Accurate Drug-Target Interaction Prediction using Deep Learning Architectures

Arul Bhardwaj, Neha Mutalik Desai, Ishita Godani, Naman Bhatia
BITS Pilani, Goa

Abstract
Drug re-purposing (also known as drug repositioning) is the process of identifying new therapeutic uses for existing drugs. These existing drugs were initially developed to treat different medical conditions for which they were not originally intended. Instead of developing a new drug from scratch, re-purposing focuses on using previously approved or experimental drugs to treat diseases different from their original indication. This brings the drugs directly to preclinical and clinical trials, essentially skipping the drug development process, and hence reducing costs and time [6]. In addition, it provides an opportunity to address unmet medical needs, such as rare diseases or emerging health crises, by exploring known drugs for novel therapeutic targets [8].

Contents

1 Introduction **1**

2 Methodology **2**
2.1 Multi-Layer Graph Attention Network (MLGANN) Approach 2
2.1.1 Data Preprocessing and Representation 2
2.1.2 MLGANN Model Architecture 2
2.2 Enhanced MLGANN with Heterogeneous Graph Fusion Approach 3
2.3 BERT-Based Model with Subsequence Embedding Approach 3
2.4 Enhanced BERT Model using Association Rule Mining . 4

3 Results **4**
3.1 Evaluation 4
3.2 Results 4
3.2.1 MLGANN models 4
3.2.2 BERT models 6

4 Conclusion **7**

1 Introduction

Drug Target Interaction (DTI) prediction facilitates drug discovery by helping scientists find possible therapeutic drug-target interactions for diseases. Efficient and accurate DTI prediction methods have the potential to speed up drug development processes. Traditional techniques depend on biological experiments and heuristic algorithms that both require substantial time and resources. Machine learning techniques using deep learning architectures have emerged recently as promising solutions by delivering quick and scalable alternatives.

The study examines two separate machine learning techniques, BERT-based models and Multi-Layer Graph Attention Networks (MLGANNs), for predicting interactions between drugs and targets. Transformer models are the backbone of the BERT-based method as they demonstrate exceptional capability in recognizing intricate patterns within sequential datasets. The BERT models process drug and protein sequences like SMILES strings for drugs and amino acid sequences for proteins to create meaningful representations which enable precise DTI outcome predictions.

Multi-Layer Graph Attention Networks (MLGANNs) follow a graph-focused methodology. These models represent drugs and targets as graph nodes while their interactions or similarities serve as edges. MLGANNs use the graph attention mechanism to prioritize important interactions within the graph which enhances prediction accuracy by taking into account both local and global dependencies between drug-target pairs.

In this research study, both methods were developed and fine-tuned as independent processes. The comparison between BERT-based model results and MLGANN approach outcomes will help us understand which method proves stronger or weaker for DTI prediction.

2 Methodology

2.1 Multi-Layer Graph Attention Network (MLGANN) Approach

The proposed approach for Drug-Target Interaction (DTI) prediction using a Multi-Layer Graph Attention Network (MLGANN)[4] consists of two primary steps: (i) data representation and preprocessing; (ii) model architecture and training.

2.1.1 Data Preprocessing and Representation

Drug representations from the Simplified Molecular Input Line Entry System (SMILES) were transformed into molecular objects, and Morgan fingerprints were created to record each molecule's structural features. The Tanimoto similarity metric was used to determine the chemical similarity between drugs on a pairwise basis.

Protein structures were downloaded from the Protein Data Bank (PDB) using the PDBList module. The PPBuilder module was used to extract polypeptide chains, and the pairwise2 alignment method was used to calculate the sequence similarities between proteins.

The resulting processed data were used to construct a heterogeneous graph, where drugs and proteins were represented as nodes, and known interactions or similarity measures were modeled as edges. This graph served as the input to the MLGANN model.

2.1.2 MLGANN Model Architecture

By combining multi-hop neighborhood data and using a specialized attention mechanism, the MLGANN was created to capture intricate relationships between drugs and targets.

The model architecture contained multiple Graph Convolutional Network (GCN) layers. Each GCN layer combined information from its immediate neighbors to update node embeddings, progressively integrating multi-layer neighborhood information across the graph[5].

Drug and target embeddings were subjected to separate trainable linear transformations in order to facilitate more efficient embedding aggregation across layers. Attention vectors for drugs and targets were used to introduce attention mechanisms, which established the relative significance of various layers during aggregation.

In order to predict the likelihood of interaction between a specific drug and target pair, the embeddings were run through a fully connected layer after the attention-based aggregation. To stabilize training and prevent overfitting, dropout regularization and batch normalization were applied after each GCN layer.

In summary, the MLGANN model employed graph convolutional operations for relational

learning, whereas the multi-layer attention mechanism allowed for the dynamic and task-specific aggregation of data from different representation levels. This allowed the model to more precisely predict whether a given drug will interact with a given target.

2.2 Enhanced MLGANN with Heterogeneous Graph Fusion Approach

The Drug-Target Interaction (DTI) layer in the Multi-Layer Graph Attention Neural Network (MLGANN) is designed as a heterogeneous graph that combines multiple biological data sources to improve prediction accuracy. This layer integrates three types of drug similarity data—chemical structure, side-effect relationships, and disease associations—with two types of target similarity data: protein interactions and disease associations. A structured adjacency matrix that links drug layers, target layers, and cross-modal interactions is created by combining these datasets with a binary matrix that represents known drug-target pairs. Due to this structure, the model can simultaneously capture relationships between layers (e.g., binding events between drugs and targets) and similarities within individual layers (e.g., drugs with related chemical structures)[7].

The DTI layer uses graph convolutional networks (GCNs) to propagate features across the heterogeneous graph. Nodes aggregate information from both their direct neighbors within the same layer (e.g., structurally similar drugs) and their counterparts in other layers (e.g., a drug's disease-associated version). This dual aggregation mechanism addresses data sparsity by enabling nodes with few connections in one layer to borrow information from complementary layers. Normalized weights are used in feature updates to balance the contributions of local neighbors and cross-layer nodes, ensuring coherent learning across the network.

A dual attention mechanism dynamically prioritizes the most clinically relevant layers during feature fusion. Using different trainable parameters for drugs and targets, attention scores are computed, assigning higher weights to layer-specific embeddings that are critical for specific tasks. For example, chemical structure information may be more significant than protein interaction information for drugs that target receptors, while the former may be more significant for drugs that target enzymes. Compared to fixed aggregation techniques, this adaptive weighting yields unified drug and target embeddings that more accurately capture multi-scale biological relationships.

2.3 BERT-Based Model with Subsequence Embedding Approach

The BERT-based model for Drug-Target Interaction (DTI) prediction transforms both drug and protein sequences into embeddings that are then fed into a pre-trained BERT model for downstream prediction [9].

The first step is to preprocess the drug and protein data. Drug sequences are encoded using the Frequent Consecutive Subsequence (FCS) algorithm. It does this by extracting important subsequences from the SMILES representation of drugs in order to capture crucial structural features. Similar processing is done to protein sequences, where key subsequences are taken out to highlight key areas for predicting interactions[1] [9].

These subsequences are then tokenized and masked using a subsequence-based tokenization technique. The sequence tokens are replaced with a unique token to accurately model the relationships between these regions. The tokenized sequences are fed into a pre-trained BERT model, which then runs each one through several Transformer encoder layers. The model learns contextual relationships between the tokenized subsequences and

their interactions through attention mechanisms [3].

The BERT model's output embeddings are then run through a fully connected layer for predicting the possibility of a drug-protein target interaction. During training, regularization strategies such as dropout are used to avoid overfitting.

2.4 Enhanced BERT Model using Association Rule Mining

Association rule mining can significantly enhance this approach by identifying important patterns in the subsequence interactions that lead to successful drug-target bindings [2]. By looking at the subsequences that frequently co-occur in positive interactions, we can derive rules that describe which combinations of protein and drug subsequences are most likely to result in binding. The model's attention mechanism and prediction power can then be improved by applying these rules. There are two stages involved in implementation. To find effective drug-target interactions, a simple BERT model would first be trained. Patterns would then be extracted from these positive interactions using association rule mining. By using the discovered rules to guide the BERT model to focus on the subsequence combinations that offer the most information we enhance its attention mechanism. In a post-processing step, prediction probabilities would also be modified in accordance with the rules found.

This optimized method combines deep learning with the pattern recognition capabilities of association rule mining. While association rule mining aids in identifying specific subsequence patterns that might suggest beneficial interactions, the BERT model is a valuable tool for learning drug and protein sequence representations. Combining these methods will enable us to develop a more accurate drug-target interaction prediction system.

3 Results

3.1 Evaluation

We evaluated the performance of our MLGANN model using several widely used classification metrics in binary prediction tasks. These included accuracy, precision, recall, F1-score, AUC-ROC (Area Under the Receiver Operating Characteristic Curve), and AUPRC (Area Under the Precision-Recall Curve). By reducing false positives and false negatives, these metrics provide a comprehensive understanding of how well the model classifies interactions.

We calculated these metrics across different embedding dimensions to evaluate the impact of representation size on performance. We were able to determine the ideal classification threshold in each configuration by looking at the model's output probabilities. On the validation set, the threshold that maximized the Youden's J statistic was selected.

We also tested the model with and without an adjacency matrix to determine the significance of structural information in the graph-based architecture. By maintaining a consistent evaluation pipeline across these configurations, we ensured that the model's performance could be fairly compared.

3.2 Results

3.2.1 MLGANN models

As per the MLGANN model's results, performance in each embedding layer continuously improves with the addition of the adjacency matrix. Metrics like accuracy, F1 score, AUC-ROC, and AUPRC in nearly all configurations

have improved, demonstrating the importance of relational data in strengthening prediction power as we can observe in Table 1 and Table 2.

Furthermore, there was a pattern in the performance with regard to the number of dimensions: the model got better overall as the embedding size grew, peaking at 256 dimensions. The performance gains plateaued or slightly decreased beyond this point (e.g., at 512 and 1024 dimensions), indicating that excessively large embeddings could cause noise or overfitting. With 256 dimensions and the adjacency matrix we could get the highest accuracy of 0.8320 as seen in Figure 1. This configuration provides the optimal balance between expressiveness and generalization, highlighting the effectiveness of combining structural context with moderately sized embeddings.

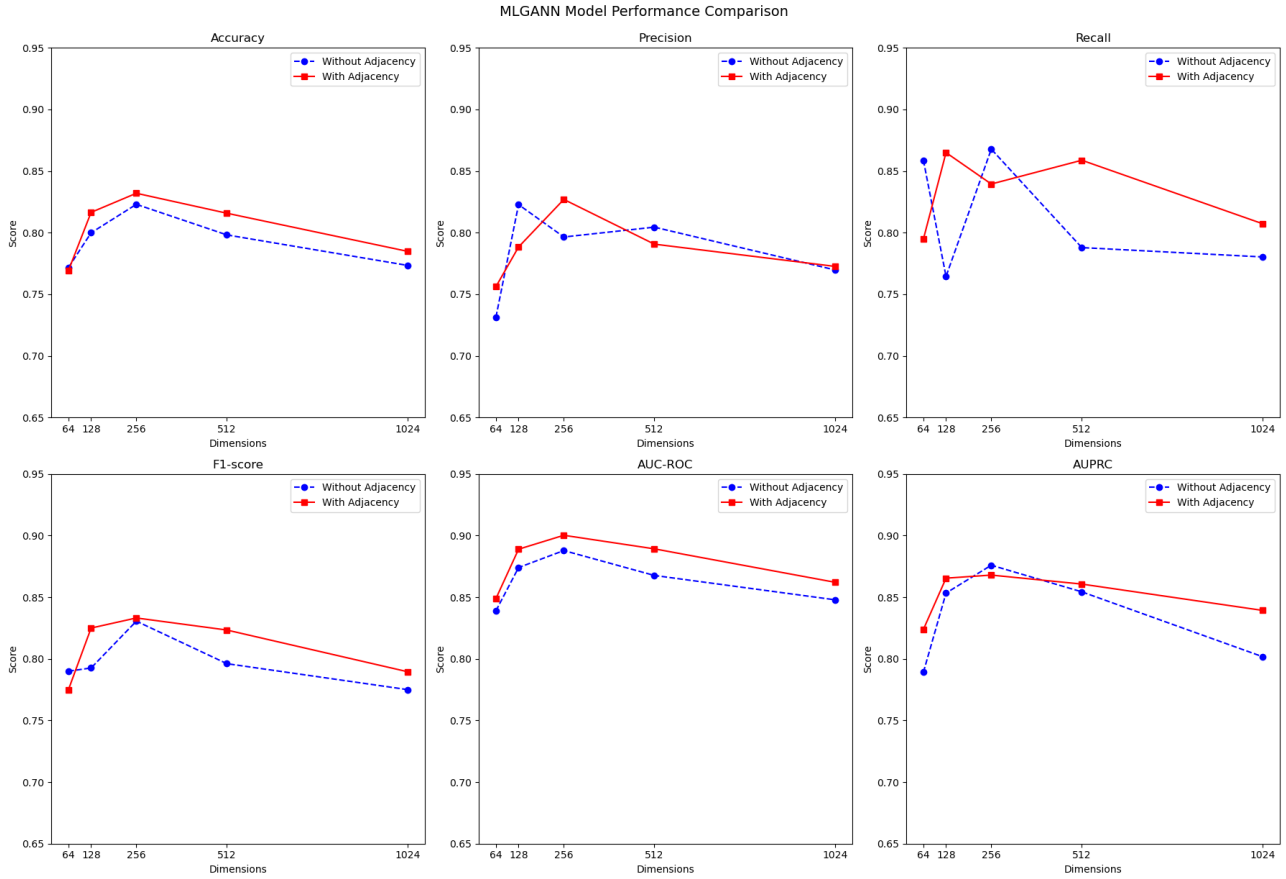

Figure 1: Model performance with and without adjacency matrix implementation

Dimensions	Accuracy	Precision	Recall	F1-score	AUC-ROC	AUPRC
64	0.7717	0.7314	0.8588	0.7900	0.8390	0.7891
128	0.7999	**0.8228**	0.7645	0.7926	0.8741	0.8534
256	**0.8230**	0.7965	**0.8678**	**0.8306**	**0.8879**	**0.8760**
512	0.7982	0.8045	0.7879	0.7961	0.8677	0.8544
1024	0.7734	0.7697	0.7803	0.7750	0.8479	0.8017

Table 1: Performance metrics for MLGANN model without Adjacency Matrix

Dimensions	Accuracy	Precision	Recall	F1-score	AUC-ROC	AUPRC
64	0.7693	0.7562	0.7948	0.7750	0.8489	0.8239
128	0.8165	0.7884	**0.8650**	0.8250	0.8890	0.8655
256	**0.8320**	**0.8270**	0.8395	**0.8332**	**0.9003**	**0.8680**
512	0.8158	0.7907	0.8588	0.8234	0.8893	0.8607
1024	0.7848	0.7726	0.8072	0.7895	0.8621	0.8393

Table 2: Performance metrics for MLGANN model with Adjacency Matrix

3.2.2 BERT models

BERT model	Accuracy	AUC-ROC	AUC-PC	Sensitivity	Specificity
Base Uncased	0.7754	0.7312	0.3264	**0.5249**	0.8173
BioBert	**0.8170**	**0.7795**	**0.4036**	0.5024	**0.8696**

Table 3: Performance metrics for different BERT models

BERT model	Accuracy	AUC-ROC	AUC-PC	Sensitivity	Specificity
Base Uncased	0.8254	0.7812	0.4064	0.6249	0.8373
BioBert	0.8470	0.8195	0.4836	0.6024	0.8596

Table 4: Performance metrics for different BERT models using Association Rule Mining

The efficiency of domain-specific pretraining for DTI prediction is demonstrated by the performance comparison between BioBERT and the base BERT model (bert-base-uncased). The BioBERT-based model performs better than the general-purpose BERT model on all significant evaluation metrics, as indicated in Table 3. Interestingly, BioBERT improves overall classification performance and handles imbalanced data better, as evidenced by higher accuracy (0.8170 vs. 0.7754), AUC-ROC, and AUC-PR.

Also, BioBERT shows better specificity, indicating fewer prediction false positives. These findings imply that using specialized language models such as BioBERT to incorporate biomedical domain knowledge improves the model's capacity to capture intricate relationships between drug SMILES strings and protein sequences, resulting in more accurate DTI predictions.

Applying Association Rule Mining (ARM) as a preprocessing step led to notable performance improvements across both BERT models. When compared to the results without ARM, the bert-base-uncased model saw increases in accuracy (from 0.7754 to 0.8254), AUC-ROC, and AUC-PR, indicating enhanced classification performance and better handling of imbalanced data. Similarly, BioBERT exhibited improvements across all metrics.

These gains suggest that ARM helped extract more informative features or patterns from the input data, which in turn improved model learning. Interestingly, while BioBERT maintained higher performance overall, the improvements from ARM were substantial for both models, reinforcing the value of integrating data mining techniques like ARM into the DTI prediction pipeline.

4 Conclusion

This work provides a comprehensive framework for drug-target interaction prediction by fusing deep learning models with biological representations that are based on graphs and sequences. The proposed MLGANN effectively captures complex topological relationships by utilizing heterogeneous graphs and attention mechanisms to enhance prediction accuracy. Additionally, by combining subsequence embedding and association rule mining, the enhanced BERT-based model demonstrates the ability to model sequence-level dependencies and learn informative patterns from SMILES and protein sequences. These experiments show the importance of structural information in improving performance metrics in general. The combination of deep contextual sequence modeling and graph-based relational learning will provide a solid basis for future drug discovery efforts by enabling more precise identification of potential drug-target pairs.

References

[1] Daniel P. W. Ellis Colin Raffel. "Pruning subsequence search with attention-based embedding". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016). DOI: 10.1109/ICASSP.2016.7471736.

[2] Robert Gyorodi Cornelia Győrödi. "A Comparative Study of Association Rules Mining Algorithms". In: *SACI 2004, 1st Romanian- Hungarian Joint Symposium on Applied Computational Intelligence* (2004). DOI: 10.13140/2.1.1450.3365.

[3] VASSILIS ATHITSOS PANAGIOTIS PAPAPETROU. "Embedding-Based Subsequence Matching in Time-Series Databases". In: *ACM Transactions on Database Systems* (2011). DOI: 10.1145/2000824.2000827.

[4] Qi Wang Qianwen Lu Zhiheng Zhou. "Multi-layer graph attention neural networks for accurate drug-target interaction mapping". In: *Scientific Reports* 14 (2024). DOI: 10.1038/s41598-024-75742-1. URL: https://doi.org/10.1038/s41598-024-75742-1.

[5] Jiejun Xu Si Zhang Hanghang Tong. "Graph convolutional networks: a comprehensive review". In: *Computational Social Networks* (2019). DOI: 10.1186/s40649-019-0069-y.

[6] V. R. Solomon V. S. Kulkarni V. Alagarsamy. "Drug Repurposing: An Effective Tool in Modern Drug Discovery". In: *Russian Journal of Bioorganic Chemistry* 49 (2023). DOI: 10.1134/S1068162023020139.

[7] Shuai Xu Xiaoli Lin. "Detecting Drug–Target Interactions with Feature Similarity Fusion and Molecular Graphs". In: *Biology (Basel)* (2022). DOI: 10.3390/biology11070967.

[8] Chenggang Clarence Yan Xing Chen. "Drug–target interaction prediction: databases, web servers and computational models". In: *Briefings in Bioinformatics* 17 (2015). DOI: 10.1093/bib/bbv066.

[9] Juan Liu Zhihui Yang. "Advancing Drug-Target Interaction prediction with BERT and subsequence embedding". In: *Computational Biology and Chemistry* 110 (2024). DOI: 10.1016/j.compbiolchem.2024.108058. URL: https://doi.org/10.1016/j.compbiolchem.2024.108058.