공학석사 학위논문

Computational Drug Combination Prediction using Biomedical Knowledge Graph

Enhancement with Drug-Drug Interaction Data and Supervised Contrastive Learning

약물 상호작용 데이터와 지도 대조 학습을 이용한 생명 의학 지식 그래프 기반 약물 조합 예측

2023년 8월

서울대학교 대학원

협동과정 인공지능전공

구 정 현

Computational Drug Combination Prediction using Biomedical Knowledge Graph

Enhancement with Drug-Drug Interaction Data and Supervised Contrastive Learning

약물 상호작용 데이터와 지도 대조 학습을 이용한
생명 의학 지식 그래프 기반 약물 조합 예측

지도교수 김 선

이 논문을 공학석사 학위논문으로 제출함

2023 년 6 월

서울대학교 대학원

협동과정 인공지능전공

구 정 현

구정현의 공학석사 학위논문을 인준함

2023 년 6 월

위 원 장	황대희
부위원장	김 선
위 원	오민식

Abstract

Computational Drug Combination Prediction using Biomedical Knowledge Graph

Enhancement with Drug-Drug Interaction Data and Supervised Contrastive Learning

Jeonghyeon Gu
Interdisciplinary Program in Artificial Intelligence
College of Engineering
Seoul National University

Combination therapies have brought significant advancements to the treatment of various diseases in the medical field. However, searching for effective drug combinations remains a major challenge due to the vast number of possible combinations. The utilization of biomedical knowledge graphs, which encompass intricate relationships among biomedical entities, has demonstrated promising potential in predicting effective combinations for a wide range of diseases. However, the absence of reliable negative samples has posed challenges for machine learning models to establish robust decision boundaries, thereby limiting prediction performance. Additionally, previous methods have relied on raw and general drug embedding vectors extracted from the knowledge graph, which is suboptimal and leaves considerable room for improvement.

To address this issue, I propose a novel framework that leverages existing Drug-Drug Interaction (DDI) data as a reliable negative dataset and employs Supervised Contrastive Learning (SCL) to transform drug embedding vectors to be more suitable for drug combination prediction. DDI data and SCL technique not only improved the performance metrics but also helpful in building tight decision boundaries for predicting drug combinations. To demonstrate the effectiveness of this approach, I conducted extensive experiments using various network embedding algorithms, including random walk and graph neural networks, on a biomedical knowledge graph called multiscale interactome (MSI) network. I also provide t-SNE plot of drug pair embedding vectors to visualize the decision boundaries between drug combination and DDI. Lastly, case study results of drug combination and DDI are also provided. In summary, this work highlights the potential of using DDI data and SCL in finding tighter decision boundaries for predicting effective drug combinations. All source codes are available on the GitHub repository (`https://github.com/gujh14/DC_with_DDI_SupCon.git`)

Keywords: Drug combination, Knowledge graph, Drug-drug interaction, Supervised contrastive learning, Random walk, Graph neural network

Student Number: 2021-28284

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Study Background

Combination therapy involves using two or more medications together to achieve a desired treatment outcome. This includes combining drugs with different mechanisms of action to target multiple aspects of a disease or to increase effectiveness while minimizing side effects. The selection and design of drug combinations require careful consideration of factors such as drug interactions, safety profiles, dosing schedules, and patient-specific characteristics. Research and computational approaches play a crucial role in identifying and optimizing effective drug combinations. Drug combination is commonly used in the treatment of various medical conditions such as cancer (Crystal *et al.*, 2014), infectious diseases (Zheng *et al.*, 2018), cardiovascular diseases (Giles *et al.*, 2014), and autoimmune diseases (Smilek *et al.*, 2014). An example of this is the use of Twynsta™, a medication that combines telmisartan and amlodipine, which is used to lower blood pressure by relaxing blood vessels and reducing the workload on the heart, respectively. Combining these two drugs in a single tablet can provide additive or synergistic blood pressure-lowering effects and can also help minimize the side effects that might be associated with taking high doses of each drug alone (Chalmers, 1999).

Despite the importance of modern drug combination therapy, identifying the combinations that effectively treat a condition while minimizing side effects is often a matter of intuition and experience, rather than following established principles. There are over 10,000 ongoing clinical trials in the US to study combination therapies, but these numbers are quite modest to cover the tremendous number of possible combinations of drugs. The search for adequate drug pairs is time consuming and requires lots of clinical experience. It is also difficult to predict the complex interactions between different drugs and their potential targets. They can involve multiple mechanisms of action and can vary depending on the specific cell type or tissue being targeted. Therefore, it is important to develop powerful computational technologies that can facilitate the identification of drug combination therapies and narrow down the search space.

There have been several studies conducted on computational drug combination predictions. However, they frequently encounter limitations such as the narrow scope of the drugs covered and/or unreliable negative data. In this article, I introduce a novel framework that addresses these challenges and can be applied to a wide range of drugs. This framework utilizes drug-drug interaction (DDI) data as a reliable negative dataset and employs supervised contrastive learning (SCL) technique for pretraining. By adopting this data-centric approach and leveraging an appropriate pretraining technique, the model significantly enhances its ability to establish robust and effective decision boundaries for predicting drug combinations.

1.2 Related works & Limitations

Recently, several methods have proposed computational approaches to narrow down the broad search space of drug combination prediction problem. One approach relies on the experimental high-throughput screening (HTS) data. HTS is a method used in drug discovery and biology to rapidly test a large number of compounds or substances against biological targets, such as enzymes, receptors, or cells, in order to assess the effects of the molecules to the biological system. HTS data typically includes information on the biological activity of compounds, such as their ability to inhibit or activate a target, as well as information on compound structure, concentration, and assay conditions. These data are obtained using automated systems that can process hundreds of thousands or even millions of compounds in a relatively short period of time. The studies that use HTS data focused exclusively on cancer drugs or anti-bacterial drugs, as their therapeutic efficacy can be easily measured at the cell line level using metrics like IC50 (half maximal inhibitory concentration), although this measure does not always align with clinical response. The degree of synergy in this framework is typically quantified by its deviation from that simulated according to a theoretical model. There are a few quantitative metrics to define the degree of synergy, such as Loewe additivity (Loewe, 1953), Bliss independence (Bliss, 1939), highest single agent (Berenbaum, 1989), and zero interaction potency (Yadav *et al.*, 2015). DeepSynergy (Preuer *et al.*, 2018) is one of the pioneering works that use deep learning to predict drug combination synergies and takes chemical and genomic information as input information. Features of two drugs and one cell line are fed into a deep neural network, and a synergy score is predicted. DeepSynergy used a large-scale oncology screen data produced by Merck & Co (O'Neil *et al.*, 2016). This dataset provides a screen of 22,737 experiments of 583 doublet combinations in 39 diverse cancer cell lines using a four-by-four dosing regimen. Sidorov et al. (Sidorov *et al.*, 2019) suggested an *in-silico* modeling with NCI-ALMANAC (Holbeck *et al.*, 2017) dataset, a large phenotypic drug combination HTS dataset and contains synergy measures of pairwise combinations of drugs on cell lines. Each cell line and drug were modeled using Random Forest (Breiman, 2001) and Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016). PRODeepSyn(Wang *et al.*, 2022) is a method that leveraged Graph Convolution Network (GCN) to integrate protein-protein interaction network with omics data and constructed embeddings for cell lines. Each drug is represented by molecular fingerprints and descriptors. Cell line features and drug features are fed into the final classifier layer to predict synergy score. And Jin et al. (Jin *et al.*, 2021) proposed a neural network architecture that jointly learns drug-target interaction and drug-drug synergy to identify synergistic drug combinations against SARS-CoV-2.

On the other hand, knowledge graph (KG)-based approaches were used to predict drug combination pairs on general drugs and diseases. A biomedical KG is a structured representation of biomedical knowledge that captures relationships between different entities, such as genes, proteins, diseases, drugs, pathways, and their associated attributes. It serves as a knowledge base that organizes and links various types of information. More detailed description about biomedical KG is provided in Section 2.1. They can be further divided into mining and learning approaches. For instance, Cheng et al. (Cheng *et al.*, 2019) mined and analyzed a drug-protein-disease KG by quantifying the distance between drug-target modules for drug combination prediction. They showed that there are some typical overlapping patterns in the target protein set of the combinatorial drugs. Recent learning approaches use network embedding algorithms to embed various entities in the biomedical KG and apply machine learning techniques to infer drug combination scores. For example, Liu et al. (Liu *et al.*, 2019) used drug similarity network, protein similarity network and known drug-protein associations to build a drug-protein heterogeneous network. They then employed the random walk with restart algorithm to get the feature vectors of each drug combinations, which are later used for the gradient tree boosting classifier. Another example is the NEWMIN (Yu *et al.*, 2022) method, which constructed multiplex drug-drug similarity network, including chemical, side effect, Anatomical Therapeutic and Chemical (ATC) codes, text-mining, protein, and category-based similarity networks. Random walk was performed on the networks and the embedding vector of each drug was built with the word2vec algorithm (Mikolov *et al.*, 2013). Finally, drug combination score was predicted by applying a Random Forest classifier on the concatenated embedding vectors of drug pair. Although the prediction performance of these works varies and depends on the selection of the KG, they have shown that biomedical KGs are useful resources for predicting drug combination.

While both HTS data-based and KG-based approaches exhibited encouraging results on drug combination predictions, but they have their own limitations. When utilizing public HTS data-based approaches, one can only perform research for anti-cancer or anti-bacterial drugs as their outcome is cell survival and/or death. However, there are often no suitable quantitative measures of drug response at the cell line level for other indications. In other words, it is hard to use those public HTS data to infer novel drug combinations for diseases like hypertension or diabetes for my research.

Alternatively, the KG-based approaches can alleviate this problem and be easily applied for various diseases and drugs. But these methods often rely on unreliable negative data, which entails randomly selecting pairs of drugs from drug lists. A machine learning model typically require negative data to establish decision boundaries and improve its ability to distinguish between classes, leading to better performance metrics. While the random negative approach has been used in tasks related to drug-target interactions and drug-disease associations, these randomly selected drug pairs are not true negatives, but rather unlabeled pairs that carry the possibility of actually being positive samples. Additionally, there is possibility that they are easy negatives, as it is highly likely to sample two drugs with obviously distinct indications. Furthermore, the KG-based approaches often use raw drug embeddings learned only from the network topology for combination prediction. Although this raw vectors from the biomedical network can serve as good initial representations, they are too general and need to be projected into a more specific embedding space that is appropriate to predict drug combination scores.

1.3 Proposed Approach

In this article, I propose a novel strategy that utilizes existing DDI data as a reliable negative dataset for predicting drug combinations. The proposed framework also employs SCL to transform the raw embedding vectors into drug combination-specific embeddings. To elaborate, this framework is a type of KG-based learning approach and can cover any drug doublet pairs with the target protein information in biomedical KG. As DDI refers to a phenomenon in which the effectiveness or toxicity of one drug is affected when taken with another drug and is usually avoided, it can be used as negative dataset to improve predictions of combination therapies (Güvenç Paltun *et al.*, 2021). SCL is a type of contrastive learning technique that leverages label information to draw embedding vectors closer to one another if they are of the same class, while pushing them away if they are not (Khosla *et al.*, 2020).

Overall training scheme is illustrated in Figure 1.1, which can be regarded as a three-step process: initial, pretraining, and final stage. The initial stage is obtaining raw drug embedding vectors from the KG using network embedding algorithms such as random walk or graph neural networks (GNNs). Then the SCL technique is applied at the pretraining stage to transform each drug embedding vectors into more suitable representations for drug combination prediction. At the final stage, the embedding vectors are fed into a classifier to predict drug combination scores.

To the best of my knowledge, the proposed approach is the first work to explicitly use the DDI dataset as a negative dataset and to employ SCL technique for drug combination prediction task. Through comprehensive experiments, I show that this approach has substantially benefited the deep learning model, not only in terms of enhancing overall predictive performance but also in establishing tighter decision boundaries for predicting drug combination pairs. This is clearly demonstrated in the visualization results of the embedding space, which indicate that the SCL pretraining effectively transformed the initial drug embedding vectors from the biomedical KG into more appropriate and specific embeddings for downstream drug combination prediction tasks. Furthermore, I have included some case studies to illustrate the performance of the approach in specific scenarios.

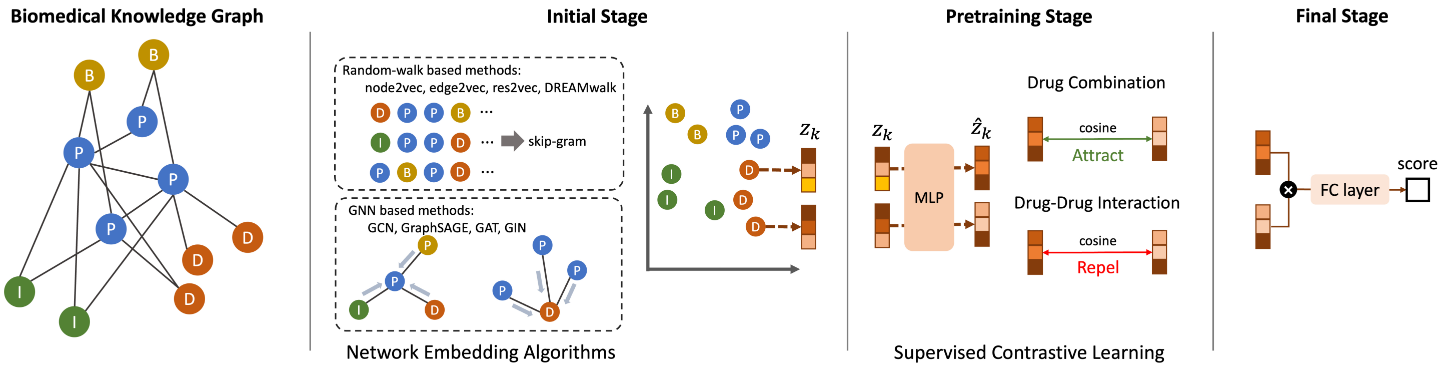

Figure 1.1: An overview of the proposed framework. It consists of three stages: initial, pretraining, final stage. At initial stage, various network embedding algorithms constructs drug embeddings from the biomedical KG called MSI network. The initial raw embeddings are pretrained with SCL technique at pretraining stage. At the final stage, the embedding vectors of drug pairs are multiplied and fed into fully connected layer to predict drug combination scores. z_k is the initial raw drug embedding vector, while $\hat{z}_k$ is the pretrained drug embedding vector. KG: knowledge graph; MSI: multi-scale interactome; SCL: supervised contrastive learning; In the figure, MLP: multi-layer perceptron; FC layer: fully connected layer; I: indication; D: drug; P: protein; B: biological function.

Chapter 2

Materials & Methods

2.1 Biomedical Knowledge Graph

KGs are a way of representing knowledge concepts and their relationships as nodes and edges. In the biomedical field, KGs have been widely used for integrating entities, such as genes, proteins, drugs, and diseases at various levels, ranging from molecular to clinical. By structuring assay results, mechanisms of action, and target protein information in a graph format, computational methodologies can be applied to infer semantics and uncover unknown associations. Biomedical KGs thus provide a powerful framework for exploring the relationships between biomedical entities and unlocking new insights. They can be used to predict drug combinations due to their ability to capture and represent complex relationships between drugs, target proteins, and diseases. For example, KGs can link drugs to their target genes or proteins and capture their interactions within biological pathways. This information enables the identification of drugs that target the same pathways or share common targets.

Node Type	Number of Nodes
Drug	1,661
Protein	17,660
Indication (disease)	840
Biological function (GO term)	9,798

Table 2.1: Node types and number of nodes in the MSI network. MSI: multi-scale interactome; GO: Gene Ontology.

Edge Type	Number of Edges
Drug-Protein	8,568
Disease-Protein	25,212
Protein-Protein	387,626
Protein-Biological function	34,777
Biological function-Biological function	22,545

Table 2.2: Edge types and number of edges in the MSI network. MSI: multi-scale interactome.

The multi-scale interactome (MSI) network is one of the well-known biomedical KGs that contains drug, disease, gene, and biological function annotations as its nodes. It contains a total of 29,959 nodes and 478,728 edges, including 1,661 drug and 840 disease entities (Ruiz *et al.*, 2021). There are five types of edge relations in the MSI network: drug-protein, disease-protein, protein-protein, protein-biological function, and biological function-biological function interactions. Detailed numerical information about the MSI network is provided in Table 2.1 and Table 2.2. The authors of the MSI network demonstrate that the inclusion of biological function entities in a molecular-scale drug-gene-disease network offers both better drug-disease treatment prediction performance and biological interpretability. I used the MSI network to embed drugs into vectors for predicting drug combinations.

2.2 Drug Databases

2.2.1 Drug Combination Databases

There are various databases that collect and organize information on drug combinations. DCDB 2.0 (Liu *et al.*, 2014) is a curated database from more than 140,000 clinical studies and the Food and Drug Administration (FDA) Orange Book (Home, 2013). It includes 1,363 drug combination pairs, consisting of 904 unique components. There are three types of combinations in DCDB 2.0, which are ‘Efficacious’, ‘Need further study’, and ‘Non-efficacious’.

Continuous Drug Combination Database (CDCDB) (Shtar *et al.*, 2022) is a comparable database that is continuously updated and comprises of 17,107 distinct drug combinations composed of over 4,129 individual drugs. It is curated from ClinicalTrials.gov (Zarin *et al.*, 2011), the FDA Orange Book, and Integrity (Clarivate Analytics)™.

I processed and merged the database to curate drug combination pairs as follows. First, I used only ‘Efficacious’ type of drug pairs from DCDB 2.0. And the drugs not included in the MSI network were excluded. Then the pairs from both databases were merged to include as much positive data as possible, resulting in 4,344 pairs. And they were used as the positive dataset for drug combination prediction. Detailed numerical information about both databases is provided in Table 2.3.

	DCDB 2.0	CDCDB	TWOSIDES
Number of unique drugs	904	4,129	645
Number of pairs (before curation)	1,363	17,107	4,649,442
Number of pairs (after curation)	455	4,221	16,157

Table 2.3: Number of unique drugs and pairs in DCDB 2.0, CDCDB, and TWOSIDES databases.

2.2.2 Drug-Drug Interaction Database

DDIs can occur when two or more drugs are administered simultaneously, leading to changes in the pharmacokinetics or pharmacodynamics of the drugs involved. For example, taking warfarin (a blood thinner) and aspirin (anti-inflammatory and anti-platelet drug) together is not recommended as they can severely increase the risk of bleeding. Although there are some rare cases, such as the combination of ritonavir (a potent CYP3A4 enzyme inhibitor) and lopinavir (CYP3A4 substrate) (Cvetkovic and Goa, 2003), where DDIs are beneficial and used intentionally to enhance the efficacy, DDIs are generally avoided.

I utilized TWOSIDES database (Tatonetti *et al.*, 2012) as a reliable negative dataset. The DDI data in this database were curated from the large Adverse Event Reporting Systems (AERS) developed by FDA, World Health Organization, and Health Canada. The TWOSIDES database contains 4,649,442 DDI pairs between 645 drugs and is included in the Therapeutics Data Commons (TDC) (Huang *et al.*, 2021), which is a coordinated initiative to access and evaluate AI models across therapeutic modalities and stages of discovery. I chose drug pairs whose entity exists in the MSI network, resulting in a 16,157

DDI pairs. Detailed numerical information about the TWOSIDES database is provided in Table 2.3.

2.3 Initial Stage: Network Embedding Algorithms

At the initial stage, the initial drug embedding vectors z_k is obtained with various network embedding algorithms.

2.3.1 Random walk-based algorithms

Random walk-based algorithms have been a popular method for graph mining since their introduction in the seminal work of DeepWalk (Perozzi *et al.*, 2014). The random walk algorithm generates node sequences $p = n_1, n_2, ..., n_l$ of length l from a graph $G = (V, E)$ of node set V and edge set E from following distribution:

$$P_{(n_i=x|n_{i-1}=v)} = \begin{cases} \pi_{vx}/Z & \text{if } (v,x) \in E \\ 0 & \text{otherwise} \end{cases}$$

where π_{vx} is the unnormalized transition probability from node v and x, and Z is the normalizing constant. The final transition probability varies depending on the various biased traversal strategies. The sequences generated are then passed through a shallow neural network, such as Skip-gram or C-BOW, to generate node embedding vectors. The network is trained to maximize the similarity of the co-occurring neighbors within a given window length. These embeddings can be used for various downstream tasks, such as node classification and link prediction. I tested four random walk-based algorithms, node2vec (Grover and Leskovec, 2016), edge2vec (Gao *et al.*, 2019), res2vec (Kojaku *et al.*, 2021), and DREAMwalk (Bang *et al.*, 2022) to embed drug nodes in the MSI network and compared their performance.

Node2vec (Grover and Leskovec, 2016) is a variation of the DeepWalk algorithm that allows for more flexibility in generating node sequences by introducing the parameters p and q. These parameters enable a balance between Breadth-First Sampling (BFS) and Depth-First Sampling (DFS) during the random walk sampling strategy. In our experiments, we set p and q to 1, resulting in a uniform random walk sampling strategy. This approach allows for the examination of various neighborhood structures and the creation of more informative node embeddings for downstream tasks.

Edge2vec (Gao *et al.*, 2019) is a method specifically designed for mining graphs from the biomedical domain, where different scales of associations exist between different types of nodes. To account for these varying edge types, edge2vec first trains a novel edge-type transition matrix using an Expectation-Maximization (EM)-like iterative approach. This allows for the assignment of transition weights to the reachable edge types. During the random walk, the algorithm traverses the network proportionally to these transition weights, resulting in a more targeted exploration of the biomedical knowledge graphs.

Residual2vec (Kojaku *et al.*, 2021) is another extension of the DeepWalk algorithm that addresses the bias towards high-degree nodes in random walk sampling. The algorithm does this by comparing the random walk sampling results with those of randomly generated graphs. The Residual2vec algorithm can be applied in two modes: homogeneous and heterogeneous. In the homogeneous setting, all nodes are treated as the same type, while in the heterogeneous setting, different types of nodes are assigned different transition probabilities. In our framework, the algorithm was used in both homogeneous and heterogeneous settings using the default parameters as suggested by the authors.

DREAMwalk (Bang *et al.*, 2022) is a random walk-based algorithm that is designed for mining biomedical graphs. It uses a teleport-guided random walk framework to generate node embedding vectors. The aim of this framework is to guide the traversal of the biomedical network by taking into account the semantic similarities of drug and disease entities, thereby reducing bias in the learning process from large and dense protein-protein interaction networks. In our experiments, we set the semantic similarity cut-off and teleport factor τ to 0.5 for all experiments.

2.3.2 Graph Neural Network Algorithms

Graph Neural Networks (GNNs) are specialized type of neural network designed to handle graph-structured data. They are able to learn the topological structure of a graph in an end-to-end fashion by updating the features of both nodes and edges based on the characteristics of their neighboring entities. I used message passing GNNs, which updates node embeddings by *aggregate* and *combine* step per each layer as follows:

$$
\begin{aligned}
a^{(t)}_{\mathcal{N}(v)} &= \mathrm{AGGREGATE}^{(t)}(\{h^{(t-1)}_u, \forall u \in \mathcal{N}(v)\}) \\
h^{(t)}_v &= \sigma(W^{(t)} \cdot \mathrm{COMBINE}(h^{(t-1)}_v, a^{(t)}_{\mathcal{N}(v)}))
\end{aligned}
$$

where $h^{(t)}_v$ is the feature vector of node v at time step t (t-th layer). $\mathcal{N}(v)$ is the set of neighbor nodes of a node v. At each time step t, a differentiable aggregation function AGGREGATE collects neighbors' representation vectors, and they are combined by the COMBINE function and a weight matrix W is multiplied and non-linear activation function σ is applied to update the hidden representation of node v.

The message passing scheme allows GNNs to effectively capture dependencies between nodes in the graph. Multiple convolutional methods have been proposed to aggregate and combine messages from different perspectives. I conducted several experiments using four well-known GNN architectures: GCN (Kipf and Welling, 2016), Graph Sample and Aggregate Network (GraphSAGE) (Hamilton *et al.*, 2017), Graph Attention Network (GAT) (Veličković *et al.*, 2017), and Graph Isomorphism Network (GIN) (Xu *et al.*, 2018).

GCN (Kipf and Welling, 2016) is an efficient variant of Convolutional Neural Networks (CNNs) on graphs. It is a basic form of message passing neural networks, which applies a local neighborhood aggregation with learned first-order spectral filters followed by nonlinear activation function to learn node representations. In GCN, the AGGREGATE and COMBINE steps are integrated as follows:

$$h_v^{(t)} = \sigma(W^{(t)} \cdot \text{MEAN}\{h_u^{(t-1)}, \forall u \in \mathcal{N}(v) \cup \{v\}\})$$

GraphSAGE (Hamilton *et al.*, 2017) was originally designed to inductively learn node embeddings in large graphs. It first samples a fixed number of nodes in a node's local neighborhood, then aggregates feature information from the sampled neighbor nodes with element-wise max pooling. The AGGREGATE and COMBINE step of GraphSAGE can be formulated as:

$$\begin{aligned} a_v^{(t)} &= \text{MAX}(\{\sigma(W^{(t)} \cdot h_u^{(t-1)}), \forall u \in \mathcal{N}(v)\}) \\ h_v^{(t)} &= \sigma(W^{(t)} \cdot \text{CONCAT}(h_v^{(t-1)}, h_{\mathcal{N}_v}^{(t)})) \end{aligned}$$

GAT (Veličković *et al.*, 2017) is another type of GNN that uses attention mechanisms to learn node feature representations. This layer learns different attention coefficients for each neighbor of a node and aggregates them using these coefficients to update the final representation of the node. The AGGREGATE and COMBINE step can be merged as follows:

$$h_v^{(t)} = \sigma(\sum_{u \in \mathcal{N}(v) \cup \{v\}} \alpha_{vu} W^{(t)} h_u^{(t-1)})$$

where α_{vu} is the learned attention weight between node v and u.

GIN (Xu *et al.*, 2018) is a GNN layer that is intended to generalize the WL test and use MLPs to model injective multiset functions for the neighborhood aggregation. The AGGREGATE and COMBINE step is unified as follows:

$$h_v^{(t)} = \mathrm{MLP}^{(t)}((1+\epsilon^{(t)}) \cdot h_v^{(t-1)} + \sum_{u \in \mathcal{N}(v)} h_u^{(t-1)})$$

ϵ can be a learnable parameter or a fixed scalar. I set all ϵ to be zero for simplicity.

The number of layers that achieved the highest performance for GCN, GraphSAGE, GAT, and GIN are 3, 3, 2, and 2, respectively.

2.4 Pretraining Stage: Supervised Contrastive Learning

At the pretraining stage, the initial drug embedding vectors z are transformed into more specific vectors $\hat{z}$ by SCL with multilayer perceptron (MLP) layers.

Contrastive learning is a machine learning technique that enables models to differentiate between data points in a certain embedding space by identifying their similarities and differences. This approach can be employed in both supervised and unsupervised settings, and it involves various types of loss functions such as contrastive loss (Chopra *et al.*, 2005), triplet loss (Schroff *et al.*, 2015), and InfoNCE loss (Oord *et al.*, 2018).

Meanwhile, multiple works have pointed out shortcomings of the simple cross entropy loss, such as its susceptibility to noisy labels (Zhang and Sabuncu, 2018) and poor margins (Liu *et al.*, 2016) that can lead to reduced generalization performance. To address this issue, Khosla et al. (Khosla *et al.*, 2020) proposed SCL framework that demonstrated improved performance in robustness benchmarks and less sensitivity to changes in hyperparameters.

Since the initial drug embeddings from the knowledge graph were too general for drug combination prediction, I applied SCL during pretraining phase with training dataset to transform them into a more appropriate embeddings for downstream drug combination prediction. To achieve this, I implemented a loss function similar to the supervised contrastive loss (Khosla *et al.*, 2020) to effectively maximize the cosine similarity between drug combinations while minimizing the similarity between DDI pairs. The loss function can be expressed as follows:

$$\mathcal{L} = -\log \frac{\sum_{\forall y_{ij}=DC} \exp(\hat{z}_i^\top \cdot \hat{z}_j)}{\sum_{\forall y_{ij}=DC} \exp(\hat{z}_i^\top \cdot \hat{z}_j) + \sum_{\forall y_{ij}=DDI} \exp(\hat{z}_i^\top \cdot \hat{z}_j)}$$

where $\hat{z}_i$ and $\hat{z}_j$ are the transformed drug embedding vectors. If the two drugs are drug combinations, their label y_{ij} is DC and if they are DDI pairs, the label is DDI. By minimizing this loss function, the MLP layers can be trained to effectively transform the embedding vectors which can be further used for final drug combination prediction.

2.5 Final Stage: Drug Combination Prediction

Once pretraining the drug embedding vectors is done, I employed them to finally predict actual drug combination scores. Two drug embeddings are element-wise multiplied, and the resulting vector is fed into a fully connected classifier layer to output a score ranging from 0 (DDI) and 1 (drug combination). The classifier layer consists of one hidden layer and one output layer with dimensions 128 and 1, respectively. Since this is binary classification task, I used binary cross entropy loss as a loss function at this stage.

2.6 Experimental Setup

To ensure the robustness of the results for performance evaluation, I repeated each experiment with 10 different random seeds. I used the Adam optimizer and early stopping technique with a patience of 20 for all experiments. To determine the best hyperparameter set, I performed grid search to find the optimal learning rates for both the pretraining and final stages for each algorithm.

Chapter 3

Results

In this section, I present the experimental results of my framework, along with visual representations of the drug embedding vectors, comparison of robustness to class imbalance, and case studies.

3.1 Performance Evaluation

As the performance of KG-based drug combination prediction depends on the quality of the biomedical KG, I set the baseline framework to use random sampling when composing the negative dataset. I conducted experiments for all the aforementioned network embedding algorithms, and the results are presented in the form of an ablation study in Table 3.1 and Table 3.2. The first row of each algorithm is my framework, which uses the TWOSIDES database as negative dataset and applied SCL during pretraining. The second row presents the results without SCL pretraining. The last row represents the baseline framework, which uses a randomly sampled negative dataset without SCL pretraining. In these experiments, I used the same number of negative data as positive data.

The SCL pretraining stage resulted in enhancements for nearly all metrics across all algorithms, except for the precision metric of GIN. Additionally, the use of the TWOSIDES dataset led to significant improvements for most metrics across all algorithms, with the exception of the recall metric of GraphSAGE and GAT. Given that comparing the performance of two different negative datasets can lead to debatable issues, I have presented some opinions in the Chapter 4 (Discussion & Conclusion section).

3.2 Visualization of Drug Pair Embedding Vectors

The use of supervised contrastive learning improved model performance substantially. I hypothesized that this is due to the tight boundary learned to distinguish between drug combination pairs and DDI pairs. To visualize the distribution of embedding vectors, I applied t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm to the multiplied embedding vectors of drug pairs of three stages: initial, pretraining, and final stage (Figure 2a, b, c). This experiment is performed with the best performing algorithm DREAMwalk, and the shown vectors are test data, i.e. they were not used as training data for the model used in the experiment.

Figure 3.1a illustrates the initial raw embedding vectors of the drug pairs. Combination pairs and DDI pairs are distributed evenly in the embedding space without any obvious boundaries or clusters, which implies that those embeddings might be too general and naïve to perform drug combination prediction. While in Figure 3.1b, the two separate classes seem to be well-separated, and the combination pairs even developed some clear clusters. It is easy to observe the effect of the SCL to transform the embeddings of drug combinations to

Algorithm	Negative dataset	SCL pretraining	Accuracy	Precision	Recall	F1
node2vec	TWOSIDES	O	**0.898 ± 0.007**	**0.925 ± 0.016**	**0.868 ± 0.008**	**0.895 ± 0.006**
	TWOSIDES		0.858 ± 0.009	0.882 ± 0.013	0.828 ± 0.016	0.854 ± 0.010
	random		0.716 ± 0.018	0.720 ± 0.019	0.707 ± 0.024	0.713 ± 0.019
edge2vec	TWOSIDES	O	**0.919 ± 0.011**	**0.945 ± 0.018**	**0.889 ± 0.014**	**0.916 ± 0.012**
	TWOSIDES		0.875 ± 0.010	0.904 ± 0.020	0.839 ± 0.015	0.870 ± 0.009
	random		0.744 ± 0.020	0.750 ± 0.024	0.733 ± 0.021	0.741 ± 0.019
res2vec_homo	TWOSIDES	O	**0.896 ± 0.006**	**0.932 ± 0.017**	**0.856 ± 0.010**	**0.892 ± 0.006**
	TWOSIDES		0.854 ± 0.010	0.871 ± 0.013	0.831 ± 0.014	0.851 ± 0.010
	random		0.745 ± 0.019	0.750 ± 0.022	0.737 ± 0.020	0.743 ± 0.019
res2vec_hetero	TWOSIDES	O	**0.884 ± 0.006**	**0.916 ± 0.012**	**0.846 ± 0.009**	**0.880 ± 0.006**
	TWOSIDES		0.853 ± 0.011	0.885 ± 0.018	0.812 ± 0.015	0.847 ± 0.011
	random		0.724 ± 0.017	0.733 ± 0.016	0.705 ± 0.028	0.718 ± 0.019
DREAMwalk	TWOSIDES	O	**0.923 ± 0.003**	**0.946 ± 0.009**	**0.898 ± 0.008**	**0.921 ± 0.003**
	TWOSIDES		0.884 ± 0.009	0.914 ± 0.008	0.849 ± 0.020	0.880 ± 0.011
	random		0.793 ± 0.012	0.802 ± 0.012	0.777 ± 0.020	0.789 ± 0.014

Table 3.1: Performance (mean ± std) of random walk-based algorithms. The first row of each algorithm is my framework, which uses the TWOSIDES database as negative dataset and applied SCL during pretraining. The second row presents the results without SCL pretraining. The last row represents the baseline framework, which uses a randomly sampled negative dataset without SCL pretraining. 'res2vec_homo' and 'res2vec_hetero' are residual2vec algorithms performed with homogeneous setting and heterogeneous setting, respectively. The best results are in bold, while second-best ones are underlined. The ratio of positive:negative dataset is 1:1. SCL: supervised contrastive learning.

Algorithm	Negative dataset	SCL pretraining	Accuracy	Precision	Recall	F1
GCN	TWOSIDES	O	**0.909 ± 0.008**	**0.932 ± 0.018**	**0.882 ± 0.020**	**0.906 ± 0.009**
	TWOSIDES		0.905 ± 0.007	0.927 ± 0.014	0.880 ± 0.020	0.903 ± 0.008
	random		0.658 ± 0.081	0.622 ± 0.068	0.876 ± 0.066	0.722 ± 0.034
GraphSAGE	TWOSIDES	O	**0.921 ± 0.013**	**0.947 ± 0.023**	**0.892 ± 0.015**	**0.918 ± 0.012**
	TWOSIDES		0.739 ± 0.017	0.812 ± 0.042	0.627 ± 0.034	0.706 ± 0.019
	random		0.632 ± 0.013	0.601 ± 0.012	0.788 ± 0.040	0.681 ± 0.014
GAT	TWOSIDES	O	**0.902 ± 0.015**	**0.935 ± 0.021**	**0.865 ± 0.042**	**0.898 ± 0.019**
	TWOSIDES		0.836 ± 0.026	0.893 ± 0.025	0.764 ± 0.042	0.823 ± 0.030
	random		0.667 ± 0.022	0.640 ± 0.023	0.766 ± 0.028	0.697 ± 0.018
GIN	TWOSIDES	O	**0.915 ± 0.017**	0.949 ± 0.021	**0.878 ± 0.020**	**0.912 ± 0.017**
	TWOSIDES		0.912 ± 0.016	**0.957 ± 0.026**	0.864 ± 0.025	0.908 ± 0.017
	random		0.780 ± 0.015	0.785 ± 0.022	0.772 ± 0.017	0.779 ± 0.013

Table 3.2: Performance (mean ± std) of GNN-based algorithms. The first row of each algorithm is my framework, which uses the TWOSIDES database as negative dataset and applied SCL during pretraining. The second row presents the results without SCL pretraining. The last row represents the baseline framework, which uses a randomly sampled negative dataset without SCL pretraining. The best results are in bold, while second-best ones are underlined. The best results are in bold, while second-best ones are underlined. The ratio of positive:negative dataset is 1:1. SCL: supervised contrastive learning.

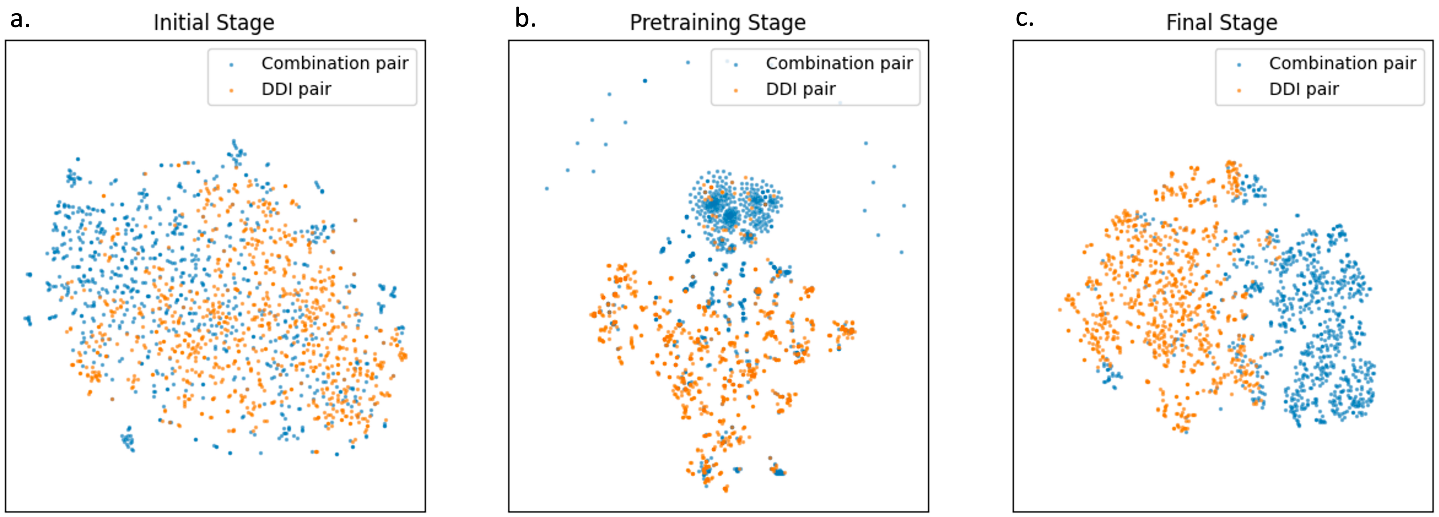

Figure 3.1: The t-SNE plot of multiplied drug pair embedding vectors at three different training stages: initial, pre-training, and final stage. The combination pairs are represented by blue dots, while the DDI pairs are represented by yellow dots. All drug pair embedding vectors are the test data (not used in training). DDI: drug-drug interaction.

be similar and the embeddings of DDIs to be dissimilar. Shown in Figure 3.1c is the final drug pair embedding vectors and they show much clear boundary suitable for drug combination binary classification.

3.3 Robustness to Class Imbalance

Since using the TWOSIDES dataset as negative samples showed significant improvement in all the performance metrics, I hypothesized that my framework is more robust in class imbalanced settings. This is a situation where one class has much more examples than the other class, which can lead to biased or ineffective models. In my case, the number of negative samples in the TWOSIDES dataset was more than three times larger than the positive drug combination samples. So, I conducted an experiment to verify the robustness of my framework by gradually increasing the size of the negative set, using the Area Under the Precision-Recall Curve (AUPRC) metric on the test set. For all algorithms, SCL pretraining was performed.

As seen in Table 3.3, I found that the AUPRC decreased substantially when the model was trained with randomly sampled negative datasets. However, when using the TWOSIDES dataset as the negative dataset, the AUPRC either decreased less or didn't decrease at all. I believe that this result provides strong evidence that utilizing DDI data as negative dataset can guide the deep learning model in building an appropriate and robust decision boundary for drug combination prediction.

Algorithm	Negative dataset	1:1 ratio	1:2 ratio	1:3 ratio
node2vec	TWOSIDES	**0.940 ± 0.007**	**0.928 ± 0.005**	**0.934 ± 0.004**
	random	0.804 ± 0.020	0.753 ± 0.017	0.766 ± 0.012
edge2vec	TWOSIDES	**0.951 ± 0.006**	**0.942 ± 0.007**	**0.944 ± 0.003**
	random	0.835 ± 0.019	0.771 ± 0.016	0.785 ± 0.015
res2vec_homo	TWOSIDES	**0.940 ± 0.005**	**0.921 ± 0.006**	**0.921 ± 0.004**
	random	0.828 ± 0.015	0.763 ± 0.019	0.747 ± 0.007
res2vec_hetero	TWOSIDES	**0.937 ± 0.012**	**0.921 ± 0.006**	**0.920 ± 0.005**
	random	0.805 ± 0.017	0.739 ± 0.018	0.738 ± 0.015
DREAMwalk	TWOSIDES	**0.955 ± 0.006**	**0.943 ± 0.008**	**0.949 ± 0.003**
	random	0.876 ± 0.010	0.830 ± 0.017	0.809 ± 0.019
GCN	TWOSIDES	**0.964 ± 0.008**	**0.955 ± 0.009**	**0.945 ± 0.009**
	random	0.786 ± 0.034	0.740 ± 0.038	0.710 ± 0.033
GraphSAGE	TWOSIDES	**0.831 ± 0.018**	**0.838 ± 0.013**	**0.882 ± 0.015**
	random	0.715 ± 0.022	0.740 ± 0.021	0.725 ± 0.017
GAT	TWOSIDES	**0.918 ± 0.025**	**0.926 ± 0.013**	**0.934 ± 0.011**
	random	0.732 ± 0.015	0.725 ± 0.024	0.740 ± 0.023
GIN	TWOSIDES	**0.967 ± 0.004**	**0.955 ± 0.008**	**0.926 ± 0.037**
	random	0.848 ± 0.020	0.791 ± 0.019	0.757 ± 0.020

Table 3.3: AUPRC (mean ± std) in class imbalanced settings. The ratio is positive:negative ratio of the dataset used. The first row of each algorithm is my framework, which uses the TWOSIDES database as negative dataset. The second row presents the results of the baseline framework, which uses a randomly sampled negative dataset. For all algorithms, SCL pretraining was performed. The best results are in bold.

3.4 Case Studies: Drug Combination & Drug-Drug Interaction

In order to investigate specific cases, I present two types of case study results: the predicted scores of well-known drug combination pairs and DDI pairs. For simplicity, I refer to the model trained with a randomly sampled negative dataset as the *random model*, and the model trained with the TWOSIDES dataset as the *TWOSIDES model*. For case study experiments, I utilized the best-performing algorithm, DREAMwalk and applied SCL pretraining. I used an equal number of negative and positive data, and the drug pairs studied in the case studies were not included in the training data.

To begin, I obtained the prediction scores of previously known drug combination pairs from both the random and TWOSIDES models, as shown in Table 3.4. One widely used combination is statins (atorvastatin, fluvastatin, rosuvastatin) and fenofibrate, which is often prescribed together to reduce cardiovascular risk in patients with dyslipidemia (Jacobson and Zimmerman, 2006; Davidson *et al.*, 2009; Farnier *et al.*, 2000; Biswas *et al.*, 2021). While the random model also predicted scores above 0.5 in these cases, the TWOSIDES model predicted higher scores close to 1.0. In other cases, the TWOSIDES model demonstrated a more distinct score difference compared to the random model. For instance, the combination of dutasteride and tamsulosin showed optimal control of male lower urinary tract symptoms associated with benign prostatic hyperplasia (Dimitropoulos and Gravas, 2016). Additionally, the fixed combination of latanoprost-timolol therapy has been found to be safe and effective for lowering intraocular pressure in patients with ocular hypertension or glaucoma (Higginbotham *et al.*, 2010). And the combination of milrinone and esmolol has shown promising results in clinical trials for treating acute myocar-

Drug 1	Drug 2	Prediction Score (↑)		Indication	Reference
		random	TWOSIDES		
Atorvastatin	Fenofibrate	0.969	**0.999**	Dyslipidemia	(Davidson *et al.*, 2009)
Fluvastatin	Fenofibrate	0.522	**0.996**	Dyslipidemia	(Farnier *et al.*, 2000)
Rosuvastatin	Fenofibrate	0.921	**0.996**	Dyslipidemia	(Biswas *et al.*, 2021)
Dutasteride	Tamsulosin	0.329	**0.986**	Benign Prostatic Hyperplasia	(Dimitropoulos and Gravas, 2016)
Latanoprost	Timolol	0.312	**0.959**	Ocular Hypertension, Glaucoma	(Higginbotham *et al.*, 2010)
Milrinone	Esmolol	0.521	**0.958**	Acute Myocardial Infarction, Severe Sepsis	(Huang *et al.*, 2011; Poh *et al.*, 2014)

Table 3.4: Prediction scores of the random model and TWOSIDES model for drug combination cases. All models used DREAMwalk algorithm to embed drugs and pretrained with SCL. The ratio of positive:negative dataset is 1:1. The higher the better. SCL: supervised contrastive learning.

dial infarction or severe sepsis (Huang *et al.*, 2011; Poh *et al.*, 2014). The test dataset for this case study included 400 drugs and the average of prediction scores of the test dataset in random model and TWOSIDES model were 0.796 and 0.906, respectively.

Then, I also observed the TWOSIDES model is more effective at identifying DDIs than the random model. As shown in Table 3.5, the predicted combination scores of major DDI pairs were significantly lower in the TWOSIDES model than in the random model. For example, the coadministration of phenytoin and ondansetron is generally avoided because the former is a strong inducer of the enzyme CYP 3A4, and the latter is metabolized by it (Zhou, 2008). Similarly, the DDI between ketoconazole and simvastatin is also related to CYP3A, and the former drug inhibits the enzyme, increasing the risk of myopathy and rhabdomyolysis (Gilad and Lampl, 1999). Celecoxib is a moderate CYP2D6 inhibitor and clonidine is metabolized by CYP2D6. Concomitant administration of these two drugs may decrease the metabolism of the latter drug and lower potassium levels in the blood (VandenBrink *et al.*, 2012). And modafinil is an inducer of various CYP enzyme (1A2, 2C9) and can decrease the blood level of the corresponding CYP substrate, duloxetine (Rendic, 2002). Furthermore, lansoprazole is an OAT3 (organic anion transporter 3) inhibitor, which can inhibit the excretion of mercaptopurine (an OAT3 substrate). Since mercaptopurine has a narrow therapeutic index, it is usually not recommended to administer these two drugs together (Duan *et al.*, 2012). Lastly, the combination of theophylline and formoterol, both used to treat asthma or COPD, is not recommended due to the increased risk of hypokalemic effects of formoterol, which can be potentially caused by theophylline (Van den Berg *et al.*, 1999). The test dataset for this case study included 400 drugs and the average of prediction scores of the test dataset in random model and TWOSIDES model were 0.446 and 0.078,

Drug 1	Drug 2	Prediction Score (↓)		DDI type	Reference
		random	TWOSIDES		
Phenytoin	Ondansetron	0.778	**0.004**	CYP3A4 interaction	(Zhou, 2008)
Ketoconazole	Simvastatin	0.955	**0.164**	CYP3A4 interaction	(Gilad and Lampl, 1999)
Celecoxib	Clonidine	0.964	**0.253**	CYP2D6 interaction	(VandenBrink *et al.*, 2012)
Modafinil	Duloxetine	0.994	**0.158**	CYP1A2, 2C9 interaction	(Rendic, 2002)
Lansoprazole	Mercaptopurine	0.849	**0.004**	OAT3 interaction	(Duan *et al.*, 2012)
Theophylline	Formoterol	0.904	**0.040**	Increased risk of hypokalemia	(Van den Berg *et al.*, 1999)

Table 3.5: Prediction scores of the random and TWOSIDES model for DDI cases. All models used DREAMwalk algorithm to embed drugs and pretrained with SCL. The ratio of positive:negative dataset is 1:1. The lower the better. SCL: supervised contrastive learning.

respectively.

I also provide the visualization of the target maps of the drug pairs in Appendix (Section 5.1). These maps include the protein targets of the drugs and corresponding biological functions of the proteins.

Chapter 4

Discussion & Conclusion

The goal of drug combination prediction is to identify effective drug pairs that work together to treat a disease while minimizing adverse effects. Given that the protein target sets of the drugs in combination often exhibit overlapping patterns (Cheng *et al.*, 2019), I believe a machine learning model can learn these patterns and represent them as embedding vectors for predicting drug combinations. Additionally, a typical machine learning model requires high-quality negative data to effectively learn decision boundaries and enhance its ability to differentiate between classes. Biomedical knowledge graph-based methods can predict drug combinations for various types of diseases, but previous studies utilized unlabeled drug pairs as negative dataset, which is unreliable. We proposed a new approach that utilizes DDI pairs as a more reliable source of negative data for drug combination prediction. To project and transform the initial drug embeddings from the network into another vector space that is much more suitable for drug combination prediction, we applied SCL technique at pretraining stage.

Our study yielded some noteworthy findings. One of the most significant was that using the TWOSIDES DDI dataset as a negative samples significantly improved performance in various settings compared to using randomly sampled negative pairs from drug lists. And we found that the supervised contrastive learning was helpful in drug combination prediction task, and it not only improved the performance of our models, but also helped the model to build better decision boundaries which was visible in the embedding space visualization using t-SNE. The robustness of the prediction performance in class imbalanced settings further demonstrated the effectiveness of our approach.

There are several important points to mention regarding this research. First, the utilization of KGs can occasionally lead to knowledge leakage problems when there are shared edges between the training dataset and the test dataset. However, it should be noted that in the MSI network, there were no drug-drug edges present. And it is essential to share the protein target layer (protein-protein edges) in order to embed each drug into vectors using network embedding algorithms. Moreover, many previous studies have also used the same process, as they form the essence of using KGs.

Second, comparing performance metrics between two different negative datasets, namely TWOSIDES and random sampling, is not a straightforward task. It can be argued that it is unfair to directly compare perfomance when the datasets are different. Despite this issue, it was necessary to provide the results for the baseline framework that use random sampling in order to demonstrate the effectiveness of utilizing DDI data as negative dataset for drug combination prediction.

Furthermore, the recall metric, which measures the ratio of true positive predictions to the sum of true positives and false negatives, serves as a fair indicator of the power of using strong negatives for predicting true positives.

As indicated in Table 3.1 and Table 3.2, the use of the TWOSIDES dataset as a negative data significantly improved the recall score in most of the algorithms except GraphSAGE and GAT. This result demonstrates that leveraging strong negatives such as DDI data contributes to the establishment of more accurate decision boundaries for predicting true positives, specifically drug combinations.

Lastly, training GNNs on my framework was not easy, occasionally leading to unstable performance results. I believe this is due to the sparse connection in the KG, which might made it difficult for an end-to-end neural networks like GNNs to effectively propagate information across the graph, leading to instability during training. Also, GNNs are prone to gradient explosion or vanishing, where the gradients either become too large or diminish rapidly during back propagation. Techniques such as gradient clipping can help mitigate these problems.

There are several limitations in my research that leave room for future research directions. Firstly, my framework does not currently incorporate precise dosing plans and the assessment of potential side effects, which are crucial factors in predicting drug combinations for practical situations. To enhance the applicability of predictions from cellular or systems biology levels to clinical settings, it would be valuable to integrate clinical trial information, thus increasing the likelihood of success. Secondly, similar to previous methods, my framework focuses on predicting scores between two drugs, while there are quite a few combinations that involve more than three drugs. Therefore, it would be beneficial to extend the model to encompass interactions among multiple drugs, enabling the modeling of complex drug combinations. Lastly, my research does not explicitly consider disease entities when predicting drug combinations. Although the MSI network utilized disease nodes implicitly during the initial embedding process using random walk or graph neural network (GNN) algorithms, it would be more informative to incorporate disease embeddings, as drug combinations are typically designed for specific indications. By incorporating disease embeddings, the framework can better capture the specificity and relevance of drug combinations to particular diseases, enhancing the overall expressiveness of the predictions.

Bibliography

Bang, D., Lim, S., Lee, S., and Kim, S. (2022). Multi-layer guilt-by-association-based drug repurposing by integrating clinical knowledge on biological heterogeneous networks. *bioRxiv*, pages 2022–11.

Berenbaum, M. C. (1989). What is synergy? *Pharmacological reviews*, **41**(2), 93–141.

Biswas, K., Tiwari, A., Jadhav, P., Goel, A., and Chanukya, G. (2021). Rosuvastatin and fenofibrate combination in the treatment of mixed hyperlipidemia: A narrative review. *Journal of Current Medical Research and Opinion*, **4**(03), 867–877.

Bliss, C. I. (1939). The toxicity of poisons applied jointly 1. *Annals of applied biology*, **26**(3), 585–615.

Breiman, L. (2001). Random forests. *Machine learning*, **45**, 5–32.

Chalmers, J. (1999). The importance of drug combinations for effective control of hypertension. *Clinical and Experimental Hypertension*, **21**(5-6), 875–884.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Cheng, F., Kovács, I. A., and Barabási, A.-L. (2019). Network-based prediction of drug combinations. *Nature communications*, **10**(1), 1197.

Chopra, S., Hadsell, R., and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Crystal, A. S., Shaw, A. T., Sequist, L. V., Friboulet, L., Niederst, M. J., Lockerman, E. L., Frias, R. L., Gainor, J. F., Amzallag, A., Greninger, P., *et al.* (2014). Patient-derived models of acquired resistance can identify effective drug combinations for cancer. *Science*, **346**(6216), 1480–1486.

Cvetkovic, R. S. and Goa, K. L. (2003). Lopinavir/ritonavir: a review of its use in the management of hiv infection. *Drugs*, **63**(8), 769–802.

Davidson, M. H., Rooney, M. W., Drucker, J., Griffin, H. E., Oosman, S., Beckert, M., Investigators, L.-A., *et al.* (2009). Efficacy and tolerability of atorvastatin/fenofibrate fixed-dose combination tablet compared with atorvastatin and fenofibrate monotherapies in patients with dyslipidemia: a 12-week, multicenter, double-blind, randomized, parallel-group study. *Clinical therapeutics*, **31**(12), 2824–2838.

Dimitropoulos, K. and Gravas, S. (2016). Fixed-dose combination therapy with dutasteride and tamsulosin in the management of benign prostatic hyperplasia. *Therapeutic advances in urology*, **8**(1), 19–28.

Duan, P., Li, S., Ai, N., Hu, L., Welsh, W. J., and You, G. (2012). Potent inhibitors of human organic anion transporters 1 and 3 from clinical drug libraries: discovery and molecular characterization. *Molecular pharmaceutics*, **9**(11), 3340–3346.

Farnier, M., Dejager, S., Group, F. F. S., *et al.* (2000). Effect of combined fluvastatin-fenofibrate therapy compared with fenofibrate monotherapy in severe primary hypercholesterolemia. *The American journal of cardiology*, **85**(1), 53–57.

Gao, Z., Fu, G., Ouyang, C., Tsutsui, S., Liu, X., Yang, J., Gessner, C., Foote, B., Wild, D., Ding, Y., *et al.* (2019). edge2vec: Representation learning using edge semantics for biomedical knowledge discovery. *BMC bioinformatics*, **20**(1), 1–15.

Gilad, R. and Lampl, Y. (1999). Rhabdomyolysis induced by simvastatin and ketoconazole treatment. *Clinical neuropharmacology*, **22**(5), 295–297.

Giles, T. D., Weber, M. A., Basile, J., Gradman, A. H., Bharucha, D. B., Chen, W., and Pattathil, M. (2014). Efficacy and safety of nebivolol and valsartan as fixed-dose combination in hypertension: a randomised, multicentre study. *The Lancet*, **383**(9932), 1889–1898.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Güvenç Paltun, B., Kaski, S., and Mamitsuka, H. (2021). Machine learning approaches for drug combination therapies. *Briefings in bioinformatics*, **22**(6), bbab293.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, **30**.

Higginbotham, E. J., Olander, K. W., Kim, E. E., Grunden, J. W., Kwok, K. K., Tressler, C. S., Group, U. S. F.-C. S., *et al.* (2010). Fixed combination of latanoprost and timolol vs individual components for primary open-angle glaucoma or ocular hypertension: a randomized, double-masked study. *Archives of ophthalmology*, **128**(2), 165–172.

Holbeck, S. L., Camalier, R., Crowell, J. A., Govindharajulu, J. P., Hollingshead, M., Anderson, L. W., Polley, E., Rubinstein, L., Srivastava, A., Wilsker, D., *et al.* (2017). The national cancer institute almanac: A comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activitynci almanac of approved cancer drug combinations. *Cancer research*, **77**(13), 3564–3576.

Home, F. (2013). Orange book: approved drug products with therapeutic equivalence evaluations. *USA: US Food & Drug Administration*.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. (2021). Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*.

Huang, M.-H., Wu, Y., Nguyen, V., Rastogi, S., McConnell, B. K., Wijaya, C., Uretsky, B. F., Poh, K.-K., Tan, H.-C., and Fujise, K. (2011). Heart protection by combination therapy with esmolol and milrinone at late-ischemia and early reperfusion. *Cardiovascular drugs and therapy*, **25**, 223–232.

Jacobson, T. A. and Zimmerman, F. H. (2006). Fibrates in combination with statins in the management of dyslipidemia. *The Journal of Clinical Hypertension*, **8**(1), 35–41.

Jin, W., Stokes, J. M., Eastman, R. T., Itkin, Z., Zakharov, A. V., Collins, J. J., Jaakkola, T. S., and Barzilay, R. (2021). Deep learning identifies synergistic drug combinations for treating covid-19. *Proceedings of the National Academy of Sciences*, **118**(39), e2105070118.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, **33**, 18661–18673.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kojaku, S., Yoon, J., Constantino, I., and Ahn, Y.-Y. (2021). Residual2vec: Debiasing graph embedding with random graphs. *Advances in Neural Information Processing Systems*, **34**, 24150–24163.

Liu, H., Zhang, W., Nie, L., Ding, X., Luo, J., and Zou, L. (2019). Predicting effective drug combinations using gradient tree boosting based on features extracted from drug-protein heterogeneous network. *BMC bioinformatics*, **20**(1), 1–12.

Liu, W., Wen, Y., Yu, Z., and Yang, M. (2016). Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*.

Liu, Y., Wei, Q., Yu, G., Gai, W., Li, Y., and Chen, X. (2014). Dcdb 2.0: a major update of the drug combination database. *Database*, **2014**.

Loewe, S. (1953). The problem of synergism and antagonism of combined drugs. *Arzneimittelforschung*, **3**, 285–290.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, **26**.

O'Neil, J., Benita, Y., Feldman, I., Chenard, M., Roberts, B., Liu, Y., Li, J., Kral, A., Lejnine, S., Loboda, A., *et al.* (2016). An unbiased oncology compound screen to identify novel combination strategies. *Molecular cancer therapeutics*, **15**(6), 1155–1162.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710.

Poh, K.-K., Xu, X., Chan, M. Y., Lee, C.-H., Tay, E. L., Low, A. F., Chan, K. H., Sia, W., Tang, L.-Q., Tan, H. C., *et al.* (2014). Safety of combination therapy with milrinone and esmolol for heart protection during percutaneous coronary intervention in acute myocardial infarction. *European journal of clinical pharmacology*, **70**, 527–530.

Preuer, K., Lewis, R. P., Hochreiter, S., Bender, A., Bulusu, K. C., and Klambauer, G. (2018). Deepsynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, **34**(9), 1538–1546.

Rendic, S. (2002). Summary of information on human cyp enzymes: human p450 metabolism data. *Drug metabolism reviews*, **34**(1-2), 83–448.

Ruiz, C., Zitnik, M., and Leskovec, J. (2021). Identification of disease treatment mechanisms through the multiscale interactome. *Nature communications*, **12**(1), 1796.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, **13**(11), 2498–2504.

Shtar, G., Azulay, L., Nizri, O., Rokach, L., and Shapira, B. (2022). Cdcdb: A large and continuously updated drug combination database. *Scientific data*, **9**(1), 263.

Sidorov, P., Naulaerts, S., Ariey-Bonnet, J., Pasquier, E., and Ballester, P. J. (2019). Predicting synergism of cancer drug combinations using nci-almanac data. *Frontiers in chemistry*, **7**, 509.

Smilek, D. E., Ehlers, M. R., and Nepom, G. T. (2014). Restoring the balance: immunotherapeutic combinations for autoimmune disease. *Disease models & mechanisms*, **7**(5), 503–513.

Tatonetti, N. P., Ye, P. P., Daneshjou, R., and Altman, R. B. (2012). Data-driven prediction of drug effects and interactions. *Science translational medicine*, **4**(125), 125ra31–125ra31.

Van den Berg, B., Derks, M., Koolen, M., Braat, M., Butter, J., and Van Boxtel, C. (1999). Pharmacokinetic/pharmacodynamic modelling of the eosinopenic and hypokalemic effects of formoterol and theophylline combination in healthy men. *Pulmonary Pharmacology & Therapeutics*, **12**(3), 185–192.

VandenBrink, B. M., Foti, R. S., Rock, D. A., Wienkers, L. C., and Wahlstrom, J. L. (2012). Prediction of cyp2d6 drug interactions from in vitro data: evidence for substrate-dependent inhibition. *Drug Metabolism and Disposition*, **40**(1), 47–53.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, X., Zhu, H., Jiang, Y., Li, Y., Tang, C., Chen, X., Li, Y., Liu, Q., and Liu, Q. (2022). Prodeepsyn: predicting anticancer synergistic drug combinations by embedding cell lines with protein–protein interaction network. *Briefings in Bioinformatics*, **23**(2), bbab587.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2018). How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Yadav, B., Wennerberg, K., Aittokallio, T., and Tang, J. (2015). Searching for drug synergy in complex dose–response landscapes using an interaction potency model. *Computational and structural biotechnology journal*, **13**, 504–513.

Yu, L., Xia, M., and An, Q. (2022). A network embedding framework based on integrating multiplex network for drug combination prediction. *Briefings in bioinformatics*, **23**(1), bbab364.

Zarin, D. A., Tse, T., Williams, R. J., Califf, R. M., and Ide, N. C. (2011). The clinicaltrials. gov results database—update and key issues. *New England Journal of Medicine*, **364**(9), 852–860.

Zhang, Z. and Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, **31**.

Zheng, W., Sun, W., and Simeonov, A. (2018). Drug repurposing screens and synergistic drug-combinations for infectious diseases. *British journal of pharmacology*, **175**(2), 181–191.

Zhou, S.-F. (2008). Drugs behave as substrates, inhibitors and inducers of human cytochrome p450 3a4. *Current drug metabolism*, **9**(4), 310–322.

Chapter 5

Appendix

5.1 Target maps of case study drug pairs

In this section, I present target maps of the drug pairs which were previously introduced in the case study section (Section 3.4). These target maps display the protein targets of both drugs and the biological function entities connected to the respective proteins. By referring to these target maps, one can easily examine the functional role of the proteins, identify overlapping patterns between the two drugs, and utilize this information to conduct further studies. These studies can investigate the possible reasons for synergistic effects in drug combinations or adverse effects in DDIs. The target maps are visualized using Cytoscape (Shannon *et al.*, 2003).

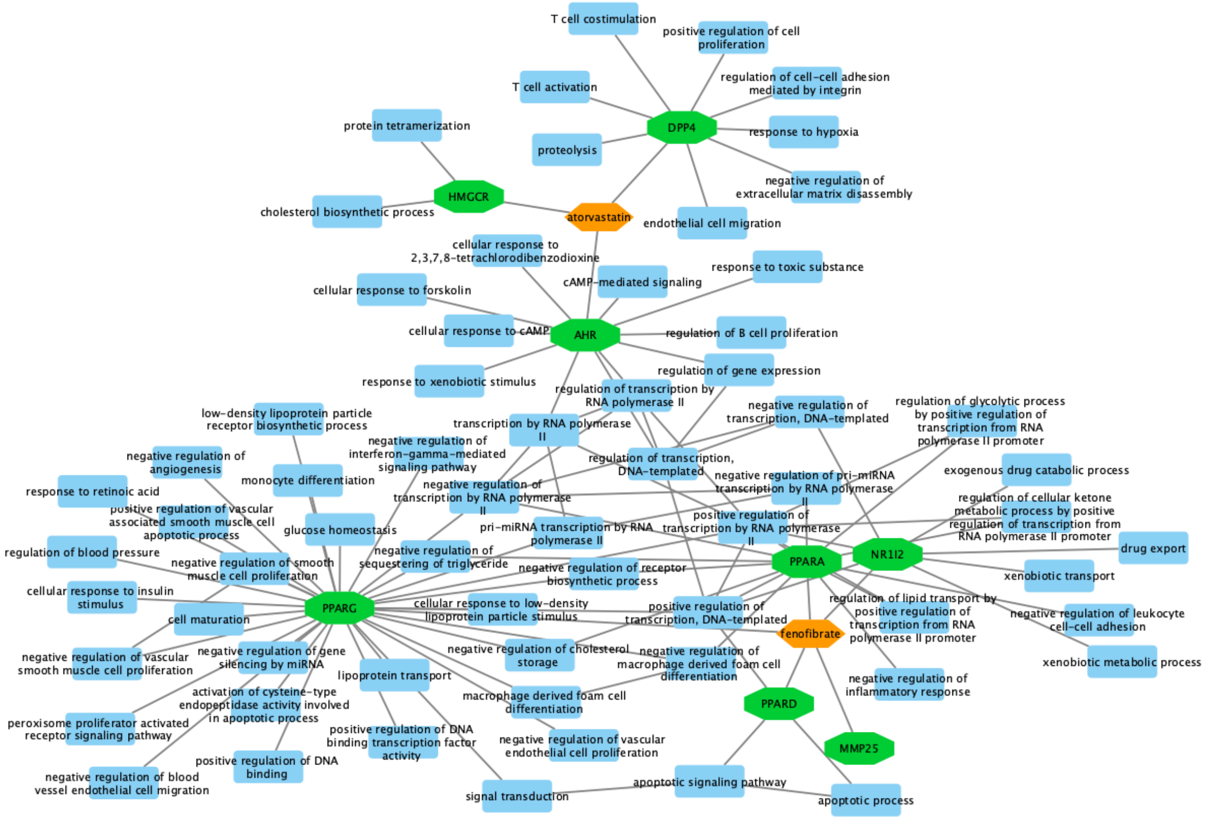
T cell costimulation
positive regulation of cell proliferation
regulation of cell-cell adhesion mediated by integrin
T cell activation
protein tetramerization
DPP4
response to hypoxia
proteolysis
negative regulation of extracellular matrix disassembly
HMGCR
cholesterol biosynthetic process
atorvastatin
endothelial cell migration
cellular response to 2,3,7,8-tetrachlorodibenzodioxine
response to toxic substance
cAMP-mediated signaling
cellular response to forskolin
cellular response to cAMP
AHR
regulation of B cell proliferation
regulation of gene expression
response to xenobiotic stimulus
regulation of transcription by RNA polymerase II
negative regulation of transcription, DNA-templated
regulation of glycolytic process by positive regulation of transcription from RNA polymerase II promoter
low-density lipoprotein particle receptor biosynthetic process
transcription by RNA polymerase II
negative regulation of interferon-gamma-mediated signaling pathway
negative regulation of angiogenesis
regulation of transcription, DNA-templated
exogenous drug catabolic process
monocyte differentiation
negative regulation of pri-miRNA transcription by RNA polymerase II
response to retinoic acid
negative regulation of transcription by RNA polymerase II
regulation of cellular ketone metabolic process by positive regulation of transcription from RNA polymerase II promoter
positive regulation of vascular associated smooth muscle cell apoptotic process
glucose homeostasis
pri-miRNA transcription by RNA polymerase II
positive regulation of transcription by RNA polymerase II
regulation of blood pressure
drug export
NR1I2
negative regulation of smooth muscle cell proliferation
negative regulation of sequestering of triglyceride
PPARA
negative regulation of receptor biosynthetic process
xenobiotic transport
cellular response to insulin stimulus
PPARG
cellular response to low-density lipoprotein particle stimulus
regulation of lipid transport by positive regulation of transcription from RNA polymerase II promoter
positive regulation of transcription, DNA-templated
cell maturation
negative regulation of leukocyte cell-cell adhesion
fenofibrate
negative regulation of cholesterol storage
negative regulation of vascular smooth muscle cell proliferation
negative regulation of gene silencing by miRNA
negative regulation of macrophage derived foam cell differentiation
xenobiotic metabolic process
lipoprotein transport
activation of cysteine-type endopeptidase activity involved in apoptotic process
negative regulation of inflammatory response
macrophage derived foam cell differentiation
PPARD
peroxisome proliferator activated receptor signaling pathway
positive regulation of DNA binding transcription factor activity
MMP25
negative regulation of vascular endothelial cell proliferation
positive regulation of DNA binding
negative regulation of blood vessel endothelial cell migration
apoptotic signaling pathway
signal transduction
apoptotic process

Figure 5.1: The target maps for atorvastatin and fenofibrate. The color scheme assigns orange, green, and cyan to represent drugs, proteins, and biological functions, respectively.

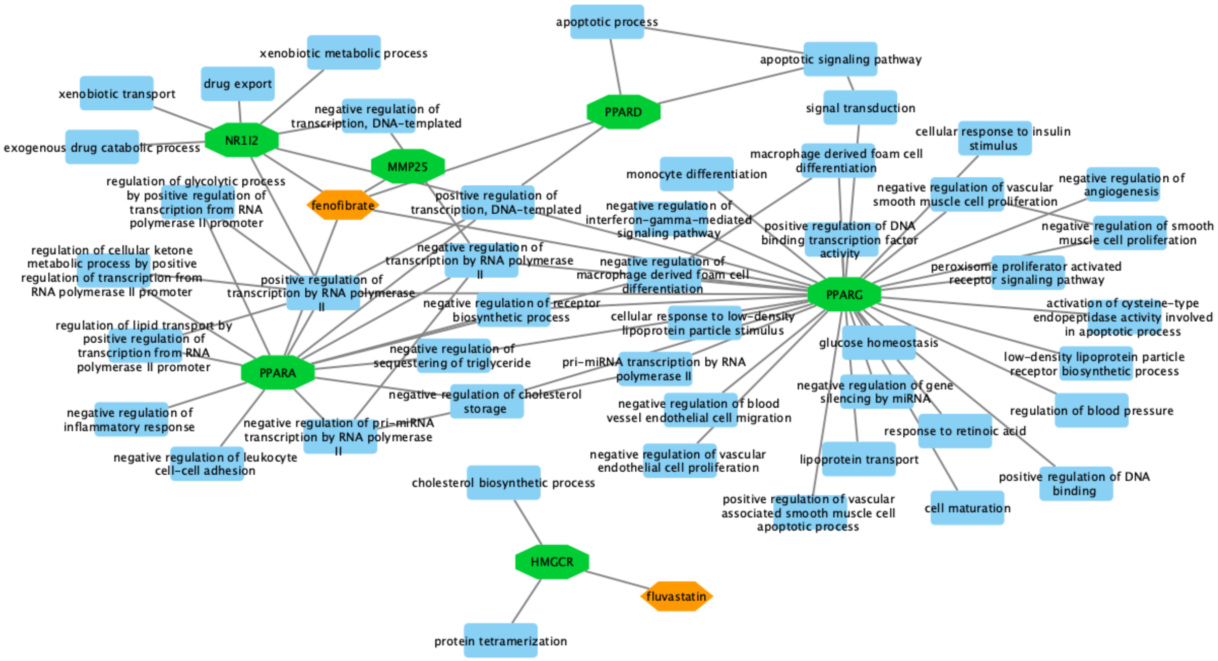
apoptotic process
apoptotic signaling pathway
xenobiotic metabolic process
drug export
xenobiotic transport
negative regulation of transcription, DNA−templated
PPARD
signal transduction
cellular response to insulin stimulus
exogenous drug catabolic process
NR1I2
MMP25
macrophage derived foam cell differentiation
monocyte differentiation
negative regulation of angiogenesis
negative regulation of vascular smooth muscle cell proliferation
regulation of glycolytic process by positive regulation of transcription from RNA polymerase II promoter
fenofibrate
positive regulation of transcription, DNA−templated
negative regulation of interferon−gamma−mediated signaling pathway
positive regulation of DNA binding transcription factor activity
negative regulation of smooth muscle cell proliferation
regulation of cellular ketone metabolic process by positive regulation of transcription from RNA polymerase II promoter
negative regulation of transcription by RNA polymerase II
negative regulation of macrophage derived foam cell differentiation
peroxisome proliferator activated receptor signaling pathway
positive regulation of transcription by RNA polymerase II
PPARG
negative regulation of receptor biosynthetic process
activation of cysteine−type endopeptidase activity involved in apoptotic process
regulation of lipid transport by positive regulation of transcription from RNA polymerase II promoter
cellular response to low−density lipoprotein particle stimulus
glucose homeostasis
negative regulation of sequestering of triglyceride
low−density lipoprotein particle receptor biosynthetic process
PPARA
pri−miRNA transcription by RNA polymerase II
negative regulation of gene silencing by miRNA
negative regulation of cholesterol storage
negative regulation of blood vessel endothelial cell migration
regulation of blood pressure
negative regulation of inflammatory response
negative regulation of pri−miRNA transcription by RNA polymerase II
response to retinoic acid
negative regulation of leukocyte cell−cell adhesion
negative regulation of vascular endothelial cell proliferation
lipoprotein transport
positive regulation of DNA binding
cholesterol biosynthetic process
positive regulation of vascular associated smooth muscle cell apoptotic process
cell maturation
HMGCR
fluvastatin
protein tetramerization

Figure 5.2: The target maps for fluvastatin and fenofibrate. The color scheme assigns orange, green, and cyan to represent drugs, proteins, and biological functions, respectively.

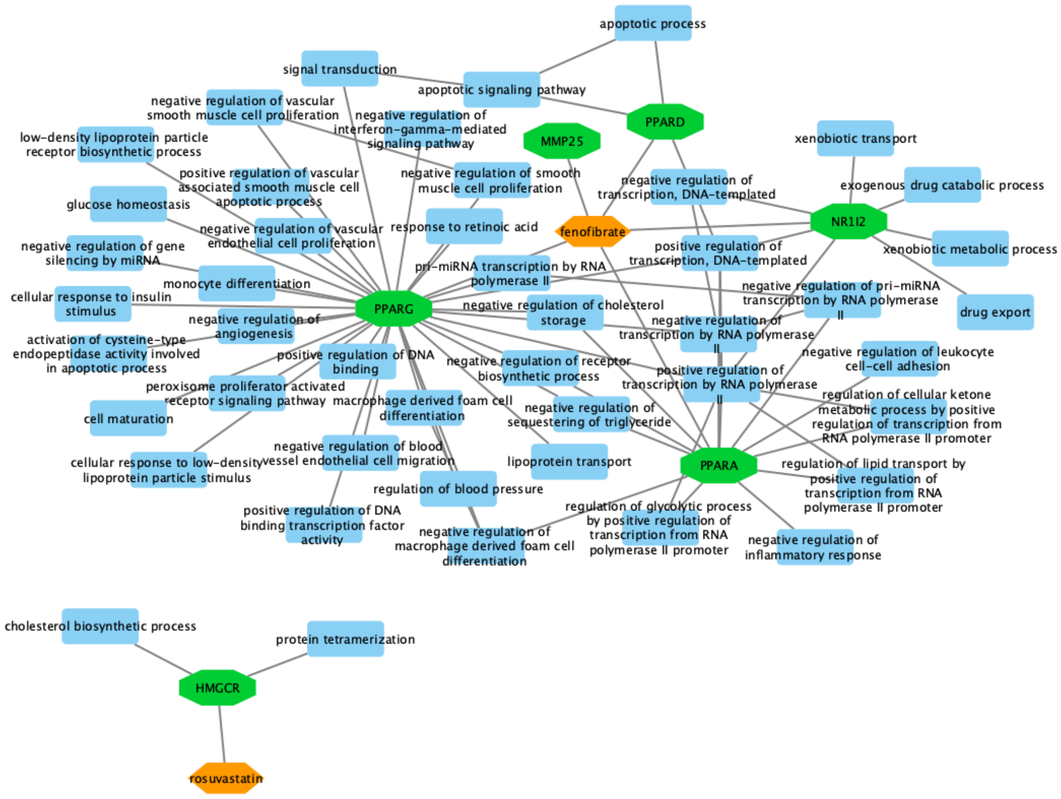
apoptotic process
signal transduction
apoptotic signaling pathway
PPARD
MMP25
negative regulation of vascular smooth muscle cell proliferation
negative regulation of interferon−gamma−mediated signaling pathway
xenobiotic transport
low−density lipoprotein particle receptor biosynthetic process
negative regulation of smooth muscle cell proliferation
negative regulation of transcription, DNA−templated
exogenous drug catabolic process
positive regulation of vascular associated smooth muscle cell apoptotic process
glucose homeostasis
NR1I2
negative regulation of vascular endothelial cell proliferation
response to retinoic acid
fenofibrate
xenobiotic metabolic process
negative regulation of gene silencing by miRNA
positive regulation of transcription, DNA−templated
pri−miRNA transcription by RNA polymerase II
monocyte differentiation
negative regulation of pri−miRNA transcription by RNA polymerase II
drug export
cellular response to insulin stimulus
PPARG
negative regulation of cholesterol storage
negative regulation of angiogenesis
negative regulation of transcription by RNA polymerase II
activation of cysteine−type endopeptidase activity involved in apoptotic process
positive regulation of DNA binding
negative regulation of receptor biosynthetic process
negative regulation of leukocyte cell−cell adhesion
positive regulation of transcription by RNA polymerase II
peroxisome proliferator activated receptor signaling pathway
macrophage derived foam cell differentiation
regulation of cellular ketone metabolic process by positive regulation of transcription from RNA polymerase II promoter
cell maturation
negative regulation of sequestering of triglyceride
negative regulation of blood vessel endothelial cell migration
lipoprotein transport
PPARA
cellular response to low−density lipoprotein particle stimulus
regulation of lipid transport by positive regulation of transcription from RNA polymerase II promoter
regulation of blood pressure
positive regulation of DNA binding transcription factor activity
regulation of glycolytic process by positive regulation of transcription from RNA polymerase II promoter
negative regulation of macrophage derived foam cell differentiation
negative regulation of inflammatory response
cholesterol biosynthetic process
protein tetramerization
HMGCR
rosuvastatin

Figure 5.3: The target maps for rosuvastatin and fenofibrate. The color scheme assigns orange, green, and cyan to represent drugs, proteins, and biological functions, respectively.

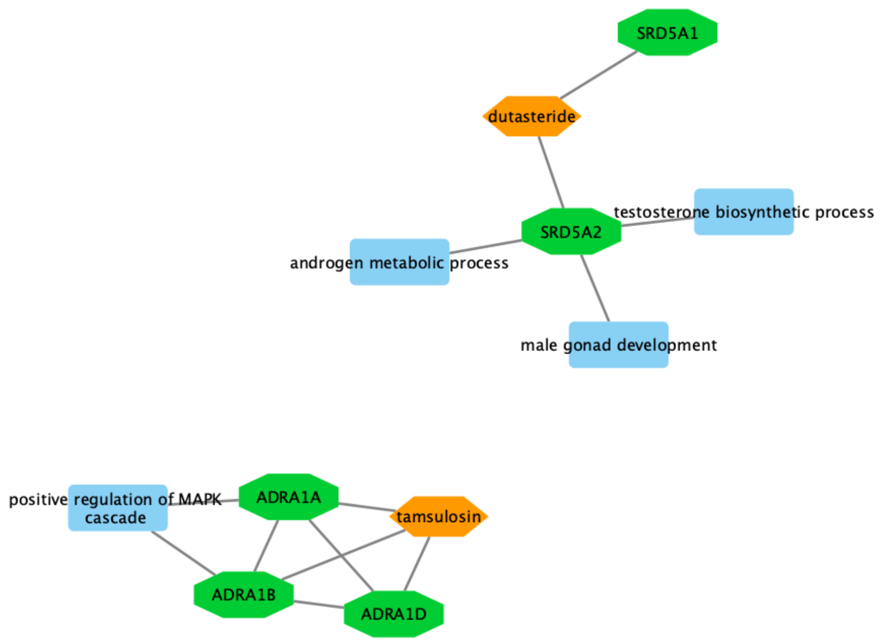
SRD5A1
dutasteride
SRD5A2
testosterone biosynthetic process
androgen metabolic process
male gonad development
positive regulation of MAPK cascade
ADRA1A
tamsulosin
ADRA1B
ADRA1D

Figure 5.4: The target maps for dutasteride and tamsulosin. The color scheme assigns orange, green, and cyan to represent drugs, proteins, and biological functions, respectively.

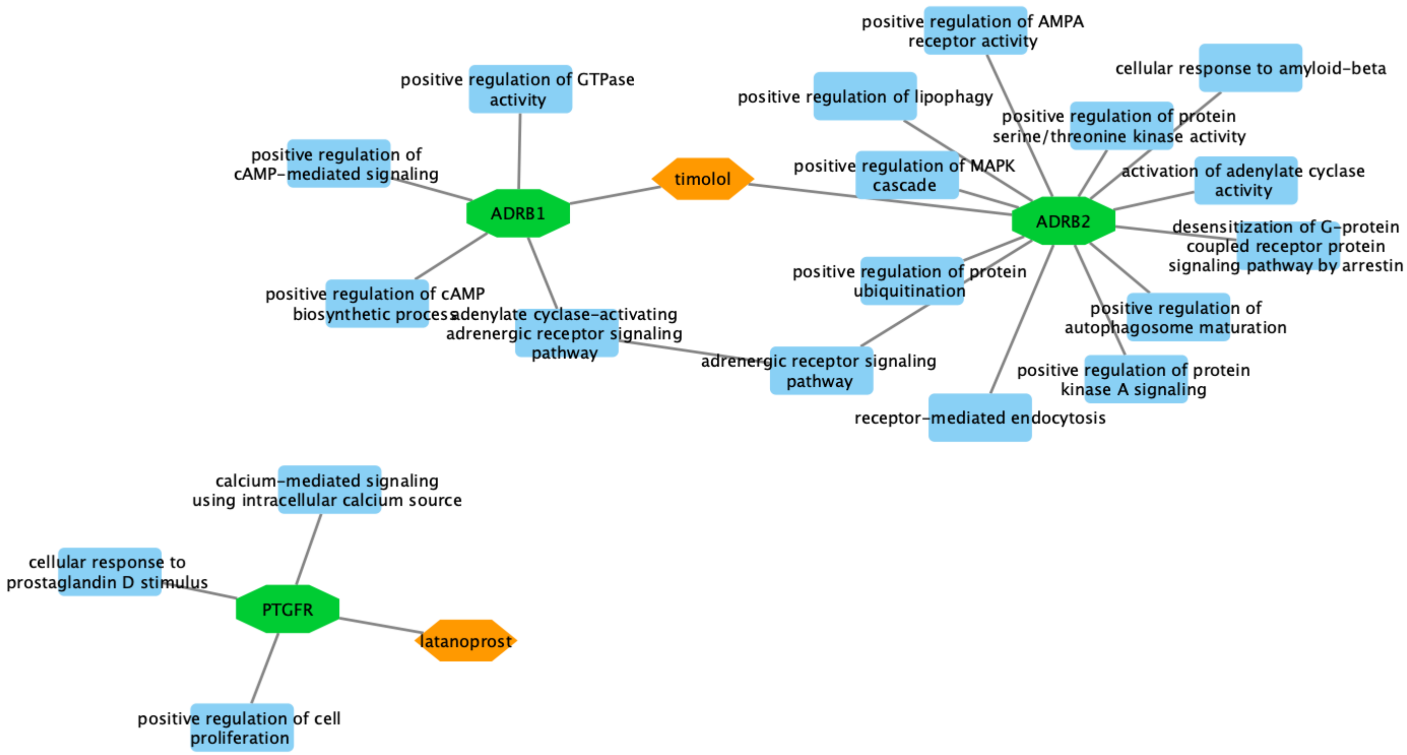
positive regulation of AMPA receptor activity
cellular response to amyloid-beta
positive regulation of GTPase activity
positive regulation of lipophagy
positive regulation of protein serine/threonine kinase activity
positive regulation of cAMP-mediated signaling
timolol
positive regulation of MAPK cascade
activation of adenylate cyclase activity
ADRB1
ADRB2
desensitization of G-protein coupled receptor protein signaling pathway by arrestin
positive regulation of protein ubiquitination
positive regulation of cAMP biosynthetic process
adenylate cyclase-activating adrenergic receptor signaling pathway
positive regulation of autophagosome maturation
adrenergic receptor signaling pathway
positive regulation of protein kinase A signaling
receptor-mediated endocytosis
calcium-mediated signaling using intracellular calcium source
cellular response to prostaglandin D stimulus
PTGFR
latanoprost
positive regulation of cell proliferation

Figure 5.5: The target maps for latanoprost and timolol. The color scheme assigns orange, green, and cyan to represent drugs, proteins, and biological functions, respectively.

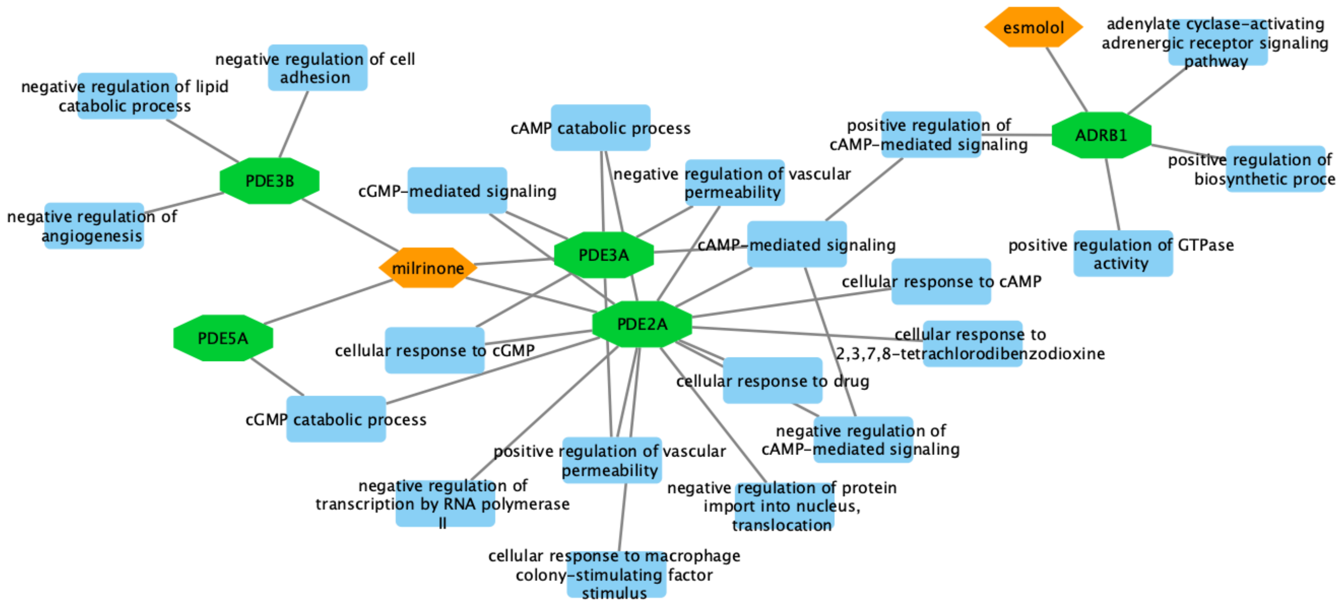
esmolol
adenylate cyclase−activating adrenergic receptor signaling pathway
negative regulation of cell adhesion
negative regulation of lipid catabolic process
cAMP catabolic process
positive regulation of cAMP−mediated signaling
ADRB1
positive regulation of cAMP biosynthetic process
PDE3B
cGMP−mediated signaling
negative regulation of vascular permeability
negative regulation of angiogenesis
PDE3A
cAMP−mediated signaling
positive regulation of GTPase activity
milrinone
cellular response to cAMP
PDE2A
PDE5A
cellular response to cGMP
cellular response to 2,3,7,8−tetrachlorodibenzodioxine
cellular response to drug
cGMP catabolic process
negative regulation of cAMP−mediated signaling
positive regulation of vascular permeability
negative regulation of transcription by RNA polymerase II
negative regulation of protein import into nucleus, translocation
cellular response to macrophage colony−stimulating factor stimulus

Figure 5.6: The target maps for milrinone and esmolol. The color scheme assigns orange, green, and cyan to represent drugs, proteins, and biological functions, respectively.

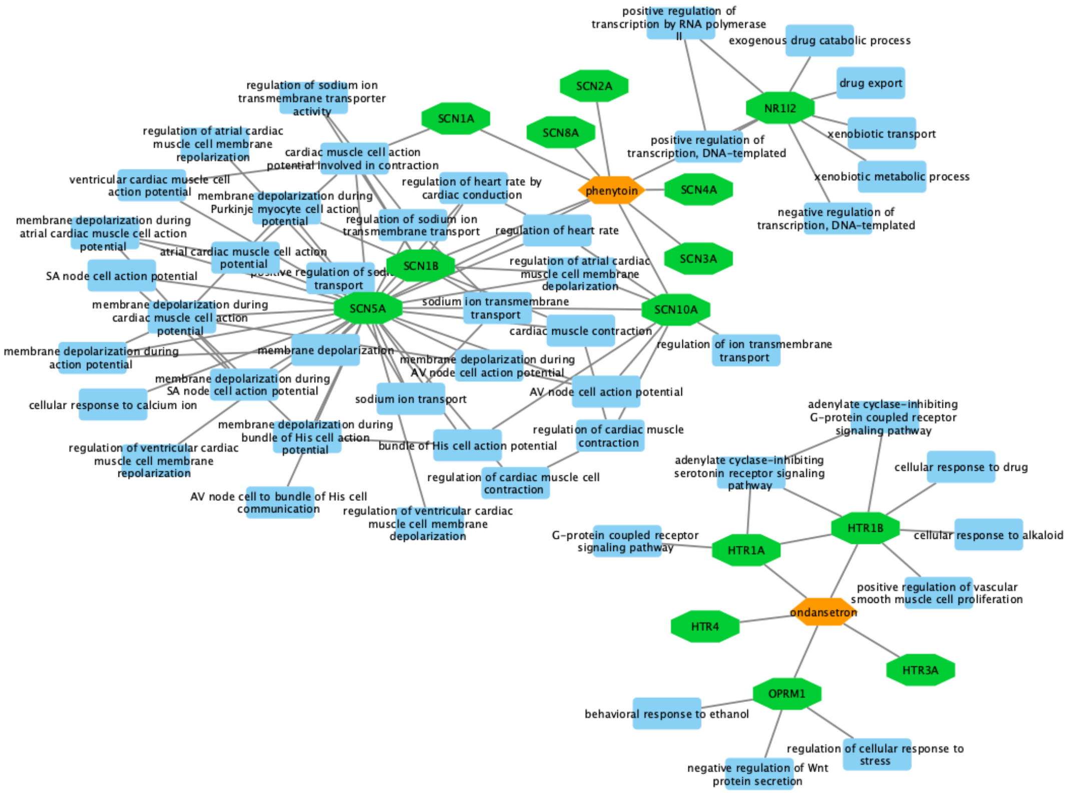
positive regulation of transcription by RNA polymerase II
exogenous drug catabolic process
drug export
NR1I2
xenobiotic transport
xenobiotic metabolic process
negative regulation of transcription, DNA-templated
positive regulation of transcription, DNA-templated
SCN2A
SCN8A
SCN1A
SCN4A
SCN3A
phenytoin
regulation of sodium ion transmembrane transporter activity
regulation of atrial cardiac muscle cell membrane repolarization
cardiac muscle cell action potential involved in contraction
ventricular cardiac muscle cell action potential
regulation of heart rate by cardiac conduction
membrane depolarization during Purkinje myocyte cell action potential
membrane depolarization during atrial cardiac muscle cell action potential
regulation of sodium ion transmembrane transport
regulation of heart rate
atrial cardiac muscle cell action potential
SA node cell action potential
positive regulation of sodium ion transport
SCN1B
regulation of atrial cardiac muscle cell membrane depolarization
SCN5A
sodium ion transmembrane transport
SCN10A
membrane depolarization during cardiac muscle cell action potential
cardiac muscle contraction
regulation of ion transmembrane transport
membrane depolarization during action potential
membrane depolarization
membrane depolarization during AV node cell action potential
membrane depolarization during SA node cell action potential
cellular response to calcium ion
sodium ion transport
AV node cell action potential
membrane depolarization during bundle of His cell action potential
bundle of His cell action potential
regulation of cardiac muscle contraction
regulation of ventricular cardiac muscle cell membrane repolarization
regulation of cardiac muscle cell contraction
AV node cell to bundle of His cell communication
regulation of ventricular cardiac muscle cell membrane depolarization
adenylate cyclase-inhibiting G-protein coupled receptor signaling pathway
adenylate cyclase-inhibiting serotonin receptor signaling pathway
cellular response to drug
HTR1B
cellular response to alkaloid
G-protein coupled receptor signaling pathway
HTR1A
positive regulation of vascular smooth muscle cell proliferation
ondansetron
HTR4
HTR3A
OPRM1
behavioral response to ethanol
regulation of cellular response to stress
negative regulation of Wnt protein secretion

Figure 5.7: The target maps for phenytoin and ondansetron. The color scheme assigns orange, green, and cyan to represent drugs, proteins, and biological functions, respectively.

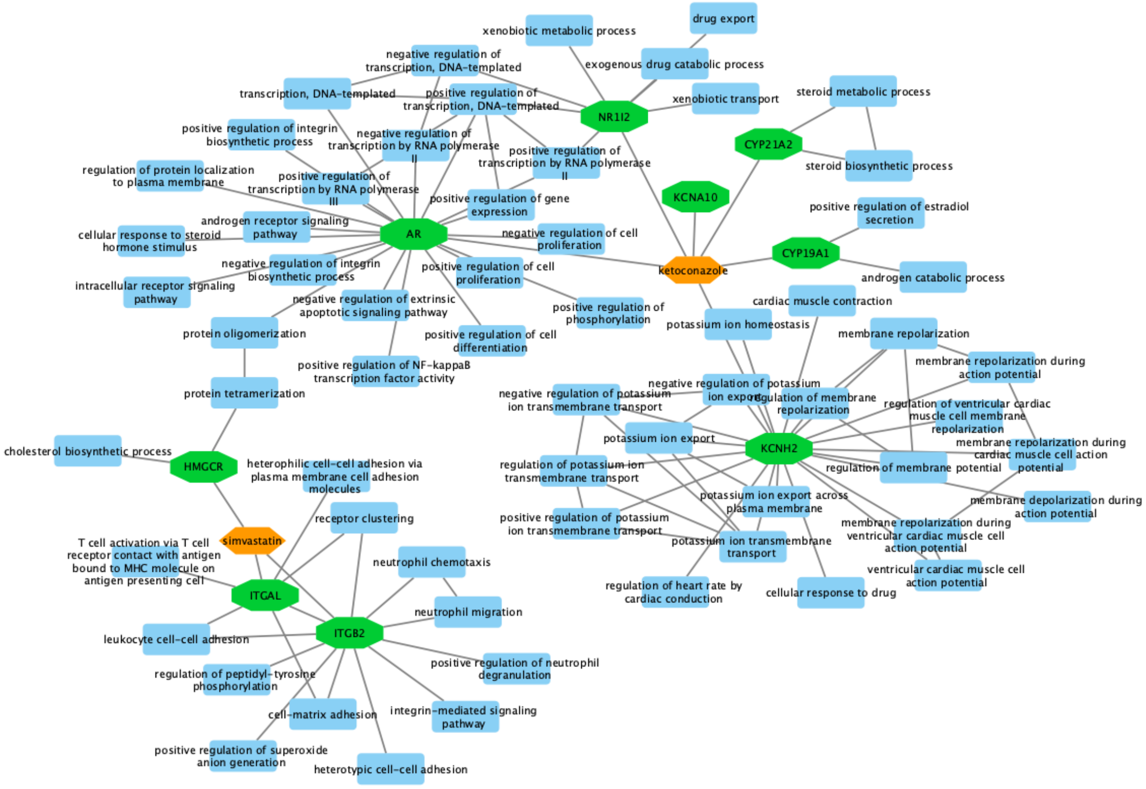
drug export
xenobiotic metabolic process
exogenous drug catabolic process
negative regulation of transcription, DNA-templated
transcription, DNA-templated
positive regulation of transcription, DNA-templated
xenobiotic transport
steroid metabolic process
NR1I2
CYP21A2
positive regulation of integrin biosynthetic process
negative regulation of transcription by RNA polymerase II
positive regulation of transcription by RNA polymerase II
steroid biosynthetic process
regulation of protein localization to plasma membrane
positive regulation of transcription by RNA polymerase III
KCNA10
positive regulation of gene expression
positive regulation of estradiol secretion
androgen receptor signaling pathway
AR
negative regulation of cell proliferation
cellular response to steroid hormone stimulus
CYP19A1
negative regulation of integrin biosynthetic process
positive regulation of cell proliferation
ketoconazole
androgen catabolic process
intracellular receptor signaling pathway
negative regulation of extrinsic apoptotic signaling pathway
cardiac muscle contraction
positive regulation of phosphorylation
protein oligomerization
positive regulation of cell differentiation
potassium ion homeostasis
membrane repolarization
membrane repolarization during action potential
positive regulation of NF-kappaB transcription factor activity
negative regulation of potassium ion export
protein tetramerization
negative regulation of potassium ion transmembrane transport
regulation of membrane repolarization
regulation of ventricular cardiac muscle cell membrane repolarization
potassium ion export
membrane repolarization during cardiac muscle cell action potential
cholesterol biosynthetic process
KCNH2
HMGCR
heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules
regulation of potassium ion transmembrane transport
regulation of membrane potential
potassium ion export across plasma membrane
membrane depolarization during action potential
receptor clustering
positive regulation of potassium ion transmembrane transport
membrane repolarization during ventricular cardiac muscle cell action potential
simvastatin
potassium ion transmembrane transport
T cell activation via T cell receptor contact with antigen bound to MHC molecule on antigen presenting cell
neutrophil chemotaxis
ventricular cardiac muscle cell action potential
ITGAL
regulation of heart rate by cardiac conduction
cellular response to drug
neutrophil migration
ITGB2
leukocyte cell-cell adhesion
positive regulation of neutrophil degranulation
regulation of peptidyl-tyrosine phosphorylation
cell-matrix adhesion
integrin-mediated signaling pathway
positive regulation of superoxide anion generation
heterotypic cell-cell adhesion

Figure 5.8: The target maps for ketoconazole and simvastatin. The color scheme assigns orange, green, and cyan to represent drugs, proteins, and biological functions, respectively.

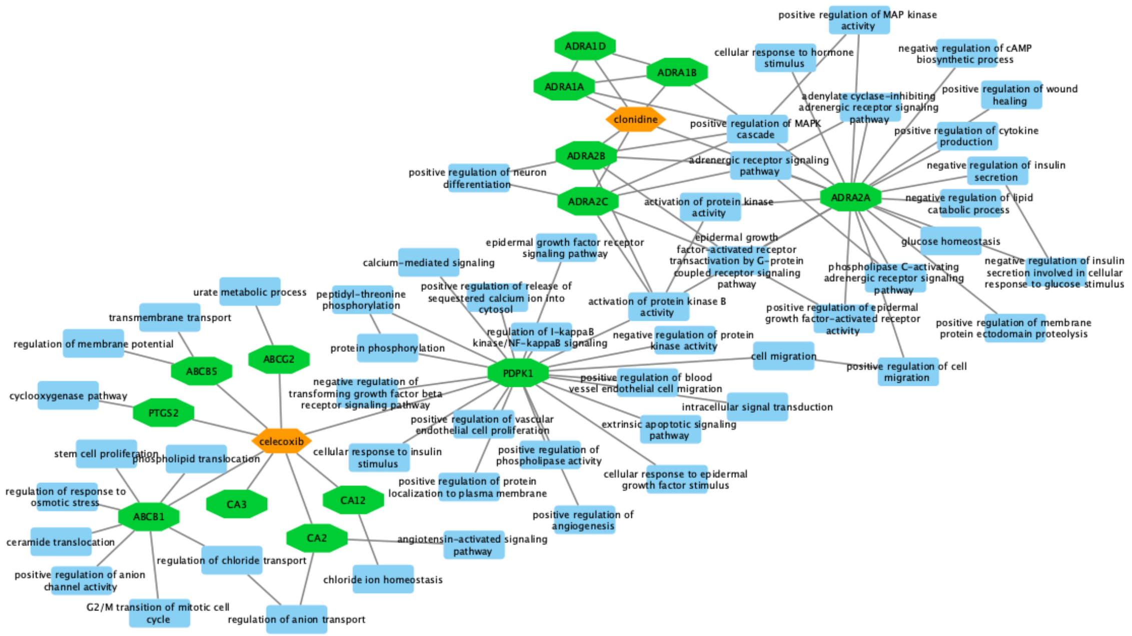
positive regulation of MAP kinase activity
negative regulation of cAMP biosynthetic process
ADRA1D
cellular response to hormone stimulus
ADRA1B
ADRA1A
adenylate cyclase-inhibiting adrenergic receptor signaling pathway
positive regulation of wound healing
clonidine
positive regulation of MAPK cascade
positive regulation of cytokine production
ADRA2B
adrenergic receptor signaling pathway
negative regulation of insulin secretion
positive regulation of neuron differentiation
ADRA2C
activation of protein kinase activity
ADRA2A
negative regulation of lipid catabolic process
epidermal growth factor receptor signaling pathway
epidermal growth factor-activated receptor transactivation by G-protein coupled receptor signaling pathway
glucose homeostasis
calcium-mediated signaling
phospholipase C-activating adrenergic receptor signaling pathway
negative regulation of insulin secretion involved in cellular response to glucose stimulus
urate metabolic process
peptidyl-threonine phosphorylation
positive regulation of release of sequestered calcium ion into cytosol
activation of protein kinase B activity
positive regulation of epidermal growth factor-activated receptor activity
transmembrane transport
positive regulation of membrane protein ectodomain proteolysis
regulation of I-kappaB kinase/NF-kappaB signaling
negative regulation of protein kinase activity
regulation of membrane potential
protein phosphorylation
ABCG2
cell migration
ABCB5
positive regulation of cell migration
PDPK1
negative regulation of transforming growth factor beta receptor signaling pathway
positive regulation of blood vessel endothelial cell migration
cyclooxygenase pathway
PTGS2
intracellular signal transduction
positive regulation of vascular endothelial cell proliferation
extrinsic apoptotic signaling pathway
celecoxib
cellular response to insulin stimulus
stem cell proliferation
phospholipid translocation
positive regulation of phospholipase activity
cellular response to epidermal growth factor stimulus
positive regulation of protein localization to plasma membrane
regulation of response to osmotic stress
CA3
CA12
ABCB1
positive regulation of angiogenesis
ceramide translocation
CA2
angiotensin-activated signaling pathway
regulation of chloride transport
positive regulation of anion channel activity
chloride ion homeostasis
G2/M transition of mitotic cell cycle
regulation of anion transport

Figure 5.9: The target maps for celecoxib and clonidine. The color scheme assigns orange, green, and cyan to represent drugs, proteins, and biological functions, respectively.

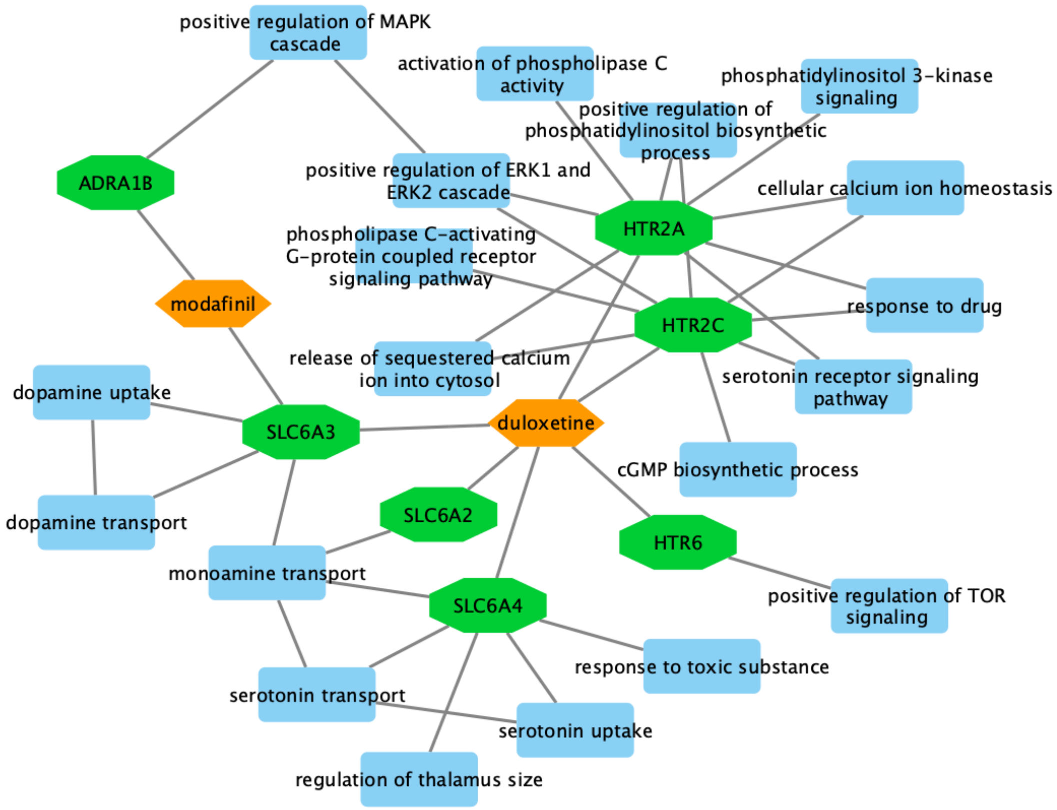
positive regulation of MAPK cascade
activation of phospholipase C activity
phosphatidylinositol 3–kinase signaling
positive regulation of phosphatidylinositol biosynthetic process
ADRA1B
positive regulation of ERK1 and ERK2 cascade
cellular calcium ion homeostasis
HTR2A
phospholipase C–activating G–protein coupled receptor signaling pathway
modafinil
HTR2C
response to drug
release of sequestered calcium ion into cytosol
serotonin receptor signaling pathway
dopamine uptake
SLC6A3
duloxetine
cGMP biosynthetic process
dopamine transport
SLC6A2
HTR6
monoamine transport
SLC6A4
positive regulation of TOR signaling
response to toxic substance
serotonin transport
serotonin uptake
regulation of thalamus size

Figure 5.10: The target maps for modafinil and duloxetine. The color scheme assigns orange, green, and cyan to represent drugs, proteins, and biological functions, respectively.

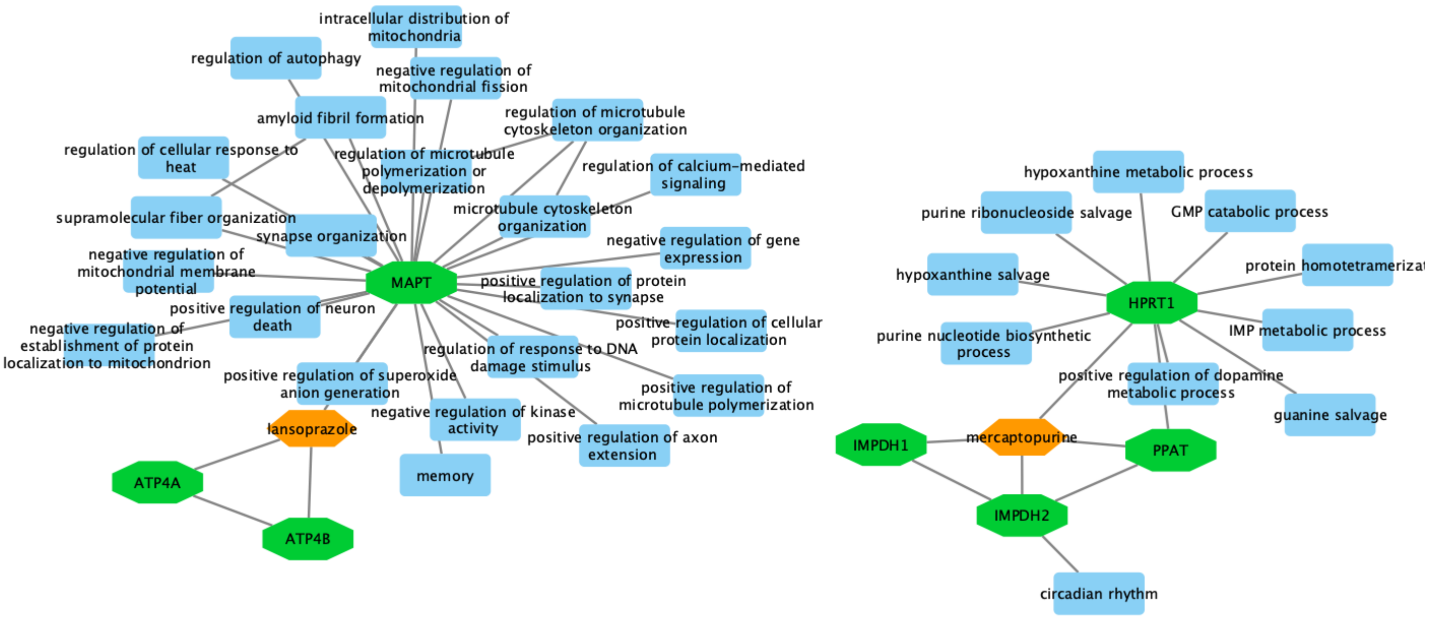
intracellular distribution of mitochondria
regulation of autophagy
negative regulation of mitochondrial fission
amyloid fibril formation
regulation of microtubule cytoskeleton organization
regulation of cellular response to heat
regulation of microtubule polymerization or depolymerization
regulation of calcium-mediated signaling
supramolecular fiber organization
microtubule cytoskeleton organization
synapse organization
negative regulation of gene expression
negative regulation of mitochondrial membrane potential
MAPT
positive regulation of protein localization to synapse
positive regulation of neuron death
positive regulation of cellular protein localization
negative regulation of establishment of protein localization to mitochondrion
regulation of response to DNA damage stimulus
positive regulation of superoxide anion generation
positive regulation of microtubule polymerization
negative regulation of kinase activity
lansoprazole
positive regulation of axon extension
ATP4A
memory
ATP4B
hypoxanthine metabolic process
purine ribonucleoside salvage
GMP catabolic process
protein homotetramerizat
hypoxanthine salvage
HPRT1
purine nucleotide biosynthetic process
IMP metabolic process
positive regulation of dopamine metabolic process
guanine salvage
IMPDH1
mercaptopurine
PPAT
IMPDH2
circadian rhythm

Figure 5.11: The target maps for lansoprazole and mercaptopurine. The color scheme assigns orange, green, and cyan to represent drugs, proteins, and biological functions, respectively.

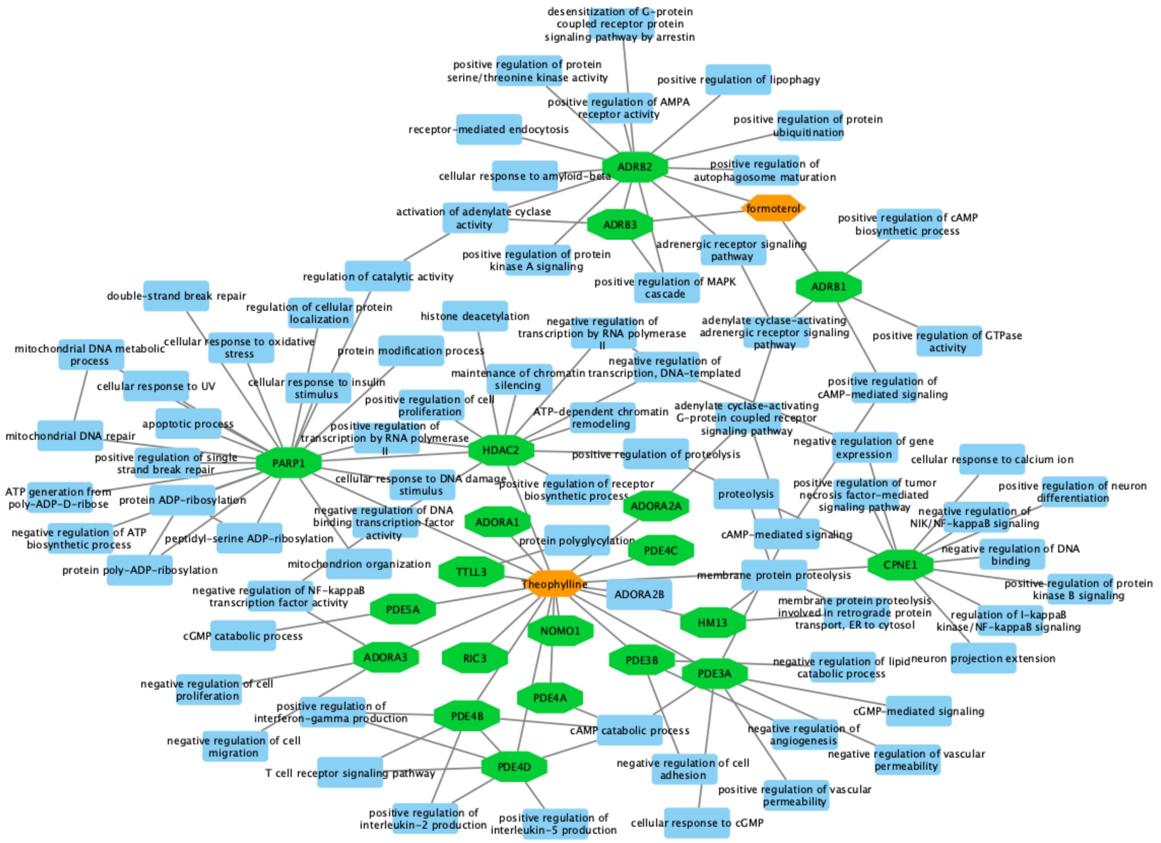
desensitization of G-protein coupled receptor protein signaling pathway by arrestin
positive regulation of protein serine/threonine kinase activity
positive regulation of lipophagy
positive regulation of AMPA receptor activity
positive regulation of protein ubiquitination
receptor-mediated endocytosis
positive regulation of autophagosome maturation
ADRB2
cellular response to amyloid-beta
formoterol
activation of adenylate cyclase activity
ADRB3
positive regulation of cAMP biosynthetic process
adrenergic receptor signaling pathway
positive regulation of protein kinase A signaling
ADRB1
regulation of catalytic activity
positive regulation of MAPK cascade
double-strand break repair
regulation of cellular protein localization
histone deacetylation
negative regulation of transcription by RNA polymerase II
adenylate cyclase-activating adrenergic receptor signaling pathway
positive regulation of GTPase activity
mitochondrial DNA metabolic process
cellular response to oxidative stress
protein modification process
negative regulation of chromatin transcription, DNA-templated
maintenance of chromatin silencing
cellular response to UV
cellular response to insulin stimulus
positive regulation of cAMP-mediated signaling
positive regulation of cell proliferation
ATP-dependent chromatin remodeling
adenylate cyclase-activating G-protein coupled receptor signaling pathway
apoptotic process
positive regulation of transcription by RNA polymerase II
negative regulation of gene expression
mitochondrial DNA repair
HDAC2
positive regulation of proteolysis
cellular response to calcium ion
positive regulation of single strand break repair
PARP1
cellular response to DNA damage stimulus
positive regulation of receptor biosynthetic process
proteolysis
positive regulation of tumor necrosis factor-mediated signaling pathway
positive regulation of neuron differentiation
ATP generation from poly-ADP-D-ribose
protein ADP-ribosylation
ADORA2A
negative regulation of NIK/NF-kappaB signaling
negative regulation of DNA binding transcription factor activity
ADORA1
negative regulation of ATP biosynthetic process
peptidyl-serine ADP-ribosylation
cAMP-mediated signaling
protein polyglycylation
PDE4C
negative regulation of DNA binding
CPNE1
protein poly-ADP-ribosylation
mitochondrion organization
TTLL3
membrane protein proteolysis
positive regulation of protein kinase B signaling
Theophylline
ADORA2B
negative regulation of NF-kappaB transcription factor activity
membrane protein proteolysis involved in retrograde protein transport, ER to cytosol
regulation of I-kappaB kinase/NF-kappaB signaling
PDE5A
HM13
NOMO1
cGMP catabolic process
ADORA3
RIC3
PDE3B
negative regulation of lipid catabolic process
neuron projection extension
PDE3A
negative regulation of cell proliferation
PDE4A
cGMP-mediated signaling
positive regulation of interferon-gamma production
PDE4B
cAMP catabolic process
negative regulation of angiogenesis
negative regulation of cell migration
negative regulation of vascular permeability
PDE4D
negative regulation of cell adhesion
T cell receptor signaling pathway
positive regulation of vascular permeability
positive regulation of interleukin-2 production
positive regulation of interleukin-5 production
cellular response to cGMP

Figure 5.12: The target maps for theophylline and formoterol. The color scheme assigns orange, green, and cyan to represent drugs, proteins, and biological functions, respectively.

국문초록

약물 병용 요법은 의료 분야에서 다양한 질병의 치료에 중요한 발전을 가져왔다. 그러나 가능한 약물 조합의 수가 매우 많기 때문에 효과적인 약물 조합을 찾는 것은 여전히 주요한 과제로 남아있다. 생명 의학 지식 그래프 기반의 방법들은 다양한 질병에 대한 효과적인 조합을 예측하는데 있어 잠재력을 보여주지만, 신뢰할 수 있는 음성 데이터의 부재로 기계학습 모델의 예측 성능이 제한되고 있다. 또한, 기존의 방법들은 지식 그래프에서 얻은 그대로의 약물 임베딩 벡터를 사용하고 있고, 이는 충분한 성능 개선 여지를 남겨두고 있다. 이 문제를 해결하기 위해, 기존 약물-약물 상호작용 데이터를 신뢰할 수 있는 음성 데이터셋으로 활용하고, 지도 대조 학습을 사용해 약물 임베딩 벡터를 약물 조합 예측에 더 적합하게 변환하는 새로운 프레임워크를 제안한다. 약물-약물 상호작용 데이터와 지도 대조 학습 기법은 성능 향상에 도움이 되었을 뿐만 아니라, 약물 조합을 찾는데 적합한 결정 경계를 구축하는데 도움을 주었다. 이 접근 방식의 구체적인 효과를 증명하기 위해, 랜덤 워크와 그래프 신경망을 포함한 다양한 네트워크 임베딩 알고리즘을 생명 의학 지식 그래프에 사용하여 광범위한 실험을 수행했다. 또한, 임베딩 공간 시각화를 통해 해당 접근 방식의 효과를 추가적으로 입증하고 결정 경계를 시극화하였다. 마지막으로, 약물 조합과 약물-약물 상호작용의 실제 사례 연구를 제공하였다. 정리하면, 이 연구는 약물-약물 상호작용 데이터와 지도 대조 학습을 사용해 효과적인 약물 조합 예측을 위한 보다 엄격한 결정 경계를 찾는데 효과적인 방법을 제안하고 있다. 연구에 활용된 소스코드는 Github (`https://github.com/gujh14/DC_with_DDI_SupCon.git`) 에서 확인할 수 있다.

주요어: 약물 조합, 지식 그래프, 약물-약물 상호작용, 지도 대조 학습, 랜덤 워크, 그래프 신경망

학번: 2021-28284

감사의 글

먼저, 제 석사 학위 논문심사위원을 맡아주신 황대희 교수님, 김선 교수님, 오민식 교수님께 감사드립니다. 바쁘신 와중에도 귀중한 시간을 내어 진심 어린 조언을 해주신 덕분에 제 졸업 논문에서 부족했던 부분들에 대한 고찰과 수정을 할 수 있었습니다.

2020년 여름부터 2023년 여름까지 인턴 및 석사 과정 동안 제가 생물정보학과 컴퓨터과학을 공부할 수 있도록 많은 지원과 응원을 아끼지 않으시고, 본 논문의 전반적인 구성 및 심사에 많은 조언을 해주신 김선 교수님께 다시 한 번 감사의 말씀을 올리고 싶습니다.

그리고 본 논문의 아이디어 구성 및 랜덤워크, 그래프 신경망 알고리즘을 구현하고 실행하는데 많은 도움을 주었던 방동민 연구원과 아이디어 및 데이터셋 구성에 도움을 주었던 이정섭 연구원, 또한 지도 대조 학습 아이디어의 구상과 논문의 전체적인 구성 및 원고 수정 논의를 도와주셨던 이상선 박사님께도 감사의 말씀을 드립니다. 또한 연구실 생활 동안 동고동락했던 연구실 동료분들께도 감사합니다. 연구실에서의 경험이 제 커리어 전환과 미래에 정말 큰 도움이 되었습니다. 사회에 나가서도 대학원 생활을 밑거름 삼아 발전된 모습으로 살아가겠습니다.

마지막으로, 제 학위 과정 중 물심양면으로 지원해주셨던 부모님과 늘 제 곁에서 희망을 주었던 친구들에게도 진심으로 감사의 말씀을 표합니다.