Predicting safe drug combinations with Graph Neural Networks (GNN)

Amirhossein Amanzadi

Master of Science Program in Pharmaceutical Modelling 120.0 hp
Degree project within Bioinformatics 45.0 hp, June 2021
Center for Molecular Medicine (CMM), Karolinska Institute
Supervisors: Dr. Narsis A. Kiani, Karolinska Institute
Examiner: Prof. Ola Spjuth, Uppsala University
External opponent: Jiaxi Zhao
# Table of Contents

Abstract ............................................................................................................................................. 4

List of abbreviations ......................................................................................................................... 4

1 Background .................................................................................................................................... 5
  1.1 Importance of polypharmacy safety ......................................................................................... 5
  1.2 Safety versus efficacy ............................................................................................................. 5
  1.3 Challenges of identifying safe drug combinations ................................................................. 5

2 Introduction ..................................................................................................................................... 6
  2.1 How side effects occur? ........................................................................................................... 6
  2.2 DTI data bias ........................................................................................................................... 7
  2.3 Graph Neural Networks (GNN) ............................................................................................. 7
    2.3.1 GNN architecture ........................................................................................................... 7
    2.3.2 GNN characteristics ......................................................................................................... 8
  2.4 Relative work and state of the art ......................................................................................... 8
  2.5 Aim of the study ..................................................................................................................... 9

3 Materials and Methods ..................................................................................................................... 9
  3.1 Pipeline overview ................................................................................................................... 9
  3.2 Software and data used ......................................................................................................... 11
  3.3 Affinity calculation ................................................................................................................. 11
  3.4 Cutoff value .......................................................................................................................... 12
  3.5 Enrichment ............................................................................................................................ 12
    3.5.1 Enrichment modules ....................................................................................................... 13
  3.6 Model ....................................................................................................................................... 14
    3.6.1 Siamese Graph Convolutional Networks (SiGCN) ..................................................... 14

4 Results ............................................................................................................................................. 16
  4.1 DTI calculations (cutoff value) ............................................................................................... 16
  4.2 Enrichment ............................................................................................................................ 16
    4.2.2 PSA subgraphs ............................................................................................................... 17
  4.3 SiGCN optimal setup ............................................................................................................. 17
  4.4 SiGCN evaluation ................................................................................................................... 19

5 Analysis and discussion .................................................................................................................... 20
  5.1 Scalability of the model .......................................................................................................... 20
  5.2 Interpretability of the model .................................................................................................... 21
  5.3 Data sparsity and imbalance ................................................................................................... 21
5.4 Epochs versus accuracy ................................................................. 21
5.5 Absence of regularization .............................................................. 22

6 Suggestions for future studies ............................................................ 22

7 Conclusion ........................................................................................ 23

Acknowledgments ............................................................................... 23

Bibliography ....................................................................................... 23

Appendix .............................................................................................. 25
   A. Popular science abstract ................................................................. 25
   B. GitHub .......................................................................................... 25
Abstract

Many people -especially during their elderly- consume multiple drugs for the treatment of complex or co-existing diseases. Identifying side effects caused by polypharmacy is crucial for reducing mortality and morbidity of the patients which will lead to improvement in their quality of life. Since there is immense space for possible drug combinations, it is infeasible to examine them entirely in the lab. In silico models can offer a convenient solution, however, due to the lack of a sufficient amount of homogenous data it is difficult to develop both reliable and scalable models in its ability to accurately predict Polypharmacy Side Effect. Recent advancement in the field of representational learning has utilized the power of graph networks to harmonize information from the heterogeneous biological databases and interactomes. This thesis takes advantage of those techniques and incorporates them with the state-of-the-art Graph Neural Network algorithms to implement a Deep learning pipeline capable of predicting the Adverse Drug Reaction of any given paired drug combinations.

Keywords: Drug Combination, Polypharmacy Side Effect (PSE), Graph Neural Networks (GNN)

List of abbreviations

ADR – Adverse Drug Reaction
ANN – Artificial Neural Networks
AUROC – Area Under the Receiver Operating characteristic Curve
AUPRC – Area Under the Precision-Recall Curve
DDI – Drug-Drug Interaction
DTI – Drug-Target Interaction
GAT – Graph Attention Network
GCN – Graph Convolutional Network
GNN – Graph Neural Network
GPU – Graphics Processing Unit
HPC – High-Performance Computing
MPNN – Message Passing Neural Networks
PPI – Protein-Protein Interaction
PSA – Protein Side effect Association
PSE – Polypharmacy Side Effect
SiGCN – Siamese Graph Convolutional Networks
WHO – World Health Organization
1 Background

1.1 Importance of polypharmacy safety

Side effects caused by polypharmacy are a significant and rising public health issue that affects all parts of the healthcare sector around the world. Although the true extent of polypharmacy usage is unknown, it is expected to grow because of a variety of factors. To begin with, the global population is undergoing a demographic change, with the proportion of older people gradually increasing. Second, epidemiological evidence suggests that multimorbidity rises with age. According to the WHO, polypharmacy misuse accounts for 4% of all avoidable costs worldwide and its safe usage could save a total of 18 billion US equivalent dollars to %0.3 of global total health expenditure\(^1\). An ideal solution would have been developing medications that interact with multiple targets and could ultimately treat multiple diseases simultaneously. There are many challenges - especially in safety and toxicity - that polypharmacology must overcome to be recognized as a feasible solution\(^2\). Therefore, predicting the safety outcomes of combining drugs which is critical to maximizing its benefits, is a more viable option. This necessitates a methodical approach that considers the patient's medical conditions, comorbidities, allergy profiles, possible drug-drug, and drug-risk interactions, and recommends the best drug combinations\(^3\). Hence, it is indispensable to differentiate between safe and hazardous polypharmacy.

1.2 Safety versus efficacy

Drug combinations are beneficial not only in managing co-existing diseases (with polypharmacy), but also in eliminating a variety of complex diseases including cancer and infectious and resistant diseases to monotherapy such as AIDS, tuberculosis, and malaria\(^4\). They can improve treatment’s effectiveness and efficacy by targeting an illness through several pathways referred to as “Synergy”\(^5\). A substantial body of literature on various aspects of synergy is beyond the scope of this thesis. What matters the most is that, despite all its advantages, an increase in Synergy does not imply a safer treatment\(^6\). As a result, this thesis is primarily focused on the safety of drug combinations by predicting their ADR, as this is a more vital subject and a fundamental concern when treating comorbidity or complex diseases. To put it another way, the goal of this research is to determine the safety of pharmacological combinations by identifying their synergistic side effects.

1.3 Challenges of identifying safe drug combinations.

As previously mentioned, the safety of drug combinations is a global health issue, so why a widely accepted solution has not been developed accordingly? The answer is straightforward; when it comes to drug safety, lack of data remains often the problem. Less than 1500 of the 5000 (%30) commonly used drugs on the market\(^7\) maintain an accessible public registry of their off-label side effects\(^8\). Our predictive models would be redundant even though there
was a comprehensive database of side effects. As the data was compiled from patients with a broad range of conditions it would be extremely difficult to determine whether the medication or the disease triggered a particular side effect. In the literature, this topic is referred to as data being "heterogeneous". We must be able to integrate relevant aspects of patient conditions for harmonization and provide a consistent model.

The problem of data availability is indeed worse for drug combinations. There are 25 million possible paired drug combinations, but we only have public data for approximately 200 thousand of them, which covers less than 0.8%. Furthermore, there is currently no side effect database for drug combinations involving three or more medications. Thus, we cannot solely rely on the available data to build a general-purpose model capable of predicting any number of drug combinations; rather, we must develop a framework that mimics the reality of the side effect phenomenon and then fine-tune its performance using the data. Using a graph representation as a data structure would be a good fit in this case. They can not only handle and unify heterogeneous data but they can also be used to build advanced deep learning models with Graph Neural Networks (GNN) that are scalable. GNNs can be used to solve problems using "inductive reasoning". This implies that if we create a model for monotherapy and then use a GNN to predict PSE for paired drug combinations, we can infer that they can also predict PSE for polypharmacy (three or more drug combinations).

2 Introduction

2.1 How side effects occur?

The human body contains approximately 400 thousand different structural proteins/genes which constitute the biochemical foundation of our physiology. Since proteins carry out a key role in many aspects of human biology, their interactions referred to as Protein-Protein Interactions (PPI), have been extensively studied, published publicly, and regularly updated. Theoretically, every one of those proteins can be a potential drug target but in reality, the number of the druggable genome is drastically lower. Nevertheless, interaction with non-druggable protein can have a physiological effect. That is why proteins are typically the biological target of drug discovery projects. There is an extensive scientific exploration of the relationship between drugs and their targeted protein commonly known as Drug-Target Interactions (DTI). However, we must remember that the human body is very dynamic, as much as we like to drugs do not only interact with our protein targets but most of the time they bind with off-target protein which causes side effects. No matter how selective our small molecule drugs are a noticeable portion of them inevitably end up in an undesirable place. That is why there always side effects when using a medication, the importance is they should not be severe or life-threatening.
2.2 DTI data bias

DTI is super useful for the identification of novel therapeutic targets and biomarkers, but it is significantly unbalanced when it comes to off-target drug interactions. The root of this data bias is in the way DTIs are constructed. In the process of drug development, the drug agent only gets tested against few targets of interest. Even for toxicology screening, it only gets examined with few critical proteins. Ergo, it is realistic to say that DTIs are unsuitable data for a model that predicts side effects since it wholly contains a limited portion of potential drug-protein interactions and offers no off-target interactions. This is one of the key reasons why so many models failed to accurately predict Drug-Drug Interactions (DDI) or side effects. This thesis took and data-centric approach to deal with this data bias and instead of using DTI, we calculated the binding affinity of each drug with the entire human PPIs. This approach will make our model highly flexible and scalable as its no longer bound to any external DTIs.

2.3 Graph Neural Networks (GNN)

GNN models are currently one of the most heated topics in the deep learning community. They may seem overly complex but the fundamental ideas behind them are rather axiomatic. The world-class attraction of GNN algorithms lies in their ability to handle non-sequential data and precisely predict various features from them. Unlike many other deep learning methods that rely thoroughly on sequential data such as text, images, audios, and videos, GNN can work with non-sequential data whose spatial and temporal states can be considered arbitrary such as graph networks, molecules, interactomes, connectomes, and any other data format that describes the relationship between entities. This obtains GNNs with high versatility as most of our current data sources from social networks to knowledge graphs, gene regulations, and recommendation systems are or can be converted to unstructured data.

2.3.1 GNN architecture

In GNNs instead of defining the Artificial Neural Networks (ANN) ourselves, the input graph network’s nodes (named A-F in figure 1) will be considered as the artificial neurons and their relationship with other neighboring nodes (edges) will produce the neural layer. In figure 1 For instance, node A includes three neighbors B, C, and D which their relations are illustrated via edges. Then B, C, and D will establish the first layer. Their values times the weight of the edges plus the baseline will aggregate information from the first neural layer to node A and this is how it can predict features of both nodes and edges using GNNs. The second layer of the GNN is consists of neighbors of the neighbors of node A, commonly referred to as second-degree neighbors. Throughout the GNN training process, these layers get made for every node.
2.3.2 GNN characteristics

Typically, GNN extremely shallow layers rarely extend to more than 6 layers, because according to the rule of “six degrees of separation”, any two nodes in the graph network can get connected with a maximum of six edges. In other words, in the worst-case scenario, all the data in the graph network can be abstracted by reaching the sixth-degree neighborhood of a single node. In summary, we will acquire overlapping and duplicated data if go deeper than six layers, Thus the majority of GNNS have two or three layers. GNNs can be classified based on two prominent characteristics, first their aggregation algorithms, and the second which part of the graph predicts its features. They are three mainstream techniques for GNNS algorism, Message Passing Neural Networks (MPNN), Graph Convolutional Networks (GCN), and Graph Attention Networks (GAT). This thesis will only utilize GCN for the PSE prediction. The dominant purpose of GNN predictions is either node classification, link prediction, or graph classification which is the focus of this thesis.

2.4 Relative work and state of the art

There has been a constant improvement in both capacity and accessibility of high-performance computing (HPC). With this increase in computing power, it became possible to confront a comprehensive range of problems using computational methods. Biomedical research had substantially benefitted from this and there has been an absolute boom of exciting computational research in this field. This paradigm shift in biomedical and pharmaceutical research and development is profoundly influenced by their cost-reducing nature, scalability, reproducibility, and their exceeding performance. Drug combinations are no exception and they have been the focus of many publications in recent years. They have been a variety of mathematical models, Synergy estimators, machine learning, and deep learning approaches to predict the outcome of a drug combination. This thesis will focus exclusively on the representation learning approaches like GNNs as they are the most
prominent, have produced the most accurate results, and are thereby considered as state-of-the-art models.

DECAGON\textsuperscript{17} (Zitnik et al., 2018) remains the current state-of-the-art model that takes a link prediction approach and utilizes the Graph Convolutional Network (GCN) to predict 964 adverse PSE from a multimodule graph consisting of PPI, DTI, and DDI networks as shown in figure 2A. It uses a double convolutional layer architecture. The Heterogeneous graphs are first transformed into an adjacency matrix (shown in an orange rectangle in figure 2B) and then a GCN layer with a Leaky RELU activation function is applied, followed by regularization (0.1 dropouts) and then a second GCN layer that predicts the PSE between the DDI nodes. DECAGON has reported AUROC of 0.872 and AUPRC of 0.832 as performance metrics. Each one of the 964 side effects is an independent edge feature that makes the DECAGON computationally intensive model even for HPCs and GPUs. Moreover, it is limited to DDI nodes that are already in the training data which as a result, makes the model unscalable.

![Figure 2](image.png)

**Figure 2.** Overview of the current state-of-the-art models DECAGON (Zitnik et al., 2018). A) Multimodule graph network of the DECAGON, consisting of DDIs, DTIs and PPIs. B) Model architecture of DECAGON’s double GCN layers.

### 2.5 Aim of the study

Although the focus of this thesis is to predict the PSE of any given drug pairs, the primary goal of this research is to establish a holistic method that addresses fundamental problems with current approaches (see sections 1-2) and lays the foundation for the acceleration of drug combination research and development. This research aims to provide a versatile and scalable framework that can be customized to meet the needs of modern R&D pipelines.

### 3 Materials and Methods

#### 3.1 Pipeline overview

Human PPI is an invariable graph network structure that stays homogeneous since mutations can only alter the expression of proteins but do not influence their interactions with other proteins. Because of this key insight, the heterogeneous data problem (see section 1.3) was
resolved by embedding all different aspects of the datasets into the PPI. An overview of this pipeline is presented in figure 3. All the tasks performed in this pipeline, have been fully automated and its source code can be accessed from the public repository of this thesis (see appendix).

In the first step, to simulate interactions of a drug entering the body, it was assumed that the drug can reach all the proteins in the PPI network. In other words, every protein in the PPI can be a target. Next, to identify proteins that the drug can potentially bind to, the binding affinity (pKs) between the drug and the entire protein was calculated using external APIs (see section 3.3). For the drug to make a binding, it must push away the solvents surrounding the protein and have a high enough binding affinity to stay bound. This thermodynamical precondition was also estimated based on the affinity between individual proteins and their solvents and is parametrized as a cut-off value (see section 3.4).

Moving forward, affinities were filtered based on the cut-off value of each protein. As a result, all possible proteins that the drug can interact with are identified regardless of them being a target or off-target which represents a tremendous accomplishment since the DTI data bias has been completely avoided (see section 2.2). Afterward, all 964 selected side effects of the drug can be assigned to the bonded protein as a node feature (indicated as colorful boxes). This makes it possible to link side effects to proteins. The theory is that this protein could be responsible for the development of a side effect somewhere in the pathway; it is not a causal relationship, but rather a reasonable association. The Protein-Side Effect Association (PSA) cannot, of course, be built solely based on a particular drug. Thus, in the final step of the pipeline, after creating PSAs of every drug in the SIDER data set, a final PSA is established with

![Diagram](image)

**Figure 3.** Overview of the data pipeline that establishes the Protein-Side effect Association (PSA). At first it is assumed that the drug can reach the entire PPI and its respective binding affinity is calculated. With cut-off filtering the number of possible interactions is significantly reduced and remaining proteins considered to be associated with recorded side effect of the drug. Enriching this data for all drugs will result in establishment of the Protein-Side effect Association (PSA).
an Enrichment process that will be discussed further in the thesis (see section 3.5). PSA will later provide the inputs for the GNN model.

### 3.2 Software and data used.

The SIDER 4.1 database was selected for drug side effects because it includes 1430 medications and over 5000 branded and off-labeled side effects. The STRING database, which contains 19,567 human proteins with their sequences and interactions, was chosen for the PPI network. For the cut-off value, SIB-Expasy was used to measure the isoelectric point (pI) of the entire PPI network from their sequences. The binding affinities were predicted by DeepPurpose API which contains 15 different pre-trained ML models for various DTI calculations. PubChem serves as the primary source for providing drug SMILES data. TWOSIDES, that dataset was used for the final GNN model, and it contained approximately 200,000 unique drug combinations. 104,000 of the drug combinations in TWOSIDES were from the drugs in the SIDER dataset and these combinations were used to train the final GNN model. 10’000 of the combinations did not contain any of the drugs in the pipeline, so they were utilized for the evaluation of the trained GNN model. A broad variety of Python libraries were used, including Numpy and Pandas for data handling, Pytorch for model training and optimization, DGL, and NetworkX for GNN construction. All tests were carried out within singularity containers and were powered by the CMM internal server, which features eighty Intel Xeon CPUs, four Nvidia Tesla V100 GPUs, 780 GB of RAM, and 30 TB of hybrid storage.

### 3.3 Affinity calculation

The binding affinities were predicted using DeepPurpose. DeepPurpose is an open-sourced deep learning toolkit developed by Harvard University during the height of the pandemic to accelerate drug development. The DTI module was used to predict the dissociation constant (pKd) or binding affinity between drugs and protein. The modules took the protein sequence and drug SMILES as inputs and then encodes them into one of the pre-trained models (Morgan_CNN_BindingDB in our case) and output the pKd. With all its advantages, the API was not designed to offer a scalable solution. The pipeline needed to predict 27 million affinities, which necessitated extensive code optimization for it to run in a reasonable amount of time. The time it takes to run a single prediction has been reduced from 1100 ms to 1.2 ms because of the optimization and GPU parallelization. This thousand-fold increase meant that instead of 12 months, the entire DTIs could be predicted in just 12 hours.

\[
pK_d = -\log \left( \frac{|P|\cdot|D|}{|PD|} \right)
\]

\[
pK_a = pH - \log \left( \frac{[Base]}{[Acid]} \right)
\]

**Figure 4.** Definitions of dissociation constant (pKd) and acid dissociation constant (pKa).
3.4 Cutoff value

In theory, a drug should be able to replace itself with bound solvents if it wishes to bind to a protein. This thermodynamical requirement does not ensure binding; on the contrary, it demonstrates the possibility of binding. The dissociation constant does not represent the genuine value of binding affinity which is in the forms of the free energy or $\Delta G$, as it is indicated in equation 1. This implies that the cut-off value should be in the same dimensions. The appropriate parameter to describe the cut-off value is the protein pK$_a$ (acid dissociation constant) since it formulates the dissociation constant of the solvent (figure 4). As a result, the pK$_d$ of each protein would be the cut-off value, and the binding affinity would have to be larger than it to pass the filtration. On the other hand, determining the pK$_a$ of a protein is difficult from a development standpoint because it requires not only the protein’s 3D structure but also the pH of the environment in which it is expressed.

To resolve the matter, isoelectric point (pI) was adopted as pH, where the protein is in its purest form. This approach will significantly simplify the situation as pI calculations are quick and require only the protein sequence. Because the protein is electrically neutral at the isoelectric point, the concentrations of base and acid in the buffer are nearly equal, resulting in a cut-off value becoming equal to pI. (equation2).

$$\Delta G = RT \ln \frac{K_d}{c^\circ}, \text{ Standard reference concentration } c^\circ = 1 \text{ mol/L.} \quad \text{Equation 1.}$$

$$pK_d > pK_a = pH - \log \left( \frac{[\text{Base}]}{[\text{Acid}]} \right) \rightarrow (\text{Neutral}); pK_a = pI \rightarrow pK_d > pI \quad \text{Equation 2.}$$

3.5 Enrichment

The enrichment procedure is critical in the establishment of the final PSA. For a protein to be linked to a side effect, two conditions must be fulfilled. To start with, the protein must interact with the drug identified by cut-off filtering, and second, the drug must have that side effect. If both criteria are met, it is safe to assume there is a link between a protein and a specific side effect. Matrix multiplications obtain the most efficient way to accomplish this. As shown in figure 5, the first matrix contains all drug-protein interactions, while the second matrix contains all drug side effects. The multiplication of these two matrices will calculate the association score for all proteins. Each row in the first matrix (outlined green) represents a specific protein, each column in the second matrix (outlined blue) represents a specific side

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad 0\times0 + 1\times1 + 1\times1 = 2$$

Figure 5. Enrichment process that calculated the association scores of PSA.
The association score of given protein and a side effect is calculated by multiplying these numbers. It is notable to remember that the association score only represents the probability of an association; PSA is not a deterministic model in general. In other words, the association score can be considered as the quantitative potential of synergistic side effects. All the data in the matrix was purposefully kept binary, and matrix multiplication was kept in its most rudimentary linear form. The GNN models are developed in such a way that they boost the performance of the PSA.

### 3.5.1 Enrichment modules

The fundamental advantage of this method is it is incredibly versatile, and numerous modules can be added to the enrichment process. In the protein-drug matrix, the binary values can be replaced by dose values. The calculated binding affinities assume a 1:1 mole interaction between each drug molecule and the protein which is unrealistic. There are different amounts of proteins in the body and not all of them have the same expressions. Also, there is a difference between consuming the standard dose of a drug and the much higher life-threatening dose. With the combinations of equation 1 and the chemical potential of the drug ($\mu_d$), proof 1 indicates that by multiplying the molar values of proteins ($N_p$) and drugs ($N_d$) with the calculated pK$_d$ it is possible to accurately represent protein expression and dose values. Thus, a new improved binding affinity value ($pK_d^*$) can introduce the influence of dose and gene expression in the enrichment process.

### Proof 1.

Combining the rewritten affinity equation (1) and chemical potential of the drug (2) will prove that by multiplying molar values of protein ($N_p$) and drug ($N_d$) with the calculated pK$_d$ it is possible to accurately represent protein expression and dose values. Thus, a new improved binding affinity value ($pK_d^*$) can introduce the influence of dose and gene expression in the enrichment process.

\[
\begin{align*}
(1) \quad \Delta G &= RT \ln \frac{K_d^*}{c^0} \rightarrow pK_d = -\frac{c^0 \log e}{RT} \Delta G \\
(2) \quad \mu_d &= \left( \frac{\partial G}{\partial N_d} \right)_{T,P,N_p} \rightarrow \partial G = \int_0^{N_d} \mu_d \times N_p \partial N_d \rightarrow \Delta G = \mu_d N_p (N_d - 0) = \mu_d N_p N_d
\end{align*}
\]

\[
\begin{align*}
(1\&2) \quad pK_d &= -\frac{c^0 \log e}{RT} \mu_d N_p N_d \rightarrow pK_d^* = pK_d N_p N_d
\end{align*}
\]

Additionally, the values in the drug side effect matrix can be altered and replaced by the occurrence frequency of the side effects. This will affect the efficiency of the enrichment and ultimately enhance the performance of the GNN model. Other modules such as tissue expressions can be added to the enrichment process, but it is important to mention that it must be in a dimension of the proteins. To achieve this, the protein-tissue matrix should first be multiplied by its transposed matrix and then it can be added to the enrichment process (Equation 3).

\[
M_{protein \times tissue} \times M_{tissue \times protein}^T \times M_{protein \times drug} \times M_{drug \times SE} = M_{protein \times SE}
\]

### Equation 3.

Indicated how tissue expression matrix can be correctly added to the enrichment
3.6 Model

The model had to overcome two considerable obstacles to be successful. First, based on its graph representation, it should determine the side effect of any drug, and second, predict the PSE of any given drug pairs. Put differently, the model should be adequate to receive paired graph inputs, as opposed to traditional models that only received one input and predicted a single outcome. The model attains this by employing a deep learning technique known as Siamese Neural Network. This method is frequently employed to create one-shot learning models. The goal of one-shot learning is not to classify given inputs, but rather to determine whether the inputs are the same or not. Siamese requires very little input and performs incredibly well and quickly\(^\text{18}\), which linger why it is exerted in the industry as an identification algorithm. As indicated in figure 6, given images first go through the same neural network (shared weights) and then get aggregated to a fully connected layer. The loss function is defined based on the differences between the output of the neural network rather than the external data set. In our case, the Siamese method was implemented in conjunction with two GCN layers for the establishment of the final model (see section 3.6.1).

![Figure 6. Architecture of Siamese Neural Network utilized for a one-shot learning image recognition. (Subramaniam et al., 2016)](image)

3.6.1 Siamese Graph Convolutional Networks (SiGCN)

The PSA produces 964 side effects as its node feature, meaning each protein in the PSA has an assigned tensor that contains 964 cells in binary values (see section 3.5). The Siamese Graph Convolutional Networks (SiGCN) extracts the two PSA subgraphs as inputs which are the corresponding DTIs of each drug that were built during the pipeline. And then these two subgraphs will go through the same double-layer GCN with ReLU as an activation function which outputs a tensor with 964 values that are between 0 and 1. These values represent the single-use side effect of the corresponding drug which resolves the first obstacle (see section 3.6). Right away, the two tensors get aggregated together representing the polypharmacy side effect (PSE) which overcomes the second obstacle (see section 3.6). Naturally, the predicted PSEs are not correct, and they must be trained using the true labels and get...
optimized by the loss functions. All the above steps have in illustrated in SiGCN architecture (figure 7).

Various configuration of loss functions, aggregations, epoch, and batch size of the model was examined and the most accurate one was designated as the final model (see section 4.3). Labels came from the TWOSIDES dataset (see section 3.2) which contains the actual PSE of the drug combinations, 80% were used for training and 20% for testing. The data was delivered in batches to speed up the training process. Since the inputs of the model are different graphs and the output is a single tensor, the SiGCN can be categorized as a graph classification model. To see how well the model performed with data outside the training set, its predictions were evaluated by 10,000 unique combinations of drugs that were not present in the training and the entire pipeline (see section 3.2).

Figure 7. Architecture of Siamese Graph Convolutional Network (SiGCN) for constructed to predict of the PSE of given PSA subgraphs.
4 Results

4.1 DTI calculations (cutoff value)

A considerable fraction of the interactions were removed after applying the cut-off filtering. The filtration omitted 78% of all interactions. In figure 8, The left heatmap represents 27 million DTIs, and the left one indicates the DTIs that have been filtered. As the result of the filtration, only the strong interactions remained, and many proteins and drugs were removed which appear as white dots or lines. What's fascinating is that, despite the loss of 78 percent of the data, DTI's structure integrity remained intact, implying that the filter eliminated a significant percentage of the noise while keeping the signals in place. The same process was repeated for the evaluation set and the corresponding filtration rate was 83%.

![Figure 8. Cut-off filtration result of the pipeline. Left, unfiltered DTIs and right filtered DTIs.](image)

4.2 Enrichment

The established PSA contained 5085 unique drug side effects as illustrated in the treemap (figure 9). It was intriguing to see a handful of side effects are associated with more than half of the PPI proteins. In other words, many proteins can be responsible for the occurrence of those side effects. The highest-ranked side effects, based on their normalized association scores, were the systematic side effects that can be caused by many proteins. To name a few, vomiting, nausea, rash, pain, and headache were among those top-ranked side effects. In contrast, there was a multitude of side effects that were associated with a few proteins. These proteins can be used to identify new biomarkers and even potential targets that make the management of rare side effects possible in the future. Among all the side effects, only 964 adverse side effects (ADR) were chosen for the final GNN model. The selection of these ADRs were also based on the benchmark data of the DECAGON model.
4.2.2 PSA subgraphs

With the calculated DTI during the pipeline, the PSA subgraph of each drug can be extracted from the PSA. To get some idea about the extension of these subgraphs, the PSA graph network of “Glyburide” is demonstrated in figure 10 as an example. As is indicated, there a high order of interactions between each node thus for simplicity the PSA subgraphs were illustrated with a small number of nodes and edges throughout the thesis report.

4.3 SiGCN optimal setup

The SiGCN is a graph classification model, and the acceptable metric for measuring the performance of these models is accuracy. Since the model outputs the 964 long tensors, the accuracy would represent the ratio of PSE values that was predicted correctly (true positive
and true negative. Single epoch runs on a sample of the training data were performed to determine which exact configurations are best suited for model training (see Table 1). Runs 8 and 9 both showed promising results, so it was decided to execute them on full-sized training data with different aggregation modes to decide the final SiGCN configurations. The result of runs 10-13 (Table 2) indicated that the leading model (run 11) has binary cross-entropy as loss function and mean pooling as aggregation method. This configuration was applied on run 14 which is the full-scaled run trained over 50 epochs. With some code optimization, the memory usage of the model was reduced by 50%. The optimal model’s hyperparameters were examined to be a learning rate of 0.01 with Adam optimizer, GCN double layer with (964, 200, 964) convolutions margin.

**Table 1.** List of different configurations of SiGCN with their corresponding performances, tested on a small sample of the training set.

<table>
<thead>
<tr>
<th>Run no.</th>
<th>Accuracy %</th>
<th>Loss function</th>
<th>Layers</th>
<th>Batch size</th>
<th>Time-per epoch (h)</th>
<th>Epochs</th>
<th>Memory (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36.50 ± 6</td>
<td>Triplet distance</td>
<td>2-GCN</td>
<td>5</td>
<td>18.5</td>
<td>1</td>
<td>250</td>
</tr>
<tr>
<td>2</td>
<td>85.73 ± 2</td>
<td>Binary cross-entropy</td>
<td>2-GCN</td>
<td>5</td>
<td>19.1</td>
<td>1</td>
<td>230</td>
</tr>
<tr>
<td>3</td>
<td>86.56 ± 3</td>
<td>Cosine similarity</td>
<td>2-GCN</td>
<td>5</td>
<td>18.9</td>
<td>1</td>
<td>225</td>
</tr>
<tr>
<td>4</td>
<td>89.32 ± 1</td>
<td>Binary cross-entropy</td>
<td>2-GCN</td>
<td>20</td>
<td>3.2</td>
<td>1</td>
<td>320</td>
</tr>
<tr>
<td>5</td>
<td>87.19 ± 1</td>
<td>Cosine similarity</td>
<td>2-GCN</td>
<td>20</td>
<td>3.5</td>
<td>1</td>
<td>327</td>
</tr>
<tr>
<td>6</td>
<td>NA</td>
<td>Binary cross-entropy</td>
<td>2-GCN</td>
<td>100</td>
<td>NA</td>
<td>1</td>
<td>750+</td>
</tr>
<tr>
<td>7</td>
<td>84.01 ± 4</td>
<td>Binary cross-entropy</td>
<td>3-GCN</td>
<td>25</td>
<td>8.9</td>
<td>1</td>
<td>423</td>
</tr>
<tr>
<td>8</td>
<td>91.27 ± 2</td>
<td>Binary cross-entropy</td>
<td>2-GCN</td>
<td>35</td>
<td>4.6</td>
<td>5</td>
<td>346</td>
</tr>
<tr>
<td>9</td>
<td>91.13 ± 3</td>
<td>Cosine similarity</td>
<td>2-GCN</td>
<td>35</td>
<td>4.5</td>
<td>5</td>
<td>345</td>
</tr>
</tbody>
</table>

**Table 2.** List of different configurations of SiGCN with their corresponding performances, tested on the full-sized training set.

<table>
<thead>
<tr>
<th>Run no.</th>
<th>Accuracy %</th>
<th>Loss function + Aggregation mode</th>
<th>Layers</th>
<th>Batch size</th>
<th>Time-per epoch (h)</th>
<th>Epochs</th>
<th>Memory (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>87.63 ± 1.4</td>
<td>BCE + Max pooling</td>
<td>2-GCN</td>
<td>35</td>
<td>6.3</td>
<td>10</td>
<td>143</td>
</tr>
<tr>
<td>11</td>
<td>89.03 ± 0.8</td>
<td>BCE + Mean pooling</td>
<td>2-GCN</td>
<td>35</td>
<td>6.4</td>
<td>10</td>
<td>145</td>
</tr>
<tr>
<td>12</td>
<td>79.16 ± 1.2</td>
<td>COS + Max pooling</td>
<td>2-GCN</td>
<td>35</td>
<td>6.2</td>
<td>10</td>
<td>142</td>
</tr>
<tr>
<td>13</td>
<td>79.95 ± 1.1</td>
<td>COS + Mean pooling</td>
<td>2-GCN</td>
<td>35</td>
<td>6.3</td>
<td>10</td>
<td>146</td>
</tr>
<tr>
<td>14</td>
<td>85.68 ± 0.6</td>
<td>BCE + Mean pooling</td>
<td>2-GCN</td>
<td>35</td>
<td>10</td>
<td>50</td>
<td>174</td>
</tr>
</tbody>
</table>
4.4 SiGCN evaluation

The fully trained SiGCN was given the evaluation dataset in batches (see section 3.2). Throughout the 301 batches, the average accuracy obtains 91.43 percent (figure 11). This means the model correctly predicted the PSE of 9143 drug combinations out of the total 10,000 instances in the data. The results indicate a normally distributed prediction with an elevated concentration in the (92, 96.5] group. Despite high accuracy, the model has inferior average precision and F1 score which are respectively 48.03% and 30.09% (Figure 12 and 13).

![Figure 11. Histogram of evaluation dataset accuracy vs. the number of the batches.](image1)

![Figure 12. Histogram of evaluation dataset precision vs. the number of the batches.](image2)

![Figure 13. Histogram of evaluation dataset F1 score vs. the number of the batches.](image3)
5 Analysis and discussion

We created a representation learning algorithm that can accurately (88% accuracy) predict the PSE of given paired drug combinations. This was achieved by combining all heterogeneous data into a single entity of the PPI network tailored for side effect predictions (PSA). By providing the molecule’s SMILES, our pipeline made it easy to construct this representation for any drug. On other hand, the performance of SiGCN on the evaluation set (91% accuracy) engendered confidence and confirmed the superior capability of our model. Few key insights need to be discussed in the following.

5.1 Scalability of the model

The Siamese technique allows scaling of the input to any number of drug combinations. Because the model weights were trained based on the aggregation of the inputs, not their genuine values, and since mean pooling endure a non-hierarchical operation/aggregation (meaning is not sensitive to change in order or scale), we can confidently infer that model is scalable. Also, since the model first constructed a tensor-based on singular used side effect of the PSA and then aggregates them with another drug to predict PSE; based on inductive reasoning (see section 1.3) we can infer the prediction of the scaled model is meaningful and will be producing accurate results. 

Apart from scalability in the number of input drugs, the model is also capable of predicting the PSE of any drug just from their SMILES structures. Since the model was developed to represent chemical interactions thus can reliably make predictions based on different interactions. As indicated in figure 14, with the SMILES string of any desired drugs the pipeline will generate the PSA subgraphs and the SiGCN will predict the PSE. This process is automated and no adjustment to the code is required.

![Diagram of automated procedure](image-url)

**Figure 14.** Schematic of automated procedure of predicting PSE of unknown drugs only with their SMILE structure. Providing the SMILES will allow the pipeline to generate the PSA subgraphs which will be fed to the SiGCN for PSE prediction.
5.2 Interpretability of the model

The outcome of our model is also highly explainable. Every prediction can be backtracked to the proteins associated with those side effects, and proteins that appear in both drug’s PSA subgraphs can be interpreted as the source of the PSE. The other advantage of the model is the common protein (figure 15) in the drug combination can be examined in terms of their tissue expression and pathways and we can conclude which exact proteins are causing the PSE. For none of the investigations, we need to conduct wet-lab experiments, all those analyses can be performed in modern and thoroughly reliable software. Constructing this PSA embedding, allowed our model and pipeline to not only be interpretable but fit the existing R&D frameworks.

Figure 15. Interpretability of the SiGCN were by identifying the common proteins (right side graph) in the input PSA subgraphs (left side graphs), they can be considered the source of the predicted PSE.

5.3 Data sparsity and imbalance

A perfect model does not exist and our SiGCN model is no exception. The model suffers from a limitation that it cannot regulate but can improve. The problem lies in the reality of PSE data. It is rare for a drug combination to consist of most of the 964 PSEs, so our training data suffer from scarcity. During training, the data forces model to predict 0 for many values of the tensor since it provides a higher probability of obtaining a valid answer. Consequently, despite our high accuracy, the model suffers from mediocre precision of 48.03%. Put differently, the model predicts what PSE will not occur because of the drug combination, much better than what PSE would occur. Furthermore, the low 30.09% F1 score indicates that there is an imbalance in the data This may be improved with the addition of new modules to the pipeline but is ultimately the drawback of the general model of SiGCN. A sophisticated SiGCN that is focused on a specific type of disease can also improve the precision of the model and equalize the imbalance of the data.

5.4 Epochs versus accuracy

As the number of epochs increases the accuracy of the model slowly declines. There is an adequate explanation for this. Because SiGCN has utilized the Siamese technique, it can be
considered as a one-shot learning model. In one-shot learning, the model reaches its peak performance with few epochs. Training the model with more epochs forces the optimizer to move farther away from the local minimum it had found. Therefore runs 8 and 9 with few epochs (table 1) demonstrated a considerably higher accuracy compared to runs 10-14 which were executed on more epochs (table 2). Furthermore, this explains why the evaluation set accuracy resembles runs 8 and 9 instead of runs 14. The best way to overcome this is to find the optimum epoch size. In addition, this problem can be eliminated by finding a suitable optimizer and an optimum learning rating.

5.5 Absence of regularization
For each prediction, the SiGCN model employs a graph classification architecture, which uses different graphs as inputs. Because each graph builds its own neural network (see section 2.3.1), the model comprehends nothing about the graph itself but rather how to best convolute its features. The applicability of any regulation is severely limited by the dynamical input of graph networks. There are three ways to apply regulation criteria such as dropout, but all of them are either impossible or ineffective to implement. First, any regulation on the nodes will transform the PSA subgraph into an unknown drug representation. Second, omitting any edges will reconstruct the key structure of the PPI network and compromise the scalability of the model. Lastly, node features that represent the probability of PSEs, already suffer from data sparsity, and applying regularization on top of that will add even more strain on the model precision. Therefore, no regulations were introduced at this level of model development due to the aforementioned reasons.

6 Suggestions for future studies
For prospective studies, it is recommended to add modules such as tissue expressions, side effect frequencies, drug dose, gene expressions, and examine how they can contribute to the general performance (especially on the precision) of the model. The addition of a new module not only enhances the performance of the SiGCN model but also the improved PSA can be utilized to discover new biomarkers and drug targets. As it was discussed in section 5.3, narrowing down the application of the SiGCN to specific diseases can ultimately improve the model precision. Cancers such as leukemia, lymphoma, melanoma, and lung cancer are ideal places to start because they have an abundance of public databases and well-written literature, but they are not extensively studied in terms of combinational treatments. Furthermore, drug-resistant infectious diseases such as HIV, Hepatitis, AIDS, Malaria, and COVID-19, may be a good fit for a much more in-depth investigation and applicability of the concept. The conversion of a model architecture to GAT and how it handles the sparsity of drug combination data would be another recommendation. Finally, since the model mimicked the true nature of the chemical interactions its application can also be explored with other types of drugs such as immunotherapy agents, large peptides, mRNAs, and vaccines. The only
precondition would be to find a suitable model that can reliably predict the binding affinity of those biological drugs.

7 Conclusion

The idea behind this thesis was not to develop a deep learning model that performs better than others, but rather to establish a completely scalable, and reliable framework that addresses many of the fundamental problems in the research of the drug combination. Embracing this mission, inspired us to constantly come with innovative and multidisciplinary ideas to resolve problems that were unaddressed in the community (discussed in sections 1&2). To begin with, we adopted a data-centric approach and designed a pipeline that only requires the SMILES of a drug to construct its graph representation or the PSA subgraph. Second, we developed a representation learning model by coupling graph convolutional networks (GCN) and Siamese neural networks together which have performed surprisingly well compared to the current state-of-the-art models. The high accuracy of the model is perhaps influenced by sparsity and imbalance of the training dataset. There are few methods for suppressing the impact of the data limitation. First by narrowing down the domain usage of the model to specific diseases and second to add more modules in the enrichment process. Apart from all of that, our model is highly interpretable to the point that its results can be directly given to biomedical researchers, and they can start examining its biological significance with no extra effort required. Having the R&D mindset, inspired us to make the framework as versatile as possible and many modules can be easily added to the enrichment process. This flexibility makes the SiGCN applicable for predicting the safety of disease-specific drug combinations such as cancer and infectious diseases. We hope our efforts assist further acceleration of drug development and improve the well-being of our fellow human beings.

Acknowledgments

Special thanks to Dr. Narsis Kiani for her guidance and Hossein Fooladi for his professional expertise. This passionate project would not have been possible without your relentless encouragement and support, for which I am grateful. I would also like to thank, Prof. Ola Spjuth, Prof. Jesper Tegnér, and Ms. Jiaxi Zhao for their constructive feedback on my thesis presentation and report. This thesis and its experiments were performed in the Algorithmic Dynamics Lab of the Center for Molecular Medicine (CMM) at Karolinska Institute.

Bibliography


Appendix

A. Popular science abstract

Scientific breakthroughs occur in a variety of settings and under unusual circumstances. Despite the scientific community’s righteous emphasis on well-structured and theoretically sound research, many important discoveries have been made when experimental results contradict a well-established scientific framework. Science is dynamic and forward-thinking by its very nature and its ability to interpret, and formulized factual data has progressed it forward. With the recent advancements in computational power, Artificial Intelligence (AI) is promptly used to bridge the gap between what data truly indicates and what can be inferred from it. But not all methods have equal inference capabilities. “Representation learning” is one of the most prominent techniques with the potential to transform the field of interpretable AI. This study has utilized the power of Graph Neural Networks — as one of the most promising representation learning frameworks — and established a model which accurately predicts polypharmacy side effects of any given paired drug combinations. To achieve this a graph network that associates side effects to proteins has been established. Because all humans share the same proteins, the model’s predictions may be applied to any individual and explain which proteins may produce specific adverse effects when receiving multiple medicines. This thesis has sought to develop an alternative approach and framework, which proposes innovative solutions for drug combination research, and not just a reliable model.

B. GitHub

Code and resources of the thesis can be found in the following GitHub repository. 
https://github.com/amanzadi/MS_Thesis