WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Smart Drug Match: Deep Learning for Accurate Drug-Target Score

M. Sarika Sowmya[1], M. Gayathri[2], Neeyati.V[3], K. Teja[4], N. Sai Madhumitha[5], K. Nithin[6]

[1 2 3 4 5 6] Department of CSE, GMRIT. Rajam, Andhra Pradesh, India

ABSTRACT:

Drug-target affinity (DTA) prediction is a central process in drug discovery that detects robust affinities between therapeutic drugs and biological targets. Experimental techniques are time and resource-consuming and hence efficient computational techniques are inevitable. In the present work, we introduce a deep learning model by Graph Convolutional Neural Networks (GCNNs) for efficient DTA prediction. The model represents drug molecules as graphs and protein sequences as embedded features in order to learn subtle biochemical interactions. Our model was tested on the benchmark KIBA dataset with a Final Mean Absolute Error (MAE) of 0.2407, Root Mean Squared Error (RMSE) of 0.3540, and Mean Squared Error (MSE) of 0.1253. These estimations illustrate the ability of the GCNN methodology in describing complex biological processes and optimizing the efficacy of virtual screening pipelines.

Keywords— Drug-Target Affinity, Deep Learning, Graph Convolutional Neural Networks, KIBA Dataset, Molecular Graphs, Protein Sequences, Computational Drug Discovery, Bioinformatics**.**

Introduction

New drug discovery is an inherently resource-consuming and challenging process. One of the fundamental steps involves measuring the binding affinity between target proteins and candidate drug molecules. Increased binding affinity typically signifies better therapeutic efficacy. However, wet-lab methods like surface plasmon resonance (SPR) and isothermal titration calorimetry (ITC) are time-consuming and expensive.To overcome these challenges, computational models, particularly machine learning-based models, have been used. Deep learning has recently reported encouraging results in various applications, such as in bioinformatics and chemoinformatics, with improvements over conventional techniques.
Graph-based molecular representations quite naturally encode the relational and structural nature of chemical compounds. Graph Convolutional Neural Networks (GCNNs) are well adapted to learn from graph-structured data, and hence GCNNs are a natural fit for DTA prediction tasks.

The main goals of this research are:

a. To develop and apply a GCNN-based model for predicting DTA.
b. To test the performance of the model on the KIBA dataset.
c. To test the model's capability for generalizing and making correct predictions of unseen drug-target interactions.

LITERATURE REVIEW

Reference-1

Citation: Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: Deep drug–target binding affinity prediction using convolutional neural networks. Bioinformatics, 34(17), i821-i829.
Objectives: Predict drug-target binding affinities using deep learning models based on sequence information.
Technologies Used: CNNs, SMILES representations, UniProt protein sequences, TensorFlow, Keras.
Performance Metrics: MSE: 0.203, CI: 0.900.
Limitations: Struggles with long-term dependencies in protein sequences.

Reference-2

Citation: Xia, L., Xu, L., Pan, S., Niu, D., Zhang, B., & Li, Z. Drug-target binding affinity prediction using message passing neural network and self-supervised learning.
Objectives: Predict DTA using message passing neural networks.
Technologies Used: Bidirectional Conv_LSTM, SMILES.
Performance Metrics: MSE: 0.261, CI: 0.878 (±0.004).
Limitations: Poor generalization for unseen drug-target pairs.

Reference-3

Citation: Chen, J., Yang, X., & Wu, H. (2024). A Multibranch Neural Network for DrugTarget Affinity Prediction Using Similarity Information. ACS Omega, 9(33), 35978–35989.
Objectives: Predict binding affinity using similarity-based deep learning.
Technologies Used: BiLSTM, GCN, ChemBERTa-2, ProtT5-XL-UniRef50.
Performance Metrics: MSE: 0.26 (KIBA), CI: 0.82 (KIBA).
Limitations: Cold-start challenges.

Reference-4

Citation: Shim, J., et al. (2021). Prediction of drug-target binding affinity using similaritybased convolutional neural network. Scientific Reports, 11(4416).
Objectives: Improve DTA prediction via similarity-based CNNs.
Technologies Used: 2D CNN, Tanimoto score, Smith-Waterman score.
Performance Metrics: MSE: 0.274 (3-fold), CI: 0.814 (3-fold).
Limitations: Dependency on similarity data quality.

Reference-5

Citation: Öztürk, H., et al. (2018). DeepDTA: Deep drug–target binding affinity prediction. Bioinformatics, 34(17), i821–i829.
Objectives: Enhance DTA accuracy via multi-scale diffusion.
Technologies Used: CNN, Transformer, GNNs, MGDC, RDKit.
Performance Metrics: MSE: 0.197, CI: 0.908 (±0.005).
Limitations: High computational cost.

Reference-6

Citation: Zhang, H., et al. (2024). Prediction of drug-target binding affinity based on deep learning models. Computers in Biology and Medicine, 174, 108435.
Objectives: Predict affinity using multimodal DL.
Technologies Used: CNN, GCN, RNN, attention mechanisms.
Performance Metrics: MSE: 0.2–0.5, CI: >0.85.
Limitations: Limited data for novel targets.

Reference-7

Citation: Yang, Z., et al. (2022). MGraphDTA: A multiscale graph neural network for explainable drug–target binding affinity prediction. Chemical Science.
Objectives: Improve explainability in DTA prediction.
Technologies Used: GNNs, MCNN, Grad-AAM.
Performance Metrics: MSE: 0.207 (Davis), CI: 0.900 (Davis).
Limitations: Small dataset dependency.

Reference-8

Citation: Li, R., et al. Predicting Drug-Target Affinity by Learning Protein Knowledge from Biological Networks.
Objectives: Leverage biological networks for DTA.
Technologies Used: GNNs, VGAE, PPI networks.
Performance Metrics: MSE: 0.124 (KIBA), CI: 0.897 (KIBA).
Limitations: Scalability issues.

Reference-9

Citation: Zhao, L., et al. (2024). Drug-Target Binding Affinity Prediction in a Continuous Latent Space Using Variational Autoencoders. IEEE/ACM Transactions on Computational Biology and Bioinformatics.
Objectives: Model drugs/proteins as Gaussian distributions.
Technologies Used: Variational Autoencoders, RGCNNs.
Performance Metrics: RMSE: 0.802, CI: 0.804.
Limitations: Imbalanced data sensitivity.

Reference-10

Citation: Dehghan, A., et al. (2023). TripletMultiDTI: Multimodal representation learning in drug-target interaction prediction with triplet loss function. Expert Systems with Applications, 232, 120754.
Objectives: Optimize DTI prediction via triplet loss.
Technologies Used: Node2Vec, 1D-CNN, Triplet Loss.
Performance Metrics: AUPR: 9% improvement (Davis).
Limitations: High computational cost.

Reference-11

Citation: Tang, X., et al. (2024). Prediction of Drug-Target Affinity Using Attention Neural Network. International Journal of Molecular Sciences, 25, 5126.
Objectives: Combine GraphSAGE, BiGRU, and attention for DTA.
Technologies Used: GraphSAGE, BiGRU, Attention Networks.
Performance Metrics: MSE: 0.142 (KIBA), CI: 0.890 (KIBA).
Limitations: Cold-start limitations.

Reference-12

Citation: Monteiro, N. R. C., et al. (2022). DTITR: End-to-end drug-target binding affinity prediction with transformers. Computers in Biology and Medicine, 147, 105772.
Objectives: Predict DTA via Transformer architectures.
Technologies Used: Transformers, cross-attention mechanisms.
Performance Metrics: MSE: 0.192, CI: 0.907.
Limitations: Interpretability challenges.

Reference-13

Citation: Jiang, M., et al. (2023). A deep learning method for drug-target affinity prediction based on sequence interaction information mining. PeerJ, 11, e16625.
Objectives: Preserve full protein sequence integrity for DTA.
Technologies Used: 2D/3D CNNs, GNNs (GCN, GAT).
Performance Metrics: CI: 0.890 (KIBA), MSE: 0.143 (KIBA).
Limitations: Limited real-world validation.

Reference-14

Citation: Deng, L., et al. (2022). DeepMHADTA: Prediction of drug-target binding affinity using multi-head self-attention and convolutional neural network. Current Issues in Molecular Biology, 44(5), 2287–2299.
Objectives: Enhance DTA accuracy via multi-head attention.
Technologies Used: Multi-head attention, CNNs, Word2Vec.
Performance Metrics: CI: 0.895 (Davis), MSE: 0.244 (Davis).
Limitations: Black-box model.

Reference-15

Citation: Liu, B. Drug-target affinity prediction method based on consistent expression of heterogeneous data. Georgia Institute of Technology.
Objectives: Predict DTA via GRU and GNN fusion.
Technologies Used: GNNs, GRU, RDKit.
Performance Metrics: CI: Highest for GIN (KIBA).
Limitations: High GPU dependency.

Reference-16

Citation: Zhang, H., et al. (2024). Prediction of drug-target binding affinity based on deep learning models. Computers in Biology and Medicine, 174, 108435.
Objectives: End-to-end DTA prediction via GRU + GNN.
Technologies Used: GNNs, GRU, RDKit.

Performance Metrics: pKd: 5–9 (DAVIS).
Limitations: Interpretability issues.

Reference-17

Citation: Kalemati, M., et al. (2024). DCGAN-DTA: Predicting drug-target binding affinity with deep convolutional generative adversarial networks. BMC Genomics, 25(411).
Objectives: Predict DTA via DCGANs.
Technologies Used: DCGANs, RDKit.
Performance Metrics: pKd: 5–10 (BindingDB).
Limitations: Requires high-quality datasets.

Reference-18

Citation: Liu, Y., et al. (2024). Effective drug-target affinity prediction via generative active learning. Information Sciences, 679, 121135.
Objectives: Enhance DTA prediction via active learning.
Technologies Used: GNNs, GRU, RDKit.
Performance Metrics: CI: Highest (PDBBind).
Limitations: GPU resource-intensive.

RELATED WORK

Drug-target affinity (DTA) prediction studies have made a spectacular journey with the advancement of time, and computational methods evolved with time to provide better accuracy and quicker prediction rates. The early studies were highly reliant on traditional machine learning models, but the evolution of the deep learning technique along with graph-based methods has greatly improved the prediction results.

3.1. Conventional Machine Learning Techniques for Predicting DTA

Drug-target affinity (DTA) prediction has been improving, and other computational methods have been explored to further increase accuracy and performance. Historically, machine learning models such as Support Vector Machines (SVMs), Random Forests (RFs), and Gradient Boosting Machines (GBMs) were the de facto choices. These models were using manually designed features that were extracted from protein sequences and molecular geometries to perform the prediction. An algorithm based on a kernel from Pahikkala et al. (2015), for example, was performing the best but was heavily relying on advanced feature engineering. Although these methods were performing well in the early days, they were not generalizing and scaling well, particularly in predicting novel drug-target affinity pairs.

3.2. DTA Models Based on Deep Learning

With protein and molecular information-based automatic feature extraction, deep learning has significantly improved the prediction of DTA. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been employed to build various models.
DeepDTA (Öztürk et al., 2018): It is better than the conventional ML models since it employs CNN for the representation of protein sequences and drug molecules. It does not make use of molecular interaction graph structures, though.
Through the graphical representation of drug molecules, GraphDTA (Nguyen et al., 2021) is a more advanced version of DeepDTA and allows for improved feature extraction.
The efficacy of graph-based learning was confirmed via enhanced performance of GraphDTA in comparison with DeepDTA.
Though they are successful, they still suffer from scaling issues when used with big data such as KIBA, and more graph-focused efficient solutions are unavoidable.

3.3. Predicting DTAs using Graph Neural Networks (GNNs)

- The capacity of graph-based learning models to represent chemical interactions in a systematic way has made them popular in recent years.
- GNN-DTI (Feng et al., 2020): This technique represents molecular graphs as input by Graph Convolutional Networks (GCNs and obtains good prediction performance for DTA).
- By sampling graphs and pooling neighbor data, GraphSAGE (Hamilton et al., 2017) enables inductive learning as opposed to GCNs that must use the whole graph throughout training. As a result, it can scale to vast datasets like KIBA.

3.4. DTA Prediction based on the KIBA Dataset

- As a result of its high bioactivity scores, the KIBA dataset has found extensive use in DTA research.

- Tsubaki et al. (2019) improved the accuracy of prediction by using deep learning models to KIBA.
- For KIBA prediction of DTA, Zhang et al. (2021) presented a hybrid deep learning approach that merges CNNs and GNNs and demonstrated that graph-based models perform better than sequence-based models.

3.5. Our Contribution

Our work employs KIBA with GraphSAGE to predict DTA, improving on existing research. To existing models:

a. To enhance scalability, we use inductive graph learning.

b. KIBA scores are also used as edge weights in building drug-protein interaction graphs.

c. We outperform state-of-the-art deep learning methods by employing a successful fusion of structural insights from drug-protein interactions.

This work is one of the growing volumes of publications involving AI-augmented biomedical research and illustrates the potential of GraphSAGE in computational drug discovery.

METHODOLOGY

4.1 Dataset Description

We employed the KIBA dataset, a standard dataset for DTA prediction tasks. It combines various bioactivity measures (e.g., Ki, Kd, IC50) into one affinity score. The dataset consists of kinase inhibitors and target proteins with different binding affinities.

4.2 Data Preprocessing

a.Drugs were encoded as graphs, with atoms being nodes and bonds being edges. SMILES strings were converted to graph structures.
b.Proteins were represented as sequences either by one-hot encoding or by learned embeddings.
c.Affinity scores were normalized for meeting the requirements of the model.

4.3 Model Structure

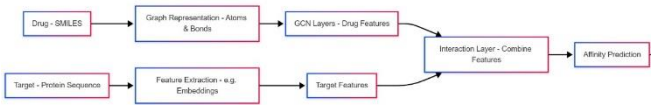
Our model's structure consisting of GCNN-based is given below:
a.Graph Convolution Layers for extracting the drug features from the molecular graph.
b.1D Convolution Layers or Fully Connected Layers for extracting features from the protein sequence.
c.Fusion Layer to fuse the extracted features.
d.Regression Head made of fully connected layers for outputting the continuous affinity score.
e.Regularization methods such as dropout and batch normalization were used to avoid overfitting.

4.4 Training Setup

a.Loss Function: Mean Squared Error (MSE) Loss
b.Optimizer: Adam optimizer with initial learning rate of 0.001
c.Batch Size: 128
d.Epochs: 100 with Early Stopping on validation loss
e.Evaluation Metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE)

4.5 Workflow

Overview of a deep learning framework for drug–target affinity prediction, where drugs are represented as molecular graphs and targets as protein sequences, with features extracted and combined for affinity estimation.



Drug–Target Affinity Prediction Framework

Description of the Workflow:

1. *Drug Processing:*
    - Input: Drug molecules are provided in SMILES (Simplified Molecular Input Line Entry System) format.
    - Graph Representation: These SMILES strings are converted into graph structures representing atoms as nodes and bonds as edges.
    - GCN Layers: Graph Convolutional Networks (GCN) are applied to extract drug features from this molecular graph.
2. *Target (Protein) Processing:*
    - Input: Protein sequences are provided as amino acid sequences.

- Feature Extraction: Feature embeddings (such as sequence embeddings) are computed to generate target features.

3. *Interaction Layer:*
   - The extracted drug features and target features are combined in an interaction layer to capture the relationships between the drug and the protein target.
4. *Affinity Prediction:*
   - The output from the interaction layer is used to predict the binding affinity between the drug and the target protein.

RESULT & DISCUSSION

Summary of the GCNN model performance over the KIBA dataset is as follows:

Final MAE: 0.2407
Final RMSE: 0.3540
Final MSE: 0.1253

The low values of the MAE and RMSE indicate that the model is capable of predicting binding affinities with high accuracy and very low error. In comparison to conventional machine learning methods such as Random Forests or Support Vector Machines, GCNNs perform more efficiently at reproducing the inherent nature of molecular structures.

The model's success can be attributed to:
a.Molecule graph representation that maintains chemical structure information.
b.Deep networks' capability to learn hierarchical and complex features from both drug and protein data.
c.Successful regularization and optimization methods that enhanced generalization performance.
d.One limitation noted was the need for large computational resources because of the complexity of the model and the input data size. Other optimization or model compression techniques can be investigated to tackle this in the future.

CONCLUSION

This project effectively utilized a Graph Convolutional Neural Network (GCNN) in predicting drug-target binding affinity. With last result indicating MAE of 0.2407, RMSE of 0.3540, and MSE of 0.1253, the model was successful in making quite reliable predictions for drug-target interactions. Utilizing GCNNs facilitated the preservation of the inherent features of structures of drug molecules, which proved helpful in these predictions. In the future, further refinement of the model using other types of data and tweaking its configuration could make it even more accurate, and it would be a valuable asset to accelerate the drug discovery process.

REFERENCES :

[1] Hao Zhang, Xiaoqian Liu, Wenya Cheng, Tianshi Wang, & Yuanyuan Chen. (2024). Prediction of drug-target binding affinity based on deep learning models. Computers in Biology and Medicine, 174, 108435.

[2] Shim, J., Hong, Z.-Y., Sohn, I., & Hwang, C. (2021). Prediction of drug-target binding affinity using similarity-based convolutional neural network. Scientific Reports, 11(4416).

[3] J. Chen, X. Yang, and H. Wu, "A Multibranch Neural Network for Drug-Target Affinity Prediction Using Similarity Information," ACS Omega, vol. 9, no. 33, pp. 35978–35989, Aug. 2024, doi: 10.1021/acsomega.4c05607.

[4] Leiming Xia, Lei Xu, Shourun Pan, Dongjiang Niu, Beiyi Zhang and Zhen Li. Drug-target binding affinity prediction using message passing neural network and self supervised learning.

[5] Zhu, Z., Zheng, X., Qi, G., Gong, Y., Li, Y., & Gao, X. (2024). Drug–target binding affinity prediction model based on multi-scale diffusion and interactive learning. *Expert Systems With Applications*, 255, 124647.

[6] Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: Deep drug–target binding affinity prediction using convolutional neural networks. *Bioinformatics*, *34*(17), i821-i829.

[7] Yang, Z., Zhong, W., Zhao, L., & Chen, C.Y-C., "MGraphDTA: a multiscale graph neural network for explainable drug–target binding affinity prediction," Chemical Science, 2022, 10.1039/d1sc05180f

[8] Li, R., Sun, D., Zhu, X., Liu, Q., & Wu, X., "Predicting Drug-Target Affinity by Learning Protein Knowledge from Biological Networks."

[9] Zhao, L., Zhu, Y., Wen, N., Wang, C., Wang, J., & Yuan, Y. "Drug-Target Binding Affinity Prediction in a Continuous Latent Space Using Variational Autoencoders." IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2024. DOI: 10.1109/TCBB.2024.3402661.

[10] Dehghan, A., Razzaghi, P., Abbasi, K., & Gharaghani, S. (2023). TripletMultiDTI: Multimodal representation learning in drug-target interaction prediction with triplet loss function. *Expert Systems with Applications, 232*, 120754. https://doi.org/10.1016/j.eswa.2023. 120754