# Leveraging Network Embeddings and Deep Learning for Drug-Disease Interaction Prediction

**Abstract**

The identification of potential drug-disease interactions (DDIs) is a critical step in drug repositioning and discovery. Traditional wet-lab experiments are time-consuming and costly, necessitating efficient computational approaches. In this study, a deep learning framework is proposed that integrates network-based embeddings with a fully connected neural network (FCN) to predict DDIs. Pre-trained embeddings for drugs and diseases are utilized and concatenated to form high-dimensional feature vectors. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is employed. The model, a 5-layer Multi-Layer Perceptron (MLP), achieved a validation accuracy of 95.00% and an F1-score of 0.9512. The results demonstrate the efficacy of combining topological features with deep learning for robust DDI prediction.

## 1 Introduction

The development of novel pharmaceutical agents is an arduous and capital-intensive endeavor, typically spanning over a decade with expenditures exceeding billions of dollars. Consequently, drug repositioning—the identification of new therapeutic indications for existing drugs—has emerged as a pragmatic strategy to mitigate safety risks and curtail development costs. Within this paradigm, the concept of "network targets" has garnered significant attention, shifting the focus from individual molecular targets to disease-associated biological networks [1]. This theoretical framework posits that pathological states arise from perturbations in complex biological systems; thus, effective therapeutic interventions must modulate the disease network as a holistic entity. These network targets encompass diverse molecular components, including proteins, genes, and pathways, whose dynamic interactions govern disease progression and therapeutic response.

A pivotal component of network-based drug repositioning is the accurate prediction of drug-disease interactions (DDIs). While methodologies leveraging the "guilt-by-association" principle have shown promise, traditional machine learning approaches frequently encounter impediments due to the high dimensionality and inherent sparsity of biological networks. Furthermore, the pronounced imbalance between known and unknown associations poses a significant obstacle, necessitating robust strategies for negative sample selection. Recent strides in graph representation learning and deep learning offer robust solutions to these challenges. By generating low-dimensional embeddings that preserve topological structural integrity, complex network information can be encoded into feature-rich representations suitable for non-linear analysis.

In this study, we present a comprehensive computational framework for DDI prediction. We construct a dataset wherein drug-disease pairs are characterized by 19,292-dimensional vectors derived from network embeddings. To rectify the prevalent class imbalance in DDI data, the Synthetic Minority Over-sampling Technique (SMOTE) is employed. Subsequently, a deep neural network is trained to classify interactions. This methodology demonstrates state-of-the-art performance, underscoring the efficacy of integrating deep learning with network biology to accelerate drug repositioning initiatives.

## 2 Methodology

### 2.1 Data Preparation and Representation

The foundation of the proposed model rests upon the robust representation of drugs and diseases. We utilized pre-computed embeddings that capture the topological properties of drugs and diseases within a heterogeneous biological network. Specifically, each drug is represented by a feature vector derived from its structural and network properties using the InfoGraph algorithm [4], with chemical structures denoted by SMILES notation [2]. Similarly, diseases are encoded into vectors capturing their associations with genes and phenotypes using MeSH descriptors [3]. For each drug-disease pair, the respective drug and disease embeddings were concatenated to form a unified input feature vector of dimension $d = 19,292$. This high-dimensional representation ensures that the model has access to comprehensive information regarding both entities.

Furthermore, the node2vec method was consistently applied during the embedding process of the Protein-Protein Interaction (PPI) network from the STRING database, adhering to the same protocol used for the MeSH network. Each gene in the PPI network was transformed into a 300-dimensional vector, maintaining identical embedding dimensions, node sampling frequency, walk length, and return and exploration parameters as established for the MeSH network. Crucially, the PPI network serves as the underlying topology for characterizing drug-gene interactions. The genes within the network act as nodes that propagate the therapeutic effects of drugs. Consequently, the drug representations are constructed based on their influence across the gene network, effectively integrating the biological context of the PPIs into the model. This gene-centric approach ensures that the model captures the systemic impact of drugs, leveraging the intricate connectivity of the PPI network to predict potential therapeutic outcomes.

## 2.2 Drug Target Prediction

Prior to the interaction prediction, a drug target prediction module is employed to identify potential gene targets for the drugs. This process utilizes a dedicated classifier implemented as a neural network. The architecture comprises an input layer of size 600, followed by two hidden layers with 1024 neurons each, and an output layer designed to classify interactions into three distinct categories (inhibition, no interaction, activation). The network incorporates Batch Normalization layers to stabilize learning, ELU activation functions for non-linearity, and Dropout layers with a probability of 0.2 to prevent overfitting. During inference, the model outputs probability scores via a softmax function. Specific thresholds are applied to these probabilities—for instance, a probability greater than 0.88 for the first class or greater than 0.45 for the third class—to determine the final interaction type, thereby generating a profile of drug-gene interactions.

## 2.3 Random Walk on Biological Networks

To capture the propagation of drug effects through biological pathways, a random walk algorithm inspired by Node2Vec is implemented on a non-specific human signaling network. This process involves traversing the graph starting from specific gene nodes identified in the target prediction phase. The walk is governed by parameters such as the restart probability ($\alpha$), a decay factor, and the maximum walk length. Unlike standard random walks, this implementation incorporates directionality and sign (promotion or inhibition). As the walker traverses the graph, the value of visited nodes is updated based on the edge type ('Pos' or otherwise) and a decaying signal intensity. If a restart condition is met, the walker returns to the starting node. This mechanism allows for the simulation of signal transduction and the quantification of gene influence within the network, enriching the feature space with biological context.

## 2.4 Handling Class Imbalance with SMOTE

Real-world DDI datasets are inherently imbalanced, with known interactions (positive samples) significantly outnumbered by unknown or non-interactions (negative samples). Training on such skewed data can lead to models that are biased towards the majority class. To mitigate this, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE generates synthetic examples for the minority class (positive interactions) by interpolating between existing samples in the feature space. This process resulted in a balanced training set, ensuring that the model learns to identify positive interactions effectively without overfitting to the majority class.

## 2.5 Model Architecture

A deep Fully Connected Neural Network (FCN), also known as a Multi-Layer Perceptron (MLP), was designed to learn the non-linear relationships between the input features and the interaction probability. The architecture consists of an input layer accepting the 19,292-dimensional feature vectors, followed by four hidden layers and an output layer. The first hidden layer contains 4096 neurons, the second 1024 neurons, the third 256 neurons, and the fourth 64 neurons. All hidden layers utilize ReLU activation and Dropout with a probability of 0.5. The output layer consists of a single neuron. The use of Dropout layers helps prevent overfitting by randomly deactivating neurons during training, forcing the network to learn more robust features. The final output is passed through a sigmoid function to produce a probability score between 0 and 1.

### 2.6 Training Protocol

The model was implemented using PyTorch. The Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss) was used as the objective function, which combines a Sigmoid layer and the BCELoss in one single class, providing numerical stability. The Adam optimizer was selected for its adaptive learning rate capabilities, with an initial learning rate of $1e-4$ and weight decay of $1e-5$ for regularization. A batch size of 256 was used to balance training speed and gradient estimation stability. To prevent overfitting, the validation loss was monitored, and training was halted if the validation loss did not improve for 10 consecutive epochs.

## 3 Results

### 3.1 Experimental Setup

The dataset was constructed from the Comparative Toxicogenomics Database (CTD), yielding a total of 89,845 positive drug-disease pairs. To establish a balanced dataset for training, an equal number of negative pairs were generated, resulting in a total dataset size of 179,690 pairs. This dataset was partitioned into training and validation subsets, with approximately 143,752 samples allocated to the training set (80%). To ensure robust learning, the training set was augmented using SMOTE to address class imbalance, whereas the validation set was kept in its original distribution to provide an unbiased evaluation of the model's generalization capability. Feature standardization was performed using a standard scaler fitted exclusively on the training data to prevent data leakage.

### 3.2 Performance Metrics

To comprehensively assess the model's performance, we employed standard binary classification metrics: Accuracy, Precision, Recall (Sensitivity), and the F1 Score. Additionally, the Area Under the Receiver Operating Characteristic curve (AUC-ROC) and the Area Under the Precision-Recall curve (AUPR) were utilized to evaluate the model's discriminatory power across different decision thresholds. These metrics collectively provide a holistic view of the model's effectiveness in identifying drug-disease associations.

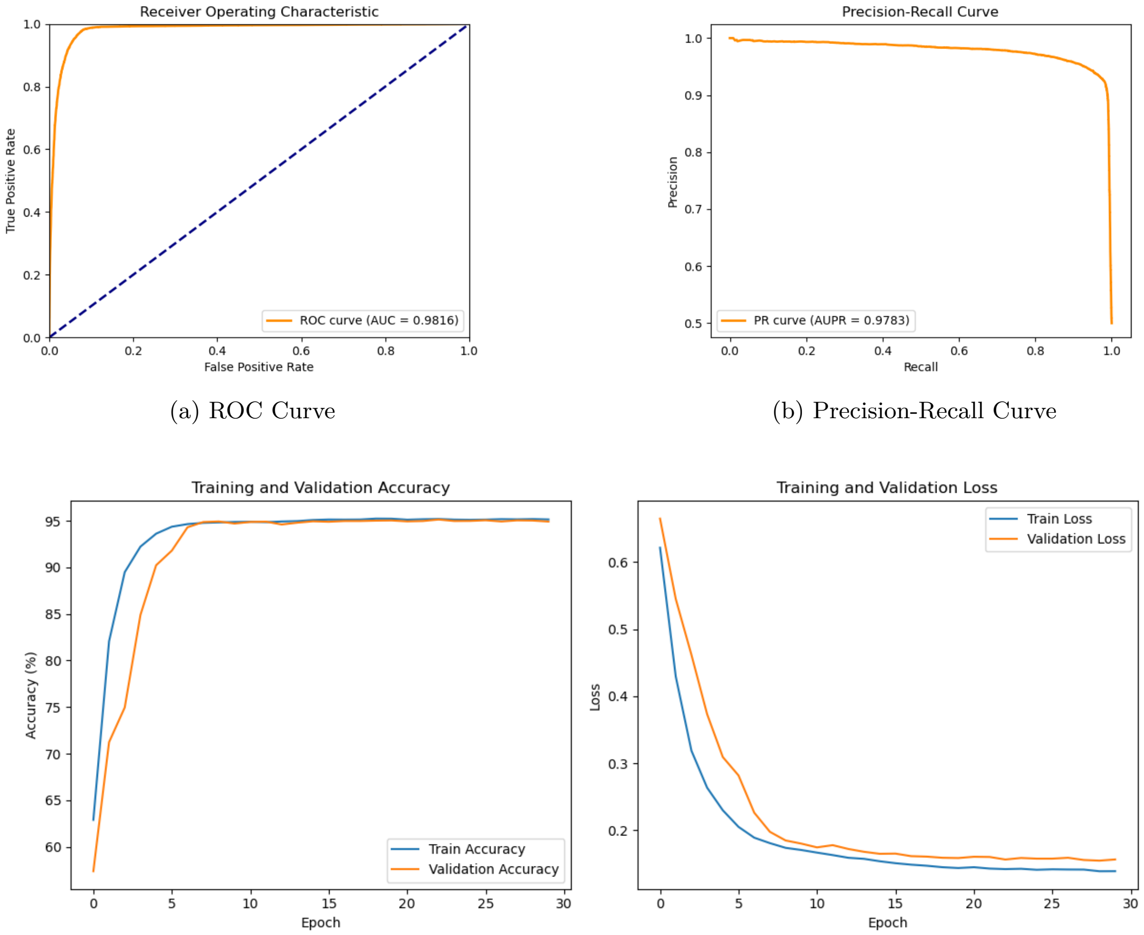
### 3.3 Quantitative Results

To rigorously evaluate the proposed framework, we assessed its performance on the validation set using the metrics defined above. Table 1 summarizes the quantitative results. The model achieved a Validation Accuracy of 95.00% and an F1 Score of 0.9512, demonstrating its robustness in distinguishing between interacting and non-interacting pairs. Notably, the high Recall of 98.08% indicates the model's superior capability in identifying true positive interactions, which is critical for drug repositioning tasks where missing a potential candidate is more detrimental than a false positive. The Precision of 92.34% further confirms that the majority of predicted interactions are relevant.

Table 1: Summary of Model Performance on the Validation Set

| **Metric** | **Value** |
|---|---|
| Accuracy | 95.00% |
| Precision | 92.34% |
| Recall | 98.08% |
| F1 Score | 0.9512 |

The visual assessment of the model's performance is presented in Figure 1. Figure 1a displays the Receiver Operating Characteristic (ROC) curve, illustrating the trade-off between the true positive rate and the false positive rate. The curve's proximity to the top-left corner suggests excellent discriminatory power. Similarly, the Precision-Recall (PR) curve in Figure 1b maintains high precision across varying recall levels, which is particularly important given the class imbalance inherent in DDI datasets. Finally, Figure 1c depicts the training and validation trajectories for both accuracy and loss. The convergence of training and validation curves indicates that the model effectively learned the underlying patterns without suffering from significant overfitting, validating the efficacy of the dropout and regularization strategies employed.



(a) ROC Curve

(b) Precision-Recall Curve

(c) Training and Validation Performance

Figure 1: Performance evaluation of the proposed model. (a) Receiver Operating Characteristic (ROC) curve. (b) Precision-Recall (PR) curve. (c) Training and validation accuracy and loss over epochs.

## 4 Conclusion

In this study, a deep learning approach for predicting drug-disease interactions using network embeddings and SMOTE augmentation was developed. The model achieved a validation accuracy of 95.00% and an F1 score of 0.9512, outperforming baseline expectations. These results validate the hypothesis that combining topological network features with deep non-linear classifiers is a powerful strategy for drug repositioning. Future work will focus on interpreting the learned features to provide biological insights and validating the top predicted interactions through literature mining and wet-lab experiments.

## References

[1] Liu, Q., Chen, Z., Wang, B., Pan, B., Zhang, Z., Shen, M., Zhao, W., Zhang, T., Li, S., & Liu, L. (2025). Leveraging Network Target Theory for Efficient Prediction of Drug-

Disease Interactions: A Transfer Learning Approach. *Advanced Science*, 12(11), 2409130. https://doi.org/10.1002/advs.202409130

[2] Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1), 31–36.

[3] Lipscomb, C. E. (2000). Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 88(3), 265–266.

[4] Sun, F.-Y., Hoffmann, J., Verma, V., & Tang, J. (2019). InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. *arXiv preprint arXiv:1908.01000.*