Drug-Target Interaction Prediction Proposal Using Deep Learning

1st Ahmad Shahbaz
department of systems and computer engineering
Carleton University
Ottawa, Canada
Ahmad.Shahbaz@carleton.ca

2nd Arushan Sinnadurai
department of systems and computer engineering
Carleton University
Ottawa, Canada
Arushan.Sinnadurai@carleton.ca

***Abstract*—Drug-Target Interaction is crucial for the purposes of drug discovery. Using deep learning drug repurposing can be preformed, which essentially means checking if an existing drug can be used to treat other diseases. Using the DeepPurpose framework two parallel convolutional neural networks can be developed, which will take in a encoded SMILE string as well as the protein sequence in order to screen a potential drug that targets the proteins of the coronavirus.**

I. Project Overview

A. Drug-Target Interaction (DTI) Prediction

The Drug-Target interaction (DTI) prediction is a crucial technique that is used for drug discovery and drug repurposing [1]. DTI prediction measures the drug's binding affinity to the protein target using equilibrium dissociation constant Kd. The Kd is ranked by the strength of the interaction between the protein and drug in which the lower the Kd, the higher the binding affinity of the drug for the specific protein target and vice versa [2]. The binding affinity of protein-drug interaction can be described by the "Lock and Key" theory [3]. The geometric shape of the protein act like a "lock" that requires a specific geometric shape or "key" from the drug to interact with the protein.

B. How DTI can be Leveraged for COVID-19

Drugs and proteins are made of chemical bonds that can be characterized in sequences such as simplified molecular-input line-entry system (SMILES) and amino acids [4]. For example, COVID-19 proteins and drugs are translated into amino acid and SMILES can be used as an input feature for DTI machine learning algorithm. In addition to the sequence, Kd can be incorporated into a machine learning algorithm such as a convolutional neural network to predict a drug's bind affinity. Moreover, the DTI machine learning algorithm can be used for finding a drug that can be repurposed for COVID-19 proteins.

C. Examples of DTI Methods

DTI can be implemented by a machine learning algorithm in many ways such as AutoDNP, DeepWalk, or DeepDTI [5]. AutoDNP is an ensemble classifier that uses an autoencoder for the input feature to produces a prediction for drug sequence. DeepWalk is a deep learning algorithm that predicts the DTI by several topologies of drugs and proteins. DeepDTI is another deep learning method that uses an encoder-decoder convolutional neural network algorithm to predict the binding affinity of proteins and drugs. These are a few examples of DTI machine learning methods are being used in the real world for drug repurposing.

II. Proposed Data and Methods

A. Datasets

The datasets to be used for this project will be the DAVIS and BindingDB. The DAVIS dataset contains the wet lab assays of various proteins and the relevant inhibitors. Along with this their respective dissociation constant (Kd) values are also contained in the dataset. The DAVIS dataset has information for 68 drugs and 379 proteins. The BindingDB database contains experimentally determined binding affinities (Kd) for 10,665 drugs and 1,413 proteins. Using these datasets for the model will aid in the screening of potentially determining a drug that targets the proteins of the coronavirus (SARS-CoV-2).

B. Data Transformations and Feature Representation

Both the SMILE strings as well as the protein sequence will be encoded using the CNN representation as the model plan to be used will have two parallel CNN blocks for the SMILE string and the protein sequence.

C. Model Architecture and Model Improvements

The DeepPurpose framework, which is a deep learning toolkit will be used for the purposes of drug repurposing. The model architecture planned to be used is a deep learning convolutional neural network (CNN). CNNs contain multiple convolutional layers usually followed by a pooling layer. A pooling layer progressively reduces the spatial size of the representation in order to reduce the amount of parameters and computation in the network [6]. This is done by down-sampling the output of the previous layer [6]. A fully connected (FC) layer will be used which will encompass the CNN and the pooling layers. The purpose of this is to take the results of the convolution and pooling process and use them to perform classification. One major benefit of using a CNN model is it can capture the local dependencies with the help of filters, meaning the size and number of filters used is very important as it determines the type of features the model learns from the input [7].

In order to design the model two separate CNN blocks will be used. Both the SMILE strings as well as the protein sequence will be encoded and then be passed into two CNN blocks. Each block will contain three layers and will be fed into a max pooling layer. The filters for each layer will need to be tuned based on experimentation. The results from the max pooling layers of the two CNN blocks will concatenated and fed into a FC block. The FC block will also have three layers. The nodes for each layer will also be tuned through experimentation. In between each layer for the FC block there will be a dropout layer which will help promote model generalization [7]. The value for the dropout layer will be tuned through experimentation.

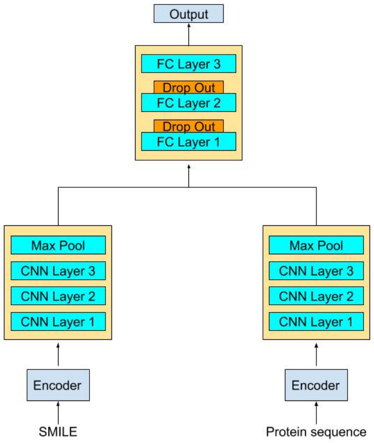

Fig. 1. Proposed Model Architecture

D. model Generalization Strategies

As mentioned previously adding a pooling layer promotes generalization as it reduces the amount of parameters and computation in the network by performing down sampling. Another strategy which promotes generalization is adding a dropout layer in between the FC layers. Dropout layers are regularization techniques that are used to reduce the affects over-fitting by setting the activation of some of the neurons to 0.

III. Expected Results

The model architecture is similar to the DeepDTI machine learning algorithm but with a few changes such as hyperparameter tuning and the addition of the dropout layer parameter. With these changes, the model is expected to perform better than the DeepDTI algorithm since the model will be more generalized and will have less chance to overfit to the input dataset. We expect the model to have a high area under the curve for receiver operating characteristic (ROC) will range around 90% and precision-recall curves around 50% on the testing dataset. This will ensure that model can accurately and confidently screen a potential drug that targets the proteins of the coronavirus.

References

[1] Z.-H. Chen, Z.-H. You, Z.-H. Guo, H.-C. Yi, G.-X. Luo, and Y.-B. Wang, “Prediction of drug–target interactions from multi-molecular network based on deep walk embedding model,” *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 338, 2020.

[2] S. Hunter and J. Cochran, “Cell-binding assays for determining the affinity of protein–protein interactions: technologies and considerations,” in *Methods in enzymology*. Elsevier, 2016, vol. 580, pp. 21–44.

[3] A. Tripathi and V. A. Bankaitis, “Molecular docking: From lock and key to combination lock,” *Journal of molecular medicine and clinical applications*, vol. 2, no. 1, 2017.

[4] N. R. Monteiro, B. Ribeiro, and J. P. Arrais, “Deep neural network architecture for drug-target interaction prediction,” in *International Conference on Artificial Neural Networks*. Springer, 2019, pp. 804–809.

[5] M. Bagherian, E. Sabeti, K. Wang, M. A. Sartor, Z. Nikolovska-Coleska, and K. Najarian, “Machine learning approaches and databases for prediction of drug–target interaction: a survey paper,” *Briefings in bioinformatics*, 2020.

[6] J. Brownlee, “A gentle introduction to pooling layers for convolutional neural networks,” *Machine Learning Mastery*, vol. 22, 2019.

[7] H. Öztürk, A. Özgür, and E. Ozkirimli, “Deepdta: deep drug–target binding affinity prediction,” *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.