Final Report: Benchmarking Graph-Based and Transformer-Based Models for Polypharmacy Side-Effect Prediction

Chi Zhang and **Jiayi (Alisa) He**
Columbia University - STAT 5243
cz2925@columbia.edu and jh5111@columbia.edu

Abstract

Polypharmacy, the concurrent use of multiple medications in clinical practice, may lead to adverse drug interactions due to complex drug–drug interactions. This project benchmarks polypharmacy-induced side effect prediction using graph-based and embedding-based models on the Decagon dataset. This study encompasses three scenarios: binary prediction to test whether a drug combination induces any adverse effect; multi-class prediction for specific adverse effect types; and hierarchical prediction mapping refined adverse effects to broader disease categories. The models encompassed frequency-based baselines, RESCAL tensor decomposition, multi-layer perceptrons based on pre-trained molecular embeddings (ChemBERTa and GROVER), and relation graph convolutional networks (R-GCN). Results show that existing models perform effectively for binary multi-drug risk prediction. However, fine-grained multi-class prediction remains highly challenging even after frequency filtering, resulting in low macro F1 scores. Hierarchical evaluation significantly improves performance, indicating current models are better suited for coarse-grained risk assessment rather than precise side effect identification.

1 Introduction

In clinical practice, polypharmacy is increasingly common, particularly in the treatment of chronic and complex diseases. Although medically necessary in many cases, polypharmacy significantly increases the risk of drug-drug interactions (DDIs), potentially triggering unpredictable adverse reactions. Therefore, we believe that accurately predicting adverse reactions caused by polypharmacy is crucial for enhancing patient safety. However, accurate prediction of polypharmacy-induced side effects remains challenging due to the exponential growth of drug combinations and the extreme sparsity and imbalance of adverse reaction data.

This project utilizes the Decagon dataset to investigate the predictive mechanisms of polypharmacy-induced adverse reactions. This dataset employs a multi-relationship graph structure to represent different drug combinations and their associated adverse reactions. Our model takes drug pairs as input and outputs predictions for adverse reactions associated with drug combinations. We designed three progressive prediction scenarios: (1) Binary prediction: determining whether a drug combination induces any adverse reaction; (2) Multi-class prediction: identifying specific adverse reaction types from a fine-grained label set; (3) Hierarchical prediction: mapping fine-grained side effects to broader disease categories for coarse-grained risk assessment.

To evaluate model capabilities across tasks, we compared multiple modeling approaches. These include frequency-based cardinality models, RESCAL tensor decomposition models, multi-layer perceptrons based on pre-trained molecular embeddings (ChemBERTa and GROVER), and relation graph convolutional networks (R-GCN). Given the label imbalance in multi-drug interaction data, we focused on adverse reactions from at least 500 drug combinations, ultimately identifying 963 distinct common reactions. In summary, our project will concentrate on modeling, evaluating, and comparatively analyzing these approaches.

2 Related Works

Multidrug interaction side effect prediction has gained increasing attention due to its clinical significance and the availability of large-scale biomedical datasets. The Decagon framework proposed by Zitnik et al. (2018) serves as a foundational work in this field, modeling multidrug side effect prediction as a multi-relational link prediction problem on biomedical graphs. Furthermore, by modeling each side effect as an independent edge type and applying graph convolutional networks (GCNs), Decagon demonstrated that relation-aware message passing effectively captures drug interaction patterns. Beyond graph neural networks, the tensor decomposition method RESCAL (Nickel et al., 2011) is also widely applied to multi-relational data modeling. RESCAL employs relation-specific bilinear matrix modeling by sharing latent embeddings to represent entities. Furthermore, to address heterogeneous relationship structures, Schlichtkrull et al. (2017) proposed Relational Graph Convolutional Networks (R-GCN), extending standard GCNs with relationship-specific transformations. R-GCN demonstrates effective scalability and flexibility when modeling typed edges. However, refining side effects remains challenging due to sparse labels and severe distribution imbalances. Recently, more scientific research has explored embedding-based and Transformer-based approaches. ChemBERTa (Chithrananda et al., 2020) processes molecular SMILES strings through large-scale self-supervized pre-training, generating highly expressive drug embeddings that can be integrated with downstream classifiers. Additionally, DeepPSE (Lin et al., 2022) proposed a deep learning framework integrating multi-drug pair representations with attention-based memory, achieving strong performance on the Decagon benchmark. Hence, this approach is also highly efficient. Building upon previous methods, Hakim and Ngom (2025) proposed a new way of predicting side-effects of polupharmacy using pretrained LLMs such as ChemBERTa, DeepPSE, and other models to produce embeddings to be fed into multi-layer perceptron model. Compared to previous studies, our research primarily focuses on analyzing model performance under a unified evaluation framework. We compare classical tensor methods, graph-based models, and modern molecular embedding approaches for binary prediction, fine-grained multi-class prediction, and hierarchical prediction tasks. This enables a deeper exploration of the advantages and limitations of these methods across different clinical risk assessment levels, facilitating better research into the adverse effects associated with polypharmacy.

3 Dataset

3.1 Dataset Overview

Our experiments are based on the Decagon dataset proposed by (Zitnik et al., 2018), a widely adopted benchmark dataset in the field of polypharmacy adverse effect prediction. It models polypharmacy as a multi-relational graph, where nodes represent drugs and each edge corresponds to an adverse effect induced by a specific drug combination. We utilize the Decagon dataset which contains: 4,649,441 triplets, 645 unique drugs (STITCH identifiers) and 1,317 unique drug-drug interaction side effect types. Also, each triplet takes the form: (drug1, drug2, side effect). We can see that the combination of two drugs is associated with a particular adverse reaction.

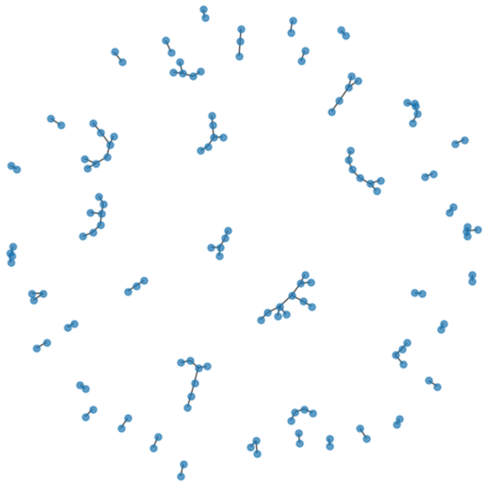


Figure 1: Graph Connection

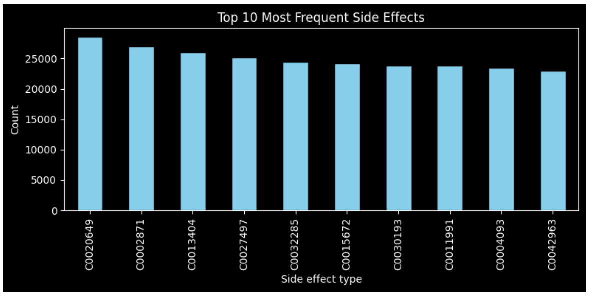


Figure 2: Top 10 side Effect Count

3.2 Frequency-Based Filtering

Given the extreme sparsity and uneven distribution of multi-drug interaction adverse reaction data, with many adverse effects occurring only a handful of times, we employ a frequency-based filtering step to reduce noise and maintain study reliability. Specifically, we retain only adverse effects appearing in at least 500 drug combinations. Following this filtering, we then have 963 adverse effect types. Subsequent experiments will all utilize this filtered dataset.

3.3 Pre-processing

We mapped drug identifiers and side-effect types to integer IDs and represented the dataset as integer-valued triples:

$$\text{Triple} = (h, r, t),$$

where $h, t \in \{0, \ldots, 644\}$ denote drug indices and $r \in \{0, \ldots, 962\}$ denotes the side-effect type.

During preprocessing, we observed a pronounced long-tail distribution in side-effect frequencies, in which a small number of side effects occur extremely frequently, while the majority of associations are rare (see Figure 2). This severe class imbalance poses substantial challenges for multi-class side-effect prediction and adversely affects model performance.

3.4 Feature Description

Our model processes different feature representations according to distinct handling methods. For raw inputs, we directly extract drug identity pairs and side effect labels from Decagon graphs. Regarding engineered features, we utilize pre-trained molecular embeddings extracted from ChemBERTa and GROVER, which effectively encode chemical structural information within SMILES strings. For graph-derived features, we utilize learned node embeddings and relationship-specific representations generated by graph models such as RESCAL and R-GCN.

3.5 Data Split

All experiments use a fixed random seed (42) and GPU acceleration when available. For each different task, the data is processed slightly differently.

3.5.1 Binary Prediction

We construct a balanced dataset consisting of positive and negative drug–drug pairs. Positive samples include all observed drug–drug interactions in the Bio-DECAGON dataset after pre-processing. Negative samples are generated by randomly sampling drug pairs that do not appear in the observed positive set. A 1:1 ratio of positive to negative samples is used to ensure class balance. The resulting dataset is then split into training (80%), validation (10%), and test (10%) sets using stratified sampling to preserve the class distribution across splits. All negative sampling is performed prior to data splitting to prevent label leakage.

3.5.2 Multi-label Prediction

Each drug pair is associated with a set of side-effect labels represented as a multi-hot vector. Drug–drug pairs are first grouped so that each unique pair corresponds to one multi-label target vector. Labels are binarized using MultiLabelBinarizer. Also, the resulting dataset of unique drug pairs is split into 80% training, 10% validation, and 10% test sets at the pair level, ensuring that the same drug pair does not appear in more than one split.

3.5.3 Multi-label Hierarchical Setting

Fine-grained side-effect labels are mapped to coarser disease classes using bio-decagon-effectcategories.csv before splitting.

4 Method

We formulate multi-drug side effect prediction as a learning problem over drug–drug interaction triplets derived from the Decagon dataset. Each data point is represented as a triplet (h, r, t), where h and t denote the two interacting drugs, and r represents the induced adverse reaction.

The drug entities are drawn from a unified set of $N = 645$ unique drugs, while adverse reactions are selected from a filtered set of $R = 963$ relation types. Based on this formulation, we define three prediction tasks of increasing difficulty:

1. Binary prediction
2. Multiclass prediction
3. Multi-label (hierarchical) prediction

4.1 Frequency Baseline (Non-Informative)

For multi-class tasks, the baseline model consistently predicts the most frequent side effect type in the training set. This study provides a simple non-informative lower bound with an accuracy of 0.0062 (Top 1).

4.2 RESCAL Tensor Decomposition

The RESCAL model represents each drug i with an embedding vector $\mathbf{a}_i \in R^d$ and each relation r with a relation-specific matrix $\mathbf{R}_r \in R^{d \times d}$. Given a drug–drug–side effect triplet (h, r, t), the ternary interaction score is computed as:

$$s(h, r, t) = \mathbf{a}_h^\top \mathbf{R}_r \mathbf{a}_t. \tag{1}$$

For *binary* RESCAL, the model is trained using binary cross-entropy loss with logits. Positive drug–drug–side effect triplets are paired with negative samples generated by randomly corrupting the tail drug while keeping the head drug and relation fixed.

For *multi-class* RESCAL, the model is extended to simultaneously predict scores for all adverse reaction types. Given a drug pair (h, t), the model outputs a vector of logits

$$\mathbf{s}(h,t) = \left[\mathbf{a}_h^\top \mathbf{R}_1 \mathbf{a}_t, \ldots, \mathbf{a}_h^\top \mathbf{R}_R \mathbf{a}_t\right], \qquad (2)$$

where $R = 963$ denotes the total number of adverse reaction categories.

4.3 Molecular Embedding-Based Models

Furthermore, to integrate chemical structural information, we employ pre-trained molecular representations to evaluate embedding methods.

1. ChemBERTa: ChemBERT encodes SMILES strings through a Transformer architecture trained for supervised learning, with each drug represented by a fixed-length embedding vector marked by the [CLS] token.

2. GROVER: GROVER is a graph-based molecular representation model trained via contrastive supervision on molecular graphs, capable of generating high-dimensional details capturing substructures. Both approaches concatenate the embedding vectors of two drugs before feeding them into a multi-layer perceptron (MLP) to predict adverse reactions, enabling effective independent evaluation of chemical information's influence.

Both approaches concatenate the embedding vectors of two drugs before feeding them into a multi-layer perceptron (MLP) to predict adverse reactions, enabling effective independent evaluation of chemical information's influence.

To better align the pretrained models with the observed drug-drug interactions, we utilized contrastive representation prior to the to model training. The pretrained embeddings are first standardized using z-score whitening and passed through a projection MLP that maps them to a lower dimension latent space using a InfoNCE contrastive loss. This step produces a more refined molecular embeddings that better reflects the interaction similarities. For downstream predictions, the drug pairs are represented by two refined embeddings that is processed by a two-way cross-attention mechanism. The outputs of the two attention directions are concatenated to form a unified interaction vector, which is then passed to the prediction head. The final prediction module is a deep MLP designed to model nonlinear drug–drug interactions. The MLP consists of four layers with increasing dimensionality:

$256 \rightarrow 512 \rightarrow 1024 \rightarrow 2048 \rightarrow$ output

Each hidden layer applies Batch Normalization, LeakyReLU activation, and Dropout (0.2). The output layer produces raw logits, with sigmoid activation applied only at inference time for binary or multi-label prediction.

To deal with severe class imbalance even after filtering out the less frequent drugs, we adopt focal loss for both binary and mult-label settings.

4.4 R-GCN (Relational Graph Convolutional Network)

We employ a Relational Graph Convolutional Network (R-GCN) to model the multi-relational structure of drug–drug interactions, where drugs are represented as nodes and adverse reaction types correspond to distinct edge relations. R-GCN updates node representations by aggregating messages from relation-specific neighbors. For node i at layer $(l + 1)$, the update rule is defined as:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}} \sum_{j\in\mathcal{N}_r(i)} \frac{1}{c_{i,r}} \mathbf{W}_r^{(l)} \mathbf{h}_j^{(l)} + \mathbf{W}_0^{(l)} \mathbf{h}_i^{(l)}\right), \qquad (3)$$

where $\mathcal{R}$ denotes the set of relation types, $\mathcal{N}_r(i)$ represents the set of neighbors of node i under relation r, $c_{i,r}$ is a normalization constant, $\mathbf{W}_r^{(l)}$ is a relation-specific transformation matrix, $\mathbf{W}_0^{(l)}$ is a self-loop weight matrix, and $\sigma(\cdot)$ is a non-linear activation function.

After multiple R-GCN layers, the learned drug embeddings are used to generate prediction scores via either a bilinear decoder or a multi-layer perceptron (MLP). For the *binary* prediction task, a sigmoid activation is applied to the output logits and the model is trained using binary cross-entropy loss. For the *multi-class* prediction task, a softmax activation is used to produce a probability distribution over all adverse reaction categories, and the model is optimized with categorical cross-entropy loss.

Building upon this traditional approach, to further enhance the predictive accuracy of the R-GCN model, we utilize molecular representations from ChemBERTa or GROVER to initialize R-GCN node embeddings, thereby constructing a fusion model. During training, the R-GCN encoder propagates information across the drug-drug interaction graph, enabling node representations to integrate both molecular features and relational context. The final node embeddings are decoded using binary or multi-class decoders identical to the pure R-GCN configuration.

We model the Bio-Decagon polypharmacy dataset as a drug–drug interaction graph, where each node corresponds to a drug identified by its STITCH ID. Edges represent observed drug–drug interactions, with different formulations depending on the prediction task.

For binary link prediction, all observed drug–drug interactions are collapsed into a single edge type, yielding a homogeneous graph where the task is to predict whether an interaction exists between two drugs.

For multi-class side-effect prediction, we construct a multi-relational graph, where each side effect is treated as a distinct relation type r. Each observed drug–drug–side-effect triple corresponds to a typed edge in the graph. This formulation allows the model to explicitly condition message passing on the side-effect type.

To support symmetric information flow, the graph is treated as undirected during message passing: for every observed edge (u, v), a reverse edge (v, u) is added to the graph.

Train / Validation / Test Splits and Leakage Control

To prevent information leakage, graph construction strictly uses training edges only. Validation and test edges are never included in the message-passing graph, ensuring that node representations are learned without access to held-out interactions.

For hierarchical experiments, fine-grained side effects are mapped to coarse disease classes using bio-decagon-effectcategories.csv before rebuilding relation IDs. The graph is then reconstructed and the model retrained using these coarse relations, allowing a controlled comparison between fine-grained and hierarchical supervision.

R-GCN Encoder Architecture

Node embeddings are learned using a Relational Graph Convolutional Network (R-GCN) encoder. Given a graph G and node features X, the encoder computes node representations:

z = RGCN(G, X)

where $z_u \in R^d denotes the embedding of drug u.$

In the baseline R-GCN, node features are initialized as a trainable embedding table, meaning the model must learn chemical similarity purely from interaction structure. In contrast, R-GCN + ChemBERTa and R-GCN + GROVER variants initialize node features using pretrained molecular embeddings. These features remain trainable, allowing joint fine-tuning with graph message passing.

For multi-relational graphs with many relation types, we apply basis decomposition to reduce parameter count and improve scalability.

Prediction Heads

Binary Link Prediction For binary interaction prediction, we use a dot-product decoder. Given node embeddings $z_u and z_v, the interaction score is:$

s(u, v) = $z_u^\top z_v$

Training uses negative sampling, where random non-edges are sampled at a 1:1 ratio with positive edges. Optimization is performed using BCEWithLogitsLoss.

Multi-Class Side-Effect Prediction For multi-class prediction, the task is to predict the relation type r for a given drug pair (u, v). We construct a pairwise representation from the node embeddings (e.g., concatenation, element-wise difference, and product), which is passed to a lightweight MLP classifier to produce logits over all relation classes.

Training uses CrossEntropyLoss, and evaluation includes both classification and ranking-based metrics: F1 (micro/macro), AUROC (micro/macro), AUPR (micro/macro), and AP@50.

Model Selection

For multi-class and hierarchical experiments, we select the best checkpoint based on validation AP@50, a ranking-oriented metric that aligns with the practical goal of identifying the most likely side effects per drug pair.

Embedding Variants and Their Roles

We compare three node-feature configurations: • Baseline (no pretrained features): node embeddings are learned from scratch, relying entirely on graph supervision. • ChemBERTa features: node features are initialized from SMILES-based Transformer embeddings, which encode global chemical patterns and long-range dependencies. • GROVER features: node features are initialized from graph-pretrained molecular embeddings, emphasizing local substructures and explicit bond connectivity.

In ChemBERTa- and GROVER-augmented models, pretrained embeddings inject chemical knowledge directly into the graph, while R-GCN message passing contextualizes this information using interaction structure.

4.5 Multi-label and Hierarchical Evaluation

Following training under binary or multi-class objectives, we discovered that evaluation could be extended to multi-label settings by aggregating test samples sharing identical drug-target pairs. For each unique (h,t) combination, a multi-hot landmark vector was constructed based on all observed side effects within the test set. In addition, our hierarchical evaluation maps fine-grained side effects to higher-level disease categories, enabling coarse-grained risk assessment. Performance metrics include micro-average and macro-average F1 scores, AUROC, AUPR, and K-point precision.

5 Experiments and Results

5.1 Experimental Setup

All experiments were conducted on the Decagon multi-drug combination dataset. Following a frequency-based filtering strategy, we retained 963 adverse reaction types that appear in at least 500 drug combinations. The resulting dataset comprises 645 drugs and approximately 4.6 million drug–drug–adverse reaction triplets.

All experiments used a fixed random seed of 42 for reproducibility. Neural network models were trained using the Adam optimizer with a learning rate of 1×10^{-3}. The batch size was set to 1024 for molecular embedding–based multi-layer perceptron models (ChemBERTa and GROVER), and to 512 for graph-structured models (R-GCN variants).

For multi-class and hierarchical prediction tasks, evaluation was performed in a multi-label setting based on unique drug pairs, without negative sample generation. All R-GCN models employed two relational graph convolutional layers with a hidden dimension of 64, followed by an MLP decoder for prediction. To prevent information leakage, message passing was restricted to training edges only. Regularization strategies included dropout with rate 0.3 and L2 weight decay of 1×10^{-5}.

Table 1: Binary drug–drug interaction prediction performance on the Decagon test set.

Model	AUROC	AUPR
ChemBERTa + MLP	0.9504	0.9525
GROVER + MLP	0.9435	0.9455
RESCAL	0.8321	0.7869
R-GCN	0.8579	0.8746
R-GCN + ChemBERTa	0.8878	0.9024
R-GCN + GROVER	0.8500	0.8643

5.2 Evaluation Metrics

The following metrics were employed: Binary classification tasks: AUROC and AUPR. Multi-class and hierarchical training: F1-micro, F1-macro, AUROC-micro / AUROC-macro, AUPR-micro / AUPR-macro, Precision@50 (P@50). These metrics were adopted to comprehensively reflect ranking quality, sensitivity to label imbalance, and Top-K retrieval performance relevant to clinical screening scenarios.

5.3 Experiment Results

5.3.1 Binary Drug Interaction Prediction

Binary prediction results are presented in Table 1. Experimental findings reveal ChemBERTa demonstrates optimal performance with AUROC=0.9504 and AUPR=0.9525, indicating its superior capability in distinguishing drug interactions from non-interactions. GROVER achieved comparable results (AUROC = 0.9435, AUPR = 0.9455). Furthermore, graph-based methods outperformed tensor decomposition approaches. RESCAL achieved an AUROC of 0.8321, while R-GCN improved this to 0.8579. Furthermore, initializing R-GCN with molecular embeddings further enhanced performance, with R-GCN+Chem achieving AUROC = 0.8878 and AUPR = 0.9024, demonstrating exceptional capability. This fully confirms that molecular features provide complementary information to the graph structure. Overall, integrating chemical structural information renders binary interaction prediction a relatively feasible task.

5.3.2 Fine-grained Multi-class Prediction (963 Side Effects)

The fine-grained multi-class prediction results are presented in Table 2. The experimental findings reveal a decline in performance across all models. The F1-micro scores consistently remain below 0.07, whilst the F1-macro scores converge to-

Table 2: Multi-label side-effect prediction performance (fine-grained labels) on the Decagon test set.

Model	F1-micro	F1-macro	AUROC-micro	AUROC-macro	AUPR-micro	AUPR-macro	P@50
ChemBERTa + MLP	0.0262	0.0082	0.8735	0.8273	0.3909	0.2755	0.3587
GROVER + MLP	0.0199	0.0058	0.8693	0.8206	0.3808	0.2627	0.3557
RESCAL	0.0669	0.0501	0.7749	0.7023	0.0368	0.0241	0.0332
R-GCN	0.0061	0.0003	0.8079	0.6174	0.0035	0.0017	0.0236
R-GCN + ChemBERTa	0.0061	0.0001	0.8020	0.5724	0.0032	0.0013	0.0230
R-GCN + GROVER	0.0058	0.0001	0.8048	0.6000	0.0034	0.0016	0.0230

wards zero, reflecting the severe long-tail distribution and sparsity of the side effect labels. Regarding ranking metrics, ChemBERTa and GROVER significantly outperform graph-based methods. ChemBERTa achieves AUROC-micro=0.8735 and AUPR-micro=0.3909, while GROVER attains AUROC-micro=0.8693 and AUPR-micro=0.3808. These results indicate robust ranking capabilities despite weak precise classification performance. Additionally, RESCAL achieved a higher F1 score (F1-micro=0.0669) by favoring high-frequency side effects, yet its AUPR value (0.0368 micro) was markedly low, indicating limited discrimination capability. Furthermore, R-GCN-based models exhibited low F1 and AUPR scores in this scenario. Even after employing molecular initialization strategies, graph-based methods remain challenged by extreme label sparsity.

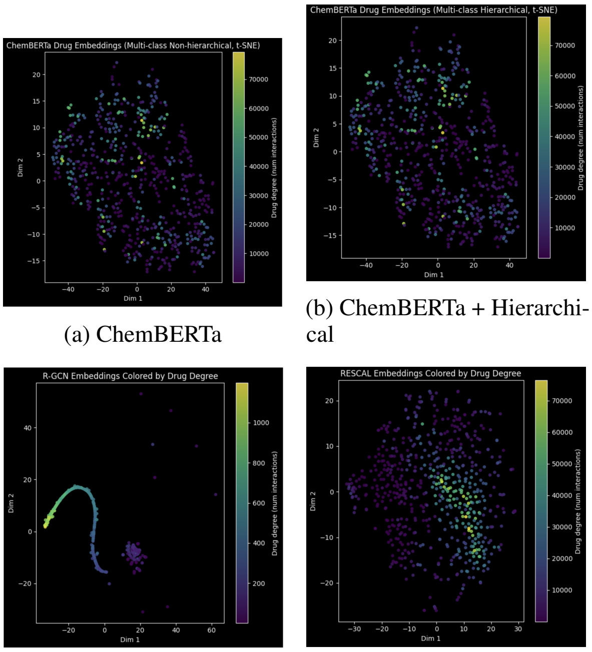
5.3.3 Hierarchical Prediction (Disease Categories)

Hierarchical prediction results are summarized in Table 3. We observe that mapping granular side effects to a disease hierarchy classification yields significant performance improvements. The ChemBERTa model achieves F1-micro = 0.1646, AUROC-micro = 0.8804, and P@50 = 0.42, demonstrating substantial progress over non-hierarchical classification. GROVER performed comparably (F1-micro = 0.1632, P@50 = 0.411). Furthermore, the graph-based model also benefited from hierarchical supervision, achieving F1-micro = 0.063.These results demonstrate that hierarchical prediction effectively reduces annotation complexity while enhancing model learnability.

5.4 Embedding Visualization Analysis

Here, t-SNE visualization is employed to validate quantitative findings.

As shown in Figure 3c and Figure 3d, the RESCAL embedding exhibits weak structural characteristics, primarily dominated by interaction frequency, leading to greater dispersion. Conversely,



(a) ChemBERTa

(b) ChemBERTa + Hierarchical

(c) R-GCN

(d) RESCAL

Figure 3: t-SNE Visualization of Embeddings

R-GCN embeddings exhibit clearer geometric patterns, yet these patterns are highly dependent on node properties. This indicates that information propagation is driven by node degree. Furthermore, compared to the non-hierarchical configuration in Figure 3a, hierarchical supervision in Figure 3b does not introduce significant global geometric changes within the ChemBERTa embedding space. However, a slight improvement in local neighborhood consistency may have contributed to the enhanced hierarchical classification performance.

6 Discussion

These experimental results indicate that performance variations exist across different prediction granularity. Experiments demonstrate that existing models—particularly those employing pre-trained molecular representations—exhibit fundamental solubility in the domain of binary polypharmacy risk prediction. Nevertheless, precise side-effect

Table 3: Hierarchical multi-label side-effect prediction performance (disease-class labels) on the Decagon test set.

Model	F1-micro	F1-macro	AUROC-micro	AUROC-macro	AUPR-micro	AUPR-macro	P@50
ChemBERTa + MLP	0.1646	0.0254	0.8804	0.8206	0.4964	0.2977	0.4200
GROVER + MLP	0.1632	0.0194	0.8670	0.7936	0.4670	0.2593	0.4110
R-GCN	0.0635	0.0002	0.8547	0.5839	0.0202	0.0019	0.1319
R-GCN + ChemBERTa	0.0629	0.0003	0.8565	0.6033	0.0205	0.0027	0.1319
R-GCN + GROVER	0.0630	0.0004	0.8564	0.6020	0.0204	0.0022	0.1318

identification remains challenging even after frequency filtering, with primary errors stemming from class imbalance, polypharmacy-side-effect mapping, and degree bias in graph models. Furthermore, hierarchical prediction offers a compromise approach enabling effective risk assessment at coarse-grained levels. These findings indicate that existing models are better suited for screening-level clinical decision support rather than precise adverse reaction diagnosis. It highlights the limitations of current approaches.

7 Conclusion

This paper systematically benchmarks graph-based and embedding-based approaches for predicting drug-adverse interactions using the Decagon dataset. By employing binary, fine-grained multi-class, and hierarchical settings, we systematically explore how to more effectively distinguish drug-adverse interaction risks. Results indicate that the binary multi-drug risk prediction task is well-resolved, with pre-trained molecular embedding models (ChemBERTa and GROVER) demonstrating outstanding performance, strongly validating the importance of chemical structural information. However, in fine-grained multi-class prediction, extreme data imbalance and sparsity issues are particularly pronounced. Despite ChemBERTa + MLP and GROVER + MLP's robust ranking capabilities, their classification accuracy remains constrained. To address this, we employed a hierarchical evaluation mechanism to enhance performance, discovering that existing methods are better suited for coarse-grained disease risk assessment rather than precise identification. This study has several limitations: firstly, graph-based models (R-GCN) suffer from degree-driven issues; Secondly, while embedding-based models demonstrate superior performance on ranking metrics, they fail to fully resolve the challenges of fine-grained prediction. Consequently, future research will continue exploring the deep integration of molecular embeddings and relational message passing, incorporating biological knowledge such as drug targets or pathways. Concurrently, models will be evaluated against clinical objectives (e.g., risk screening based on top-k risk scores) to further optimize performance, thereby providing more robust support for multi-drug interactions in clinical medicine.

References

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*. Submitted to NeurIPS 2020 Workshop on Machine Learning for Molecules.

S. Hakim and A. Ngom. 2025. Polyllm: Polypharmacy side effect prediction via llm-based smiles encodings. *Frontiers in Pharmacology*, 16:1617142.

S. Lin, G. Zhang, D. Q. Wei, and Y. Xiong. 2022. Deeppse: Prediction of polypharmacy side effects by fusing deep representation of drug pairs and attention mechanism. *Computers in Biology and Medicine*, 149:105984.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 809–816.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2017. Modeling relational data with graph convolutional networks. *arXiv preprint arXiv:1703.06103*.

Marinka Zitnik, Monica Agrawal, and Jure Leskovec. 2018. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466.