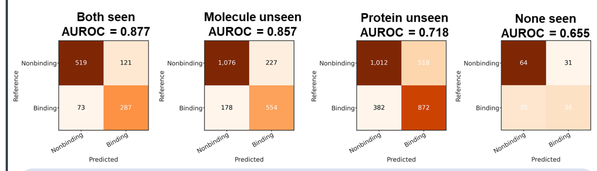


Figure 3. A randomly reserved part of the BindingDB dataset was grouped into four cases for testing, according to whether the drug molecules or the proteins are used in the training set. (Note that the drug-target pairs in the test set are excluded from the training set.)

Our model is better than the baseline method by 76%.

We compare with DeepCDA, a recently published DTI prediction model that reported to be better than previous models. DeepLPI is 76% better in accuracy metric AUROC in classification tasks.

BindingDB	AUROC	Sensitivity	Specificity	PPV	NPV
DeepLPI	0.790	0.684	0.773	0.671	0.783
DeepCDA	0.448	0.000	1.0	N/A	0.596

Test on external data (Davis dataset) proves generalizability.

The BindingDB-trained model is applied to experiment Davis data, another common benchmark in studies of drug-protein interactions. AUROC score of our model reached 0.53, 25% better than DeepCDA.

Application

Optimize model for drug developers to repurpose drugs for new targets.

The model is also optimized on doing the regression and ranking using common binding affinities (Kd, Ki, etc.), and the current Kd test showed a better performance. The regression results cannot directly show disease-curing ability of drugs because published models do not use a consensus binding affinity. Therefore, classification is recommended.

	R	$RMSE$	MAE	R^2	MSE
DeepPLI (test)	0.84	**0.80**	**0.60**	0.71	**0.64**
DeepCDA	0.84	N/A	N/A	N/A	0.808

Verification on COVID-19 data

As a preliminary attempt, model trained on BindingDB was applied to a recent COVID-19 dataset targeting the 3CL-protease, which reported 897 small molecule drugs. Our model is better than the baseline method by 25%.

	AUROC	Sens.	Spec.	PPV	NPV
DeepLPI	**0.61**	0.538	0.576	0.110	0.928
DeepCDA	**0.40**	0.0	1.0	N/A	0.911

Future Work

Pick top COVID-19 drug candidates

We will use our pretrained model to predict interactions between all types of FDA approved drugs and the 3CL protease of SARS-COV-2. Top candidates with strongest interactions will be suggested for wet lab verifications.

Conclusion

- Proposed a new model architecture for predicting drug-target interaction
- Used NLP-inspired embedding methods for higher prediction accuracy.
- Performance on the benchmark datasets are better than baseline methods
- Trained model can reach good performance on benchmark external dataset.
- Trained model performed better on COVID-19 data compare to baseline method
- Has the generalizability to apply to all diseases to speed up drug discovery.

Acknowledgement

Thanks to Dr. Xiang Gong from PRISMS for his guidance in statistics, computational modeling, and on scientific methodologies.
Thanks to Prof. Yue Zhang from University of Utah for his help on project design, on understanding drugs and proteins and on applying machine learning.