STAT 5243 Final Project Proposal: Benchmarking Graph-Based and Transformer-Based Models for Polypharmacy Side-Effect Prediction

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

Problem Description

Polypharmacy, defined as taking multiple types of drugs at once, is becoming a crucial topic in healthcare. Although it is effective in treating chronic and complex diseases, it increases risks such as adverse drug reactions (ADRs). Thus, the task of identifying high-risk drug combinations is crucial. Employing computational methods can help predict and screen high-risk drug combinations in an efficient way. Therefore, this project is dedicated to constructing and evaluating different machine learning models for detecting polypharmacy side effects.

Background and Context

Prior research demonstrates that graph-based models exhibit significant improvement in accuracy and performance. Decagon (Zitnik et al., 2018) introduced the Relational Graph Convolutional Network (R-GCN). This approach explored the construction of a heterogeneous graph comprising drug, protein target, and side effect relationships, effectively capturing the intricate connections. This model enables predictions across three domains: drug-drug interactions, protein-protein interactions, and drug-protein interactions. Another direction in this field is led by Transformer-based architectures. For instance, ChemBERTa (Chithrananda et al., 2020) utilized encoding SMILES strings and captured chemical features through its mechanism, achieving effective molecular representation learning without relying on graph structures. This project benchmarks Decagon, MLP with ChemBERTa, with baseline models (RESCAL).

Data

This project utilizes the publicly available Decagon dataset (Zitnik et al., 2018), which integrates multiple biomedical sources into a heterogeneous graph, including:

- Protein–protein interaction (PPI) network
- Drug–target protein associations
- Polypharmacy side-effect triples (drug A, side effect, drug B)
- Individual drug side effects
- Side-effect categories

Additionally, for the Transformer model ChemBERTa, although first-hand SMILES string data is unavailable, identifiers provided in the Decagon dataset will be used to retrieve corresponding SMILES strings from the DrugBank and PubChem databases. If the retrieval process fails, we would fall back to public benchmarks datasets. Finally, the potential side effects of combination drugs will be classified based on Decagon-generated triplets (drug_i, side_effect_r, drug_j) as the ground truth labels.

Methods

This project is a link prediction task.

Graph-based model approach:

- Construct a heterogeneous graph.
    - Nodes representing proteins, drugs, and side-effects.
    - Edges representing interaction (Protein-Protein, Protein-Drug, and Drug-Drug interactions).
- Evaluate RESCAL for knowledge-graph factorization, serving as non-GNN baseline.
    - RESCAL factorizes the graph's adjacent tensor (dimensions representing drug, drug, and side effect type) into drug entity embeddings and then uses them to predict the possibility of a side effect for a triplet of drugs.
- Evaluate R-GCN as the main model for prediction.
    - R-GCN used to generate drug embeddings.
    - A tensor factorization decoder (DistMult, etc.) is used to score potential likelihoods of side effects.

Transformer-based model approach:

- Use identifiers in DrugBank and PubChem to obtain SMILES strings.
- Input the strings into the pre-trained ChemBERTa to obtain embeddings.
- Concatenate the drug embeddings into a single feature vector representing a possible drug pair and feed into a classifier (MLP).
    - If time permits, we would explore different pairing strategies.

We aim to predict the interactions in two ways:

- Multi-label classification (predicting the specific types of side effects in one combination).
- Binary classification (predicting whether side effects will occur).

Evaluation Plan

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

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

Team Structure

- Chi will be working primarily with graph-based models such as Relational Graph Convolutional Network (R-GCN) and reproducing baseline RESCAL, and extend the analysis with new evaluation methods such as visualization of embeddings.
- Alisa will explore the transformer-based models, including retrieving the SMILES strings, fine-tune ChemBERTa, and feed the embeddings into the MLP model.