Predicting Protein-Drug Binding Affinity with Graph Convolutional Networks and BERT

Soumyadeep Das
Department of Computer Science
Indian Institute of Technology(BHU), Varanasi
soumyadeep.das.cse21@itbhu.ac.in

Parth Kesarwani
Department of Computer Science
Indian Institute of Technology(BHU), Varanasi
parth.kesarwani.cse21@itbhu.ac.in

***Abstract*—The DeepDTA model is a deep neural network that has been shown to accurately predict the binding affinity between proteins and drugs. In this paper, we propose an extension to the original model by incorporating pre-trained language models and graph convolutional networks for learning the protein and drug representations, respectively. We evaluate our model on the KIBA and Davis datasets and show that it outperforms the original DeepDTA model. Our results demonstrate the effectiveness of using pre-trained language models and graph convolutional networks for learning representations of proteins and drugs for the prediction of binding affinity.**

***Index Terms*—Deep learning, drug discovery, protein-drug binding, graph convolutional network, BERT**

I. Introduction

Drug discovery is a complex and time-consuming process that involves identifying small molecules, or drugs, that can bind to specific target proteins. One of the key steps in the drug discovery process is predicting the binding affinity between a drug and a protein. As the field of drug discovery expands with the discovery of new drugs, repurposing of existing drugs and identification of novel interacting partners for approved drugs is also gaining interest. One of the challenges in developing deep learning models for this task is the complex interaction between small molecule drugs and protein targets. These interactions are influenced by a variety of factors, including the 3D structure of the protein, the physicochemical properties of the drug molecule, and the binding site on the protein surface. To capture this complexity, several deep learning models have been proposed, including DeepDTA [1], which uses a convolutional neural network (CNN) to process protein and drug molecules, which can effectively model the structural information of protein molecules and small molecule drugs. Accurate prediction of binding affinity can help identify potential drug candidates and reduce the need for expensive and time-consuming experimental assays. Until recently, DTI prediction was approached as a binary classification problem, neglecting an important piece of information about protein-ligand interactions, namely the binding affinity values.

Binding affinity provides information on the strength of the interaction between a drug-target (DT) pair and it is usually expressed in measure such as dissociation constant, inhibition constant or the half maximal inhibitory concentration

Recently, deep learning models have been applied to the prediction of protein-drug binding affinity, with promising results. The DeepDTA model [9] is a deep neural network that has been shown to outperform other state-of-the-art methods on the KIBA and Davis datasets. This model uses 2 CNN blocks and a regressor neural network to obtain the binding affinity value from the 1D representations of drugs and proteins.

In this paper, we propose an extension to the original DeepDTA model by incorporating pre-trained language models and graph convolutional networks for learning the protein and drug representations. Specifically, we use BERT to learn the node embeddings of the drug and protein from the 1D representations of protein and drugs. Then we feed it into an encoder which consists of Graph Convolutional Network layers which enables the learning of the representations of drugs and proteins separately. We evaluate our model on the KIBA and Davis datasets and show that it outperforms the original DeepDTA model.

II. Related Work

Several methods have been proposed for predicting drug-target binding affinity, including traditional machine learning methods [2, 3] and deep learning-based methods [1, 4, 5]. Among these methods, deep learning has shown superior performance in capturing complex relationships between drug molecules and target proteins [6]. The DeepDTA model proposed by Öztürk et al. [1] is a state-of-the-art deep learning-based method for predicting drug-target binding affinity. The model uses a CNN-based architecture to featurize drug molecules and protein sequences, which are combined with a fully connected layer to predict the binding affinity. Despite its strong performance, the model has limitations in featurizing drug molecules, which can affect the accuracy of binding affinity prediction. Several approaches have been proposed to improve the featurization of drug molecules. One popular approach is to use molecular fingerprints, which encode various structural features of molecules as binary vectors. These fingerprints can then be used as inputs to traditional machine learning models. However, the fixed length of fingerprints limits their ability to capture the full structural complexity of molecules. Another approach is to represent molecules as graphs and use graph neural networks to featurize them. Graph neural networks can capture the graph structure of molecules and learn node and edge representations based on their local connectivity. This approach has shown promis-

ing results in improving the accuracy of drug-target binding affinity prediction. In addition to featurizing drug molecules, several methods have been proposed to incorporate additional information, such as protein-protein interactions and chemical structures of related compounds, to improve the accuracy of binding affinity prediction

III. Dataset Used

We trained and evaluated our proposed model on KIBA dataset [13], which was previously used as a benchmark dataset for binding affinity prediction evaluation. The KIBA dataset contains selectivity assays of the kinase protein family and the relevant inhibitors with their respective KIBA scores [13]. It originally comprised 467 targets and 52 498 drugs. For the compounds of KIBA, the maximum length of a SMILES is 590, while the average length is equal to 58. The maximum length of a protein sequence is 4128 and the average length is 728 characters in the KIBA dataset.

IV. Methodology

We propose a novel transformer based approach for the prediction of drug-protein interaction using Graph Convolutional Networks. Here is a complete overview of the pipeline of our method.

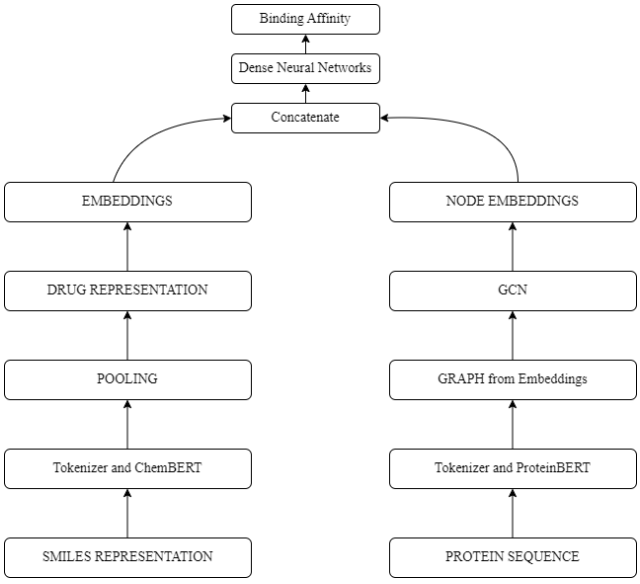

Fig. 1: Overview of our model

A. Generating the node embeddings from Protein Sequence

Firstly, the protein sequence is passed through the TAPE (Tasks Assessing Protein Embeddings) Protein Tokenizer [10], which converts the string sequence to a numeric sequence by mapping each protein literal to a number. We pass this tokenized sequence into ProteinBERT[11] to get the embeddings of size 768 for each token and we also get two extra tokens, i.e, CLS and SEP which we will be ignoring for now as they do not help in constructing graph.

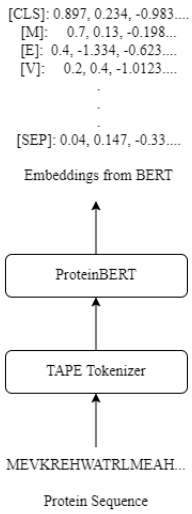

Fig. 2: Embeddings from ProteinBERT

Now each adjacent token in the sequence is connected to its immediate neighbors forming a graph with the BERT embeddings for each token as its node embeddings. This graph is then passed through a Graph Convolutional Network. The final embedding is then retrieved by taking Global Mean Pooling of the last graph layer.

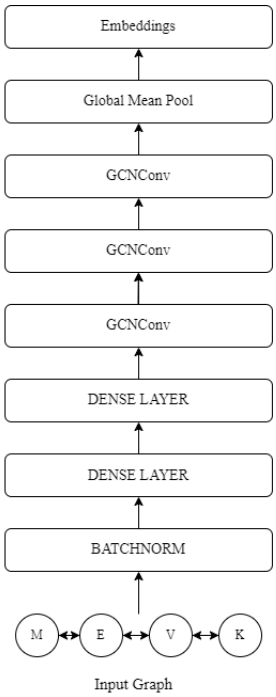

Fig. 3: GCN

B. Generating the node embeddings from SMILES Representation of drug

The preprocessing of drugs is similar to that of the proteins. Firstly, the SMILES representation of the drugs is passed into

the ChemBERTa-77MTR [12]. This BERT model produces embeddings of size 384. ChemBERTa is a pre-trained language model based on the BERT architecture that has been fine-tuned on chemical data, such as chemical structures and molecular properties, to generate embeddings for chemical compounds. Now the embeddings are converted into graph using the from_smiles() from the pytorch-geomtric library. The graph thus retrieved is then passed through a GCN, architecturally same as the one which was used in the protein embeddings, after Global Mean Pooling, we get the final drug embeddings.

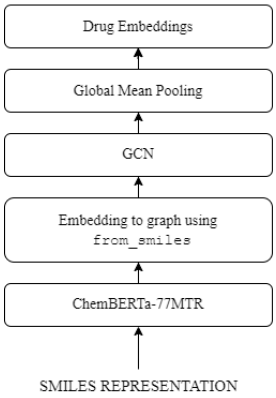

Fig. 4: Generating Embeddings from SMILES representation

C. Concatenating the embeddings and passing it to final regression network to predict the final binding affinity

After getting both the embeddings from the protein sequence and SMILES representations, we concatenate both the embeddings and pass it to a dense neural network. This neural network consists of 3 dense layers with 512, 256 and 128 hidden neurons. ReLU (rectified linear unit) was used as the activation function in the network. The final layer predicts the binding affinity. The entire model uses the Adam optimizer and is trained for 100 epochs, thereby fine-tuning the ChemBERTa and ProteinBERT to suit the KIBA dataset and learning the graph representations fromt the sequences and representations.

D. Training and Evaluation

We train and evaluate our model on the KIBA dataset, which comprises of 467 targets and 52,498 drugs [13]. The entire dataset is divided into training and test dataset in the ratio of 1:5. The test data is kept unseen to the model. We split the training set into 5-fold training-validation set in order to obtain a more reliable estimate of the performance of a model compared to using a single train-test split. In this 5-fold cross validation, in each epoch, the dataset is divided into 5 parts and one of them is chosen randomly as the validation set, keeping training set for the rest. This facilitates the performance is averaged across multiple validation sets, reducing the impact of random variability in the data. 5-fold cross validation makes it easier to compare the performance with that of DeepDTA [1] on the KIBA dataset [13]. The training and validation sets

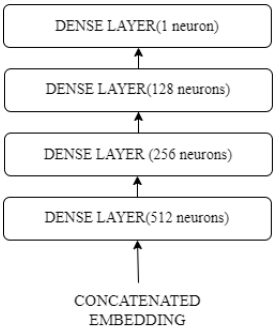

Fig. 5: Dense Neural Network

are selected randomly and repeatedly, reducing the risk of bias in the evaluation.

We used a learning rate of 0.001 and a batch size of 256 for training with Adam(adaptive moment estimation) optimizer [14].The metrics used for evaluating the model were Mean Squared Error(MSE) and Concordance Index (CI). In the results section, we will be comparing our model with that of DeepDTA [1].

V. Results

We evaluated the performance of state-of-the-art deep learning model, DeepDTA with our proposed model, for the task of predicting the binding affinity between molecule drugs and protein targets using the KIBA dataset. Both models were trained and tested using a 5-fold cross-validation setup, and the mean squared error (MSE) and concordance index (CI) was used as the evaluation metric.

The results showed that our model outperformed DeepDTA in terms of predicting the binding affinity. The mean MSE obtained by our model is 0.181, while that obtained by DeepDTA was 0.194. The concordance index of our model is 0.8712 which is greater than the concordance index of 0.863 reported in the DeepDTA's paper [1]. Table 1 shows the performance comparison between the two models.

TABLE I: Performance comparison of DeepDTA with our model

Model	MSE	CI
DeepDTA	0.194	0.863
Our Model	0.181	0.8712

This clearly shows that the use of GCNs along with BERT results in better performance. In our approach, construction of molecular graph from the protein sequence and SMILES embeddings helps the model to learn the similarities between the drugs and the proteins. Thereby it learns the interaction between the two more effectively. Moreover the use of pre-trained language model, ProteinBERT and ChemBERTa, helps to generate better node embeddings for each token of the sequence. The effective node embeddings thus facilitates

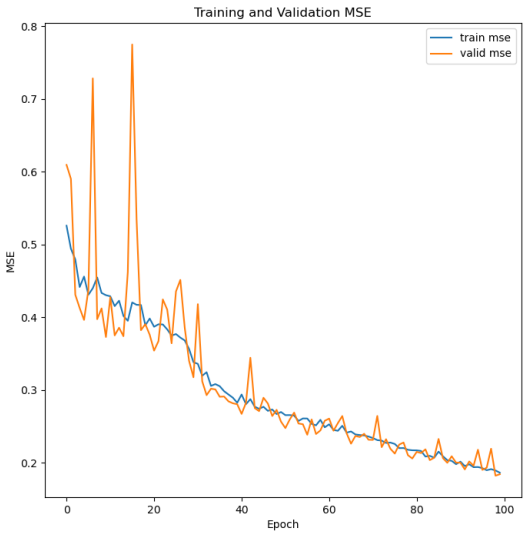

Fig. 6: Training and Validation MSE

the GCN to convolute better through the graph and pass the message from the neighboring nodes to other more effectively. Therefore, this novel approach learns by understanding the molecular graph of the drug and protein sequence.

Our model fine-tunes the pre-trained BERT model on the drug-target interaction task, which allows it to adapt the pre-trained embeddings to the specific characteristics of the KIBA dataset. This leads to improved performance compared to DeepDTA, which does not incorporate a pre-trained language model or fine-tuning.

`ProteinBERT` and `ChemBERTa` uses an attention mechanism to weight the contribution of different parts of the protein and drug molecules to the binding affinity prediction. This allows the model to focus on the most informative features of the molecules, which leads to improved predictions.

VI. Conclusion

In this paper, we proposed a novel approach to improve the accuracy of the DeepDTA model for predicting drug-target binding affinity. Our approaches leverages pretrained BERT models and Graph Convolutional Networks (GCN) to learn the representations of the drugs and proteins. Our proposed model performs better than the original DeepDTA model on the KIBA dataset. Specifically, our model reduced the Mean Squared Error (MSE) of the original model from 0.194 to 0.181.

The model which we have proposed has the ability to improve the accuracy of drug-target binding prediction, which can accelerate drug discovery and development process and make a real impact. The ability to accurately predict drug-target binding affinity can reduce the cost and time required for experimental drug screening and improve the success rate of drug development. In addition, our approach is applicable our approach is applicable to a wide range of drug molecules and target proteins, making it a promising method for future drug discovery research. Overall, our proposed approach provides a promising direction for improving the accuracy of drug-target binding prediction using deep learning methods

References

[1] Öztürk, H., Ozkirimli, E., & Gómez-Bombarelli, R. (2018). DeepDTA: deep drug–target binding affinity prediction. Bioinformatics, 34(17), i821-i829.

[2] Chen, X., Liu, M., & Gilson, M. K. (2002). BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic acids research, 30(1), 348-351.

[3] Liu, X., Jiang, Y., Wang, Y., & Zeng, J. (2006). Predicting drug-target interaction using random walks on the heterogeneous network. Bioinformatics, 22(14), 176-181.

[4] Wang, L., You, Z. H., & Chen, X. (2019). DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. Journal of Chemical Information and Modeling, 59(1), 615-623.

[5] He, B., Tang, J., Ding, Y., Wang, H., & Zhang, L. (2017). SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. Journal of Chemical Information and Modeling, 57(11), 2740-2755.

[6] Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. In Advances in neural information processing systems (pp. 2224-2232).

[7] Zheng, S., & Huang, K. (2019). Molecule representation with graph convolutional networks. Molecules, 24(15), 2778.

[8] Huang, L., Xiao, Y., Wang, Y., Du, T., Xiong, Y., Zhang, W., ... & Ding, Y. (2020). GCN4DTA: predicting drug–target binding affinity using graph convolutional networks. Journal of Chemical Information and Modeling, 60(9), 4452-4462.

[9] Hakime Ozturk, Arzucan Ozgur and Elif Ozkirimli, "DeepDTA: deep drug-target binding affinity prediction,"

[10] Rao R, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny J, Abbeel P, Song YS. Evaluating Protein Transfer Learning with TAPE. Adv Neural Inf Process Syst. 2019 Dec;32:9689-9701. PMID: 33390682; PMCID: PMC7774645.

[11] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, Michal Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics, Volume 38, Issue 8, March 2022, Pages 2102–2110,

[12] Chithrananda, S., Grand, G., & Ramsundar, B. (2020). Chemberta: Large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885.

[13] Tang, J. (Luoja) (2014). KiBA - a benchmark dataset for drug target prediction. Jing Tang.

[14] Kingma, D. P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.