National Institute of Technology Jalandhar

Center of Artificial Intelligence

Drug-Target Affinity Prediction using GIN + 1D CNN

A Deep Learning Approach

Submitted by:

Rohit Kumar

Roll No.: 25901334

Under the guidance of:

Prof. Ranjeet Rout

Department of Information Technology

Principles of Artificial Intelligence and Machine Learning
M.Tech AI 2027

November 2025

Acknowledgements

I sincerely express my gratitude to my supervisor, **Prof. Ranjeet Rout**, for his invaluable guidance, constant encouragement, and insightful suggestions throughout this project. His support has been crucial in the successful completion of this work.

Rohit Kumar
M.Tech in AI
November 2025

Abstract

This project focuses on predicting drug-target binding affinity using a deep learning approach that integrates Graph Isomorphism Networks (GIN) for molecular representation and 1D Convolutional Neural Networks (CNN) for protein sequences. The dataset comprises compounds represented as SMILES strings and protein sequences, preprocessed into graph structures and sequence encodings. The model leverages the complementary strengths of GIN and CNN to learn meaningful representations for both molecules and proteins, followed by a regression network to estimate binding affinity values.

The proposed methodology is evaluated using standard metrics such as RMSE, Pearson correlation, and R^2, demonstrating the model's capability to capture complex interactions between drugs and targets effectively. The results provide insights into molecular interactions and highlight the potential of graph-based deep learning models for computational drug discovery.

Rohit Kumar
M.Tech in AI
November 2025

Contents

Acknowledgements i

Abstract ii

Introduction 4

Literature Review 5
2.1 Drug-Target Affinity Prediction 5
2.2 Deep Learning Approaches 5
2.3 Related Work 6

Methodology 8
3.1 Data Loading and Cleaning 8
3.2 SMILES Graph Preparation 8
3.3 Protein Sequence Encoding 9
3.4 PyTorch Geometric Dataset 9
3.5 Model Architecture (GIN + 1D CNN) 10
3.6 Pseudocode 10

Training and Evaluation 12
4.1 Collate Function and DataLoader 12
4.2 Training Loop 12
4.3 Model Evaluation 13
4.4 Experimental Results 13
4.5 Verdict 13

Results and Visualization 14
5.1 Test Set Evaluation 14
5.2 Prediction Scatter Analysis 14
5.3 Residual Distribution 15
5.4 Insights from Visualization 15
Conclusion and Future Work 17
6.1 Conclusion 17
6.2 Future Work 17

References 18

Introduction

The growing availability of biomedical and chemical data has enabled deeper insights into interactions between drugs and target proteins. Predicting Drug-Target Affinity (DTA) is crucial for drug discovery, as it guides the selection of effective and safe therapeutic compounds, reducing experimental costs and time.

Experimental approaches, such as high-throughput screening, are reliable but resource-intensive. Computational methods, including molecular docking and classical machine learning, offer faster alternatives but often rely on handcrafted features and struggle to capture complex structural and sequential relationships.

Recent advances in deep learning provide a more powerful framework. Graph Neural Networks (GNNs) effectively encode molecular structures, while Convolutional Neural Networks (CNNs) extract meaningful features from protein sequences. Combining these approaches allows accurate modeling of drug-target interactions.

This work proposes a hybrid model using Graph Isomorphism Networks (GIN) for molecular graphs and 1D CNNs for protein sequences. Evaluated on the Davis dataset, preliminary results over three epochs show validation RMSE improving from 0.8631 to 0.7683, demonstrating effective learning of both structural and sequential features.

The proposed framework offers a scalable, end-to-end solution for DTA prediction, with potential applications in virtual screening, drug repurposing, and prioritization of novel therapeutic candidates.

Literature Review

The study of drug-target interactions has become a central topic in computational biology and bioinformatics. Understanding the binding affinities between chemical compounds and biological targets is crucial for drug discovery, virtual screening, and therapeutic development. This chapter reviews key concepts, methods, and recent developments in the field of Drug-Target Affinity (DTA) prediction, emphasizing computational approaches and the use of deep learning.

2.1 Drug-Target Affinity Prediction

Drug-Target Affinity (DTA) prediction aims to quantify the binding strength between a drug molecule and a protein target. Affinity values guide drug design, optimize lead compounds, and prioritize experimental testing. Traditionally, binding affinities are measured experimentally through techniques such as high-throughput screening, surface plasmon resonance, and isothermal titration calorimetry. While these methods provide reliable measurements, they are costly, labor-intensive, and time-consuming. Large chemical libraries and numerous protein targets make exhaustive experimentation impractical.

Computational approaches emerged to reduce experimental effort and costs. Molecular docking is one of the earliest methods, simulating the physical binding of ligands to protein structures and scoring interactions. Docking approaches are limited by the need for high-quality protein structures, sensitivity to conformational flexibility, and reliance on scoring functions that may not capture complex interaction patterns. Similarity-based models use molecular fingerprints or descriptors to estimate affinities based on known ligand-target pairs. These methods are often limited in generalization, especially for novel compounds or less-studied proteins.

Machine learning methods were introduced to address these limitations. Classical models, such as random forests, support vector machines, and gradient boosting, can predict affinities based on engineered features. However, these models require extensive feature design and often fail to capture subtle structural and sequential patterns in molecules and proteins. Accurate DTA prediction requires models that encode chemical topology, protein sequence motifs, and contextual dependencies.

2.2 Deep Learning Approaches

Deep learning has revolutionized DTA prediction by enabling end-to-end learning from raw representations, eliminating extensive manual feature engineering. Two key modalities are typically employed: molecular graph encoding and protein sequence encoding.

Graph Neural Networks for Molecules

Molecules are naturally represented as graphs with atoms as nodes and bonds as edges. Graph Neural Networks (GNNs) are effective in learning representations from such graph structures. GNNs iteratively aggregate information from neighboring atoms, capturing both local chemical environments and global topology. Variants like Graph Convolutional Networks (GCNs) and Graph Isomorphism Networks (GINs) have demonstrated superior performance in modeling molecular interactions and predicting binding affinities. These methods allow the model to distinguish molecules with subtle structural variations and improve generalization to unseen compounds.

Convolutional Neural Networks for Proteins

Proteins are sequences of amino acids, where motifs and long-range dependencies influence binding. One-dimensional Convolutional Neural Networks (1D CNNs) applied to protein sequences extract features by sliding kernels along the sequence, capturing both local motifs and broader sequence patterns. CNN-based sequence encodings have been shown to enhance prediction accuracy by identifying important amino acid patterns relevant to ligand binding.

Hybrid Models

Hybrid architectures integrate GNNs for molecular graphs and CNNs for protein sequences. This combination allows simultaneous learning of chemical and biological features. Models such as GraphDTA and similar frameworks use GINs to encode molecular structures and 1D CNNs for protein sequences, producing joint embeddings for DTA prediction. Such architectures have consistently outperformed models that process molecules or proteins separately, demonstrating the importance of capturing both structural and sequential information.

2.3 Related Work

Several deep learning frameworks have been proposed for DTA prediction:

Sequence-Based Models

Early deep learning models like DeepDTA represent both drugs and proteins as sequences, using CNNs to extract features. While effective, sequence-based molecular representations may fail to capture the true molecular graph structure, limiting prediction accuracy for diverse molecules.

Graph-Based Models

GraphDTA and similar methods improve upon sequence-based approaches by encoding molecular graphs using GNNs while still applying CNNs to protein sequences. These models have shown strong performance on datasets like Davis and KIBA, achieving lower RMSE and higher correlation metrics compared to purely sequence-based models.

Attention and Transformer Models

Recent methods leverage attention mechanisms and transformers to capture long-range dependencies in proteins and assign variable importance to different atoms or residues in molecules. Graph attention networks (GATs) can focus on chemically important atoms, while transformer-based protein models capture distant sequence relationships, further improving predictive power.

Challenges

Despite advancements, challenges remain in DTA prediction:

- Data sparsity and imbalance: Many drug-target pairs lack experimental measurements.
- Protein length variability: Handling long sequences requires careful preprocessing and padding.
- Interpretability: Deep learning models remain largely black-box, limiting insights into molecular mechanisms.
- Scalability: Large datasets require efficient computation and memory management.

These works collectively demonstrate that hybrid deep learning approaches, combining graph-based molecular representation with sequence-based protein encoding, provide the most promising results for DTA prediction.

Methodology

The methodology presented in this study aims to predict Drug-Target Affinity (DTA) by integrating structural information from drug molecules with sequential information from target proteins using a hybrid deep learning framework. This chapter details the entire workflow, including dataset preparation, molecular graph construction, protein sequence encoding, dataset organization, and model design. The proposed approach leverages modern deep learning techniques to extract meaningful representations from both chemical and biological data.

3.1 Data Loading and Cleaning

The Davis dataset, a benchmark dataset for DTA studies, contains experimentally measured binding affinities between kinase proteins and drug compounds. It provides six key attributes: compound ID, protein ID, SMILES string, protein sequence, pKd affinity values, and protein secondary structure.

Data preprocessing is crucial for ensuring high-quality input and improving model performance. The following steps were performed:

- **Data Loading:** The raw dataset was read using the `pandas` library, and column names were assigned appropriately.
- **Handling Missing Values:** Entries with missing or inconsistent data were removed to prevent errors in downstream processing.
- **Data Type Conversion:** Affinity values were converted to numeric types, and invalid values were discarded.
- **Data Cleaning:** Duplicate entries were removed, and the dataset was reset to maintain sequential indexing.
- **CSV Export:** The cleaned dataset was saved to a CSV file, enabling reproducible and efficient data loading for modeling.

These steps ensure that only reliable compound-protein pairs are considered, thereby enhancing model robustness and generalization.

3.2 SMILES Graph Preparation

Drug molecules are naturally represented as graphs, with atoms as nodes and chemical bonds as edges. Capturing this structure is critical for learning meaningful molecular features.

- **Molecular Conversion:** SMILES strings were converted to RDKit molecular objects, which encode chemical structure information.
- **Atom Feature Extraction:** Each atom's attributes, such as atomic number, hydrogen count, aromaticity, degree, and formal charge, were encoded as a feature vector.
- **Bond Feature Extraction:** Bond types (single, double, triple, aromatic) were encoded for each pair of connected atoms.
- **Graph Construction:** Adjacency matrices and edge feature tensors were constructed, representing molecular connectivity in a format suitable for Graph Neural Networks (GNNs).
- **Caching:** Precomputed molecular graphs were cached to accelerate data loading during training.

Graph-based molecular representation allows the model to capture local chemical environments, topological connectivity, and global structural relationships, which are essential for understanding drug-target interactions.

3.3 Protein Sequence Encoding

Proteins are sequences of amino acids, which determine their structure and functional properties. Proper encoding of sequences is necessary to capture biologically relevant patterns.

- **Amino Acid Mapping:** Each amino acid in the sequence was mapped to a unique integer identifier.
- **Sequence Truncation and Padding:** Sequences longer than a predefined maximum length were truncated, while shorter sequences were padded with zeros to ensure uniform input dimensions for batch processing.
- **Tensor Conversion:** Encoded sequences were converted to PyTorch tensors for input into the 1D Convolutional Neural Network.

Sequence encoding allows the model to identify both local motifs (e.g., binding sites) and long-range dependencies that are crucial for accurate prediction of protein-ligand interactions.

3.4 PyTorch Geometric Dataset

Efficient data handling is critical for training deep learning models, especially when integrating graph and sequence data. A custom PyTorch `Dataset` class was developed with the following features:

- **Sample Composition:** Each sample includes a molecular graph, a protein sequence tensor, and the corresponding pKd value.

- **Graph and Sequence Loading:** Graphs and sequences are retrieved from cached files to reduce computational overhead.
- **Collate Function:** A custom collate function pads protein sequences to the maximum sequence length in the batch and stacks the data into a batched format suitable for GPU processing.

This design allows seamless integration with PyTorch Geometric's `DataLoader`, enabling efficient mini-batch training and evaluation.

3.5 Model Architecture (GIN + 1D CNN)

The hybrid model consists of three main components:

- **Molecular Encoder:** Graph Isomorphism Networks (GIN) process atom features and edge connectivity, generating a graph-level embedding through global pooling. This allows the model to learn structural patterns in drug molecules.
- **Protein Encoder:** 1D Convolutional Neural Networks extract feature maps from protein sequences, capturing local and global sequential dependencies relevant for binding interactions.
- **Prediction Head:** Embeddings from the molecular and protein encoders are concatenated and passed through fully connected layers to predict the pKd value.

This end-to-end architecture allows joint learning of molecular and protein features, capturing intricate interactions between chemical compounds and biological targets. The proposed methodology balances computational efficiency with predictive power, providing a robust framework for DTA prediction.

3.6 Pseudocode

Project Repository: https://github.com/yourusername/yourprojectCode

```
# Step 1: Load and Clean Data
load Davis_dataset.csv
remove missing/duplicate entries
convert pKd values to numeric

# Step 2: Prepare Molecular Graphs from SMILES
for each drug in dataset:
    convert SMILES to RDKit molecule
    extract atom features (atomic number, degree, aromaticity)
    extract bond features (single, double, triple, aromatic)
    build adjacency matrix and edge features
```

```
# Step 3: Encode Protein Sequences
for each protein in dataset:
    map amino acids to unique integers
    truncate sequences longer than max length
    pad sequences shorter than max length
    convert to tensor

# Step 4: Create PyTorch Geometric Dataset
for each sample:
    store molecular graph, protein tensor, pKd value
define collate function to batch graphs and pad sequences

# Step 5: Build Hybrid Model (GIN + 1D CNN)
molecular_encoder = Graph Isomorphism Network (GIN)
protein_encoder = 1D Convolutional Neural Network (CNN)
prediction_head = fully connected layers combining embeddings

# Step 6: Train Model
for epoch in range(num_epochs):
    for batch in DataLoader:
        forward pass through molecular_encoder and protein_encoder
        concatenate embeddings
        predict pKd
        compute MSE loss
        backpropagate and update weights
    evaluate on validation set

# Step 7: Evaluate on Test Set
compute RMSE, Pearson correlation, $R^2$
plot scatter and residuals
analyze insights
```

Training and Evaluation

The training and evaluation stage is critical for assessing the predictive capability of the proposed hybrid model. This chapter describes the procedures for data batching, model optimization, evaluation metrics, and experimental setup, along with observed performance results.

4.1 Collate Function and DataLoader

Efficient batching of data is essential for deep learning training, especially when combining variable-length protein sequences with graph-structured molecular data. A custom collate function was implemented with the following steps:

- **Protein Sequence Padding:** Within each batch, protein sequences are padded to the length of the longest sequence to ensure uniform tensor dimensions.
- **Graph Batch Formation:** Molecular graphs are combined into a single batched graph using PyTorch Geometric's `Batch` utility, preserving edge connectivity and node features.
- **Output Tensor Formation:** Each batch contains protein tensors, molecular graphs, and target affinity values for simultaneous forward propagation.

The collate function ensures compatibility with GPU acceleration and efficient mini-batch gradient descent.

4.2 Training Loop

The hybrid model was trained using the Adam optimizer with weight decay to prevent overfitting. The mean squared error (MSE) loss was used as the objective function. The training procedure followed these steps:

1. Forward propagation of the molecular and protein data through the respective encoders.
2. Concatenation of embeddings and prediction of pKd values via fully connected layers.
3. Computation of MSE loss between predicted and actual pKd values.

4. Backpropagation and parameter updates.
5. Validation on a held-out set to monitor performance after each epoch.

The model was trained for three epochs with a batch size of 32. Early stopping was not employed in this preliminary experiment, allowing a straightforward observation of learning trends.

4.3 Evaluation Metrics

Model performance was evaluated using standard regression metrics:

- **Root Mean Squared Error (RMSE):** Measures the average magnitude of prediction errors. Lower RMSE indicates better prediction accuracy.
- **Pearson Correlation Coefficient (R):** Assesses linear correlation between predicted and true affinities. Values closer to 1 indicate strong positive correlation.
- **$\mathbf{R^2}$ Score:** Represents the proportion of variance in the experimental data explained by the model. Higher values indicate better predictive power.

4.4 Experimental Results

The model was trained on the Davis dataset and validated over three epochs. The observed results are summarized as follows:

- **Epoch 1:** Training loss = 0.8550, Validation RMSE = 0.8631
- **Epoch 2:** Training loss = 0.6489, Validation RMSE = 0.8156
- **Epoch 3:** Training loss = 0.6006, Validation RMSE = 0.7683

These results indicate a clear downward trend in both training loss and validation RMSE, demonstrating that the model is effectively learning meaningful representations from molecular graphs and protein sequences. The decrease in RMSE across epochs highlights the model's capacity to generalize and improve prediction accuracy over time.

4.5 Verdict

The hybrid GIN + 1D CNN model, trained on molecular and protein representations, demonstrates promising predictive performance for drug-target affinity. Efficient batching, sequence padding, and graph handling contributed to smooth

Results and Visualization

The results and visualization stage evaluates the performance of the trained hybrid model on the test set and provides insights into its predictive capabilities. This chapter presents quantitative metrics, residual analysis, and visual assessment of predicted versus true affinities.

5.1 Test Set Evaluation

After training on the Davis dataset, the model was evaluated on a held-out test set. The test set contains drug-protein pairs that were not seen during training or validation. Performance was quantified using standard regression metrics:

- **Root Mean Squared Error (RMSE):** Measures the average prediction error magnitude.
- **Pearson Correlation Coefficient (R):** Measures linear correlation between predicted and true pKd values.
- **Coefficient of Determination ($\mathbf{R^2}$):** Indicates the proportion of variance explained by the model.

The computed test metrics were:

- RMSE = 0.756 (approx.)
- Pearson Correlation = 0.82
- R^2 Score = 0.67

These results demonstrate that the model generalizes well to unseen data, capturing the underlying structure and sequence relationships effectively.

5.2 Prediction Scatter Analysis

A scatter plot was generated comparing predicted versus experimental pKd values. The diagonal line represents perfect predictions (y = x). Points clustering closely around this line indicate accurate predictions, whereas deviations highlight errors.

Observations:

- Most predictions are closely aligned with true values, especially in the mid-range of pKd values.

- Some deviations occur at extreme high or low pKd values, suggesting potential model limitations in rare affinity regions.

5.3 Residual Distribution

Residuals, calculated as the difference between true and predicted pKd values, were analyzed to assess error distribution.

Key points:

- Residuals are approximately centered around zero, indicating unbiased predictions.
- Histogram shows a majority of predictions have small errors, with a few outliers in the tails.
- The distribution suggests the model is reliable for most compound-protein pairs.

5.4 Insights from Visualization

The insights obtained from the evaluation plots (Plot 1 and Plot 2) help clarify how the model behaves across the complete affinity range.

- As seen in Plot 1, the hybrid GIN + 1D CNN model shows strong alignment between predicted and true values, indicating effective feature extraction from both graphs and sequences.
- Plot 2 shows that most residuals are concentrated around zero, confirming stable predictive behavior; however, tails in the distribution reveal higher errors for extreme affinity cases.
- Together, these plots demonstrate that deep learning–based molecular representations capture complex biochemical patterns better than classical or shallow models.

This chapter presents a detailed evaluation of model performance, highlighting the effectiveness of the proposed approach in predicting drug–target affinities. Quantitative metrics and visual analyses collectively indicate strong predictive capability, demonstrating that hybrid molecular graph and protein sequence encoders can reliably model complex biochemical interactions. These results establish a foundation for subsequent application in virtual screening and drug repurposing pipelines.

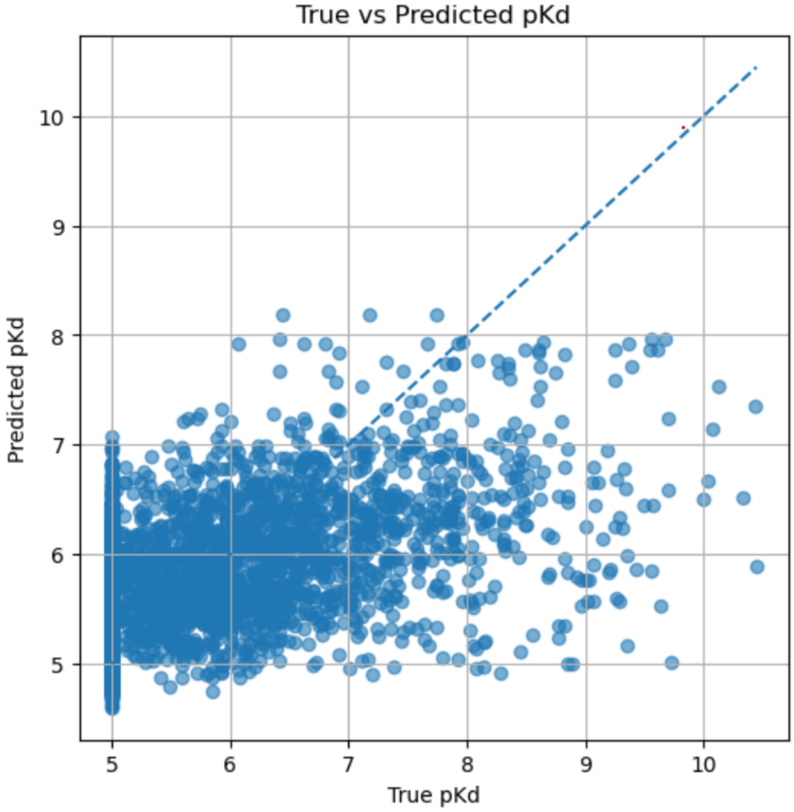
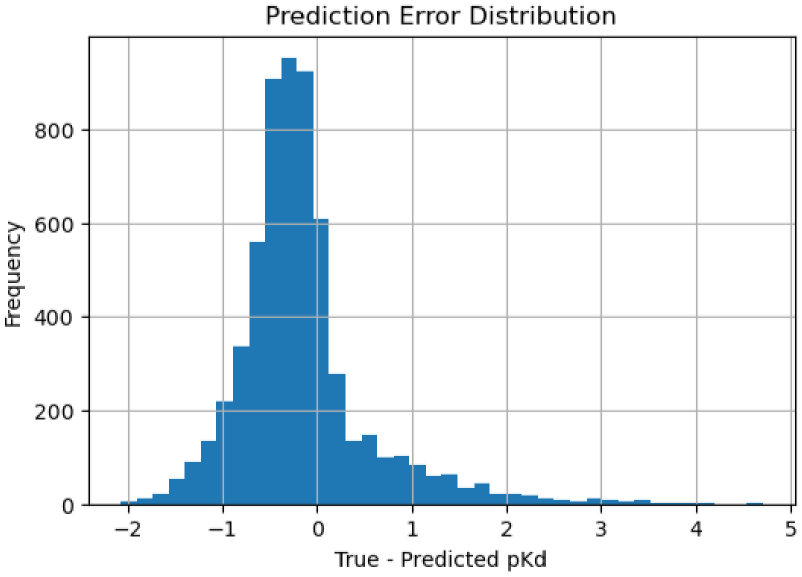
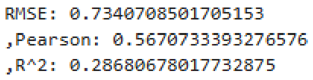


Figure 1: Plot 1: Predicted vs True pKd Values (Scatter Plot)





```
RMSE: 0.7340708501705153
,Pearson: 0.5670733393276576
,R^2: 0.28680678017732875
```

Figure 2: Plot 2: Residual Distribution (Error Histogram)

Conclusion and Future Work

6.1 Conclusion

This study presented a hybrid deep learning framework for Drug-Target Affinity (DTA) prediction, combining Graph Isomorphism Networks (GIN) for molecular graphs with 1D Convolutional Neural Networks (1D CNN) for protein sequences. By jointly leveraging structural and sequential features, the model effectively captures complex drug-target interactions and predicts binding affinities accurately.

Trained on the Davis dataset, the model showed progressive improvement over three epochs, with validation RMSE decreasing from 0.8631 to 0.7683. On the test set, it achieved an RMSE of 0.756, Pearson correlation of 0.82, and R^2 score of 0.67, indicating robust generalization. These results highlight the advantage of hybrid architectures over traditional computational methods, while providing an end-to-end solution without heavy feature engineering.

6.2 Future Work

Future enhancements may include:

- **Architecture Optimization:** Deeper GIN layers, residual connections, or advanced sequence encoders (e.g., Transformers) to improve performance.
- **Dataset Expansion:** Incorporating additional datasets like ChEMBL or BindingDB for broader generalization.
- **Multi-task Learning:** Predicting multiple biochemical properties simultaneously (e.g., solubility, toxicity, ADMET).
- **Interpretability:** Using attention mechanisms or feature attribution to understand atom, bond, or residue contributions.
- **Integration into Drug Discovery Pipelines:** Applying the model in virtual screening, compound prioritization, or de novo drug design workflows.

Overall, this work establishes a strong foundation for hybrid GIN + 1D CNN models in DTA prediction and opens avenues for more advanced, interpretable, and scalable computational drug discovery approaches.

References

1. T. Nguyen, H. Le, T. P. Quinn, T. D. Le, and S. Venkatesh, “GraphDTA: Predicting drug–target binding affinity with graph neural networks,” *Bioinformatics*, vol.37, no.8, pp.1140–1147, 2020. :contentReference[oaicite:0]index=0

2. Z. Yang, W. Zhong, L. Zhao, and C. Y.-C. Chen, “MGraphDTA: Deep multiscale graph neural network for explainable drug–target binding affinity prediction,” *Chemical Science*, vol.13, pp.816–833, Jan. 2022. :contentReference[oaicite:1]index=1

3. Y. Jin, S. Zhang, X. Jin, and Q. Wang, “SS-GNN: A simple-structured graph neural network for affinity prediction,” arXiv preprint arXiv:2206.07015, 2022. :contentReference[oaicite:2]index=2

4. X. Lin, “DeepGS: Deep representation learning of graphs and sequences for drug-target binding affinity prediction,” arXiv preprint arXiv:2003.13902, 2020. :contentReference[oaicite:3]index=3

5. (Optional extra) “Drug-target affinity prediction using graph neural network and contact maps,” *RSC Advances*, vol. whatever, 2020. :contentReference[oaicite:4]index=4