UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Dehua Chen

Modelagem de Efeitos Farmacológicos com Representações Não Supervisionadas em Grafos Multi-Relacionais

Belo Horizonte
2020

Dehua Chen

Modelagem de Efeitos Farmacológicos com Representações Não Supervisionadas em Grafos Multi-Relacionais

Versão final

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Adriano Veloso
Coorientador: Nivio Ziviani

Belo Horizonte
2020

Dehua Chen

Modeling Pharmacological Effects with Multi-Relation Unsupervised Graph Embedding

Final version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Adriano Veloso
Co-Advisor: Nivio Ziviani

Belo Horizonte
2020

Chen, Dehua

C518m Modeling Pharmacological Effects with Multi-Relation Unsupervised Graph Embedding / Dehua Chen. — Belo Horizonte, 2020.
xiii, 46 f.

Orientador: Adriano Veloso
Coorientador: Nivio Ziviani
Dissertação (mestrado) — Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação
Referências: f. 40-46.;

1. Computação – Teses. 2. Aprendizado do computador – Teses. 3. Aprendizado de Representações – Teses. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*82(043)

Ficha catalográfica elaborada pela bibliotecária Irenquer Vismeg Lucas Cruz CRB 6ª Região nº 819

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Modeling Pharmacological Effects with Multi-Relation Unsupervised Graph Embedding

DEHUA CHEN

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ADRIANO ALONSO VELOSO - Orientador
Departamento de Ciência da Computação - UFMG

Nivio Ziviani

PROF. NIVIO ZIVIANI - Coorientador
Departamento de Ciência da Computação - UFMG

PROFA. RAQUEL CARDOSO DE MELO MINARDI
Departamento de Ciência da Computação - UFMG

PROFA. DEBORAH SCHECHTMAN
Instituto de Química - USP

Belo Horizonte, 30 de Março de 2020.

To my family and friends, who have always supported me.

Acknowledgments

I would first like to thank my thesis advisor prof. Adriano Veloso and co-advisor prof. Nivio Ziviani of the Graduate Program in Computer Science of Federal University of Minas Gerais. Without their guidance and inspiration, it could not be possible this great accomplishment. They consistently allowed this paper to be my own work, but steered me in the right direction whenever they thought I needed it.

Finally, I must express my very profound gratitude to my family and to my friends for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Resumo

O reposicionamento de medicamentos (também conhecido como reaproveitamento) pode ser definido como a renovação de medicamentos não aprovados (com uso seguro comprovado, mas não demonstrou eficácia na indicação primária) e a expansão de uso dos medicamentos aprovados, desenvolvendo novos usos terapêuticos, que estão além dos seus usos originais inicialmente aprovados. Os medicamentos reposicionados representam aproximadamente 30% dos medicamentos aprovados pela Food and Drug Administration (FDA) dos EUA nos últimos anos. Um fármaco reposicionado usa compostos de menor risco, podendo ir diretamente para testes pré-clínicos e ensaios clínicos, fornecendo assim alternativas mais baratas comparando ao pipeline caro do desenvolvimento de novos fármacos.

Um efeito farmacológico de um medicamento nas células, órgãos e sistemas refere-se à interação bioquímica específica produzida por um medicamento, também chamado como mecanismo de ação. Existem várias abordagens para a identificação de novas oportunidades de reposicionamento, como correspondência de assinatura, docagem molecular (acoplamento molecular, or ancoragem molecular) e associação genética na literatura. Neste trabalho, apresentamos um novo método baseado em um modelo de representações não supervisionadas de grafos multi-relacionais que aprende representações latentes de medicamentos (mecanismo de ação) e doenças, de modo que a distância entre essas representações revele oportunidades de reposicionamento. Uma vez obtidas representações de medicamentos e doenças, aprendemos a predizer a probabilidade de novas indicações entre medicamentos e doenças. As indicações conhecidas de medicamentos são usadas para aprender um modelo que prediz potenciais novas indicações de medicamentos. Comparado com os métodos existentes de representações não supervisionadas de grafos, nosso método mostra desempenho superior em termos de área abaixo da curva ROC (*area under the ROC curve*). Também apresentamos exemplos de oportunidades de reposicionamento encontradas na literatura biomédica recente que também foram previstas pelo nosso método.

Palavras-chave: Aprendizado por Representação, Reaproveitamento de Medicamento, Reposicionamento de Medicamento.

Abstract

Drug repositioning (aka repurposing) can be defined as renewing failed drugs (proved safety but failed to show efficacy for their primary indication) and expanding successful ones by developing new therapeutic uses that are beyond their original uses or initial approved indications. Repositioned drugs account for approximately 30% of the US Food and Drug Administration (FDA) approved drugs in recent years. A repositioned drug uses de-risked compounds, going directly to preclinical testing and clinical trials, thus providing inexpensive alternatives to the costly pipeline associated with the development of new drugs.

A pharmacological effect of a drug on cells, organs and systems refers to the specific biochemical interaction produced by a drug substance, which is called its mechanism of action. There are several approaches for novel repositioning opportunities identification, such as signature matching, molecular docking and genetic association in literature. In this work, we present a novel method based on a multi-relation unsupervised graph embedding model that learns latent representations for drugs (mechanisms of action) and diseases so that the distance between these representations reveals repositioning opportunities. Once representations for drugs and diseases are obtained we learn the likelihood of new links (that is, new indications) between drugs and diseases. Known drug indications are used for learning a model that predicts potential indications. Compared with existing unsupervised graph embedding methods our method shows superior prediction performance in terms of area under the ROC curve, and we present examples of repositioning opportunities found on recent biomedical literature that were also predicted by our method.

Palavras-chave: Representation Learning, Drug Repurposing, Drug Repositioning.

List of Figures

List of Tables

Contents

Chapter 1

Introdution

The global pharmaceutical industry is facing several challenges, such as the escalating cost and length of time required for new drug development. A study of Berger et al. (2014) has showed that for every dollar spent on drug research and development, less than a dollar of value is returned on average. Furthermore, it takes 10 to 17 years to bring new drugs to the market, from the drug discover to the marketed drug (Ashburn and Thor, 2004). Accordingly, biopharmaceutical companies investigate and attempt to increase productivity through other strategies, in particular, drug repositioning.

The process of renewing failed drugs (proved safety but failed to show efficacy for their primary indication) and expanding successful ones by developing new therapeutic uses that are beyond their original uses or initial approved indications is also known as drug repositioning (repurposing, redirecting or reprofiling). Repositioned drugs account for approximately 30% of the US Food and Drug Administration (FDA) approved drugs in recent years (Jin and Wong, 2014). A repositioned drug uses derisked compounds, going directly to preclinical testing and clinical trials, thus providing inexpensive alternatives to the costly pipeline associated with the development of new drugs. One of the well-known examples is sildenafil citrate (brand name: Viagra), which was repositioned from a common hypertension drug to a therapy for erectile dysfunction (Renaud and Xuereb, 2002).

Figure 1.1 illustrates the biochemical interaction that gives rise to the pharmacological effect of a drug. This paper is motivated by the problem of finding drug repositioning opportunities by modeling the mechanisms of action of drugs (Iorio et al., 2010a). For instance, different biological solutions might be considered in order to chemically decrease the blood pressure such as removing the excess of salt from the body, thereby decreasing the tension in the vessels, or inhibiting the vasoconstrictive signalling of a hormone, or acting directly on the cells physically narrowing the vessels and preventing their unwanted action this way (Ong et al., 2007). Each of the aforementioned solutions requires a different mechanism of action. The same drug can have several mechanisms of action and therefore it can potentially play a multitude of roles by interacting with proteins involved in various biological processes, which are accountable for the drug polypharmacology (Car, 2012). Thus, drug repositioning is a

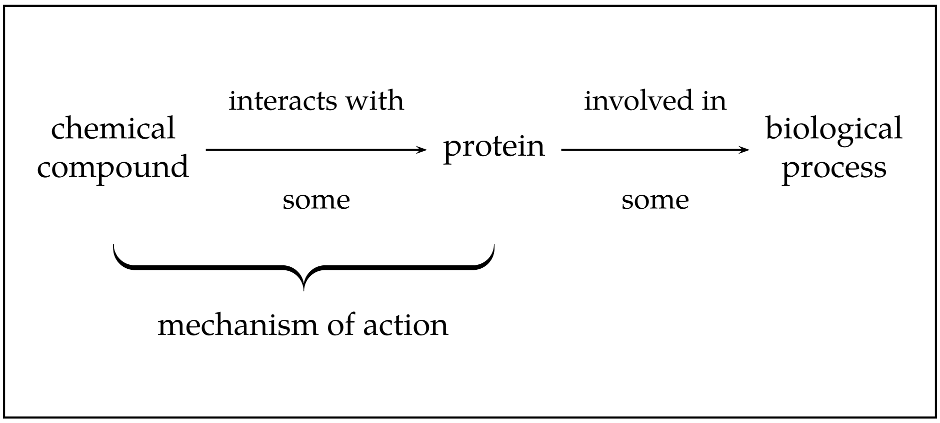

Figure 1.1. A chemical is assigned as a treatment to a disease because it exhibits a particular mechanism of action that affects biological processes associated with the disease. Repositioning opportunities exist because the same drug interacts with multiple proteins themselves involved in multiple biological processes.

direct application of drug polypharmacology (Zhang et al., 2016).

1.1 Thesis Statement

Creating new drugs is an expensive procedure, instead, repositioning an existent drug is much cheaper. In this master thesis, we introduce a novel approach for identifying drug repositioning opportunities based on unsupervised graph embedding. The main idea of our method is embedding the mechanism-of-action of drugs into vector representations, afterwards, train a classifier to predict drug repositioning candidates. Furthermore, the multi-relation graph (drug-disease interactions and drug-protein interactions) is a three-layered graph, thereby, we devised an efficient unsupervised graph embedding to exploit this specific structure. Finally, we show the efficiency of our algorithm and also provide several improvements to achieve a better performance.

1.2 Our Solution

Our main goal is to discover new relations between current drugs and diseases by utilising existing public drug-disease-protein interactions. The main three steps of our proposed method are as follows.

In the first step, we built a large and heterogeneous graph comprising drug, disease, and protein entities that are linked according to information collected from the biomedical literature, as shown in Figure 1.2. Specifically, we formulate the drug repositioning problem as a three layer multi-relation directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{R}, E)$, where $\mathcal{V}$ is the set of entities (i.e., drugs, diseases and proteins), $\mathcal{R}$ is a set of relations (i.e., drug-protein, drug-disease and protein-protein), and E is a set of edges connecting different entities in $\mathcal{V}$. In the graph, mechanisms of action are represented by relations involving drugs and proteins and repositioning opportunities are represented by (hidden) relations involving drugs and diseases. The graph also contains protein-protein interactions in order to increase connectivity and information propagation while learning node representations. The datasets used to build the graph are described in Section 3.1.

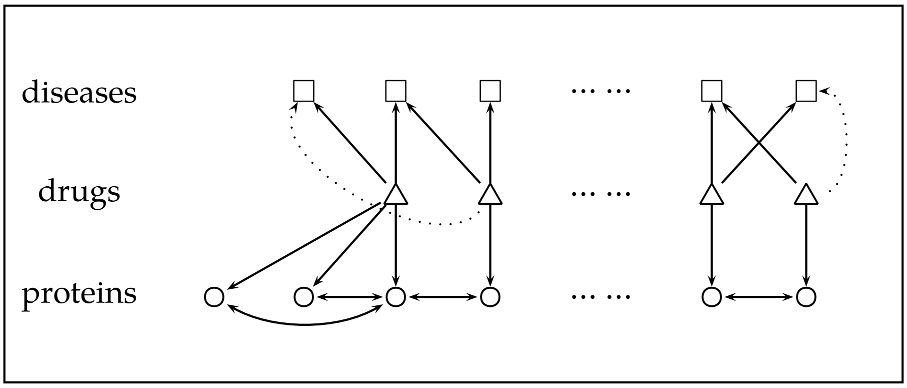

Figure 1.2. Multi-relation graph, composed of drug-protein, drug-disease and protein-protein interactions. A drug (△) interacts with some proteins (○) and this drug is indicated to certain diseases (□). A single drug may interacts with different proteins, and these proteins may also interact. Further, the same drug may be indicated to different diseases. Links may provide evidence for repositioning opportunities (i.e., dotted links).

In the second step, our goal is to find a low-dimensional latent representation for drugs and diseases, so that the latent representation embeds the relationship between mechanisms of action and drug indications. Drug-protein and drug-disease interaction graphs usually exhibit a particular structure with many isolated sub-graphs, and often a protein is linked to drugs residing in different parts of a graph. We employ a SkipGram based algorithm (Mikolov et al., 2013) to learn node representations in an unsupervised way, but instead of performing deep random walks to produce contexts (Grover and Leskovec, 2016; Perozzi et al., 2014), we employ a restricted number of permutations over the immediate neighborhood of a node as context to generate its representation (Pimentel et al., 2018). This choice is motivated by the particularly sparse structure of drug-disease and drug-protein interaction graphs. Further, we exploit the multi-relation nature of the graph by employing two types of contexts while learning node representations: contexts composed of drugs and proteins (i.e., mechanisms of action) and contexts composed of drugs and diseases (i.e., drug indications). This results in an embedding for each drug and for each disease, so that adjacent entities are placed close to each other in the vector space, while unconnected entities are pushed apart. As a result, drugs and diseases that have a similar distribution of neighbors will end up being nearby in the vector space. Details for obtaining node embeddings can be seen in Section 3.2.

In the third step, we learn the likelihood of new links between drugs and diseases, as representations for drugs and diseases were obtained in previous step. Known drug indications are used for learning a parametric model which predicts other likely indications. Our evaluation follows the typical cross-validation framework, in which a subset of the known drug uses are hidden. Details of the experimental setup can be seen in Section 4.1 and the results obtained by different embedding algorithms we used can be seen in Section 4.2.

1.3 Contributions

In the following we briefly summarize our contributions:

- We employ interaction graphs involving drugs, diseases and proteins in order to learn suitable vector representations for drugs and diseases. The input graph presents particular characteristics, such as high sparsity and low connectivity, so that contextual information based on the immediate neighborhood is likely to produce better representations than typical random walk approaches.

- Given the vector representations for drugs and diseases, we build a parametric model learned to identify if a specific drug is indicated to a specific disease.

- We evaluate the effectiveness of our model on predicting repositioning opportunities under a cross-validation framework. Our model reaches an area under the curve of +0.98, being significantly superior than predictive models built using contextual information produced by deep random walks.

- Finally, we compared our specific findings with repositioning opportunities reported in the recent biomedical literature. We present some interesting cases that were predicted by our model, including the use of Amitriptyline for relief of Fibromyalgia in adults. Amitriptyline is an antidepressant that is now being reported in the literature to treat Fibromyalgia if used at doses below those at which the drugs act as antidepressants.

1.4 Organization

The rest part of this thesis is organized as follow: in Chapter 2, we review principal related works on drug repositioning, graph embeddings approaches and graph embeddings applied in bioinformatics, such as drug repositioning and polypharmacy side effects modelling; In Chapter 3, we present our collected and compiled data, which is used in this work, and our graph embedding algorithm; In chapter 4, we describe the experimental setup and then present the results; In chapter 5, we provide several further analysis; Finally, in Chapter , we present the concluding remarks and future work.

Chapter 2

Related Work

2.1 Drug Repositioning

Drug repositioning (aka drug repurposing, reprofiling, or re-tasking) is a process of identifying new uses outside the scope of the original medical indication for existing drugs (Ashburn and Thor, 2004; Pushpakom et al., 2019). This strategy can offer a better risk-versus-reward trade-off as compared with other drug development strategies, such as drug discovery, due to by using existing drugs which are previously tested. Furthermore, repositioning drugs have a reduced time, 3 to 12 years to be available in the market, while new drugs development process can take 10 to 17 years (Ashburn and Thor, 2004).

Several strategies have been proposed for drug repositioning, such as chemical structure based, transcriptional signatures based, molecular ligand based and network mapping. Structured based methods are based on the idea that similar proteins have similar functionality, through a similarity comparison it is possible to find secondary targets of an already existing drugs (Ehrt et al., 2016). Molecular transcriptional signatures can be compared to create relations between drugs and new disease indications. These relations provide useful information for finding new uses of known drugs (Lamb et al., 2006). Ligand based approaches are based upon the concept that similar compounds tend to have similar biological properties. In drug repositioning, this method has been widely used to analyze and identify the activity of ligands for new disease indications (Liu et al., 2010). However, drug repositioning approaches can be divided into three categories, computational, experimental and mixed. Generally, computational approaches are largely data-driven, based on systematic analysis on data such as gene expression, chemical structure, genotype or proteomic data and experimental methods are based on chemical/biological experiments and analysis. While mixed approaches combine both. In this work, we focus mainly on computational approaches. In following subsections, we present main works on computational approaches in literature.

2.1.1 Signature Matching

Signature matching is a technique based on the comparison of the characteristics (or signature) of a drug against another drug, disease or clinical phenotype (Hieronymus et al., 2006; Keiser et al., 2009; Pushpakom et al., 2019). The signature of a drug can be derived from three general types of data: transcriptomic (RNA), proteomic or metabolomic data; chemical structures; or adverse event profiles.

Transcriptomic signatures. This type of signature can be obtained by observing the effect of a drug or a disease on the gene expression profile of biological material, such as a cell or a tissue. The effect (or the change) on the gene expression profile is considered as the drug's signature or the disease's signature. A example usage of these signatures is, if there is a pair of a drug and a disease shares a negative correlation, which means the drug's signature is opposite of the disease's signature, following the signature reversion principle (SRP), where it is assumed that if a drug can reverse the expression pattern of a given disease phenotype, then that drug might be able to revert the disease phenotype itself (Pushpakom et al., 2019). This method has been studied and demonstrated a promising results in the works Wagner et al. (2015); Hsieh et al. (2016). Another usage of these signatures is comparing dissimilar drugs' signatures. Two dissimilar drugs are initially designed for different sets of diseases or clinical applications, a shared transcriptomic signatures may imply that they are repositioning candidates for each other, according to the principle of guilt by association (Chiang and Butte, 2009). These works Iorio et al. (2010b,a) have demonstrated that this method is also promising in finding repositioning opportunities.

Chemical signatures. Chemical signatures are based on chemical structures of drugs and their relationship to biological activity. With the chemical signatures of drugs, comparing them to see whether there are chemical similarities and then suggest repositioning opportunities (Pushpakom et al., 2019). In order to obtain the chemical signature, a drug's chemical structure is analyzed and a set of structural features is selected as their signature. However, this is a complicated process due to the importance and the sensibility of selected features on repositioning candidate finding.

Adverse effect signatures. Since every drug present a relatively unique adverse effect profile, it raises a hypothesis that two drugs with the same adverse effects may be effective on a same target or on the same biological activity (Dudley et al., 2011; Pushpakom et al., 2019). Based on this hypothesis, it emerges adverse effect signature matching. Some works Campillos et al. (2008); Yang and Agarwal (2011) have shown the effectiveness of this method. However, there are still some problems for this approach: mining large adverse effect information of drugs, lack of well-defined adverse effect profile and causality assessment of adverse effects and drugs(Dudley et al., 2011).

2.1.2 Molecular Docking

Molecular docking is a structure-based computational method, it aims to predict binding site complementarity between the ligand, such as a drug, and the target, a receptor (Kitchen et al., 2004). This approach also depends on the prior knowledge of receptors involved in diseases. Once we have the knowledge of a disease and its corresponding receptor, it is possible to exploit docking algorithms to perform molecular fit computations to find feasible drugs. If there is a novel interaction between a drug and a disease-related receptor, it can be taken forward repositioning.

2.1.3 Genetic Association

Recent advances in genotyping technology and reduction of the cost of genotyping have contributed to the growth of genome-wide association studies (GWAS) over the past 10 years (Pushpakom et al., 2019). The goal of GWAS is to identify genetic variants associated with diseases and thereby provide insights in the biology of diseases (Pushpakom et al., 2019). This information may also help to identify repositioning opportunities, if the genotypes of two diseases are shared and one is treated by a specific drug, then this drug may also be effective with another disease. However, there still remains problems to be solved before taking fully advantage of GWAS: identification of causal gene and/or gene variants is still a complex work (Sanseau et al., 2012); another issue is that there is still unknown the right direction of the effect of the gene variant whether it has to be activator or a suppressor to control the disease (Sanseau et al., 2012).

2.1.4 Retrospective Clinical Analysis

Retrospective clinical analysis are another useful resource for drug repositioning. A systematic analysis on clinical data is increasingly suggested for identifying drug repositioning opportunities (Jensen et al., 2012). Electronic health records (EHRs) contain an enormous amount of data on patient outcomes, such as results of laboratory tests, drug prescribing data, clinical descriptions of patient symptoms and signs and imaging data (Pushpakom et al., 2019). These records could be used as a source for identifying signals for drug repositioning (Hurle et al., 2013). Additionally, the enormous amount of EHR data also provides high statistical power (Paik et al., 2015). The work Paik et al. (2015) has analyzed over 13 years of EHRs from a tertiary hospital and extracted clinical signatures, they were able to identify over $17,000$ known drug–disease associations and identified terbutaline sulfate, an anti-asthmatic, as a promising candidate for the treatment of amyotrophic lateral sclerosis (ALS).

2.2 Graph Embedding

Uncountable real-world problems can be solved using graph algorithms due to their graph-structured nature. This has made graph an important data representation structure. However, the increasing volume of available information has made graph processing a hard and costly task. Many studies have been conducted to find efficient ways to deal with large graphs. Thus, graph embeddings methods have emerged and they aim to obtain a low-dimensional vector representation of graph, while preserving their properties, and thereby solve problems with these representations.

Graph embedding algorithms can be divided into 3 categories: node embedding, edge embedding and whole-graph embedding (Cai et al., 2018). Node embedding refers to the process of mapping the nodes of a graph into low dimensional vector space, each node is represented by a unique vector. Edge embedding, different from node embedding, it aims to embed graph's edges into vector space and therefore, instead of nodes, each edge is represented by a vector. Meanwhile, whole-graph embedding intends to embed a whole graph into a single vector. In this work, we only focus on node embedding.

Historically, node embedding has been widely studied and it shows being effective in many real-world applications (Cai et al., 2018; Goyal and Ferrara, 2018), such as friendship or content recommendation in social networks (social graphs) (Liben-Nowell and Kleinberg, 2007), Protein-Protein interaction networks analysis (biology graphs) (Theocharidis et al., 2009). Nonetheless, there are three principal node embeddings strategies: factorization methods, random walk techniques and deep learning methods. Each of these strategies will be presented in following subsections.

2.2.1 Factorization Methods

Factorization (or matrix factorization) based node embedding represents graph's nodes in the form of a matrix and then factorize this matrix to obtain node representation (Cai et al., 2018). There are several types of matrices that can be used to represent graphs: adjacency matrix, Laplacian matrix, node transition probability matrix, and Katz similarity matrix, and so on. Matrix factorization methods are different depending on the representation matrix used. Nevertheless, there are two main factorization approaches: eigenvalue decomposition and gradient descent, the former can be used only when the matrix is positive semidefinite and the latter can be applied to both structured and unstructured matrices. Some representative works in factorization based node embedding are Locally Linear Embedding (Roweis and Saul, 2000), Laplacian Eigenmaps (Belkin and Niyogi, 2002), Cauchy Graph Embedding (Luo et al., 2011) and Structure Preserving Embedding (Shaw and Jebara, 2009).

2.2.2 Random Walk Based Methods

Random walk based embedding methods adopt a neural language model (Skip-Gram (Mikolov et al., 2013)) for embedding, the main difference of these is usually in the way the algorithm use to generate random walks. The idea behind of random walk based embedding is that given a node's embedding, the probability of observing its neighbourhood should be maximized, preserving second-order proximity between the nodes. Intuitively, the second-order proximity compares the similarity of the nodes' neighbourhood structures. The more similar two nodes' neighbourhoods are, the larger the second-order proximity value between them (Cai et al., 2018). Two more representative random walk based embeddings are DeepWalk (Perozzi et al., 2014) and Node2Vec (Grover and Leskovec, 2016), the former uses a normal random walk algorithm and the latter uses a biased random walk algorithm with two user-defined parameters, which takes breadth and depth of the walk into account.

2.2.3 Deep Learning Methods

With the fast growth of research in deep learning, several deep neural networks have been applied in graph embedding. One example of this is structural deep network embedding (SDNE) (Wang et al., 2016), it combines deep autoencoders (Bengio et al., 2013) and Laplacian Eigenmaps (Belkin and Niyogi, 2002), the autoencoder aims at finding an embedding for a node which can reconstruct its neighborhood and the Laplacian Eigenmaps is applied to ensure that similar nodes are mapped close from each other in the embedding space. Graph convolutional networks (GCN) (Kipf and Welling, 2016), different from SDNE that takes the whole graph (adjacency matrix) as input, is optimized for large graphs through convolution operations. GCN iteratively aggregates the embeddings of neighbors for a node and uses a function of the obtained embedding and its embedding at previous iteration to obtain the new embedding. Aggregating embedding of only local neighborhood makes it scalable and multiple iterations allows the learned embedding of a node to characterize global neighborhood (Goyal and Ferrara, 2018).

2.3 Graph Embedding in Bioinformatics

In recent years, machine learning has been vastly employed in drug repositioning. Kumar et al. (2019) used a fully connected deep neural networks for training the model using transcriptional data at gene level to predict drug therapeutics and to use them in drug repositioning. They analyzed the confusion matrix and found out that the miss-classified cases can indeed be considered as an indication of their potential in novel uses. Donner et al. (2018) has proposed ligand based approach based on the learning of embeddings of gene expression profiles using deep neural networks and considered it as a measure of compound functional similarity for drug repositioning. Hu and Agarwal (2009) created a disease-drug network using publicly available gene expression. By defining a new network component called cancer-signaling bridge, Jin et al. (2012) presented a new computational method for off-target drug repositioning.

Graphs are the typical structures used to model the relation between drugs and diseases. The major challenge is to find a way to incorporate complex structures like graphs into the existing machine learning algorithms. Therefore, we can take advantages of graph embedding algorithms (described in Section 2.2), which are able to transform graphs into low dimensional vector without losing their graph properties. In order to model the polypharmacy side effects, Zitnik et al. (2018) trained convolutional neural networks on a graph, with proteins and drugs as its nodes and drug-protein and drug-drug interaction as its edges. Deepika and Geetha (2018) used node2vec (Grover and Leskovec, 2016) representation along with bagging Support Vector Machine (SVM) to predict drug-drug interactions. Gao et al. (2018) applied Long Short-term Memory Neural Networks (LSTM) and graph-based convolutional neural network to obtain a low dimensional representation of protein and drug structures. These representations were then engaged in the prediction of drug-target interaction. Cheng et al. (2012) predicted new drug candidates using a network obtained from DrugBank (Wishart et al., 2008). Wang et al. (2014) applied an information-flow approach on a heterogeneous network of drug-drug and disease-disease similarities along with the known disease-drug relations. The algorithm updates the disease-drug relations through several iterations and finally converges to stationary scores in predicting the network connections. Yamanishi et al. (2008) introduced a bipartite graph-learning method based on kernel regression in order to learn a co-mapping of drugs and proteins into a common pharmacological space. In the pharmacological space, the correlation between compound-protein pairs can be conveniently calculated to predict their interactions for drug re-positioning. Zheng et al. (2013) proposed a method to factorize the existing drug-target relations so as to predict the new relations constrained by the drug-drug and disease-disease similarity networks. Additionally, Xia et al. (2010) proposed a manifold regularization semi-supervised learning method in which two classifiers in drug and disease space are learned and then combined together to give a final score for drug-disease interaction prediction.

Recent neural representation learning methods include neural fingerprints (Duvenaud et al., 2015), graph convolutional networks (Hamilton et al., 2017), message passing networks (Gilmer et al., 2017) etc.) are a related line of research. However, these graph embedding methods do not apply in our setting, since they solve a (supervised) graph classification task and/or embed entire graphs while we embed individual nodes.

Chapter 3

Learning Mechanisms of Action

In this Chapter, we present our novel drug repositioning method based on modelling mechanism of action of drugs. At first, we present the data that we collected. Afterwards, the unsupervised graph embedding algorithm is presented.

3.1 Data

In this section, we discuss the datasets used to build the graph presented in Figure 1.2. As in Zitnik et al. (2018), we used the human protein-protein interaction (PPI) network compiled by Menche et al. (2015); Chatr-aryamontri et al. (2015), integrated with additional PPI information from Szklarczyk et al. (2017). The PPI graph contains physical interactions experimentally documented in humans, such as metabolic enzyme-coupled interactions and signaling interactions. The network is unweighted and undirected with 19,085 proteins and 719,402 physical interactions. Table 3.1 presents statistics about the data from which we built two graphs:

1. For the graph drug-protein, we obtained relationships between drugs and proteins from the STITCH database (Chatr-aryamontri et al., 2015). This database integrates various chemical and protein networks and there were over 8,083,600 interactions present between 8,934 proteins and 519,022 chemicals. We considered only the interactions between chemicals (i.e., drugs) and proteins that had been experimentally verified, which comprises 16,546 proteins and 584 drugs, and there are 1,824,204 interactions amongst them.

2. Drugbank (Wishart et al., 2008) was used to retrieve known drug-disease links. DrugBank is a bioinformatics and cheminformatics resource that provides a knowledge-base for drugs, drug actions and drug targets. We focused on 600 drugs that were indicated to 508 diseases, resulting in a total of 2,836 drug-disease links.

Table 3.1. Basic statistics of the data.

drug-protein		drug-disease known indications	
# of drugs	584	# of drugs	600
# of proteins	16,546	# of diseases	508
# of interactions	1,824,204	# of interactions	2,836

To represent the proteins, we used Ensembl protein ID (e.g. ensp00000200652), and for the drugs, we used a unified identification (e.g. cids00441300). Finally, for the diseases, we used their name (e.g. fibromyalgia) in the Unified Medical Language System. Since we unified the identification for proteins, drugs and diseases, we were able to join the two different graphs together to build a large and heterogeneous graph.

3.2 Unsupervised Node Embedding

In this section we aim to learn representations for drugs and diseases that best preserve the original graph structure, generalizing mechanisms of action in order to find novel uses and repositioning opportunities. Graph embedding consists in finding a continuous vector space representation for entities in the set of nodes $\mathcal{V}$. The task is to learn a dictionary $Z \in \mathbb{R}^{|\mathcal{V}| \times d}$, with one $d-$dimensional embedding for each node in $\mathcal{V}$. In other words, graph embeddings are the transformation of a graph to a set of vectors, by capturing the graph structure as well as node-to-node relationship. Unsupervised learning of graph embeddings has benefited from the information contained in contexts (Pimentel et al., 2018), and thus embedding methods usually work by simulating contexts and operate in two steps:

1. They sample pair-wise relationships from the graph through random walks. Each random walk generates a sequence of nodes, simulating a context.

2. They train an embedding model, e.g. using Skipgram algorithm (Mikolov et al., 2013), to learn representations that encode pairwise node similarities.

Figure 3.1 illustrates the representation learning process that gives each node a unique embedding in the same vector space. Embedding methods differentiate mainly on the first step, as there are many possible ways to extract context from a graph. The best strategy for producing context depends on specific characteristics of the graph. In this work the contexts are based solely on the first order neighborhoods of nodes,

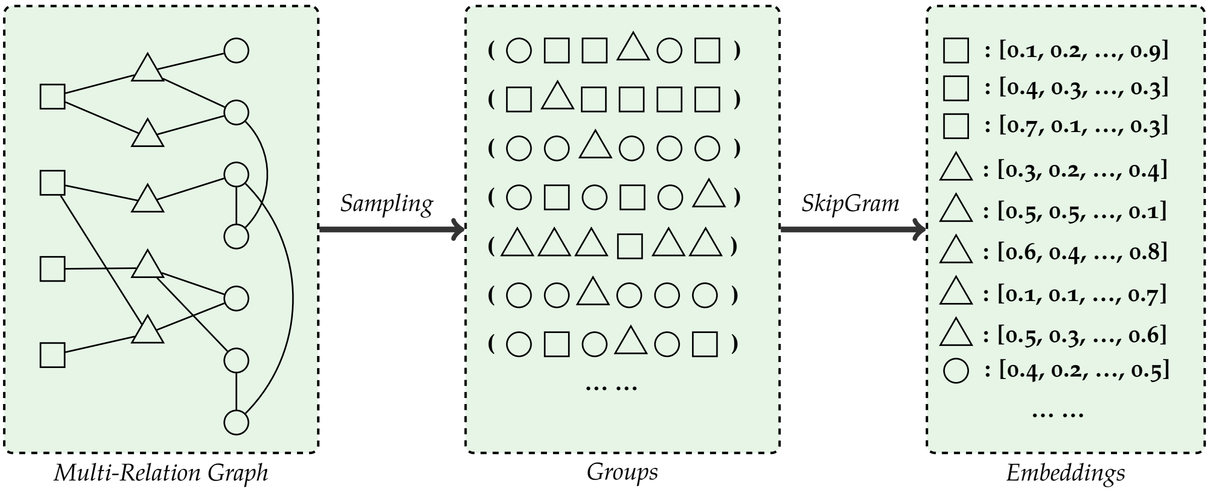

Figure 3.1. Representation learning process, it takes a multi-relation graph as input. △ represent drugs, ○ are proteins and □ are diseases. At the first step, several groups of node are sampled using random permutation. Afterwards, SkipGram is applied to learning embeddings for each node. The group size is controlled by parameter k and the embedding size by d.

defined here as the nodes that are directly connected. Consequently, nodes' representations will be mainly defined by their first order neighborhoods and nodes with similar neighborhoods (contexts) will be associated with similar representations (Pimentel et al., 2019). This results in embeddings focused mainly on the first-order proximity. More specifically, we first separate a node neighborhood in small groups and then we maximize the log likelihood of predicting a node given another in such a group (Pimentel et al., 2017).

3.2.1 Generating Contextual Groups

The first step is to group nodes based on their neighborhoods, so that context can be exploited. There are two main challenges in forming groups from neighborhoods, as follows:

- Nodes have different degrees, so groups containing all the neighbors from a node are difficult to treat.
- There is no explicit order in the nodes in a neighborhood. So there is no clear way to choose the order in which they would appear in a group.

To deal with these challenges, we create small groups with only k neighbors in each, using random permutations of their neighborhoods (Pimentel et al., 2018). The number of permutations n is specified and controls the trade-off between training time and increasing the training dataset. Selecting a higher value for n creates a more uniform distribution on possible neighborhood groups, but also increases training time.

3.2.2 Learning Representations

The first step results in a set of groups S, where each member of S is a node in the graph. Then, we learn vector representations of nodes by maximizing the log likelihood of predicting a node given another node in a group and given a set of representations r, making each node in a group predict all the others. The log likelihood to maximize is given by:

$$\max_{r} \quad \frac{1}{|S|} \sum_{s \in S} \left(\log \left(p \left(s | r \right) \right) \right) \tag{3.1}$$

where $p\left(s|r\right)$ is the probability of each group, given as:

$$\log \left(p \left(s | r \right) \right) = \frac{1}{|s|} \sum_{i \in s} \left(\sum_{j \in s, j \neq i} \left(\log \left(p \left(v_j | v_i, r \right) \right) \right) \right) \tag{3.2}$$

where v_i is a node in the graph and v_j are the other nodes in the same group. The probabilities in this model are learned using the feature vectors r_{v_i}, which are then used as the node representations. The probability $p\left(v_j|v_i, r\right)$ is given by:

$$p \left(v_j | v_i, r \right) = \frac{\exp \left({r'}^{T}_{v_j} \times r_{v_i} \right)}{\sum_{v \in V} \left(\exp \left({r'}^{T}_{v} \times r_{v_i} \right) \right)} \tag{3.3}$$

where ${r'}^{T}_{v_j}$ is the transposed output feature vector of node j, used to make predictions. The representations r'_v and r_v are learned simultaneously by optimizing Equation 3.1.Essentially, by optimizing this log probability the algorithm maximizes the likelihood of predicting a neighbor given a node, creating node embeddings so that nodes with similar neighborhoods have similar representations. Since there is more than one neighbor in each group, this model also makes connected nodes having similar representations, because they will both predict each others neighbors, resulting in representations also with first order similarities. A trade-off between first and second order proximities can be achieved by changing the parameter k, which controls the number of nodes within each group.

Chapter 4

Experiments and Results

4.1 Experimental Setup

Our data is a multi-relation graph compose of drug-protein and drug-disease interactions. Thus, in order to learn our models, it is necessary to select an efficient embedding space in order to better exploit the information within the graph. We discuss the choice for an appropriate embedding space which includes evaluating different graph-embedding algorithms and their corresponding hyper-parameters.

4.1.1 Learning the Embedding Space

We first find an efficient embedding space for the different node embedding algorithms that will be compared in our experiments. This involves 25 hyperparameter combinations that were randomly selected for each algorithm and embedding models are then learned in an unsupervised way. We considered three graph-embedding algorithms in our experiments: (i) DeepWalk (Perozzi et al., 2014), where the best hyperparameters are: window size of 12, number and length of walks were set to 7 and 25, respectively; (ii) Node2Vec (Grover and Leskovec, 2016), with window size, number and length of walks equal to 5, 57 and 73, respectively; (iii) NBNE (Pimentel et al., 2018), with hyperparameters: window size of 6 and number of permutations set to 30.

Table 4.1. Number of positive and negative examples used to learn the embedding space and to train the parametric model.

Steps	Interactions	Examples
Learning Embedding Space	1,827,040	–
Model Evaluation	–	2,836 (+) 30,196 (-)

4.1.2 Model Evaluation

We used the Multilayer Perceptron (MLP) as a binary classifier, which predicts possible links between drugs and diseases using the embedding space. Specifically, the vector of a drug and the vector of a possible indication (i.e., a disease) to the drug are concatenated, and the MLP model takes the final vector as input and makes a prediction (in this case, the output is the probability of a link existing between the drug and the disease). As shown in Table 3.1, the known indications that form the drug-disease interaction graph contains 2,836 links. We used 5-fold cross-validation to assess the embedding's quality. Thus, we divided 2,836 into five folds, each time one of them is used for the validation and the rest for the training. As the known indications data contains only positive examples, we have generated 30,196 negative examples using the complementary graph of the known indications in order to learn the MLP model (de Oliveira Jr. et al., 2014), as shown in Table 4.1. It is worth mentioning that there are more negative occurrences than positive ones in the real-world scenario, because drugs are produced for a small group of diseases or health issues, being ineffective for others. As can be seen in the table, the data is highly imbalanced, thus making our experiment close to a real-world scenario. Finaly, we used *area under the curve* (AUROC score) as the basic measure for assessing the performance of the algorithms (Marczewski et al., 2017).

4.2 Results

In this section we report results obtained by the three embedding algorithms DeepWalk, Node2Vec and NBNE. We also discuss examples of drug repositioning opportunities endorsed by recent biomedical literature.

Prediction Performance: As shown in Figure 4.1, NBNE has obtained the best result. Specifically, NBNE achieved numbers as high as 0.98 in terms of AUROC, while Node2vec and Deepwalk achieved 0.75 and 0.77, respectively. The improvement provided by NBNE compared to Deepwalk and Node2Vec is significant – 27% and 28% of improvement respectively. The main difference of NBNE from the other two algorithms is the context generation approach, as NBNE is based on the neighborhood while the other two algorithms are on random walks, as discussed in Section 3.2.1. It seems that the neighborhood based approach generates more accurate representation in the drug repositioning scenario.

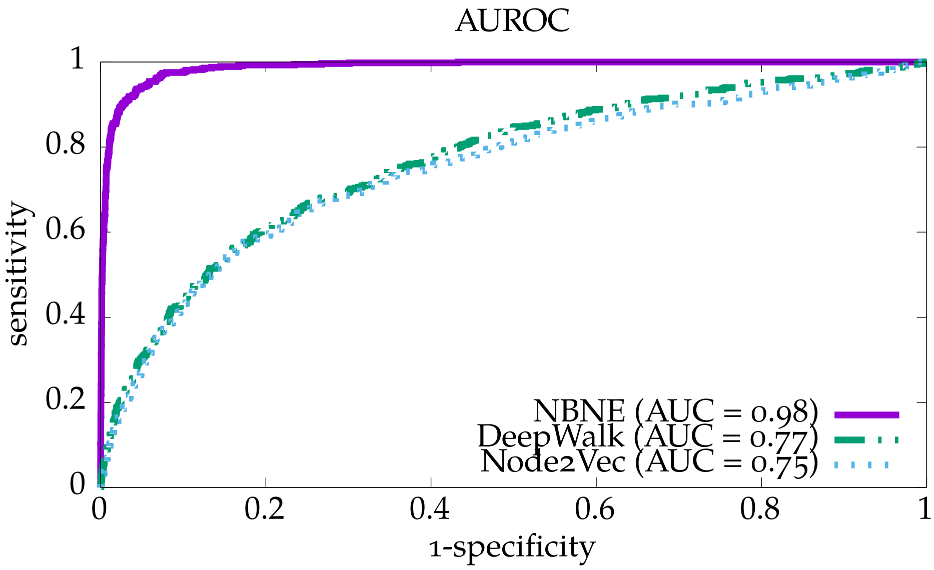

Figure 4.1. AUROC values for Deepwalk, Node2vec and NBNE.

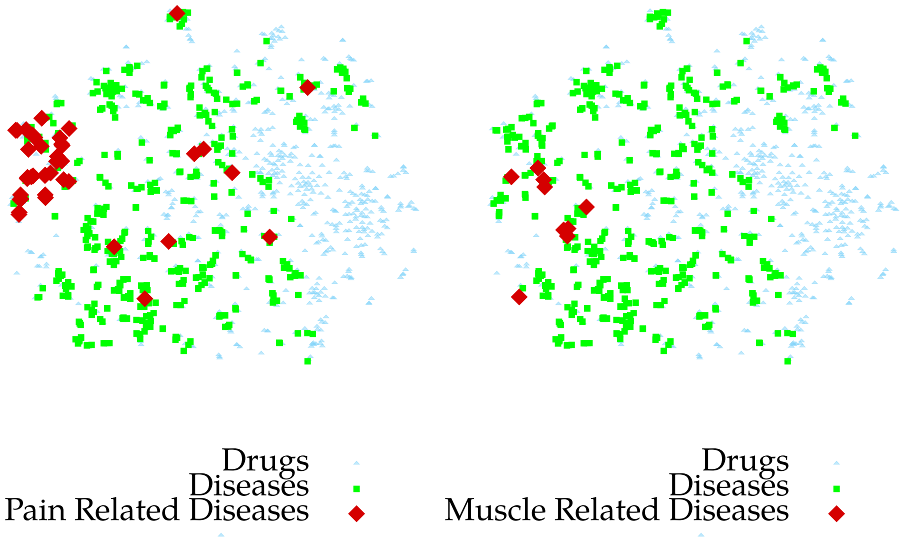

Figure 4.2. Proximity of related diseases. Left – Pain related diseases. Right – Muscle related diseases.

Related Diseases in the Embedding Space: We employ t-SNE in order to visualize

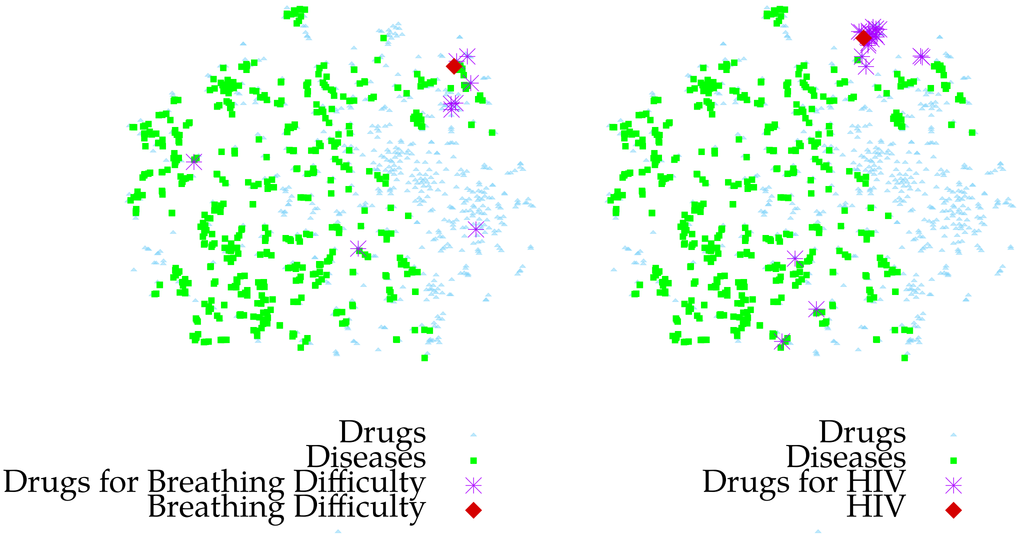

Figure 4.3. The proximity of diseases and their corresponding medications. Left – *Breathing Difficulty*. Right – *HIV*. In general, drug indications and the corresponding health problems are located closely.

the embedding space of drugs and diseases. T-SNE is a technique used to visualize high-dimensional data by giving each data point a location in a two-dimensional map (Maaten and Hinton, 2008). The visualization suggests some insights about the reasons that lead to the good performance of our method. In order to have a clear visualization, the drugs are represented by triangles (blue) and the diseases by rectangles (green). Further, some points are also highlighted in the figures to demonstrate interesting properties of the embedding space. Figure 4.2 shows that similar diseases have close vector representations. Figure 4.2 (Left) shows a cluster of diseases (red points) representing pain related diseases, while Figure 4.2 (Right) shows a cluster of muscle related diseases. These visualizations suggest that our method generates meaningful representations as related diseases are located close to each other in the embedding space.

Diseases and their Corresponding Drugs in the Embedding Space: We have also analysed the spatial relation of diseases and their corresponding drugs. Figure 4.3 (Left) highlights drugs which are used to treat *Breathing Difficulty*. In this case, most of the indicated drugs are concentrated next to the disease. The same trend is observed in Figure 4.3 (Right), where we highlighted the disease *HIV* – again, most of the indicated drugs are placed next to the disease. These visualization give us a great insight of the embedding space generated by NBNE. Furthermore, we observed the proximity of *Breathing Difficulty* to *HIV* in the Embedding Space, and it may suggest a relationship between these two diseases. One possibility is who is infected by HIV may suffer

Table 4.2. List of some possible candidatesfor drug repositioning reported in biomedical literature and found by our algorithm.

Medication	Target disease	Also appeared in
Gabapentin	Bipolar II disorder	Fullerton et al. (2010)
Naproxen	Myofascial Pain	Khalighi et al. (2016)
Amitriptyline	Fibromyalgia	Guymer and Littlejohn (2019)
Amlodipine	High blood pressure	Donato and Brown (2019)
Atorvastatin	High blood pressure	Bubnova et al. (2019)

Breathing Difficulty too.

Repositioning Opportunities and Biomedical Literature: Table 4.2 presents examples of repositioning opportunities. Our prediction model suggests Gabapentin as a candidate for bipolar II disorder, which has been confirmed by Fullerton et al. (2010) and in several other studies. While Naproxen is used in treating balance problems, it can also be used for treating Myofascial Pain (Khalighi et al., 2016), which is confirmed by our model as it places these diseases and medications in the same group. Recent studies show that fibromyalgia is associated with muscle tension and depression (Bosco et al., 2019). Recent research carried out by Guymer and Littlejohn (2019) shows that Amitriptyline, which has been used in the treatment of muscle tension, is a possible candidate for fibromyalgia. Lately, Bubnova et al. (2019) confirmed that both Amlodipine and Atorvastatin caused significant improvement in patients with high blood pressure which is in accordance with our results.

Chapter 5

Further Analysis

In this chapter, we present various performance analysis over distinctive classifiers and ensemble models, and embedding-level improvement by denoising node representation. Those analyses have revealed some interesting facts, such as embeddings generated by NBNE are more complex and needed a more powerful classifier to fully exploit its hidden information, and MLP shows an outstanding generalization power.

5.1 Classifier Analysis

In the previous experiment, we have used only one classifier, MLP. Thus, we are curious about how our algorithm will perform with other classifiers. To achieve this goal, we run the previous experiment with several different classifiers: 1. Gaussian Naive Bayes (GaussianNB), 2. Quadratic Discriminant Analysis (QDA), 3. AdaBoost, 4. Random Forest, 5. XGBoost. In case of AdaBoost, Random Forest and XGBoost, two different size of ensembles are created, one with 50 weak classifiers and another, 100.

The experiment's result is shown in Table 5.1. We found that the embedding generated by Node2Vec and by DeepWalk with the classifier GaussianNB leads to the worst performance, but in case of NBNE, the same classifier achieves a better result. Additionally, Node2Vec and DeepWalk generally have close performances with different classifiers. However, the best overall result is NBNE with MLP, which is used in the earlier experiment. The other combinations that achieve a good result (+0.85) are QDA+NBNE, RandomForest+NBNE, XGBoost+NBNE. The Random Forest with 100 Decision Trees has achieved the best result overall tested ensemble models.

We found that the combination with NBNE usually achieves a better result; this may imply that the embedding generated by NBNE is more suitable for our work context. Moreover, the embedding made by NBNE may present a more complex structure, because only more robust classifiers produce a better result with it. Evidence to explain this phenomenon is that QDA, RandomForest, XGBoost and MLP are more complex

	NBNE	DeepWalk	Node2Vec
GaussianNB	0.8106	0.7327	0.7373
QDA	0.8859	0.8130	0.8134
AdaBoost (50)	0.8342	0.7630	0.7643
AdaBoost (100)	0.8357	0.7910	0.7882
RandomForest (50)	0.9234	0.7947	0.7996
RandomForest (100)	0.9293	0.8044	0.8097
XGBoost (50)	0.8726	0.7970	0.7970
XGBoost (100)	0.9273	0.8120	0.8118
MLP	0.9800	0.7700	0.7500

Table 5.1. Best score (AUC) of each classifier on the embeddings generated by the NBNE, DeepWalk and Node2Vec. The weak learner used in AdaBoost is Decision Tree, AdaBoost (50) means an ensemble with 50 weak classifiers; the same applies for Random Forest and XGBoost. As can be seen in the table, MLP+NBNE outperforms other tested combinations. Another interesting phenomenon is that the embedding generated by Node2Vec and DeepWalk leads to close results.

learning algorithms than GaussianNB and AdaBoost. Another proof for this is when the number of weak learners of XGBoost increases from 50 to 100, which means higher the classifier's complexity, the accuracy with NBNE raises too.

5.2 Denoising Embedding

In our previous work (Chen et al., 2018), we have showed that node embedding (NE) algorithms can introduce noise while they learn node's representations due to the randomness in the generation of walks or permutations, thus preventing the effective use of all information in graphs to address real world problems. Therefore, we have proposed a novel approach to reduce noises in the NE node's representations by using denoising autoencoders.

5.2.1 Denoising Architecture

We adapted the typical denoising autoencoder's structure by using *tanh* (instead of *sigmoid*) as the activation function and added noise to the input instead of dropping out part of its features. By using *tanh* we preserve the representation's value learned by

NE algorithms between $[-1, 1]$ and by adding noise we simulate real world noise that is introduced during the process of representation learning.

The denoising steps of our approach are: (i) first corrupt the original input x by adding noises, $\tilde{x} := x + noises$, (ii) map the corrupted input into a latent space using an encoder multi layer perceptron (MLP), generating a more efficient and robust hidden representation of the input, $h := f(\tilde{x}) = tanh(W_f^n ... tanh(W_f^0 \tilde{x} + b_f^0) + b_f^n)$, where W_f^i are weight matrices and b_f^i are bias vectors, both learned during the training process, (iii) reconstruct the original input from the hidden representation using a decoder MLP, $\hat{x} := g(h) = tanh(W_g^n ... tanh(W_g^0 h + b_g^0) + b_g^n)$, where W_g^i are weight matrices, and b_g^i are bias vectors. Note that processing time increases linearly as the number of nodes rises.

	NBNE	DeepWalk	Node2Vec
MLP	0.9800	0.7700	0.7500
MLP + Denoised	0.9842	0.7809	0.7691

Table 5.2. AUC score of MLP on the embeddings and on its denoised version. The denoised embeddings have led to a small improvement. However, the increase is minimal.

Table 5.2 shows the result of our denoising experiment. The denoised version has led to a small improvement. These results are unexpected since we suppose that the embedding generated may contain noises. Nevertheless, we suspect that MLP can filter the noise due to its outstanding generalization power, so whether the embedding is noisy or not, MLP can exploit almost the whole information in embeddings.

Chapter 6

Conclusion and Future Work

In this chapter, we present the concluding remarks and several ideas for future works.

6.1 Conclusion

In this work, we proposed an multi-relation unsupervised graph embedding based approach for drug repositioning opportunities identification. Our approach takes advantages of multi-relation graph, learning drug's indications and embedding drug's mechanisms of action into a continuous low dimensional vector. Further, a classifier is trained with these vector representations on known indications. Finally, the trained classifier is used to predict other likely indications.

Additionally, we proposed various technique to improve the algorithm accuracy, such as denoising node embedding to reduce noises introduced during embedding process. We also have done a classifier analysis, comparing the performance of different classifiers and ensemble models, and have found a insight about the embedding algorithms, that the embedding of NBNE shows a more complex structure and it is necessary to make use of more powerful classifier to fully exploit its potential.

Moreover, we searched from biomedical literature and confirmed that several repositioning opportunities suggested by our model are truly new indications. This shows that our algorithm is a promising tool to find novel off-targets for existing drugs, since the first step of drug repositioning procedure is to identify potential repositioning candidates. As mentioned before, new drug development takes much more time and is usually costly, drug repositioning offers a cheaper and safer option.

6.2 Future Work

In this section, we discuss future work directions as follows.

Including disease-protein interactions: Our graph contains only drug-disease, drug-protein and protein-protein interactions. If we include disease-protein interactions, it may help to generate better representations for the drugs and diseases and consequently, leading a better and more accurate result. Therefore, a future work direction is finding those interactions and do experiments with them in our graph.

Explainability: Although the evidences that we showed about the method reliability, there is a need of a better understanding of the predictions. Deep learning algorithms usually face an explainability problem, the vector representations that graph embedding methods generated are not interpretable to human. Therefore, it is interesting to use other tools to analyse the model and its predictions, even developing a new explanation tool.

Combining more Strategies: Moreover, there are many other repositioning identification strategies as mentioned in Section 2.1, many of them are suitable for applying machine learning and deep learning algorithms. Therefore, it is an opportunity to explore more in this direction and it is possible to combine those different strategies to obtain a more accurate model.

Bibliography

Ashburn, T. T. and Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683. ISSN 1474-1784.

Belkin, M. and Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings of NIPS*, pages 585--591.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798--1828.

Berger, A. C., Olson, S., Johnson, S. G., Beachy, S. H., et al. (2014). *Drug repurposing and repositioning: workshop summary*. National Academies Press.

Bosco, G., Ostardo, E., Rizzato, A., Garetto, G., Paganini, M., Melloni, G., Giron, G., Pietrosanti, L., Martinelli, I., and Camporesi, E. (2019). Clinical and morphological effects of hyperbaric oxygen therapy in patients with interstitial cystitis associated with fibromyalgia. *BMC urology*, 19(1):108.

Bubnova, M., Aronov, D., and Persiyanova-Dubrova, A. (2019). Effects of rosuvastatin and atorvastatin on blood pressure, cerebral blood flow, endothelial function, angiotensin ii in patients with ischemic stroke-complicated hypertension. *Journal of Hypertension*, 37.

Cai, H., Zheng, V., and Chang, K. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616--1637.

Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J., and Bork, P. (2008). Drug target identification using side-effect similarity. *Science*, 321(5886):263--266.

Car, D. (2012). *Polypharmacology in Drug Discovery*. Wiley.

Chatr-aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., Reguly, T., Nixon, J., Ramage, L., Winter, A., Sellam, A., Chang, C., Hirschman, J., Theesfeld, C., Rust, J., Livstone, M., Dolinski, K., and Tyers, M. (2015). The biogrid interaction database: 2015 update. *Nucleic Acids Research*, 43:D470--D478.

Chen, D., Pimentel, T., Veloso, A., Ziviani, N., and Brandão, W. (2018). Denoising node embedding. In *Poster presented at LXAI Research, Montreal, Canada*.

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J., and Tang, Y. (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. *PLos Computational Biology*, 8(5):e1002503.

Chiang, A. P. and Butte, A. J. (2009). Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. *Clinical Pharmacology & Therapeutics*, 86(5):507--510.

de Oliveira Jr., R., Veloso, A., Pereira, A., Meira Jr., W., Ferreira, R., and Parthasarathy, S. (2014). Economically-efficient sentiment stream analysis. In *Proc. of ACM SIGIR*, pages 637--646.

Deepika, S. and Geetha, T. (2018). A meta-learning framework using representation learning to predict drug-drug interaction. *Journal of Biomedical Informatics*, 84:136--147.

Donato, A. and Brown, K. (2019). In black africans with hypertension, amlodipine-based therapy vs perindopril–hydrochlorothiazide improved bp control. *Annals of internal medicine*, 171(2):JC5--JC5.

Donner, Y., Kazmierczak, S., and Fortney, K. (2018). Drug repurposing using deep embeddings of gene expression profiles. *Molecular Pharmaceutics*, 15(10):4314--4325.

Dudley, J. T., Deshpande, T., and Butte, A. J. (2011). Exploiting drug–disease relationships for computational drug repositioning. *Briefings in bioinformatics*, 12(4):303--311.

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. (2015). Convolutional networks on graphs for learning molecular fingerprints. In *Proceedings of NIPS*, pages 2224--2232.

Ehrt, C., Brinkjost, T., and Koch, O. (2016). Impact of binding site comparisons on medicinal chemistry and rational molecular design. *Journal of Medicinal Chemistry*, 59(9):4121--4151.

Fullerton, C. A., Busch, A. B., and Frank, R. G. (2010). The rise and fall of gabapentin for bipolar disorder: a case study on off-label pharmaceutical diffusion. *Medical care*, 48(4):372.

Gao, K. Y., Fokoue, A., Luo, H., Iyengar, A., Dey, S., and Zhang, P. (2018). Interpretable drug target prediction using deep neural representation. In *Proceedings of IJCAI*, pages 3371--3377.

Gilmer, J., Schoenholz, S., Riley, P., Vinyals, O., and Dahl, G. (2017). Neural message passing for quantum chemistry. In *Proceedings of ICML*, pages 1263--1272.

Goyal, P. and Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78--94.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of KDD*, pages 855--864.

Guymer, E. K. and Littlejohn, G. O. (2019). Pharmacological treatment options for fibromyalgia. *Prevention*, 10:00.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. In *Proceedings of NIPS*, pages 1024--1034.

Hieronymus, H., Lamb, J., Ross, K. N., Peng, X. P., Clement, C., Rodina, A., Nieto, M., Du, J., Stegmaier, K., Raj, S. M., et al. (2006). Gene expression signature-based chemical genomic prediction identifies a novel class of hsp90 pathway modulators. *Cancer cell*, 10(4):321--330.

Hsieh, Y.-Y., Chou, C., Lo, H., and Yang, P. (2016). Repositioning of a cyclin-dependent kinase inhibitor gw8510 as a ribonucleotide reductase m2 inhibitor to treat human colorectal cancer. *Cell death discovery*, 2(1):1--8.

Hu, G. and Agarwal, P. (2009). Human disease-drug network based on genomic expression profiles. *PloS one*, 4(8):e6536.

Hurle, M., Yang, L., Xie, Q., Rajpal, D., Sanseau, P., and Agarwal, P. (2013). Computational drug repositioning: from data to therapeutics. *Clinical Pharmacology & Therapeutics*, 93(4):335--341.

Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., Murino, L., Tagliaferri, R., Brunetti-Pierri, N., Isacchi, A., and di Bernardo, D. (2010a). Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*, 107(33):14621--14626.

Iorio, F., Isacchi, A., di Bernardo, D., and Brunetti-Pierri, N. (2010b). Identification of small molecules enhancing autophagic function from drug network analysis. *Autophagy*, 6(8):1204--1205.

Jensen, P. B., Jensen, L. J., and Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395--405.

Jin, G., Fu, C., Zhao, H., Cui, K., Chang, J., and Wong, S. T. (2012). A novel method of transcriptional response analysis to facilitate drug repositioning for cancer therapy. *Cancer Research*, 72(1):33--44.

Jin, G. and Wong, S. (2014). Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug Discovery Today*, 19(5):637--644.

Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., Jensen, N. H., Kuijer, M. B., Matos, R. C., Tran, T. B., Whaley, R., Glennon, R. A., Hert, J., Thomas, K. L. H., Edwards, D. D., Shoichet, B. K., and Roth, B. L. (2009). Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181. ISSN 1476-4687.

Khalighi, H. R., Mortazavi, H., Mojahedi, S. M., Azari-Marhabi, S., and Abbasabadi, F. M. (2016). Low level laser therapy versus pharmacotherapy in improving myofascial pain disorder syndrome. *Journal of lasers in medical sciences*, 7(1):45.

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

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, 3(11):935–949. ISSN 1474-1784.

Kumar, A., Ramaraju, K., Singh, R., Mittal, B., Bhargava, R., and Mittal, M. (2019). Drug delivery device for pharmaceutical compositions. US Patent App. 10/238,803.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., et al. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929--1935.

Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019--1031.

Liu, A., Ouyang, S., Yu, B., Liu, Y., Huang, K., Gong, J., Zheng, S., Li, Z., Li, H., and Jiang, H. (2010). Pharmmapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic acids research*, 38(2):609--614.

Luo, D., Nie, F., Huang, H., and Ding, C. H. (2011). Cauchy graph embedding. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 553--560.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579--2605.

Marczewski, A., Veloso, A., and Ziviani, N. (2017). Learning transferable features for speech emotion recognition. In *Proc. ACM Multimedia*, pages 529--536.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347:1257601.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111--3119.

Ong, L., Cheung, B., Man, Y., Lau, P., and Lam, S. (2007). Prevalence, awareness, treatment, and control of hypertension among united states adults 1999–2004. *Hypertension*, 1(49):69--75.

Paik, H., Chung, A.-Y., Park, H.-C., Park, R. W., Suk, K., Kim, J., Kim, H., Lee, K., and Butte, A. J. (2015). Repurpose terbutaline sulfate for amyotrophic lateral sclerosis using electronic medical records. *Scientific reports*, 5:8580.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of KDD*, pages 701--710.

Pimentel, T., Castro, R., Veloso, A., and Ziviani, N. (2019). Efficient estimation of node representations in large graphs using linear contexts. In *Proc. of IJCNN*, pages 1--8.

Pimentel, T., Veloso, A., and Ziviani, N. (2017). Unsupervised and scalable algorithm for learning node representations. In *Proc. of ICLR*.

Pimentel, T., Veloso, A., and Ziviani, N. (2018). Fast node embeddings: Learning egocentric representations. In *Proceedings of ICLR*.

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., Norris, A., Sanseau, P., Cavalla, D., and Pirmohamed, M. (2019). Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug Discovery*, 18(1):41–58. ISSN 1474-1784.

Renaud, R. and Xuereb, H. (2002). Erectile-dysfunction therapies. *Nature Reviews Drug Discovery*, 1(9):663--664.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323--2326.

Sanseau, P., Agarwal, P., Barnes, M. R., Pastinen, T., Richards, J. B., Cardon, L. R., and Mooser, V. (2012). Use of genome-wide association studies for drug repositioning. *Nature Biotechnology*, 30(4):317–320. ISSN 1546-1696.

Shaw, B. and Jebara, T. (2009). Structure preserving embedding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 937--944.

Szklarczyk, D., Morris, J., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N., Roth, A., Bork, P., Jensen, L., and von Mering, C. (2017). The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, 45:D362--D368.

Theocharidis, A., Van Dongen, S., Enright, A. J., and Freeman, T. C. (2009). Network visualization and analysis of gene expression data using biolayout express 3d. *Nature protocols*, 4(10):1535.

Wagner, A., Cohen, N., Kelder, T., Amit, U., Liebman, E., Steinberg, D. M., Radonjic, M., and Ruppin, E. (2015). Drugs that reverse disease transcriptomic signatures are more effective in a mouse model of dyslipidemia. *Molecular systems biology*, 11(3).

Wang, D., Cui, P., and Zhu, W. (2016). Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225--1234.

Wang, W., Yang, S., Zhang, X., and Li, J. (2014). Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics*, 30:2923--2930.

Wishart, D., Knox, C., Guo, A., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36:D901--D906.

Xia, Z., Wu, L., Zhou, X., and Wong, S. (2010). Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Systems Biology*, 4(6).

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. (2008). Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24:232--240.

Yang, L. and Agarwal, P. (2011). Systematic drug repositioning based on clinical side-effects. *PloS one*, 6(12).

Zhang, W., Bai, Y., Wang, Y., and Xiao, W. (2016). Polypharmacology in drug discovery: A review from systems pharmacology perspective. *Curr Pharm Des.*, 22(21):3171--3181.

Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S. (2013). Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In *Proceedings of KDD*, pages 1025--1033.

Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457--i466.