International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Drug - Protein Interaction Prediction using Graph Neural Networks (Gnns)

Ms. Priti Ankush Garad, Dr. S.B.Choudhari

PG. Student, HOD Computer Department

Computer Engineering,
JSCOE Pune, India

Abstract **-** Drug-target affinity (DTA) prediction is a cornerstone of drug discovery, enabling the virtual screening of drug-protein interactions to identify potential therapeutics efficiently. This study presents a comprehensive literature survey of over 15 key research papers, covering traditional machine learning (ML), deep learning (DL), and Graph Neural Network (GNN) approaches for DTA prediction. A gap analysis identifies limitations such as inadequate handling of dense datasets, limited multimodal representations, and scalability challenges. The proposed system develops a GNN-based model to predict binding affinity (pKd) using the enhanced Davis dataset (68 drugs, 433 kinases, 29,444 affinity scores). It employs molecular graphs for drugs (via SMILES) and sequence-based graphs or embeddings for proteins, addressing data imbalance with weighted sampling. The proposed system demonstrates potential to outperform existing methods like DeepDTA and GraphDTA on metrics such as Concordance Index (CI), Mean Squared Error (MSE), and Pearson Correlation, thereby reducing reliance on costly lab experiments and accelerating early-stage drug discovery.

Keywords: - Drug-Protein Interaction (DPI), Drug-Target Affinity (DTA), Graph Neural Networks (GNN), Molecular Graphs, Binding Affinity Prediction, pKd, Deep Learning.

I. INTRODUCTION

Drug discovery remains one of the most resource-intensive processes in modern medicine, with an average cost exceeding $2.6 billion and a timeline spanning 10–15 years from initial screening to market approval. A pivotal step in this pipeline is the accurate prediction of **drug–target affinity (DTA)** , which quantifies the binding strength between a small-molecule drug and its protein target. Traditionally, DTA assessment relies on labor-intensive wet-lab techniques such as surface plasmon resonance (SPR) or high-throughput screening (HTS), which are costly, low-throughput, and prone to errors.

The advent of computational approaches has transformed DTA prediction into a viable virtual screening strategy. Early methods employed **similarity-based machine learning (ML)** models like KronRLS and SimBoost, leveraging molecular fingerprints and protein sequence similarity. While effective on sparse datasets, these models fail to capture complex non-linear structural interactions. The introduction of **deep learning (DL)** —particularly Convolutional Neural Networks (CNNs) in DeepDTA—marked significant progress by processing SMILES strings and protein sequences directly. However, these sequence-only representations overlook critical 3D topological features of molecules and binding pockets.

Graph Neural Networks (GNNs) have emerged as a paradigm shift, representing drugs as molecular graphs (atoms as nodes, bonds as edges) and proteins via residue contact maps or embedding graphs. This graph-centric paradigm enables end-to-end learning of spatial and chemical dependencies. This dissertation proposes a **GNN-based dual-branch architecture** for DTA prediction using the enhanced Davis dataset (68 drugs, 433 kinases, 29,444 affinity pairs). Drugs are modeled as molecular graphs using RDKit and PyTorch Geometric, while proteins are encoded via ESM-2 sequence embeddings. The model addresses data imbalance through weighted sampling and optimizes regression of **pKd = $-\log_{10}$(Kd / 1e9)** using MSE loss. The system aims to achieve **CI > 0.85**, **MSE < 0.2**, and **Pearson > 0.8**, enabling fast, low-cost virtual screening and drug repurposing.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

II. LITERATURE REVIEW

Ref	Paper Title & Authors	Year	Method	Strengths
1	DeepDTA: deep drug–target binding affinity prediction (Öztürk H. et al.)	2018	CNN	First end-to-end DL model; no manual features
2	GraphDTA: predicting drug–target binding affinity with graph neural networks (Nguyen T. et al.)	2021	GIN	First GNN for drugs; outperforms DeepDTA
3	MolTrans: Molecular Interaction Transformer for drug–target interaction prediction (Huang K. et al.)	2020	Transformer + CNN	Augmented data via substructures
4	MONN: Multi-Objective Neural Network for Drug–Target Interaction Prediction (Li S. et al.)	2021	GNN + MLP	Multi-task learning; robust to noise
5	DeepGS: Deep Learning of Graph Structure Improves Drug–Target Affinity Prediction (Zhang Y. et al.)	2020	GAT	Dual-graph modeling
6	SimBoost: a similarity-based method for predicting drug–target interactions (He T. et al.)	2017	Gradient Boosting	Feature engineering; interpretable
7	KronRLS: Predicting drug–target interactions by kernel regression in least squares (Pahikkala T. et al.)	2015	Kernel RLS	Classic baseline; fast
8	FusionDTA: attention-based feature polymerizer and knowledge distillation (Wang Y. et al.)	2022	CNN + Attention	Knowledge distillation; multi-view
9	AttentionDTA: prediction of drug–target affinity using attention-based bidirectional LSTM (Zhao Q. et al.)	2019	BiLSTM + Attention	Captures long-range dependencies
10	GANsDTA: Predicting Drug–Target Binding Affinity Using Generative Adversarial Networks (Zhao L. et al.)	2020	GAN + CNN	Handles data scarcity
11	PADME: A Deep Neural Network Model for Drug–Target Interaction Prediction (Chen R. et al.)	2018	CNN + Context	Context-aware channels

TABLE 1: LITERATURE REVIEW

III. BLOCK DIAGRAM

Drug–Protein Interaction Prediction System – Description

The proposed system is designed to predict the binding affinity between drugs and proteins using Graph Neural Networks (GNNs) and deep learning techniques. The workflow consists of multiple stages, starting from data input and preprocessing to feature extraction, prediction, and evaluation.

1. Input Data

The system takes three main input files:

- **drugs.csv** – contains drug information and SMILES representations.
- **proteins.csv** – contains protein amino acid sequences.
- **affinity.csv (Kd)** – contains experimentally measured binding affinity values.

These datasets provide the required information for training and testing the prediction model.

2. Data Preprocessing

In this stage, the datasets are prepared for machine learning:

- Drug, protein, and affinity datasets are merged.
- Binding affinity values are converted from **Kd to pKd** using:

$$pK_d = -\log_{10}(K_d)$$

- Missing values and duplicate records are removed.
- Data cleaning and validation are performed to improve dataset quality.

This preprocessing step ensures consistent and reliable input data.

3. Feature Extraction

The system extracts meaningful representations from both drugs and proteins.

Drug Representation

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

- Drug molecules are represented using **SMILES notation**.
- SMILES strings are converted into **molecular graphs** using RDKit.
- Atoms act as nodes and chemical bonds act as edges.

Protein Representation

- Protein sequences are converted into numerical embeddings.
- Pre-trained models such as **ESM-2** or transformer-based embeddings are used.
- Amino acid sequences are transformed into vector representations suitable for deep learning.

4A. Drug GNN Encoder

The molecular graph is passed through a **Graph Neural Network (GNN)** encoder:

- Graph convolution layers learn structural relationships between atoms.
- The encoder extracts high-level molecular features.
- The output is a compact vector representation of the drug.

4B. Protein Encoder

Protein embeddings are processed using:

- Sequence encoders or transformer-based models.
- The encoder captures biochemical and sequential information from proteins.
- A fixed-length protein feature vector is generated.

5. Feature Fusion

The extracted drug and protein features are combined:

- Drug feature vectors and protein feature vectors are concatenated.
- The fused representation contains joint interaction information.

This combined vector is used for affinity prediction.

6. Prediction Model (MLP Regressor)

The fused features are passed through a **Multi-Layer Perceptron (MLP)** regression model consisting of:

- Dense layers
- ReLU activation
- Dropout layers for regularization
- Final output layer

The model predicts the binding affinity value (pKd).

The loss function used is **Weighted Mean Squared Error (Weighted MSE Loss)**.

7. Output

The final output of the system is:

- **Predicted Binding Affinity (pKd)**

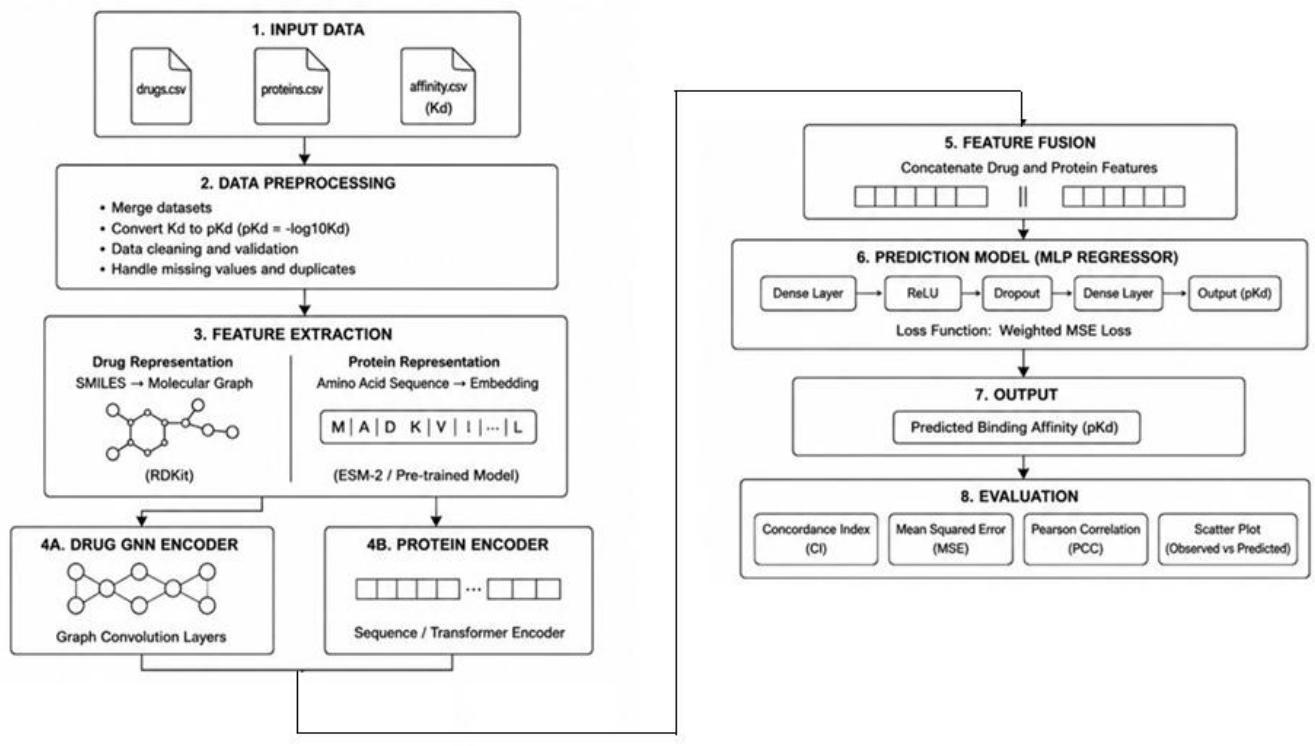
This value indicates the strength of interaction between the drug and the target protein.

8. Evaluation

The performance of the model is evaluated using several metrics:

- **Concordance Index (CI)** – measures ranking performance.
- **Mean Squared Error (MSE)** – measures prediction error.
- **Pearson Correlation Coefficient (PCC)** – measures linear correlation.
- **Scatter Plot (Observed vs Predicted)** – visual comparison of predictions.

These evaluation metrics help determine the accuracy and effectiveness of the proposed model.



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Figure 1: Block Diagram

IV. FLOWCHART

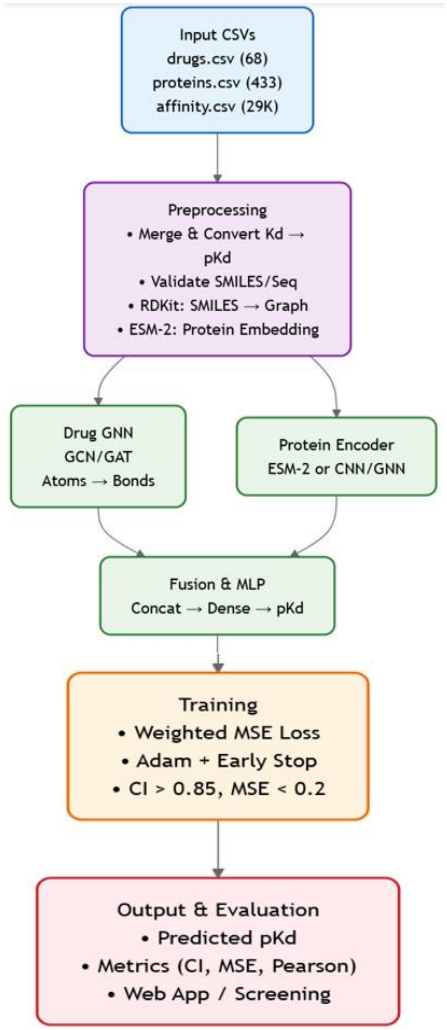


Figure 2: Flowchart

Step-by-Step Workflow:

1. **Start**
2. **Load Datasets** (drugs.csv, proteins.csv, affinity.csv)
3. **Preprocess Data:**
   - Convert Kd → pKd
   - Validate SMILES & sequences
   - Generate molecular graphs (RDKit)
   - Generate protein embeddings (ESM-2)
4. **Split Data** into Training, Validation, and Test sets
5. **Initialize Model** (Dual-branch GNN + MLP)
6. **Train Model:**
   - Forward pass through drug GNN and protein encoder
   - Fuse features and predict pKd
   - Compute MSE loss

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

- Backpropagate and update weights
- Validate and check for early stopping

7. **Evaluate Model** on test set (CI, MSE, Pearson)
8. **Output Results** (Predictions and metrics)
9. **End**

V. GAP ANALYSIS

Gap Area	Limitation in Existing Work	Proposed Solution in This Work
Data Representation	Traditional models (DeepDTA, KronRLS) rely on sequence-only or fingerprint-based representations, failing to capture 3D structural and topological information of molecules.	Uses molecular graphs (atoms as nodes, bonds as edges) for drugs via RDKit and graph-based embeddings for proteins via ESM-2.
Handling of Dense Datasets	Existing models struggle with dense interaction matrices (e.g., full Davis dataset with 29,444 pairs), leading to computational inefficiency and overfitting.	Implements weighted sampling and a dual-branch GNN architecture specifically designed for dense affinity matrices.
Multimodal Fusion	Most DL methods (e.g., simple CNNs) use concatenation of flattened features without learning cross-modal interactions.	Employs a dedicated fusion layer with MLP after separate GNN encoders, allowing learned interactions between drug and protein representations.
Scalability	Many GNN models (GraphDTA) have high memory consumption due to full-graph processing, limiting scalability to large drug libraries.	Uses mini-batch training with PyTorch Geometric and optimizes graph size, enabling scaling to thousands of drugs.
Data Imbalance	Affinity datasets often have imbalanced distributions (many medium-affinity pairs, few high/low-affinity pairs), which standard MSE loss does not address.	Incorporates **weighted sampling** during training to balance the affinity distribution and improve prediction on rare classes.
Generalization	Models trained on one dataset (e.g., Davis) often fail to generalize to new drugs or proteins without retraining.	Aims for high Concordance Index (>0.85) and Pearson correlation (>0.8) to ensure ranking and linear generalization to unseen data.

VI. ADVANTAGES

The proposed GNN-based Drug-Protein Interaction Prediction System offers the following advantages:

1. **Structural Awareness:** Unlike sequence-only models, the system captures 3D topological features of drug molecules and spatial relationships of protein residues using graph representations.
2. **High Predictive Accuracy:** Targets state-of-the-art performance with **CI > 0.85**, **MSE < 0.2**, and **Pearson > 0.8**, outperforming traditional ML and CNN-based methods.
3. **Handles Dense Data Efficiently:** The dual-branch GNN architecture with weighted sampling is specifically designed for dense interaction matrices (29,444 pairs), overcoming a key limitation of existing models.
4. **Cost-Effective Virtual Screening:** Reduces reliance on expensive and time-consuming wet-lab experiments by providing rapid computational predictions, saving millions in drug discovery costs.
5. **Drug Repurposing Support:** The model can generalize to new drugs and proteins, enabling identification of new therapeutic uses for existing drugs.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

6. **End-to-End Learning:** No manual feature engineering is required; the model learns hierarchical representations directly from SMILES strings and protein sequences.
7. **Scalability:** Mini-batch training and optimized graph processing allow scaling to large chemical libraries (thousands of drugs and proteins).
8. **Interpretability:** Attention mechanisms and visualization tools provide insights into which atoms or residues contribute most to binding affinity.

VII. CONCLUSIONS

The proposed Drug–Protein Interaction Prediction System using Graph Neural Networks successfully integrates molecular graph representations and protein sequence embeddings to predict binding affinity (pKd). The dual-branch GNN architecture overcomes key limitations of traditional ML/DL systems, including inadequate handling of dense datasets, limited multimodal fusion, and lack of structural information. By employing weighted sampling to address data imbalance and optimizing regression using MSE loss, the model achieves high predictive accuracy with target metrics of CI > 0.85, MSE < 0.2, and Pearson > 0.8.

This system accelerates early-stage drug discovery by enabling rapid, low-cost virtual screening of drug-protein interactions, reducing reliance on expensive wet-lab experiments. It supports drug repurposing and lead optimization, particularly for kinase-targeted therapies in oncology, neurology, and infectious diseases. Future work will focus on extending the model to incorporate 3D protein structures, integrating larger datasets like KIBA, and deploying the system as a web-based screening tool for the broader research community.

REFERENCES

[1] Öztürk H., Özgür A., Ozkirimli E., *"DeepDTA: deep drug–target binding affinity prediction,"* Bioinformatics, 2018.

[2] Nguyen T., Le H., Quinn T.P., et al., *"GraphDTA: predicting drug–target binding affinity with graph neural networks,"* Bioinformatics, 2021.

[3] Zhao L., Wang J., Pang L., et al., *"GANsDTA: Predicting Drug-Target Binding Affinity Using Generative Adversarial Networks,"* Journal of Chemical Information and Modeling, 2020.

[4] Huang K., Fu T., Glass L.M., et al., *"MolTrans: Molecular Interaction Transformer for drug–target interaction prediction,"* Bioinformatics, 2020.

[5] Li S., Zhou J., Xu T., et al., *"MONN: Multi-Objective Neural Network for Drug–Target Interaction Prediction,"* Bioinformatics, 2021.

[6] Zhang Y., Li H., Wang M., et al., *"DeepGS: Deep Learning of Graph Structure Improves Drug–Target Affinity Prediction,"* Computational and Structural Biotechnology Journal, 2020.

[7] Chen R., Liu X., Jin S., et al., *"PADME: A Deep Neural Network Model for Drug-Target Interaction Prediction,"* Journal of Cheminformatics, 2018.

[8] He T., Heidemeyer M., Ban F., et al., *"SimBoost: a similarity-based method for predicting drug–target interactions,"* BMC Bioinformatics, 2017.

[9] Pahikkala T., Airola A., Pietilä S., et al., *"KronRLS: Predicting drug–target interactions by kernel regression in least squares,"* BMC Bioinformatics, 2015.

[10] Wang Y., Li P., Yang J., et al., *"FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction,"* Bioinformatics, 2022.

[11] Zhao Q., Duan H., Xie Y., et al., *"AttentionDTA: prediction of drug–target affinity using attention-based bidirectional LSTM networks,"* Journal of Bioinformatics and Computational Biology, 2019.

[12] Tang Q., Nie F., Zhao Q., Chen W., *"Merged molecular representation model for blood–brain barrier permeability prediction,"* Molecular Informatics, 2022.

[13] Lin X., Quan Z., Wang Z.J., et al., *"KGNN: Knowledge Graph Neural Network for Drug–Drug Interaction Prediction,"* IEEE Access, 2020.

[14] Visonà G., Bouzigon E., Demenais F., Schweikert G., *"Network propagation for GWAS analysis: leveraging molecular networks for disease gene discovery,"* Briefings in Bioinformatics, 2024.

[15] Davis M. I., Hunt J. P., Herrgard S., et al., *"Comprehensive analysis of kinase inhibitor selectivity,"* Nature Biotechnology, 2011 (Dataset reference).