Polypharmacy Side-Effect Prediction with Graph Neural Networks

# Overview/Abstract:

Significant number of diseases directly or indirectly require combination of multiple drugs for treatment. Usage of multiple drugs can significantly alleviate the chances of adverse drug reaction event (ADE), hindering the treatment of primary underlying condition. Hence, it is utmost import to study the ADE, develop methodologies to predict the ADE and take preventative actions to avoid ADE. In this study we present a graph neural network (GNN) model that processes three types of interactions namely, drug to protein (in humans) interactions, protein to protein interactions, and drug to drug interactions, to detect an ADE corresponding to a pair of drugs. We used publicly available dataset XXX to train our model and achieved XXX% of accuracy in detection of an ADE corresponding to a given pair of drugs. The GNN model proposed in this study improves the accuracy by XXX% compared to a basic neural network classifier. With our results we conclude the proposed GNN model efficiently detects the critical patterns in the inter and intra-connections between drug and proteins, to accurately predict and ADE.

# Introduction:

Many complex diseases need combination of drugs for effective treatment. Additional comorbid conditions can also lead to addition of more drugs in a treatment plan. Such usage of multiple drugs for treatment of a single or multiple conditions is known as polypharmacy. Even though polypharmacy strategies have been effective in treating underlying conditions, they are associated with higher risk of an adverse drug reaction event (ADE). Therefore, for successful polypharmaceutical treatment it is critical predict and prevent an ADE.

For example, people living with HIV (PLWH) are treated with one of antiretroviral therapy regimes. Every regime is a specific combination of multiple antiretroviral drugs and regimes are changed if PLWH develop resistance to a regime. Hence, the entire treatment trajectory of PLWH involves consumption of combination of multiple drugs. For successful antiretroviral therapy, continuity of drug consumption and protection of patient’s hampered immunity is important. Occurrence of an ADE can disturb both, treatment continuity and immunity of PLWH, impede the treatment progress and put patient’s health at risk. Similar to HIV, treatment of many physical conditions include combination of drugs for effective treatment. Such treatments can be made more robust by developing mechanisms to predict ADEs.

Combined effect of drugs in treatment plan on human body and effect of chemical interactions between multiple drugs needs consideration for prediction of ADE. Such fundamental biological and chemical dynamics can shed light on patterns behind an ADE and if considered can improve the performance of prediction model. In this study, we consider protein-to-protein interactions and drug-to-protein interactions datasets to capture the biological and chemical dynamics [1]. We develop a prediction model based on graph neural network (GNN) structure to effectively process the available information for the prediction of occurrence of an ADE for a given pair of drugs. We also extend our model to predict type of an ADE if a given pair of drugs results in ADE.

# Problem:

In this study we specifically study the problem of occurrence of an ADE for a given pair of chemicals. Moreover, we also extend our problem to detect the type of ADE if an ADE occurs. In other words, given following information,

|  |  |
| --- | --- |
|  | A graph of protein-to-protein interactions in humans where,is set of unique proteins[[1]](#footnote-1) in humans and is the set of connections between proteins. This graph is unweighted and undirected graph. |
|  | A graph of chemical-to-chemical interactions (reported events of ADEs) where, is the set of unique chemicals[[2]](#footnote-2) and is the set ADE corresponding to two chemicals. Let . This is an undirected graph with links having multiple features. |
|  | A graph of chemical-to-protein interactions (reported in previously published studies). This is a directed and unweighted graph. |

We develop a model such that,

and

We develop such that it satisfies the following objective,

Where, is the set of model parameters and is the objective function to be minimized. We express as follows,

Equation 1: Objective function

Where,

is a negative sample of a chemical, sampled from set with a distribution

# Data:

We use data published by Zitnik et.al. [1] to develop the above proposed model, . The dataset consists of three components that are, protein-to-protein interactions, drug-to-protein interactions, and drug-to-drug interactions. We will discuss each one in detail in following sections.

## Protein-to-protein interactions:

The in the Problem section is the protein-to-protein interactions data. The data is read as a graph with nodes () representing unique proteins and edges representing the experimentally validated physical reactions between a pair of proteins (). Authors of Zitnik et.al [1] have compiled this dataset from multiple previously published and publicly available datasets [2]–[5]. This graph is undirected and unweighted. There are over 19,000 proteins and over 700,000 physical interactions documented in the graph. The node-degree (number of proteins connected to a protein in the graph i.e., number of proteins that has some validated interaction/s with a protein in the graph) distribution follows the power law with a mean value of 75 and median of 33. Eight components exist in the graph, the largest component subsume 99.9% of all the nodes while the others are small, disconnected sets of nodes.

## Drug-to-drug interactions:

The in the Problem section is the chemical-to-chemical (or drug-to-drug) interactions data. Authors of Zitnik et.al. have combined information from SIDER, OFFSIDES and TWOSIDES datasets [1], [6], [7]. SIDER dataset is created by mining ADEs from drug label text, while OFFSIDES and TWOSIDES datasets are created from reports from patients, doctors, and drug-companies on ADEs. Therefore, all datasets used for creating contains information on ADEs only. Nodes () in the final combined data represents unique chemical/drug and the edges (or ) between a pair of drugs represent the category of the side effect. If two drugs are not connected by an edge, then we assume that there is no side effect corresponding to the pair. There are 645 unique chemicals and over 63,000 side-effects documented in the final dataset. All side-effects present in the dataset belong to 561 unique categories. The number of chemicals (number of nodes) in the graph exponentially reduce with increase in the number of side-effects associated with the chemical i.e., there are very few chemicals with more than 300 documented side-effects associated to it.

## Drug-to-protein interactions:

Authors of Zitnik et.al. used STITCH dataset to create the drug to protein interaction dataset [1], [8]. The STITCH dataset consists of experimentally verified interactions between chemicals (over 500,000 in number) and proteins (little less than 9,000). The final curated dataset; in our study; consists of 645 unique chemicals and over 19,000 proteins. We refer the proteins which are connected to a chemical as target proteins for that specific chemical. The mean and the median of the number of target proteins in are 75 and 9, respectively. We also probed in the dataset to analyze the common side effects pairs of chemicals and that majority of pairs of chemicals do not have any target proteins in common. The 25th, 50th and 75th percentiles of the number of common target proteins (calculated as ratio of intersection set size to the union set size) are 0%, 0%, 0.6%. In other words, most of the pairs of chemicals do not have any target protein in common.

# Methodology:

The problem presented in the problem section is a classification problem i.e., we aim to classify a given pair of drugs into categories, say, ‘side-effects’ and ‘no-side-effects’. In this section we lay out parts of the final model and steps taken to solve the classification problem.

## Step 1: Initial node representation

In the available data, there are only two types of nodes available, drug or protein. For drug nodes we have list of side-effects associated with it (in addition to ). Hence, we utilized this information to derive initial node embeddings for drug nodes. Previously published pretrained language model; SciBERT; was used to encode the text information of side-effects into real valued vectors [9]. We specifically calculated average of all side-effect embeddings for every drug.

Therefore,

Equation : Initial drug node embeddings

Where,

|  |  |
| --- | --- |
|  | Real valued vector representation of drug node i.e., of type |
|  | Side-effects associated with drug |
|  | SciBERT, a language model [9] |
|  | Text corresponding to side-effect of drug node |
|  | A side effect in |
|  | Drug node |

Unlike drug nodes, protein nodes did not have additional text information than . To compute the initial node embeddings for protein nodes we leveraged unsupervised version of the previously published GraphSAGE method for inductive node representation [10].

Therefore,

Equation : Initial protein node embeddings

|  |  |
| --- | --- |
|  | Real valued vector representation of protein node i.e., of type |
|  | GraphSAGE, an inductive node representation learning method [10] |
|  | Protein node |

## Step 2: Classification model

# Results:

# Discussion/Observations:

# Conclusion:

# References

[1] M. Zitnik, M. Agrawal, and J. Leskovec, “Modeling polypharmacy side effects with graph convolutional networks,” *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, Jul. 2018, doi: 10.1093/BIOINFORMATICS/BTY294.

[2] J. Menche *et al.*, “Uncovering disease-disease relationships through the incomplete human interactome,” *Science*, vol. 347, no. 6224, p. 1257601, Feb. 2015, doi: 10.1126/SCIENCE.1257601.

[3] A. Chatr-Aryamontri *et al.*, “The BioGRID interaction database: 2015 update,” *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D470–D478, Jan. 2015, doi: 10.1093/NAR/GKU1204.

[4] T. Rolland *et al.*, “A proteome-scale map of the human interactome network,” *Cell*, vol. 159, no. 5, pp. 1212–1226, Nov. 2014, doi: 10.1016/J.CELL.2014.10.050/ATTACHMENT/8F86630B-9EB8-429F-B5B6-C2B8D18B237F/MMC7.XLSX.

[5] D. Szklarczyk *et al.*, “The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D362–D368, 2017, doi: 10.1093/NAR/GKW937.

[6] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, “The SIDER database of drugs and side effects,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1075–D1079, 2016, doi: 10.1093/NAR/GKV1075.

[7] N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman, “Data-Driven Prediction of Drug Effects and Interactions,” *Sci. Transl. Med.*, vol. 4, no. 125, p. 125ra31, Mar. 2012, doi: 10.1126/SCITRANSLMED.3003377.

[8] D. Szklarczyk, A. Santos, C. Von Mering, L. J. Jensen, P. Bork, and M. Kuhn, “STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data,” *Nucleic Acids Res.*, vol. 44, no. Database issue, p. D380, 2016, doi: 10.1093/NAR/GKV1277.

[9] I. Beltagy, K. Lo, and A. Cohan, “SCIBERT: A Pretrained Language Model for Scientific Text,” pp. 3615–3620, Accessed: Dec. 22, 2021. [Online]. Available: https://github.com/google-research/.

[10] W. L. Hamilton, R. Ying, and J. Leskovec, “Inductive Representation Learning on Large Graphs.”

# Appendix

# Tables:

# Figures

A screenshot of a computer

Description automatically generated with low confidence

1. : Note that the set is not exhaustive [↑](#footnote-ref-1)
2. : Note that the set is not exhaustive [↑](#footnote-ref-2)