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

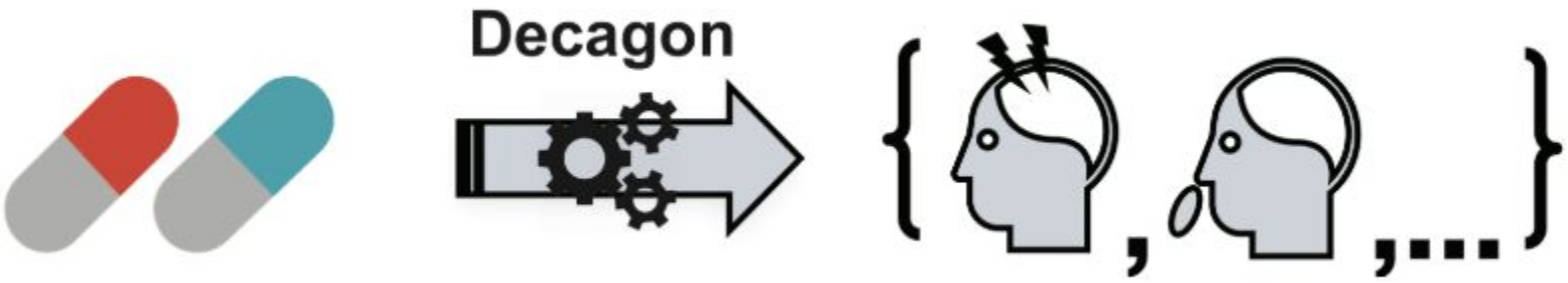
Chi Zhang (cz2925)
Jiayi (Alisa) He (jh5111)

Applied Data Science, Fall 2025, Forum B

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Introduction

- Polypharmacy is emerging as a significant issue → **holds risks** such as adverse drug reactions.
- These reactions are non-linear
  - not predictable from individual drug effects.
- Clinical trials cannot test every drug combination → computational modeling becomes essential.
- → Machine learning models for detecting polypharmacy side effects.
  - Graph-based models & Transformer-based models



Problem Statement

Objective

- Model drug–drug interactions (DDIs) as a link prediction task within multimodal biomedical networks.
- **Binary Prediction:**
  a. Predict if a drug pair (drug_i, drug_j) will cause any side effect.
- **Multi-Class Prediction:**
  a. Predict the specific side effect
  b. Each link represents a specific polypharmacy side effect connecting: Drug A – Side Effect – Drug B.

Problem Statement

Dataset

- Dataset used: **Decagon (Zitnik et al., 2018)**.
- Subset used in this project includes:
  - **4,649,441 triplets** *(Drug$_1$, Side Effect, Drug$_2$)*
  - **645 unique drugs**
  - **1,317 unique side effect types**
- Dataset provided in `.csv` format with fields:
  - **STITCH 1, STITCH 2, Polypharmacy Side Effect, Side Effect Name**

	STITCH 1	STITCH 2	Polypharmacy Side Effect	Side Effect Name
0	CID000002173	CID000003345	C0151714	hypermagnesemia
1	CID000002173	CID000003345	C0035344	retinopathy of prematurity
2	CID000002173	CID000003345	C0004144	atelectasis
3	CID000002173	CID000003345	C0002063	alkalosis
4	CID000002173	CID000003345	C0004604	Back Ache
...	...	...	...	...
4649436	CID000003461	CID000003954	C0149871	deep vein thromboses
4649437	CID000003461	CID000003954	C0035410	rhabdomyolysis
4649438	CID000003461	CID000003954	C0043096	loss of weight
4649439	CID000003461	CID000003954	C0003962	ascites
4649440	CID000003461	CID000003954	C0038999	bulging

4649441 rows × 4 columns

Problem Statement

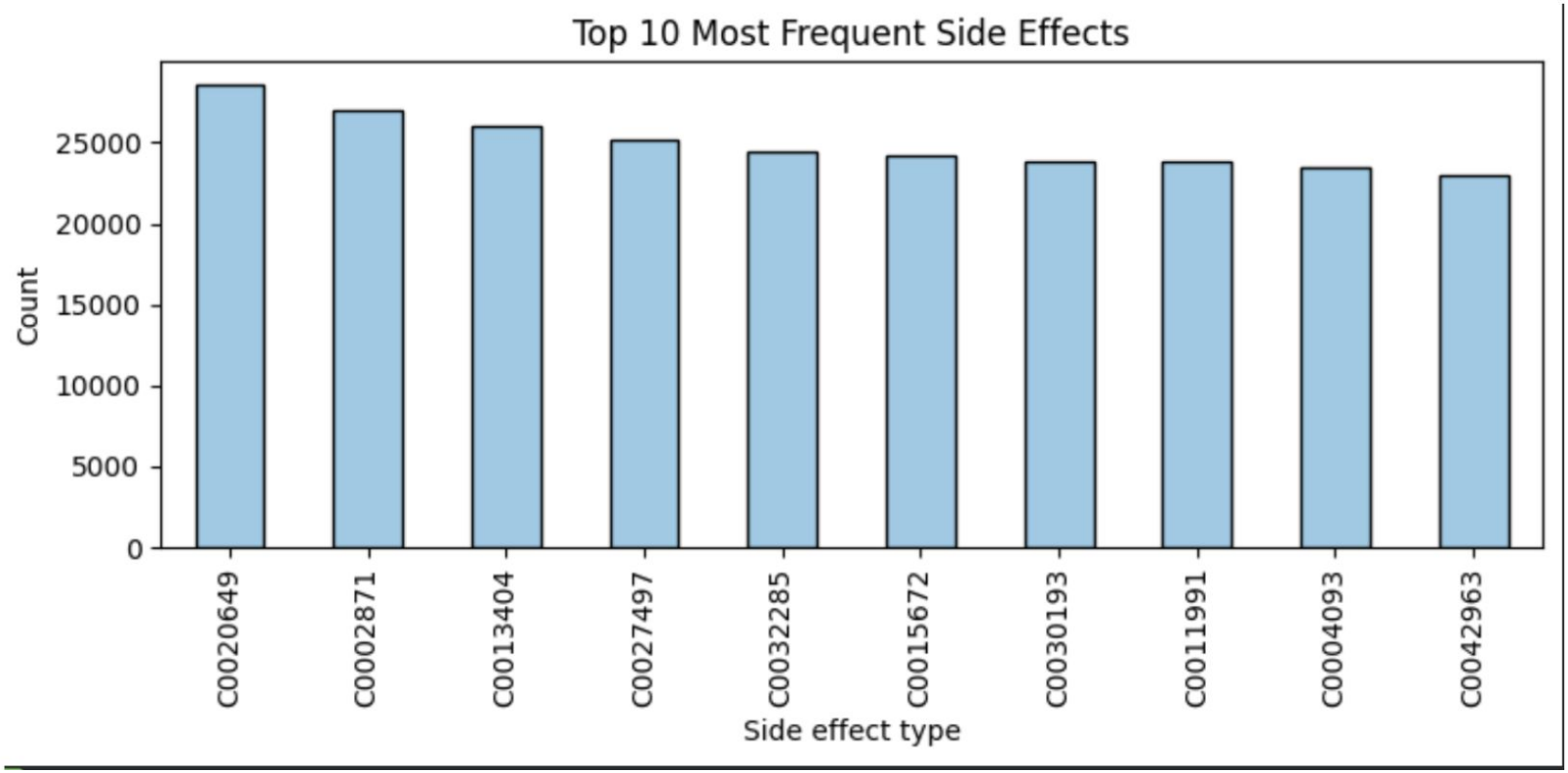
Objectives of This Stage

- Load, clean, and explore dataset characteristics.
- Quantify distributions of drugs and side effects.
- Visualize frequent side effects and drug–drug associations.
- Establish baseline models for comparison with upcoming graph-based methods.

Methodology - Preprocessing & EDA

This study effectively identified and visualized the 10 most common side effects using bar charts.

→ Highly imbalanced: Top few side effects appear >30k times, long tail dominates.

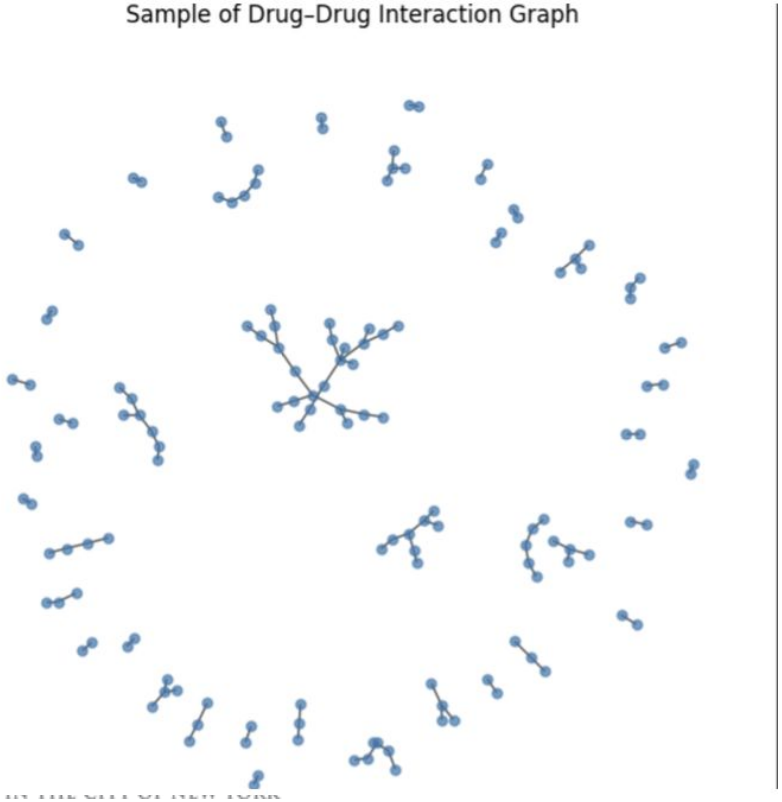


Methodology - Preprocessing & EDA

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

→ Graph is sparse and modular → ideal for graph representation.

Many drugs share similar side-effect clusters → suitable for embedding-based models.



Models: R-GCN (Main Graph Model)

We will further extend relational learning using R-GCN. The layer-wise message passing mechanism of R-GCN is defined as:

$$H^{(l+1)} = \sigma\left(\sum_{r \in R} \sum_{j \in N_r(i)} \frac{1}{c_{i,r}} W_r^{(l)} H_j^{(l)} + W_0^{(l)} H_i^{(l)}\right)$$

where $W_r^{(l)}$ denotes the transformation weight for relation r and $c_{i,r}$ is a normalization constant.

Similar to RESCAL, the link prediction score can be computed as:

$$\hat{y}_{ij}^{(k)} = h_i^\top R_k h_j$$

Pros

- Captures heterogeneous relations
- Learns drug embeddings from global structure
- Successful in Decagon benchmark

Cons

- Training cost higher
- Sensitive to relation imbalance

Models: RESCAL - Non-graph baseline

- Establish RESCAL for multi-relational link prediction.
- The factorization is obtained by minimizing

$$\min_{A,R_k} \sum_{k=1}^{K} \|X_k - AR_kA^\top\|_F^2$$

  - A (nxd): shared drug embedding
  - Rk (dxd): relation specific matrix for side-effect k
  - Xk (nxn): adjacency matrix for relation k

- The predicted likelihood for a drug pair is computed as

$$\hat{y}_{ij}^{(k)} = a_i^\top R_k a_j$$

Pros

- Fast
- Interpretable
- Strong baseline for knowledge graphs

Cons

- No message passing
- Cannot capture local graph structure
- Param-heavy for 1,317 side-effect relations

Models: ChemBERTa + MLP

- Pipeline:
  - Retrieve SMILES via STITCH CIDs $h_i = B(SMILES(d_i)) \in R^{768}$
  - Encode with pre-trained ChemBERTa → 768-dim embedding
  - Concatenate the embeddings
  - Feed the embeddings into MLP classifier $x_{ij} = [\, emb_i \,\|\, emb_j \,] \in \mathbb{R}^{1536}$

$$\hat{p}_{ij} = \sigma\left(W_2 \cdot \phi(W_1 x_{ij} + b_1) + b_2\right)$$

where ϕ is a nonlinear activation (e.g., ReLU), $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function, and W_1, W_2, b_1, b_2 are learnable parameters.

Pros

- Uses chemical structural information
- Pretrained model gives strong starting point
- Excellent binary classification performance

Cons

- Multiclass prediction extremely difficult
- Pairing function (concat) may be suboptimal
- No relational info like graphs have

Models: Metrics

We will be using the following metrics as our measure of success:

- **AUROC & AUPR & confusion matrix** for binary predictions
- **F1 score (Macro & Micro)** for multiclass prediction of drug combinations.
- **Graph embeddings (t-SNE/UMAP)** of drugs and side effects as additional visual analysis.

Models: ChemBERTa + MLP

- Binary Classification provides a strong baseline → chemical structure matters.
- Multi-class needs a little more work
- **Binary Dataset:** 63,473 real positive drug pairs + 63,473 randomly generated negative pairs
- **Multiclass Dataset:** we did no rebalancing in this experiment. Only real DDI pairs are used.

Task	Validation Error	Test Scores	Train Samples	Test Samples
Binary Class	0.21518567	AUROC = 0.9178 AUPR = 0.9214	Train = 101,556 (50,778 pos + 50,778 neg)	Test = 25,390 (12,695 pos + 12,695 neg)
Multi-Class	/	F1 (Micro) = 0.0020 F1 (Macro) = 0.0002	≈ 50,778 drug pairs (only positive pairs with ≥1 side effect)	≈ 12,695 drug pairs

Models: Graph Baseline

```
================================================
Summary:
Total triples: 4649441
Unique drugs: 645
Unique side effects: 1317
Baseline accuracy (most frequent SE): 0.0062
Figures saved to Google Drive.
================================================
```

Dataset Overview

- **Total triples: 4,649,441**
- **Drugs: 645**
- **Unique side effects: 1,317**
- **High-frequency effects (e.g., C0020649, C0002871)**

Key Observations

- **Most common side effects appear ~25k–30k times**
- **Drug-pair network is sparse and modular, reflecting limited co-labeled side effects**
- **High data imbalance → predicting without structure is difficult**

Baseline Result

- **Most frequent side effect baseline = 0.0062**
- **Indicates that blind guessing is nearly impossible**
- **Motivates using graph models to learn relationships**

Insight:

Side effects cannot be predicted from frequency alone — we must model drug interactions.

Discussion

What We Learned

- Chemical features alone → great binary predictor
- Chemical + relational structure → required for multiclass
- Graph sparsity aligns well with GNN message passing
- Side effect imbalance is the hardest challenge

Why models fail on multiclass

- 1,317 classes
- Long-tail distribution
- Drug pairs usually trigger only 1–2 effects
- Need relation-aware models (GNN, tensor factorization)

Future Work

Before End of Semester

- Finish RESCAL and R-GCN experiments
- Finetune ChemBERTa, and compare the above results to ChemBERTa
- Generate t-SNE / UMAP plots
- Error analysis by side-effect category
- Improve multiclass performance via:
  - Focal loss
  - Label smoothing
  - Hierarchical side-effect taxonomy

If Project Continues for 6 Months

- Try graph transformers (Graphormer, GROVER)
- Use cross-attention between drug structures
- Explore contrastive learning for drug representations

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