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

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

1.Introduction

Polypharmacy is emerging as a significant issue in healthcare. While effective for chronic and complex conditions, this approach holds risks such as adverse drug reactions. This project aims to develop and evaluate multiple machine learning models for detecting polypharmacy-induced side effects. Specifically, this project focuses on modeling and predicting polypharmacy-induced side effects using graph-based machine learning methods. We have demonstrated data pre-processing, preprocessing, and baseline construction using the publicly available Decagon dataset, laying a solid foundation for subsequent Relational Graph Convolutional Networks (R-GCNs).

2. Problem Statement

This study aims to model drug-drug interactions (DDIs) as a link prediction problem within multimodal biomedical networks. Each link connects two drugs and one side effect type, representing a specific polypharmacy side effect.

The analysis primarily utilizes the Decagon dataset (Zitnik et al., 2018), which integrates multiple biomedical data sources into a single heterogeneous graph. Specifically, the subset employed for this milestone includes: **4,649,441 triplets** (Drug$_1$, Side Effect, Drug$_2$), **645 unique drugs, 1,317 unique side effect types.** Each edge, or triplet, corresponds to an observed side effect triggered by a specific drug combination. The dataset is provided in .csv format, specifically including: **STITCH 1**, **STITCH 2**, **Polypharmacy Side Effect**, and **Side Effect Name.**

The primary objectives of this phase are to load, clean, and explore the dataset. We will also quantify the basic statistical characteristics of drug and side effect distributions. At the same time, we will explore visualizations of high-frequency side effects and drug-drug associations. Finally, we will establish a baseline model to prepare for subsequent graph-based method comparisons.

This study will use the accuracy of the baseline model to measure the proportion of correctly predicted side effects in the validation set. Subsequently, the full project evaluation will employ metrics including AUROC and AUPR for multi-class classification. For the multi-label task that predicts which side effect will happen, the evaluation will employ Macro and Micro F1-scores.

3. Methodology

Overall, the workflow for this milestone was implemented in Google Colab, utilizing Python libraries including pandas, matplotlib, scikit-learn, and networkx.

3.1 Data Loading and Cleaning

First, the dataset file was imported from Google Drive bio-decagon-combo.csv, and using basic descriptive statistics were calculated to determine unique drugs, side effects, and total records.

3.2. Graph Methods

3.2.1 Data Visualization

Subsequently, this study effectively identified and visualized the 10 most common side effects using bar charts.

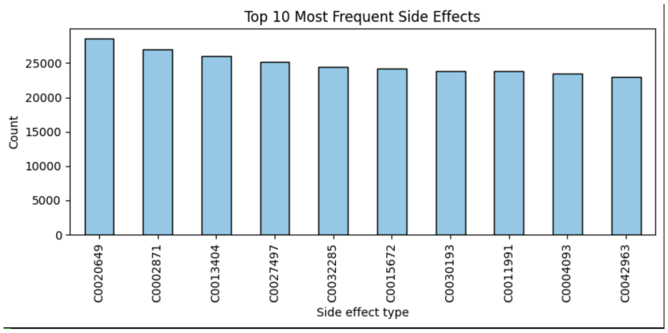


Figure 1. Visualization of the top ten most common side effects in the Decagon dataset.

Additionally, NetworkX was employed to visualize subgraph samples of 100 drug pairings, demonstrating the heterogeneous structure of drug interaction networks.

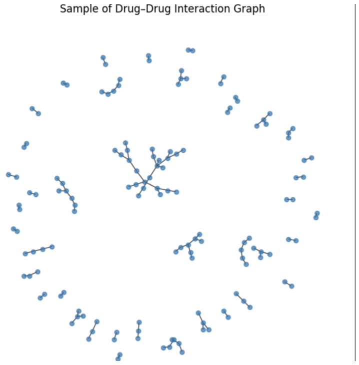


Figure 2. Example of a subgraph illustrating drug-drug interactions.

3.2.2 Building the Baseline Model

In addition to the above, this study also constructed a simple frequency-based classifier as the initial baseline model. This model predicts the most common side effect in the training data. Using this baseline model effectively establishes a lower bound for subsequent graph-based predictions.

3.3. ChemBERTa based MLP

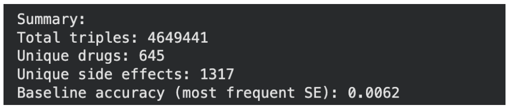
In order to establish a non-graph baseline model, we implemented a pipeline based on the chemical structure embedding derived from Transformers (*Chithrananda, S., Grand, G., & Ramsundar, B. (2020)*). We first identified all 645 unique drugs using their STITCH CIDs, and then using these to retrieve their SMILES strings, which is a textual representation of the chemical structures from the PubChem database. Then, we feed the SMILES strings into a pre-trained ChemBERTa model (**DeepChem/ChemBERTa-77M-MLM**) and extracted the 768-dimension vector from the [CLS] token output for each of the drugs to be used as embeddings. For a given drug pair (i, j), we would concatenate their embeddings together to form a unique xij. Then, we feed this into our MLP classifier. The model is then trained on minimizing the binary entropy loss as our measure of success.

4. Preliminary Findings

4.1 Graph baseline

Figure 1 reveals that the most prevalent adverse effects include `C0020649`, `C0002871`, and `C0013404`, occurring approximately 25,000 to 30,000 times across the dataset. These high-frequency effects may correspond to common physiological reactions such as pain, nausea, or even rash. Furthermore, Figure 2 reveals that the network diagram of 100 randomly selected drug pair combinations exhibits a highly sparse modular structure. This may reflect the limited known side effects for each combination, effectively validating the efficacy of using graph learning methods to explore global associations.

Through analysis, we also observed the following:



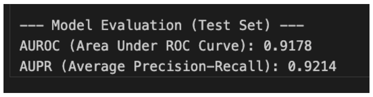
```
Summary:
Total triples: 4649441
Unique drugs: 645
Unique side effects: 1317
Baseline accuracy (most frequent SE): 0.0062
```

This result strongly confirms the large-scale and heterogeneous characteristics of the Decagon dataset, solidifying the foundation for training graph-based models. Simultaneously, the baseline accuracy of 0.0062 demonstrates the extreme difficulty of predicting adverse effects without relationship modeling, further highlighting the necessity of neural networks.

4.2 ChemBERTa-MLP Approach

The ChemBERTa based MLP model is proved to be a rather strong baseline in binary classification. We constructed a dataset with 63,473 positive pairs (from the original dataset) and 63,473 randomly generated negative pairs. 80% of those data were used in the training process, and the rest are used in the testing phase. The model trained on concatenated 1536-dimension embeddings achieved:

- **Test Set AUROC: 0.9178**
- **Test Set AUPR: 0.9214**



```
--- Model Evaluation (Test Set) ---
AUROC (Area Under ROC Curve): 0.9178
AUPR (Average Precision-Recall): 0.9214
```

However, when training on multi-class predictions, the model was performing rather poorly. The macro F1 and micro F1 were not promising

```
--- Model Evaluation (Test Set) ---

F1 Score (Micro): 0.0020 & F1 Score (Macro): 0.0002
```

5. Next Steps

(1) Graph-based modeling (Chi): Reproduce the RESCAL factorization model as a non-GNN benchmark. Additionally, implement relation graph convolutional networks (R-GCNs) Simultaneously, visualize and evaluate embedded vectors via t-SNE or UMAP to analyze drug-side effect clustering patterns.

(2) Transformer-based MLP modelling (Alisa): Refine multi-class prediction hyper-parameters.

References

1. Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(13), i457–i466.

2. Chithrananda, S., Grand, G., & Ramsundar, B. (2020). ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885