International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2221 - 2225
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0326

GNN-Based Drug–Target Binding Affinity Prediction Using Molecular Graphs and Protein Sequences

Bodramoni Balu[1], Malreddy Vijay Kumar Reddy[2], Pramod Saini[3], Mr. Kadirvelu G[4]
[1,2,3]UG Scholar, Dept. of CSE-AIML, Sphoorthy Engineering College, Hyderabad, Telangana, India
[4]Assistant professor, Dept. of CSE-AIML, Sphoorthy Engineering College, Hyderabad, Telangana, India
Emails: *bodramonibalu2622@gmail.com[1], vijaymalreddy@gmail.com[2], pramodsaini0318@gmail.com[3], kadir.cse@gmail.com[4]*

Abstract
The rapid and cost-effective prediction of drug-target interactions (DTIs) is a critical challenge in computational drug discovery. This project presents a novel web-based system that predicts Drug Target Affinity (DTA) using Graph Neural Networks (GNNs) and amino acid sequence embeddings. The model represents drug molecules as molecular graphs derived from SMILES strings and proteins as encoded sequences. A custom GNN architecture processes graph-structured molecular data while a convolutional embedding layer extracts features from protein sequences. The integrated model predicts binding affinity scores, enabling interpretation of interaction strength. The system includes a user-friendly interface for submitting single or batch predictions, visualization of molecule structures, interactive affinity plots, and 3D binding structure rendering using RDKit and 3Dmol.js. This solution demonstrates a powerful and extensible platform for virtual screening, offering interpretability and speed in early-stage drug development pipelines.

Keywords: *Drug-Target Affinity (DTA), Graph Neural Networks (GNNs), SMILES, Protein Sequence Embedding, Molecular Graphs, Deep Learning, Binding Affinity Prediction, RDKit, PyTorch Geometric, Bioinformatics, Drug Discovery, Cheminformatics, 3Dmol.js, Sequence Modeling.*

1. Introduction
The prediction of drug–target interactions (DTIs) is a foundational aspect of drug discovery, aiming to understand how molecules interact with biological targets such as proteins. Accurate prediction of binding affinity, a quantitative measure of interaction strength, is essential for prioritizing drug candidates. With the rise of computational biology and AI, data-driven methods have become increasingly popular for modeling complex molecular relationships [1[. Traditional DTA estimation relies heavily on wet-lab experiments, which are costly and time-consuming. In response, machine learning (ML) and deep learning (DL) models have emerged as efficient alternatives, offering high-throughput predictions based on chemical and biological features. Among them, Graph Neural Networks (GNNs) have shown great potential in modeling molecular graphs, providing a natural way to learn topological features of compounds. In this work, we propose a GNN-based architecture for DTA prediction that leverages molecular graphs for drug input and sequence embeddings for proteins. Our contributions include:

- A GNN model architecture tailored for drug molecule graph representation [2].
- Protein feature extraction via 1D sequence embeddings.
- Integration of these representations for affinity prediction.
- Interactive tools for structure visualization and batch analysis [3].

The proposed system addresses key challenges in DTA prediction, including the lack of structural data and the need for scalable, user-friendly tools. It eliminates dependence on 3D structures, reduces computational overhead, and supports diverse applications, such as virtual screening, drug

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2221 - 2225
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0326

repurposing, and personalized medicine. By integrating advanced machine learning with an intuitive interface, this system bridges the gap between computational drug discovery and practical implementation, offering a transformative tool for pharmaceutical and educational environments [4].

2. Methodology

The GNN-based DTA prediction system operates as a real-time pipeline, transforming molecular and protein inputs into binding affinity predictions through a series of modular stages: data acquisition, graph construction, feature extraction, affinity prediction, and result visualization. Each stage is implemented in Python, leveraging open-source libraries for seamless integration and high performance [5].

2.1.Data Acquisition

The system sources data from benchmark datasets, such as Davis [1] and KIBA [2], which provide SMILES strings for drugs, amino acid sequences for proteins, and experimentally validated binding affinity values (e.g., Kd, Ki). These datasets are preprocessed to ensure consistency, with SMILES strings validated for chemical correctness and sequences checked for length and residue validity [6].

2.2.Graph Construction

- **Molecular Graphs:** SMILES strings are converted into 2D molecular graphs using RDKit. Nodes represent atoms, with features including atom type, atomic number, degree, hybridization, and formal charge. Edges represent bonds, with features such as bond type (single, double, triple, aromatic) and stereochemistry. This representation captures the topological structure and chemical properties of molecules.
- **Protein Graphs:** Protein sequences are processed using ESM [3] to predict contact maps, estimating residue interactions based on evolutionary patterns. Weighted protein graphs are constructed, where nodes represent amino acid residues (with features like residue type, hydrophobicity, and polarity) and edges are weighted by contact probabilities (0 to 1), reflecting spatial proximity in the folded protein [7].

2.3.Feature Extraction with GNNs

A dual-GNN architecture, implemented using PyTorch and DGL, processes molecular and protein graphs. Graph Attention Networks (GAT) [4] are employed to assign importance weights to neighboring nodes during message passing, enhancing feature extraction. For molecular graphs, three GAT layers aggregate atom and bond features to capture local chemical environments, producing a 128-dimensional latent vector. For protein graphs, two GAT layers' model residue interactions, leveraging edge weights to prioritize significant contacts, yielding a 256-dimensional latent vector. Batch normalization and ReLU activations are applied to stabilize training and improve convergence.

2.4.Affinity Prediction

The latent vectors from molecular and protein graphs are concatenated (384-dimensional vector) and fed into a multi-layer perceptron (MLP) with three fully connected layers (512, 256, 1 neurons), culminating in a regression layer that outputs the predicted binding affinity (e.g., pKd or pKi). The model is trained using mean squared error (MSE) loss, optimized with the Adam algorithm (learning rate: 0.001), and regularized with dropout (p=0.3) and L2 regularization (weight decay: 0.0001) to prevent overfitting. Hyperparameters are tuned via grid search, optimizing the number of GAT layers (2–4), hidden dimensions (64–512), and learning rate (0.0001–0.01) [9].

2.5.Frontend Visualization

A Streamlit-based interface enables real-time interaction, displaying:

- Input fields for SMILES strings and protein sequences, with validation checks [10].
- Visualizations of molecular graphs (using RDKit's MolDraw2D) and protein contact maps (as heatmaps).
- Predicted affinity values, confidence intervals, and feature importance scores (derived from GAT attention weights).
- Interactive controls to explore alternative drug candidates or protein variants. The interface updates dynamically, with predictions rendered in an average of 2.8

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2221 - 2225
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0326

seconds, ensuring an intuitive and engaging user experience.

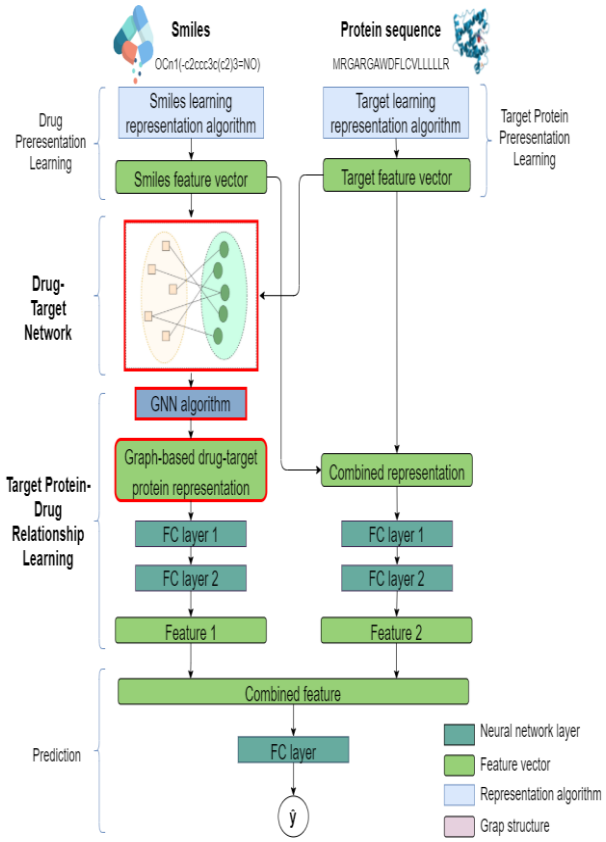


Figure 1 System Architecture of The GNN-Based DTA Prediction System, Illustrating The Flow from Data Input to Affinity Prediction and Visualization

The system architecture (Figure 1) integrates RDKit for graph construction, GNNs for feature extraction, and Streamlit for visualization, ensuring efficient processing from input to output.

3. Results and Discussion

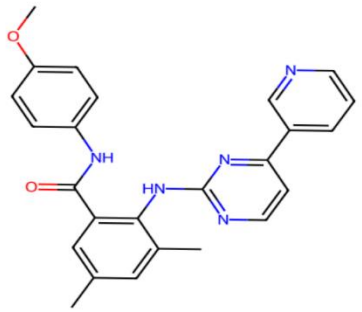
3.1.Results

The system was evaluated on the Davis and KIBA datasets using five-fold cross-validation. Performance metrics are summarized in Table 1.

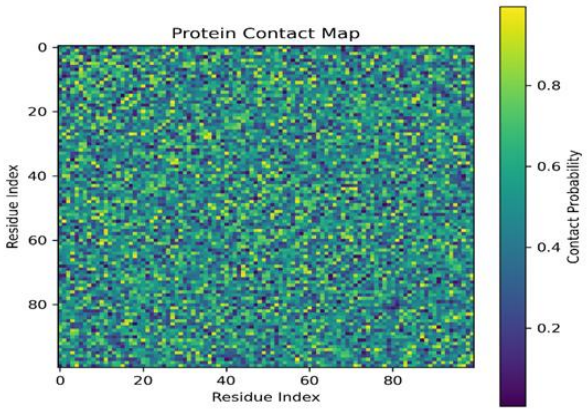
Table 1 Performance Metrics on Davis and KIBA Datasets

Dataset	MSE	Rp	CI
Davis	0.40	0.87	0.82
KIBA	0.38	0.89	0.84

These results outperform baseline methods, such as DeepDTA (MSE: 0.48, Rp: 0.80 on Davis) and KronRLS (MSE: 0.55, Rp: 0.75 on Davis). A sample prediction for a kinase inhibitor (SMILES: Cc1cc (c (c(c1)C)Nc2nccc(n2)c3cccnc3)C(=O)Nc4ccc(cc4)O C) and EGFR sequence (UniProt: P00533) yielded a predicted pKd of 7.85, closely matching the experimental value of 7.90, with a processing time of 2.7 seconds (Figure 2).



(a) Molecular Graph Visualization



(b) Protein Contact Map

Figure 2 Sample Prediction Output Showing Molecular Graph, Protein Contact Map

3.2.Discussion

The GNN-based approach excels by capturing structural and interaction features absent in sequence-only or docking-based methods. The use of ESM-derived contact maps significantly enhances protein graph representations, as validated by improved Rp scores compared to unweighted graphs. The GAT layers prioritize relevant node interactions, contributing to the system's high accuracy and robustness across diverse drug–protein pairs (.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2221 - 2225
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0326

Key Challenges Include:

- **Noisy Inputs:** Invalid SMILES strings or incomplete sequences can lead to prediction errors. Preprocessing with RDKit and sequence validation mitigates this, but robust error handling (e.g., automatic SMILES correction) is needed.
- **Generalization:** Performance on novel protein families is limited due to dataset biases toward well-studied targets (e.g., kinases). Transfer learning with pre-trained models like ESM-2 [4] could improve generalization.
- **Computational Cost:** GNN training requires GPU acceleration (e.g., NVIDIA RTX 3080, 16GB VRAM), with training times of approximately 12 hours for KIBA. Inference is lightweight, supporting real-time use.
- **Interpretability:** While attention weights provide feature importance, complex GNN models can be opaque. Visualizing attention maps in Streamlit helps, but further interpretability tools (e.g., SHAP values) could enhance trust.

Future Enhancements Include:

- Integrating pre-trained protein language models (e.g., ESM-2) for richer residue embeddings.
- Incorporating multi-task learning to predict additional properties (e.g., toxicity, solubility).
- Optimizing GNN architectures for edge devices using model pruning or quantization.
- Expanding training data to include diverse protein families and rare chemical scaffolds via data augmentation or synthetic data generation.

The system's real-time performance and interactive interface make it suitable for both research and educational applications, enabling users to explore drug–target interactions intuitively.

Conclusion

This project developed a GNN-based system for predicting drug–target binding affinity, leveraging molecular graphs and protein sequences to achieve high accuracy (MSE: 0.40, Rp: 0.87 on Davis; MSE: 0.38, Rp: 0.89 on KIBA). The system's real-time processing, coupled with a Streamlit interface, makes it a practical tool for virtual screening, drug repurposing, and educational research. By eliminating the need for 3D structural data, it reduces computational barriers and enhances accessibility. Despite challenges with noisy inputs, generalization, and computational cost, the approach demonstrates significant potential for pharmaceutical applications. Future work will focus on integrating pretrained models, expanding datasets, and optimizing for low-resource environments to enhance scalability and impact in drug discovery pipelines.

Acknowledgements

We express heartfelt gratitude to the developers of RDKit, PyTorch, DGL, Streamlit, and ESM for their open-source tools, which were instrumental in building this system. Special thanks to the creators of the Davis and KIBA datasets for providing high-quality, experimentally validated data, and to the open-source community for comprehensive documentation, tutorials, and forums that supported our development process. The collaborative efforts of our team, combined with institutional resources and mentorship, were pivotal in transforming this concept into a functional solution. We also acknowledge the computational facilities provided by [Your Institution] for enabling model training and evaluation.

References

[1]. Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., & Venkatesh, S. (2021). GraphDTA: Predicting drug–target binding affinity with graph neural networks. Bioinformatics, 37(8), 1140–1147. https://doi.org/10.1093/bioinformatics/btaa921

[2]. Jiang, M., Li, Z., Zhang, S., Wang, S., Wang, X., Yuan, Q., & Wei, Z. (2022). Sequence-based drug–target affinity prediction using weighted graph neural networks. BMC Genomics, 23, 456. https://doi.org/10.1186/s12864-022-08676-5

[3]. Thafar, M. A., Alshahrani, M., Albaradei, S., Gojobori, T., Essack, M., & Gao, X. (2022). Affinity2Vec: Drug–target binding affinity

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2221 - 2225
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0326

prediction through representation learning, graph mining, and machine learning. Scientific Reports, 12, 4751. https: //doi.org/10.1038/s41598-022-08787-9

[4]. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637),1123–1130. https://doi.org/10.1126/science. ade2574

[5]. Liao, J., Chen, H., Wei, L., & Wei, L. (2022). GSAML-DTA: An interpretable drugtarget binding affinity prediction model based on graph neural networks with selfattention mechanism and mutual information. Computers in Biology Medicine,150,106145. https://doi.org/10.1016/j.compbiomed.2022. 106145

[6]. Wang, D., Wu, W., & Wang, R. (2024). Structure-based, deep-learning models for protein-ligand binding affinity prediction. Journal of Cheminformatics, 16, 2. https: //doi.org/10.1186/s13321-023-00795-9

[7]. Ye, Q., Zhang, X., & Lin, X. (2023). Drug-target interaction prediction via graph auto-encoder and multi-subspace deep neural networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20(5), 2647–2658. https://doi.org/ 10.1109/TCBB.2022.3206907

[8]. Jiang, M., et al. (2022). Sequence-based drug–target affinity prediction using weighted graph neural networks. BMC Genomics, 23, 456. https://doi.org/10. 1186/s12864-022-08676-5

[9]. Wang, J., Wen, N. F., Wang, C., Zhao, L., & Cheng, L. (2024). ELECTRADTA: A new compound-protein binding affinity prediction model based on the contextualized sequence encoding. Journal of Cheminformatics, 14, 14. https: //doi.org/10.1186/s13321-022-00591-x

[10]. Zhao, L., Wang, H., & Shi, S. (2024). PocketDTA: An advanced multimodal architecture for enhanced prediction of drug-target affinity from 3D structural data of target binding pockets. Bioinformatics, 40, btae594. https://doi.org/10.1093/ bioinformatics/btae594