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

Chi Zhang (cz2925) & Jiayi (Alisa) He (jh5111)

1. Abstract

Polypharmacy, the concurrent use of multiple medications in hospital settings, may trigger adverse drug interactions that are difficult to predict at scale. This project benchmarks polypharmacy-induced side effect prediction using graph-based and embedding-based models on the Decagon dataset. To address potential extreme imbalance in the data, our experiments focus primarily on side effects occurring in at least 500 drug combinations, ultimately yielding 963 refined side effect categories. 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. Additionally, 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, with even ChemBERTa-based methods achieving AUROC values exceeding 0.95. 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.

2. 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.

3. Related Work

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. (2018) 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-supervised 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. So this approach is also highly efficient. 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.

4. Dataset

4.1 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.

4.2 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 utilise this filtered dataset.

4.3 Preprocessing

We mapped drug identifiers and side effect types to integer IDs, representing the data as integer triples: Triple = (h, r, t), where h, t $\in$ {0,…,644} and r $\in$ {0,…,962}. During processing, we encountered several challenges. We observed a pronounced long-tail distribution in side effect types, where a small number of side effects occur extremely frequently (Figure 1), while most associations are relatively rare. This imbalance inherently poses certain adverse effects on multi-class task performance.

4.4 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.

4.5 Data split

To stabilize learning and reduce extreme long-tail effects, we filter side effects to those appearing at least 500 times in the dataset for the non-hierarchical setup. All experiments use a fixed random seed (42) and GPU acceleration when available. For each different task, the data is processed slightly differently.

Binary prediction: We construct a balanced dataset consisting of positive and negative drug–drug pairs. Positive samples are all observed drug–drug pairs in the Bio-Decagon dataset after preprocessing. Negative samples are generated by randomly sampling drug pairs that do not appear in the observed positive set. A 1:1 positive-to-negative ratio is used. In addition, the resulting dataset is split into training (80%), validation (10%), and test (10%) sets using stratified sampling to preserve class balance across splits. All negative sampling is performed before splitting to avoid label leakage.

Multi-label evaluation: 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.

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

5. Method

We formulate multi-drug side effect prediction as a learning problem based on drug-drug interaction triplets derived from the Decagon dataset, where each data point takes the form (h, r, t). Here, h and t denote the two interacting drugs, while r represents the induced adverse reaction. Drug indices are sourced from a union of N=645 unique entities, whilst adverse reactions are extracted from a filtered set of R=963 relationship types. We define three progressively challenging prediction tasks: 1. Binary prediction, 2. Multiclass prediction, 3. Multi-label/hierarchical evaluation.

5.1 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).

5.2 RESCAL Tensor Decomposition

The RESCAL model represents drug i with an embedding vector ai $\in R^d$ and each relation r with a matrix $R_r \in R^{d*d}$. The ternary correlation formula is: s(h,r,t) = a$h^T R_r a_t$. For binary RESCAL, the model is trained using binary cross-entropy loss with logit: in binary prediction, positive drug-drug-side effect triplets are combined with negative samples generated by randomly tampering with tail drugs. For multi-class RESCAL, the model is adapted to simultaneously output logit values for all R=963 adverse reaction types.

5.3 R-GCN (Relational Graph Convolutional Network)

This study employs R-GCN to construct a multi-relational graph, where drugs serve as nodes and side effect types define edge types. R-GCN updates node representations by aggregating typed neighbor messages:

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_r(i)}\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)}\right)$$

We know that after undergoing multi-layer R-GCN processing, drug embeddings generate prediction scores through either a bilinear decoder or a multi-layer perceptron combination. Furthermore, for the binary task, a sigmoid output with cross-entropy loss is employed, while the multi-class task utilises a softmax output with categorical cross-entropy.

5.4 Molecular Embedding-Based Models

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

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. **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.

5.5 Graph-Chemistry Fusion Model (R-GCN + Molecular Embedding)

Building upon the preceding approach, to further enhance the predictive accuracy of the R-GCN model, we utilise molecular representations from ChemBERTa or GROVER to initialise 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.

5.6 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.

6. Experiments & Results

6.1 Experimental Setup

All experiments were conducted using the Decagon multi-drug combination dataset. Following our frequency-based design, this dataset retained 963 adverse reaction types appearing in at least 500 drug combinations. It comprises 645 drugs and 4.6 million drug-drug-adverse reaction triplets. In addition, all experiments employed a fixed random seed (42). Neural network models were trained using the Adam optimiser with a learning rate of 1e-3. Concurrently, the batch size for molecular embedding-based multi-layer perceptron models (ChemBERTa, GROVER) was set to 1024, while that for graph-structured models (R-GCN variants) was set to 512. Furthermore, multi-class and hierarchical tasks employed multi-label evaluation based on unique drug pairs, without negative sample sampling. All R-GCN models incorporated two convolutional layers (hidden dimension 64), followed by an MLP decoder. To prevent information leakage, message passing was restricted to training edges. Finally, regularisation strategies included dropout (0.3) and L2 weight decay (1e-5).

6.2 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.

6.3 Experimental Results

5.3.1 Binary Drug Interaction Prediction

Binary Prediction

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

Table 1

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, initialising 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)

Multi-class Prediction (Fine-grained Side Effects)

Model	F1-micro	F1-macro	AUROC-micro	AUROC-macro	AUPR-micro	AUPR-macro	P@50
ChemBERTa	0.0262	0.0082	0.8735	0.8273	0.3909	0.2755	0.3587
GROVER	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 + Chem	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

Table 2

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 towards 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 AURCO-micro=0.8735 and AUPR-micro=0.3909, while GROVER attains AURCO-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 favouring 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 initialisation strategies, graph-based methods remain challenged by extreme label sparsity.

5.3.3 Hierarchical Prediction (Disease Categories)

Hierarchical Prediction (Disease Class)

Model	F1-micro	F1-macro	AUROC-micro	AUROC-macro	AUPR-micro	AUPR-macro	P@50
ChemBERTa	0.1646	0.0254	0.8804	0.8206	0.4964	0.2977	0.4200
GROVER	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 + Chem	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

Table 3

Hierarchical prediction results are summarised 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, AURCO-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.

6.4 Embedding Visualisation Analysis

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

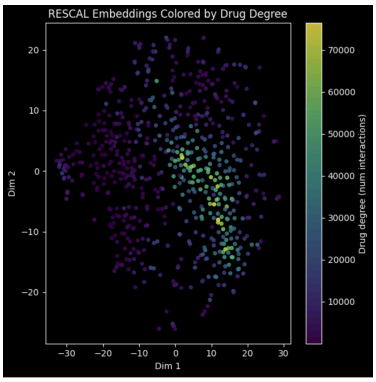

Figure 2 : RESCAL Embeddings

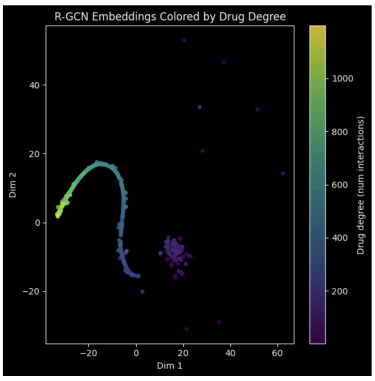

Figure 3: R-GCN Embeddings

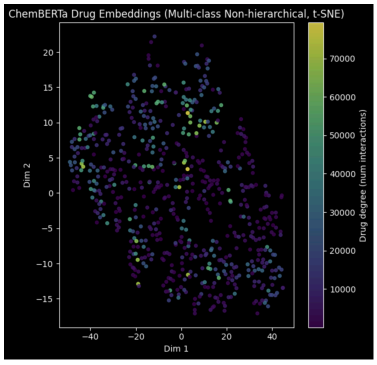

Figure 4: ChemBERTa Embeddings (multi-class)

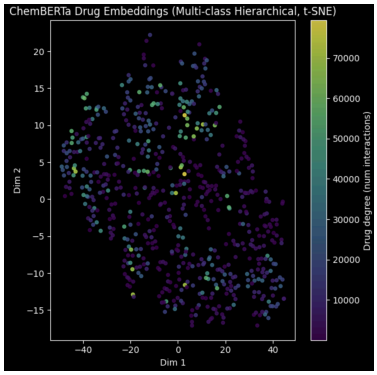

Figure 5: ChemBERTa Embeddings (Hierarchical)

We can see that RESCAL embeddings exhibit weak structural features, primarily influenced by interaction frequency pearl islands, with drugs highly dispersed within the embedding space. R-GCN embeddings reveal clearer geometric patterns yet remain highly node-correlated, indicating degree-driven message propagation. In addition, observing Figures 4 and 5, compared to the non-hierarchical setting, hierarchical supervision did not introduce significant global geometric changes in the ChemBERTa embedding space. However, a slight improvement in local neighbourhood consistency may have contributed to the observed enhancement in stratified classification performance.

6.5 Discussion

These experimental results indicate that performance variations exist across different prediction granularities. 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 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. Conclusions and Future Work

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 and GROVER'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.

8. References

1. Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics.
2. Nickel, M., Tresp, V., & Kriegel, H.-P. (2011). A Three-Way Model for Collective Learning on Multi-Relational Data (RESCAL). ICML.
3. Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., & Welling, M. (2018). Modeling Relational Data with Graph Convolutional Networks. ESWC.
4. Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar, "ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction," arXiv:2010.09885, 2020.
5. S. Hakim and A. Ngom, "PolyLLM: polypharmacy side effect prediction via LLM-based SMILES encodings," Frontiers in Pharmacology, vol. 16, p. 1617142, July 31, 2025, doi:10.3389/fphar.2025.1617142.