Bonn-Aachen International Center for Information Technology (B-IT)
University of Bonn
Master Programme in Life Science Informatics

Master's Thesis

Generation and Applications of Side Effect Embeddings in Biomedical Knowledge Discovery

Submitted by
Rana Subhi Aldisi

First Supervisor: Prof. Dr. Martin Hofmann-Apitius
Second Supervisor: Prof. Dr. Holger Fröhlich
Internal Supervisor: Charles Tapley Hoyt

In collaboration with the Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

September 15, 2019

Abstract

Side effects are unintended consequences of introducing drugs into the body. Identifying a drug candidate's side effects is an important step in drug development that can both be expensive and still result in incomplete side effects profiles. Further, side effect profiles have been used to identify drugs' targets and hypothesize new therapeutic benefits for existing drugs.

This thesis introduces a workflow that applies network representation learning to biomedical networks that contain drugs, their targets, their indications, and their side effects in an attempt to understand the mechanisms of action underlying side effects. Different network representation learning models were evaluated and optimized before selecting the best for training a predictive model for relations between different entities in the original network. It was then used to predict chemical-phenotype, chemical-target, and target-phenotype relations, which were analyzed and validated using literature.

Acknowledgements

First and foremost, I would like to thank Prof. Dr. Martin Hofmann-Apitius for taking me under his wing and giving me the opportunity to work on this project.

I would also like to thank Prof. Dr. Holger Fröhlich for his guidance and advice throughout the project.

To my supervisor Charles Tapley Hoyt, thank you for your guidance and support, and for pushing me to my limits.

To my colleagues at Fraunhofer SCAI and fellow students, thank you for taking interest in my work and encouraging me to do my best.

Lastly, to my family, thank you for your unconditional love and support.

Declaration

I hereby certify that this material is my own work, that I used only those sources and resources referred to in the thesis, and that I have identified citations as such.

Bonn, September 15, 2019

.....
Rana Subhi Aldisi

Contents

List of Figures

List of Tables

1 Biological Background

Drug development and discovery is a time-consuming and expensive process that has a low success rate, and bringing a single drug to market can take 10-15 years and billions of dollars [1, 2]. Even with the recent increase in research on drug development, the number of new therapeutic chemical and biological entities that have been approved by the United States Food and Drug Administration (FDA) has been decreasing since the late 1990s [3].

Drug repositioning (i.e., drug repurposing) is the process of discovering new therapeutic benefits for existing drugs. Repositioning offers several advantages over the drug discovery and development process. In many cases, repositioning candidates have passed many stages of drug development such as screening, chemical optimization, and even clinical development, thus they would have well-known safety and pharmacokinetic profiles. Therefore, repositioning provides a faster pathway to the market in which several years of drug design, development, and clinical stages can be removed [4]

1.1 Drugs, Targets, and Side Effects

1.1.1 Drugs

A drug is described as a chemical or substance that when applied to a physiological system can affect its function in a certain way [5]. A drug can be used for the purpose of diagnosis, relief, prevention, or cure of a pathological state [6].

1.1.2 Targets

A target refers to protein, peptide, or nucleic acid that has an activity which can be modified by a drug [7]. Ideally, one should have a proven role in the pathology of a disease, and its modulation does not have a significant role in other diseases or under normal conditions. It should also have a biomarker that can be monitored for measuring therapeutic efficacy [7].

1.1.3 Side Effects

Even though drugs are taken for their therapeutic effects, they also have the potential risk of being harmful. Since drugs' effect inside the body are not only limited to their intended targets, sometimes, they can cause unintended medical reactions in the body. These are known as a side effects, or adverse drug events [8]. While their causes generally lack a mechanistic understanding, some intrinsic risk factors have been suggested for their developments such as age, gender, weight, genetics, and state of health. They could also be affected by extrinsic factors like the dosage of the drug, the route of administration, or taking multiple drugs at the same time [8].

1.2 Biological Data Sources

Because biological data is highly heterogeneous, it can be found in numerous data sources, each of which has its own data structure and query interface [9]. Biological data sources available are abundant, ranging from gene and protein resources, like the Hugo Gene Nomenclature Committee (HGNC) [9] and UniProt [10], to biomedical literature resources and ontologies, like PubMed Central [11] and Gene Ontology (GO) [12]. However, since biological systems are complicated and biological entities are connected, using one biological data source might not be sufficient to answer complex biological questions. Thus, data integration methods need to be applied to extract knowledge from different biological resources in order to comprehensively analyze biological systems [9].

1.2.1 Chemicals Resources

DrugBank

DrugBank is a comprehensive chemical database containing molecular information on drugs, their associated mechanisms, their interactions, and their targets [13]. This database not only contains information about FDA-approved drugs, but also experimental and investigational drugs. As of 2018, DrugBank contained 11,926 drugs; more than 6,000 of which were (at the time) FDA-approved. Furthermore, the database contains physico-chemical, pharmacological, pharmacogenomic, pharmacokinetic and molecular biological data of drugs and their targets, as well as drug-drug interactions, drug-food interactions, and drug transporter data. All these information were manually extracted and curated from more than 27,000 peer-reviewed articles.

PubChem

PubChem is a United States National Center for Biotechnology Information (NCBI) resource that contains three inter-linked databases: the PubChem Substance database, which consists of chemical information submitted by data contributors, the PubChem Compound database, which stores the extracted chemical structure for the Substance database [14], and BioAssay database, which consists of biological assay experiments' descriptions and results [15]. As of its 2019 publication, PubChem contains more than 247.3 million substance descriptions, more than 96.5 million different chemical structures, and bioactivity assays covering over 10,000 proteins [15].

1.2.2 Side Effects Resources

SIDER

Side effect resource (SIDER) is a public database that contains information about drugs and their side effects. It contains 888 drugs, 1,450 side effects, and more than 62,000 drug-side effect relations. The side effect information was compiled mainly from the United States FDA, among other public resources, and the drug names

were mapped to PubChem identifiers to enable linking to other databases [16].

1.2.3 Other Resources

One challenge in data integration of biological data sources is the differences in terminology between them. Controlled vocabularies such as Medical Subject Headings (MeSH) and the Unified Medical Language System (UMLS) have been developed to standardize terminology between different resources [17] to address this challenge.

Medical Subject Headings

MeSH is a controlled vocabulary that was developed by the United States National Library of Medicine (NLM). It describes many biomedical concepts such as chemicals, drugs, and diseases in order to support indexing the Medical Literature Analysis and Retrieval System Online (MEDLINE), a database for biomedical literature [18]. The headings in MeSH are arranged in a hierarchical tree structure with main headings (e.g., Anatomy, Diseases, Organism) and branches that have many levels of sub-branches. This hierarchy permits searches of MEDLINE to include narrow terms in all the below branches when searching for a broad term [19].

Unified Medical Language System

UMLS is a freely available resource, developed by NLM, that consists of biomedical vocabularies. It includes GO, Online Mendelian Inheritance in Man [20], MeSH, NCBI taxonomy, and the Anatomist Symbolic Knowledge Base [21]. UMLS terms are inter-related and are cross-references to internal or external resources [17]. It also consists of a metathesaurus of inter-related concepts and a semantic network that categorizes the metathesaurus concepts as well as lexical resources that generate lexical variants of the concepts [17].

1.3 Related Work

Many studies have been conducted to discover the relationship between a given drug and its side effects, or the side effects of a drug and the link to indication areas, or the side effects of a drug and the link to different (off-) targets. For example, SIDER was primarily created to combine drugs and their side effects information to analyze and investigate them [16]. In an attempt to investigate side effects, Scheiber *et al.* used a chemical space to map adverse reactions by extracting chemical features that are highly correlated to a specific effect [22]. For every side effect term used, all associated molecules were extracted from PharmaPendium[1] and the similarities between the side effects were then calculated using Pearson correlation after creating the chemical space from molecular descriptors of the molecules extracted. The authors concluded that common side effects of drugs can be associated with common chemical structures [22].

1.3.1 Prediction of Side Effects

Although clinical trials can reveal drugs' side effects, they are usually expensive and may lead to incomplete reflection of all adverse reaction events. Furthermore, this does not preclude the discovery of side effects after the drug is introduced to the market, which can have severe consequences on patients [23].

Side effects are significantly responsible for drug failure during clinical trials, which is why it is essential for the commercial success of the drugs to create approaches for predicting and countering those side effects during the developmental phase [24]. Most of the developed approaches for detecting side effects of drugs are based on the fact that chemically similar drugs induce similar side effects. However, since drugs induce multiple effects on the biological system, it is challenging to discover the underlying mechanism of side effects [23].

Atias and Sharan (2011) applied a canonical correlation analysis to obtain a low dimensional subspace, which contains associations of drugs and side effects and molecular information of drugs. This allowed the identification of side effects that best correlates with a drug query that is introduced to the subspace, then side effect similarity network was used to obtain the final scores that are based on side effects of drugs that are similar to the query [25]. A similar method used sparse canonical

[1] `https://www.pharmapendium.com`

correlation analysis to find the correlation between chemical substructures of molecules and side effects [26]. A more recent approach, DrugClust, used machine learning methods to cluster drugs based on their chemical features then obtains side effects prediction based on calculated Bayesian scores[23].

1.3.2 Identification of Drugs' Targets

Many approaches have been created for identifying novel targets for a drug, usually using chemical similarities or cellular features [27]. However, the side effects of the drug can also be beneficial in such task because these side effects may be due to the drug binding an off-target causing unexpected reaction in a metabolic or signalling pathway [28]. Although these unexpected reactions caused by the off-targets are often harmful and undesirable, they can sometimes lead to beneficial discoveries, like finding new therapeutic indications for drugs [27]. For example, thalidomide, a drug that was prescribed as a treatment for morning sickness in pregnant women in the 1950s, was the cause of more that 10,000 babies born with birth deficits. However, it was later discovered that thalidomide is an angiogenesis inhibitor and it was repurposed for treating cancers like multiple myeloma [29].

Approaches using side effects to predict drug targets rely on the assumption that similar side effects of dissimilar drugs are caused by a common off-target. This is generally because drugs that have similar binding profiles tend to cause similar side effects, suggesting a direct correlation between target binding and side effects similarity. An example of that are the two drugs cisapride and astemizole, which have serotonin and histamine receptors as their primary targets, respectively. Both of those two unrelated drugs inhibit hERG, the cardiac ion channel, causing cardiac arrhythmias [27]. Campillos *et al.* inferred molecular activities of drugs by exploiting the side effects of marketed drugs rather than their chemical similarities or their known targets. They then measured the side effect similarities of the marketed drugs and analyzed their likelihood of sharing protein targets and concluded that indeed side effect similarity can indicate common protein targets of unrelated drugs [27].

In another study, a large-scale analysis was used to identify protein-side effect relations by integrating drug-target and drug-side effect relations. This approach predicted that the activation of serotonin receptor family is associated with hyperaesthesia, which is the increase in pain sensitivity. To confirm this prediction, a serotonin receptor, HTR7 (5-hydroxytryptamine receptor 7), was tested on mice

to see if it elicits hyperaesthesia, and the results suggested that it is indeed the case [30]. All these successful studies support the assumption that side effects of drugs can be used to identify drugs' targets.

1.3.3 Drug Repositioning

Side effects can also be used as phenotypic biomarkers for diseases because both indications and side effects are measurable physiological changes in response to drugs. Therefore, if drugs used for the treatment of a disease have common side effects, an underlying mechanism of action might be linking the disease and the side effects [31]. Yang and Agarwal used this reasoning to build a disease-side effect association database from drug-side effect and drug-disease data, taken from SIDER and PharmGKB respectively, which can be used for predicting new indications for marketed drugs [31]. In a slightly different approach, a drug-drug relationship network was constructed using side effect similarities, which was then used to predict new indications of drugs according to their network neighbors [32].

2 Computer Science Background

A graph, or network, is a collection of nodes which are connected by edges. A graph is denoted as $G = (V, E)$ in which $v \in V$ is a vertex, or node, and $e \in E$ is an edge [33]. If a graph has multiple types of nodes and/or edges, it is called heterogeneous. Alternatively, a graph that has only one type of node and one type of edge is called homogeneous [33].

Analysis of networks is used in many different fields to understand the community structure of the network and the relationships between entities [33]. It also helps gain insight on the hidden information and patterns in the network [33]. Network analysis have been used in many applications such as relationship prediction, entity classification, clustering, and visualization [34].

NRL is a machine learning approach that learns embeddings for nodes of a network in a latent, low-dimensional vector space, without compromising node content, network topology, and other information [35]. The topological and structural properties of a node are encoded into its embedding, and the distance between nodes in the vector space captures the relationships between them. Since each node is represented by a vector that contains its information of interest, computing mapping functions or distance matrices on the embedding can help avoid high complexity during network analysis. Thus, NRL methods have two main goals: to be able to reconstruct the original network from the learned embedding and to be able to use the learned embedding space to effectively support network interface [36].

2.1 Background on Network Biology

Relationships in systems biology are often represented as networks, which allows analysis and modeling of the data using computational methods. With the emergence of the field of systems biology, many biological networks have been created and analyzed such protein-protein interactions, gene regulatory networks, and metabolic networks. The study of such networks help in the understanding of human disease, mechanisms of action of drugs, and complex biological systems [37].

2.1.1 Biological Network Topology

The structure of a network plays an important role in analyzing and understanding its performance. The most common topological features (e.g., degree distribution, distance, clustering coefficient) are discussed below.

The degree of a node is the number of relations it has, a node with high degree has a better connection in the network, therefore it plays a more important role in preserving the network structure. Biological networks usually consist of a small number of nodes that are highly connected (hubs), and a large number of nodes that have fewer connections, which is known as "scale-free" format [38]. This network format follows a power-law distribution k^{-a} where k is the degree of a fraction of nodes and $a > 1$ [39].

The geodesic distance between two nodes is defined as the length of the shortest path between them. On the other hand, the diameter of a network is the maximum shortest path length between all pairs of nodes. The approximate distance between nodes in a network can be measured by calculating the average distance and diameter. A network with small diameter means that on average two nodes are connected by relatively short paths. This type of graph is generally known as a "small world" network [38].

The clustering coefficient of a node is the percentage of existing relations among its neighborhood. It can be calculated by the number of edges between nodes in the neighborhood divided by the total number of edges that are possible between them, which gives a value between 0 and 1. A "small world" network usually has a high clustering coefficient, which indicates that nodes in the network tend to form groups [38]. On the other hand, a network that was created randomly has a

low clustering coefficient, usually very close to 0 ($C \approx 0$) [40].

2.2 Random Walk Representation Learning Methods

A random walk is a stochastic process in which a path is created by iteratively randomly choosing a neighbor of the last node in the walk. Random walks are useful for detecting communities and capturing community information [41]. For the purpose of NRL, a stream of short random walks is used to extract the network information, and several walkers can be used to explore different parts of the same graph at the same time [41].

2.2.1 Language Models

Language models are an important part of natural language processing, which is a branch of artificial intelligence that aims to enable computers to interact, analyze, and process natural language. The objective of a language model is to assign probabilities specific sequence of words appearing in a corpus. For a sentence with finite sequence of words from a given vocabulary, the goal is to maximize the probability of a word over all the training corpus [41]. To use a language model on short random walks created from a graph, one can generalize the model by processing the random walks as a special language with nodes as words [41].

2.2.2 SkipGram and GloVe

SkipGram is a language model that attempts to predict the context of a word by maximizing the probability of co-occurrence among the words that appear within a window in a sentence [41]. This can be done by training a neural network with pairs of words within a certain window from the training corpus, the probability of the co-occurrence is then calculated by the number of times each pair appeared in the training [42]. However, since the goal is to learn latent representations rather than the probability of node co-occurrences, a mapping function is introduced to the probability calculation [41].

Global Vectors (GloVe) is another language model that uses a count matrix factorization approach. It learns word representations by calculating their ratio of the co-occurrence probability instead of raw co-occurrence probability, which is said to encode meaningful global information [43].

2.2.3 DeepWalk

DeepWalk is an unsupervised scalable representation learning model that learns latent representations by obtaining information from short random walks. The algorithm of the model consists of two main steps:

1. Perform random walks on the nodes in the graph and generate short sequences of nodes

2. Run SkipGram using the paths generated in the first step to learn features of nodes and create node embeddings.

The model also uses hierarchical softmax to approximate the probability distribution, since using softmax as an activation function would be computationally expensive, with a computation time of $O(|V|)$. In hierarchical softmax, a binary tree, with the nodes of the graph as leaves, is used to handle the computation problem by factorizing the conditional probability. The probability of a given node v_i is computed by calculating the probability of each sub-path from the root node to the node v_i, which reduces the computation time to $O(log|V|)$ [41].

2.2.4 Node2vec

Node2vec is a semi-supervised approach that learns scalable latent features in networks. This model is very similar to DeepWalk with two additional parameters: p and q. The return parameter, p, controls the probability of revisiting nodes in a walk. A high value of p means that it is less likely to revisit a node for the next two steps. If p is low, it is more likely to revisit the node immediately. Having a low p parameter ensures that the walk stays local. The q parameter deals with "inward" and "outward" nodes. If $q > 1$, the random walk is more likely to visit a node that is close to the previous node, which means that it focuses more on the local structures. If $q < 1$, then the walk is biased toward nodes that are further away from the previous node, which encourages exploration of the graph [44].

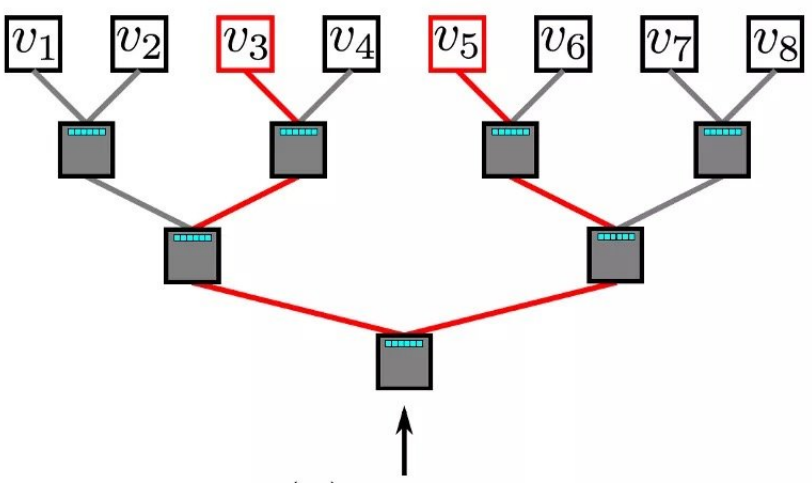


Figure 1: A schematic diagram of the hierarchical Softmax. Image adapted from [41]

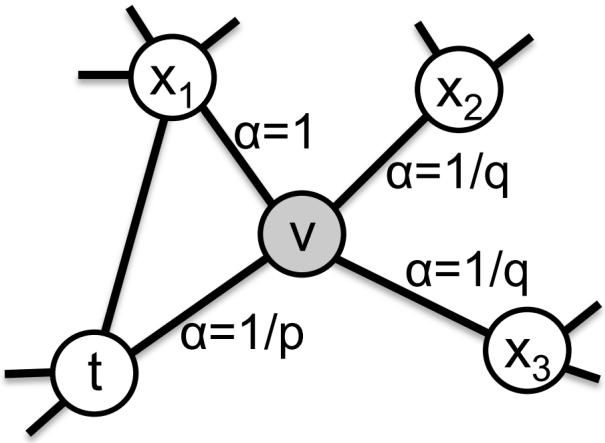


Figure 2: Random walk evaluation for next step transition for v. Edge labels indicate the search bias. Image adapted from [44]

2.3 Other Representation Learning Methods

2.3.1 Translational Distance Models

The idea of translational distance methods is to calculate the possibility of a fact by measuring the distance between two entities using distance-based scoring

functions [45]. Translational models have multiple applications in biomedical literature, including drug-drug interaction prediction from drug knowledge graph [46, 47] and disease prediction and clustering from symptom-disease network [48].

TransE

TransE is an energy-based model that represents relationships as translations in the embedding space. Assuming there is a directed graph with entities and edges in the form of (head, relation, tail), or (h, r, t), which indicated that there is a relationship between the head and tail entities. Then the embedding of the two entities h and t should be close to one another plus a translation vector r, i.e $h + r \approx t$ when (h, r, t) holds. This approach learns only one embedding for each entity and each relationship, i.e 1-to-1 relationships [49].

TransH

TransH, or translation on hyperplane, tries to solve the problem in dealing with 1-to-N/N-to-N/N-to-1 relations in TransE by interpreting the translation vector on a hyperplane so that an entity will have distributed representations when involved in different relations. For (h, r, t), the relation r will have a translation vector d_r which will be projected in the relation-specific hyperplane w_r, the embedding h and t are projected into the hyperplane as h' and t', respectively, which are expected to be connected by a translation vector d_r on the hyperplane [50].

TransR

While both TransE and TranH assume that the embeddings of entities and relations are in the same space, TransR suggests creating entities and relations in different spaces and performs the translation depending on the relation space. A a projection matrix M_r is generated for each relation r, which projects entities from the entity space to relation space. For a triple (h, r, t), h and t entity embeddings are projected into r-relation space with operation M_r [51]. Though TransR has improved compared to the previous translational models, it has a few limitations that can affect its performance. For example, even though entities linked by relations can have various types and attributes, it maps all relations to the same mapping

matrix M_r. Also, mapping matrices are determined by the relations alone, despite the fact that there is an interactive process between an entity and a relation [52].

2.3.2 Matrix Factorization Methods

As the name suggests, these algorithms factorize a matrix which is formed from connections between nodes to obtain embeddings. The matrix representing the connections can be created using different data like the adjacency matrix, the Laplacian, or the node transition probability matrix [34]. Matrix factorization methods have been utilized before for drug-target interaction prediction using similarity profiles [53, 54] and in gene-disease networks using gene similarity and disease similarity matrices [55]. One important advantage of these methods is that they can preserve the global structure of a network by considering global nodes proximity. However, they are unscalable to large graphs because they are time and space consuming [33].

HOPE

High-Order Proximity preserved Embedding (HOPE) is a graph embedding approach that attempts to preserve the asymmetric transitivity in the graph, which is important in capturing its structure. Asymmetric transitivity represents the correlation between directed edges, which is that if there is a directed edge from u to v, then there is probably a directed edge from v to u. In HOPE, an adjacency matrix is used to derive two polynomial matrices, which are then used to generate generalized singular values for each polynomial matrix and their corresponding singular vectors. The two generalized singular values vectors can be combined and then used along with the singular vectors to create the optimal embeddings [56].

GraRep

Graph Representation (GraRep) is an approach that captures the graph's global structure information using an extended version of SkipGram [57]. GraRep learns the different k-step relation information with different k values among nodes form the graph by utilizing different global transition matrices defined over the graph. In this model, nodes with common k-step neighbors should have similar latent features [57]. The method starts by creating three matrices, an adjacency matrix S,

which indicates the presence or absence of edges between given node pair, degree matrix D, which contains information about the number of connections each node has, and 1-step probability transition matrix A which indicates the probability of transition between node pair within one step.

$$A = D^{-1}S \tag{1}$$

k-step transition probability matrix for each k-step can be computed. Where $A^k_{i,j}$ refers to the transition probability between v_i and v_j in which the transition contains exactly k-step(s). For each k-step, the positive log probability matrix is produced and representations are generated separately, then all k-step representations are concatenated together to form the final representation.

2.3.3 Deep Learning Methods

While the previously described methods perform poorly on large and real world information networks and face challenges in handling non-linear data structures [58], deep learning methods solve these issues by incorporating autoencoders, which contain multiple nonlinear functions, and deep neural networks, which are robust and effective because of their multi-layered architecture [33]. These kind of models have been used in tasks like utilizing electronic health records for risk prediction [59], and predicting polypharmacological side effects [60].

LINE

Large-scale Information Network Embedding (LINE) is a model proposed to handle the issue of embedding large information networks into low dimensional vector space. LINE optimizes an objective functions that can preserve both the local (i.e. first-order proximity) and global (i.e. second-order proximity) structures of multiple types of networks (e.g. directed, undirected, and/or weighted) [61].

First-order proximity, which is the local pairwise proximity between nodes in the network, for an undirected edge(i, j) can be found by defining the joint probability between two nodes as shown in equation 2. Where u_i is the low-dimensional representation (vector) of the node v_i. The objective function is optimized by minimizing the difference between two probability distributions. First-order proximity

can only be applied to undirected edges in this model [61].

$$p_1(v_i, v_j) = \frac{1}{1 + \exp\left(-u_i^T . u_j\right)} \tag{2}$$

Second-order proximity assumes that nodes with many connections to other nodes are similar to each other. Also, each node is perceived as a "context" and nodes that have similar distributions over the "contexts" are presumed to be similar. For each edge(i, j), the conditional distribution of "context" v_i is defined in equation 3 where u_i is the representation of the node v_i itself while $u_i^{'}$ is the representation of the node as "context" to other nodes and $|V|$ is the number of nodes or "contexts". Similar to first-order proximity, the objective function of the second-order proximity is calculated by minimizing the difference between two probability distributions [61].

$$p_2(v_i|v_j) = \frac{\exp(u_j^{'T} . u_i)}{\sum_{k=1}^{|V|} \exp\left(u_k^{'T} . u_i\right)} \tag{3}$$

The model can preserve both the first-order and second-order proximity by concatenating the embedding trained by both methods for each node [61].

SDNE

Structural Deep Network Embedding (SDNE) model aims to capture the highly non-linear structure of networks and preserve their local and global structures. SDNE is a semi-supervised autoencoder model that contains multiple layers of nonlinear mapping functions to capture the nonlinear network. Then, the first-order proximity is used by the supervised component to capture the local structure while the second-order proximity is used by the unsupervised component to preserve the global structure of the network [62].

The neighborhood for each node is reconstructed to preserve the second-order proximity, and the pairwise similarities for a small portion of node pairs are obtained to preserve the first-order proximity, this creates an adjacency matrix which is the input to the autoencoder. The model also introduces a penalty on the reconstruction error for non-zero elements because while the presence of links indicate similarity between nodes, the absence of links does not necessarily

mean dissimilarity. Parameter α balances the weight between first-order and second-order proximity, when $\alpha = 0$, the model performance is dependent on second-order proximity, as α gets larger, the model focuses on first-order proximity. The β parameter controls the reconstruction weight of the non-zero elements in training set. The larger β is, the more susceptible the model is to reconstructing non-zero elements [62].

2.4 Application of Network Representation Learning

NRL methods have been used in several tasks including node classification, node ranking, node clustering and edge prediction.

In node classification, the NRL model learns latent features from labelled nodes, then assigns a class label for each node in the graph based on those features. Thus, similar nodes will have similar labels [58]. Node classification can be used for classifying proteins according to their biological functions [44].

The aim behind node ranking is to rank top k nodes of interest to a given node based on criteria like similarity. One example of an approach using such task is GuiltyTargets [63], which is a recently developed model that uses gat2vec [36], an NRL approach extending DeepWalk, to map a protein-protein interaction network that is annotated with differential gene expression, then using machine learning methods a ranking is assigned for candidate drug targets [63].

The aim of node clustering is to group similar entities together. General clustering methods like k-cluster or k-nearest neighbours are applied on the node embeddings to create clusters. This task is particularly useful for discovering related drugs or proteins [64].

The edge prediction, or link prediction, task aims to predict missing edges between nodes in a graph using the learned features. This is possible because even though the representation is low-dimensional, it preserves the structure of the graph, so these embeddings are rich with information and can be used for edge prediction. This is one of the most common tasks used in biological network analysis because biological networks are never complete and thus new edges can always be discovered [64].

3 Motivation

The goal of this thesis is to create a network that can be used to gain insight into the causes of side effects for a given drug relying on the relationships between chemicals, targets, phenotypes (Figure 3). In which case, the cause of side effects can be understood by analyzing the mechanism of action of drugs.

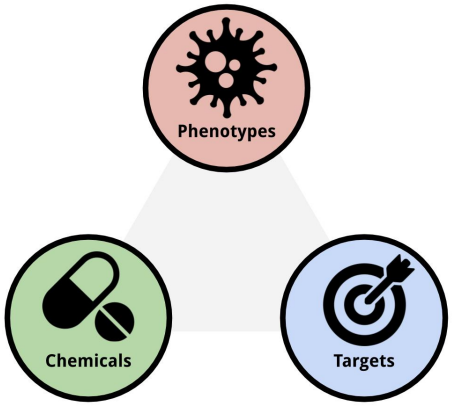


Figure 3: Chemicals-targets-phenotypes relationship triangle.

To achieve the goals of the project, first, a knowledge graph was created and enriched from multiple sources. This graph contained three different types of entities:chemicals, targets, and phenotypes. It contained three different relations: chemical-chemical, chemical-target and chemical-phenotype. Various NRL approaches were used to create embeddings of the graph, which were then trained and optimized to obtain the best predictive model. This model was used to predict new relations with different node types, which were then contextualized with additional literature. A schematic of the workflow is presented in Figure 4.

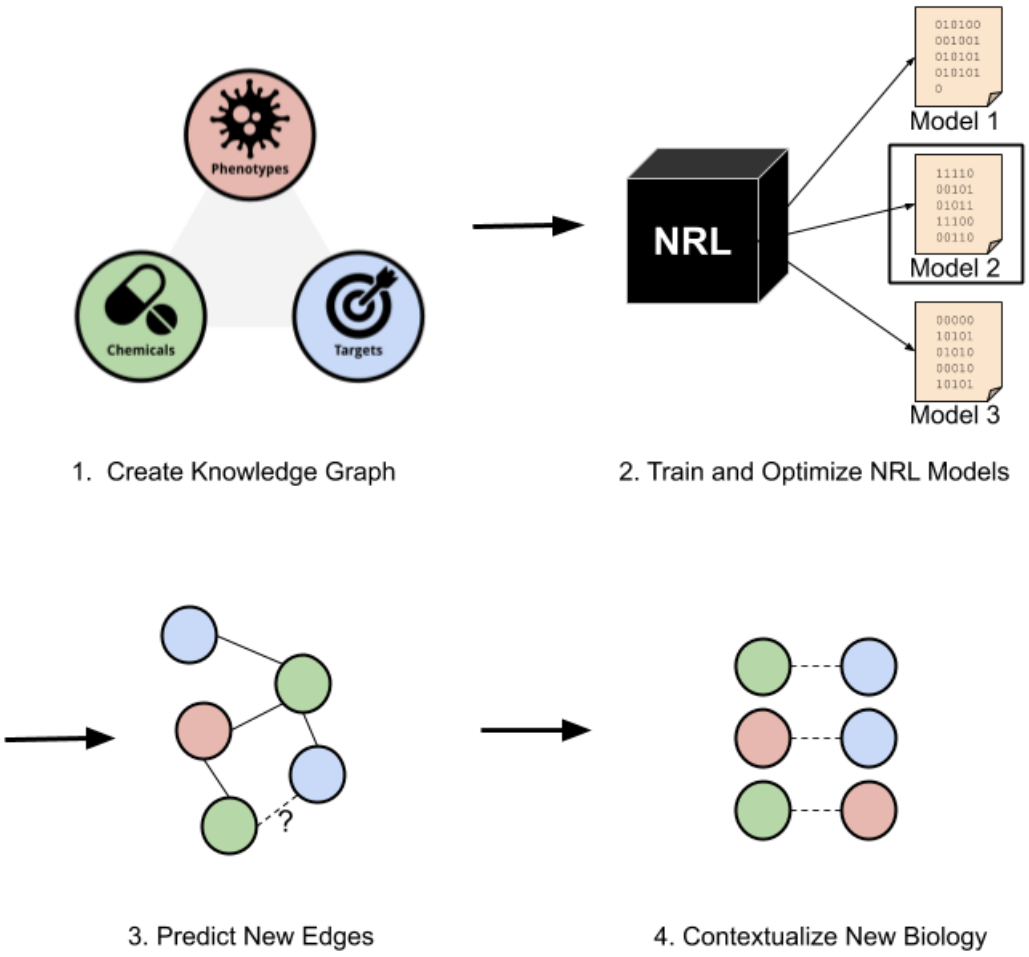


Figure 4: The workflow of this thesis.

4 Materials and Methods

This section explains the methods that have been used in the project. It describes how the network was built, the different NRL models ran on the created network, the classifier used for prediction, and the metrics used to evaluate the models.

4.1 Network Construction

For the purpose of this thesis, a network of drugs-side effects-targets was constructed using two databases: SIDER and DrugBank. The Bio2BEL framework, which integrates biological data using Biological Expression Language (BEL), was used to easily navigate and extracted information from the databases [65]. RDKit[1] was also used to calculate chemical similarities and create relation edges between drugs.

4.1.1 SIDER

The bio2bel_sider Python package [66] was used to download information from SIDER database and convert the drugs, side effects, indications, and their interrelations into a BEL graph. The resulting SIDER graph contained 1,507 chemicals, identified with PubChem identifiers, and 6,990 side effect/indication, normalized using UMLS terms. The total number of edges was 180,203.

[1] `http://www.rdkit.org`

4.1.2 DrugBank

DrugBank graph was built using bio2bel_drugbank Python package [67]. The package asks the user to register and download the full database from DrugBank website, then it uses the downloaded file to extract information and relations, which can then be converted to BEL graph. The chemicals in the graph are labeled with PubChem identifier and the proteins are identified with UniProt identifiers. The total number of chemicals in the graph was 6,386 and the total number of proteins was 4,049. The graph also consisted of 43,589 edges.

4.1.3 Chemical Similarity Graph

The chemical similarity graph was created with the help of methods from RDKit, which is an open source toolkit for chemoinformatics. The molecules and their molecular fingerprints were identified and calculated using their simplified molecular-input line-entry system (SMILES) strings. The molecular fingerprints used here were the MACCS keys, which contain 166 structural features of a molecule. Then, the Tanimoto similarity metric was used to calculate the similarity between each node pair using their MACCS keys fingerprints.

Two different chemical similarity graphs were built and tested. The first was created using pairwise similarities, in which chemical pair that have more than 50% similarity will have a relation. This resulted in a graph of 6,664 nodes and 1,738,887 edges in total. The second graph was built using clustering, in which all chemicals in the same cluster were given relations between one another. This was done by calculating the dissimilarity between fingerprints ($1 - similarity$), creating a distance matrix, and clustering using RDKit's implementation of Butina's clustering algorithm, which creates clusters that have centroids that are at least similar to every molecule in the cluster [68]. This graph had a total of 5,529 nodes and 110,228 edges.

4.1.4 A Complete Graph

To construct the complete graph, with drugs, targets, and side effects, SIDER, DrugBank, and chemical similarity graphs were combined together. Each similarity graph was combined with the complete graph seperately, creating two different

complete graphs. A mapping file was created from the DrugBank database containing the PubChem identifier, DrugBank identifier, and drug name. The mapping was also improved by adding the canonical SMILES which were taken from PubChem API. The drugs from both graphs were merged using the SMILES to remove duplicated chemicals. The resulting graph had a total of 17,720 nodes: 4,049 proteins, 6,681 chemicals, and 6,990 side effects/indications. The relations that existed were chemical-protein, chemical-side effect/indication, and chemical-chemical.

4.2 Experimental Runs

To find the best NRL model for prediction, six different NRL models (node2vec, DeepWalk, HOPE, GraRep, SDNE, and LINE) were trained and their hyperparameters were optimized. These models were used based on their application and performance in [69]. The best model was selected based on the collective evaluation results between all optimized models.

4.2.1 NRL Models

To choose the best prediction model for the network, six different models from three categories were selected and tested. HOPE and GraRep were chosen from the matrix factorization approaches, LINE and SDNE were selected from the deep learning methods, and from random walk approaches, DeepWalk and node2vec were chosen. All models were run using the BioNEV Python package [69]. To evaluate the models, the edges of the complete graph were split randomly to create the training and testing sets with a ratio of 8:2 respectively. Negative edges, which are edges that do not exist in the graph, for each set were also generated.

4.2.2 Edge Embeddings

To train the classifier to predict relations between two nodes, edge features need to be generated using the learned node embeddings. Table 1 introduces a number of binary operators that are generally used for such a task.

According to many representation learning experiments [44, 70, 71], the Hadamard

Hadamard	$\varepsilon(u,v) = \eta(u) * \eta(v)$
Concatenation	$\varepsilon(u,v) = [\eta(u), \eta(v)]$
Average	$\varepsilon(u,v) = 0.5 * (\eta(u) + \eta(v))$
Weighted L_1	$\varepsilon(u,v) = \mid \eta(u) - \eta(v) \mid$
Weighted L_2	$\varepsilon(u,v) = \mid \eta(u) - \eta(v) \mid^2$

Table 1: Example binary operators $\varepsilon(u,v)$ for embedding for edge $(u,v) \in E$. Operators adapted from [44]

operator produces the most stable and most accurate edge representation for multiple types of networks. Following that observation, the edge embeddings used for training the classifier in this thesis were calculated using this approach. The Hadamard operator is defined such that each element in the vector (u) is multiplied by the corresponding element in (v), resulting in the new vector (u,v) that is the "product" of (u) and (v). Alternatively, other approaches, such as gat2vec and GuiltyTargets, have used concatenation operator to generate edge features [36, 63].

4.2.3 Hyperparameter Optimization

All models have gone through a hyperparameter optimization process. With the help of the optuna Python package [72], ranges for the hyperparameters were selected and a number of trials were run while randomly changing parameters to find the hyperparameters that maximize the Matthews correlation coefficient, explained in 4.4. The ranges and values for the hyperparameters where chosen based on their role in the model and were taken from [69].

4.3 Binary Classification

A logistic regression model was used for binary classification to predict edges between a given pair of nodes. The logistic regression is one of the most popular statistical models used for binary classification because it is simple, easy to interpret, and proven to perform well in many classification and prediction tasks. The results of the classifier are interpreted using two measures, p-value and minus log probability. The p-value is the probability that the null hypothesis is true. In this case, it is that a given edge does not exist between two nodes in the graph.

Method	Hyperparameters	Range/Value
HOPE	Dimensions	100 - 300
GraRep	Dimensions	100 - 300
	k-step	1-10
LINE	Dimensions	100 - 300
	Proximity order	[1,2,3]
	Epochs	[5, 10, 15, 20, 25, 30]
SDNE	Proximity balance (α)	0.0 - 0.4
	Reconstruction weight (β)	0 - 30
	Epochs	[5, 10, 15, 20, 25, 30]
DeepWalk	Dimensions	100 - 300
	Walk length	$[2^3, 2^4, 2^5, 2^6, 2^7]$
	Number of walks	$[2^3, 2^4, 2^5, 2^6, 2^7, 2^8]$
	Window size	2 - 6
node2vec	Dimensions	100 - 300
	Walk length	$[2^3, 2^4, 2^5, 2^6, 2^7]$
	Number of walks	$[2^3, 2^4, 2^5, 2^6, 2^7, 2^8]$
	Window size	2 - 6
	Return parameter (p)	0.0 - 4.0
	In/out parameter (q)	0.0 - 4.0

Table 2: Hyperparameters optimized for NRL methods. HOPE's only hyperparameter is the dimensions of the embeddings. GraRep need the dimensions and k-step parameters. LINE depend on the dimensions, proximity order and the epochs. SDNE has the α, β and the epochs parameters. DeepWalk depends on the dimensions, walk length, number of walks and the window size. Node2vec uses the same parameters as DeepWalk with the addition of the p and q. All hyperparameter ranges/values were chosen based on [69].

The minus log_{10} probability (MLP) is also reported to enable easier interpretation. Generally, a p-value of less than 0.05, corresponding to an MLP of greater than 1.3, is considered to be significant. It is worth mentioning that in this thesis, the p-value is mainly for ranking purposes, and multiple hypothesis testing correction was not employed.

4.4 Evaluation Metrics

To evaluate the performance of the predictive models, three evaluation metrics were calculated: area under the receiver operating characteristic (ROC) curve (AUC-ROC), area under the precision-recall curve (AUC-PR), and Matthews Correlation Coefficient (MCC). The best performance was selected based on the overall evaluation results. The calculations depend on the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Table 3 presents the calculations used in the evaluation metrics.

Metric	**Definition**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
True Positive Rate (recall, sensitivity)	$\frac{TP}{TP+FN}$
F_1	$2 * \frac{Precision*Recall}{Precision+Recall}$
False Positive Rate	$\frac{FP}{FP+TN}$
MCC	$\frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(FN+TN)(FP+TN)(TN+FN)}}$

Table 3: Evaluation metrics definitions.

The accuracy measure could give a good evaluation of how accurate the classifier is, however, it is not as useful if the data is imbalanced. For example, if most of the data belongs to one class, the classifier can predict that all the samples belong to that one class and still calculate a high accuracy value.

Precision is the ratio of correct positive predictions (TP) to the total positive predictions. On the other hand, recall, or sensitivity, is the ratio of correct positive predictions (TP) to all predictions in the class. In other words, while the precision of a model is its ability to return only the relevant cases in the data set, the recall is the ability of the classifier to find all relevant cases in the data set. Thus, there is a trade-off between precision and recall of a model – if the recall increases, precision decreases.

The F_1 score is defined as the balance between precision and recall. It is the harmonic mean of precision and recall and it is a good alternative to accuracy. An F_1 score close to 1 indicates low false positives and false negatives.

The MCC is a more strict measure of quality of binary classification. Since MCC uses all quantities produced from the data set (TP, TN, FP, FN), it provides a

better summary of the performance of classifier. Thus, it is typically considered a balanced and robust measure which can be used even if the data is imbalance [73].

4.4.1 The Area Under the Curve

ROC curve plots the true positive rate (recall) against the false positive rate. The AUC-ROC represents the measure of separability; the closer the AUC-ROC is to 1, the better the model is at distinguishing classes. However, the AUC-ROC can be misleading for comparing predictive distribution models [74]. Alternatively, the AUC-PR plots precision against recall to provide a more realistic evaluation on imbalanced data. The major difference between the two methods is how they account for the true negatives in the data set. If the number of negative samples is much less than the number of positive samples, then it is better to use precision-recall curve for evaluation since it does not consider the true negatives in its calculation so it will not be affected by the imbalance.

5 Results and Discussion

5.1 Data Preprocessing and Network Construction

The issue with combining different data sources is that sometimes they are inconsistent and/or incomplete, which is why data preprocessing is needed to create a sufficient network that can be used for prediction tasks. However, preprocessing of data can also mean the loss of incompatible yet important information. The main preprocessing task in this thesis was the merging of chemicals from the SIDER and DrugBank graphs, although both graphs contained chemicals that were labeled with PubChem identifiers, some chemicals can have two or more different PubChem identifier, that can contain different stereochemistry or isotopes of the same compound. This resulted in having different nodes of the same chemical from the two different databases, which meant that the model would not be able to correctly predict relations for this chemical because they are not complete. To resolve this issue, the nodes of the chemicals were merged using their canonical SMILES, which focuses on the topology of the compound and disregards its stereochemistry. However, another issue could emerge from this approach, since isomers from the same compound do not always elicit the same response, which could cause a different kind of problem in the predictions.

The network was built first using two databases, SIDER and DrugBank, which contained three types of nodes, chemicals, phenotypes, and targets, and two types of relations, chemical - phenotype, and chemical - target. However, this was not enough to build a graph that is able to predict new relations, since it would not have enough relations to learn from. This is why chemical similarity, chemical - chemical, relations were added. The 50% similarity complete graph

was very well connected, because the similarity threshold was low, thus many of the chemicals had relations with one another. On the other hand, the clustered chemicals complete graph followed a power-law degree distribution, shown in Figure 5, which is a property of scale-free network format. The average clustering coefficient of 0.165 indicates that while the nodes are not in highly clustered, the network is not random.

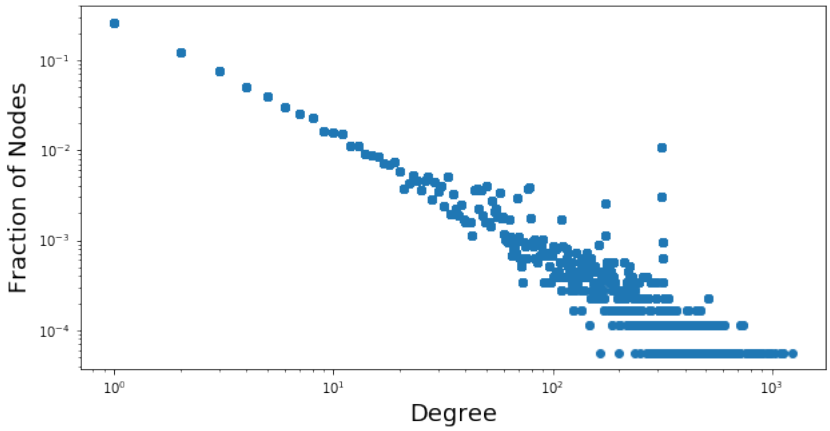


Figure 5: The degree distribution of the constructed network showed power-law distribution.

5.2 Model Evaluation and Selection

Six different NRL models (node2vec, DeepWalk, GraRep, HOPE, LINE, and SDNE) were evaluated and the approach that works best with the data set created for this thesis was selected. Translational distance-based models, TransH and TransR, were also trained and evaluated, however, they performed poorly on this kind of biological network. This might be because the network created for this thesis contains a large number of relations compared to the entities. Moreover, this network contains unbalanced relationships, e.g. many-to-one, one-to-many, and many-to-many, which might create a challenge for translational models to map n-side entities to suitable positions in the embedding space [75].

Biological networks are usually sparse, noisy, and incomplete [76], leading to challenges in identifying and understanding structures, patterns, and dynamics of the network [77]. Furthermore, they can be heterogeneous and high dimensional, making the embedding task much more complicated. Hence, understanding the

topology of the network to be analyzed is important to decide the preserved structural properties in the embeddings. In the case of biological networks, preserving both local and global structural properties seem to be the most fitting for a better analysis of the network [78]. This is true because biological networks are complex and usually contain essential information both locally and globally. Thus, the NRL models were selected based on the fact that they can capture the local and global structures of the network to insure that they perform well enough.

5.2.1 Trained Graph Selection

Both the complete graph with 50% similarity relations and the complete graph with clusters relations were tried, and both performed well. However, the graph with 50% similarities produced a lot of predictions that do not make sense, which could be related to the fact that the threshold for making an edge between two chemicals was low, creating many unnecessary relations that could contribute to an inaccurate prediction. The graph could have performed well because of its high inter-connectivity, which means that there is a high chance that an edge exists between any two given nodes. Thus, the complete graph with clusters relations was chosen for training and prediction.

All the models went through 100 trials, hyperparameters were randomly selected at each run. For each method, a set of parameters that play a role in the embeddings generation were selected to be optimized. All the models, with the exception of SDNE, have embeddings dimensions, which is an important parameter to optimize since it will contain all the low-dimensional representations of the graph. In GraRep, k-step parameter indicates length of the longest path that the embeddings will contain. The selection of order in LINE determines if the model will preserve the local structure (order=1), global structure (order=2), or both (order=3). The epochs in LINE and SDNE control the number of times the model will go through the training set. The α and β parameters determine the proximity and reconstruction weight, respectively, as mentioned previously in 2.3.3. Random walk methods (e.g., DeepWalk and node2vec) use the walk length to control length of path, number of walks to determine how many paths are learned for each node, and the window size which determines the number of nodes that are captured on either sides of the target node. Additionally, node2vec has p parameter, which determines the return probability and q parameter, which controls if the path is going inward or outward, both parameters were dicussed previously in 2.2.4.

The training set was used to generate the embeddings for each NRL model and

to train the logistic regression, then the model was evaluated using the testing set. The best trial for each model was selected based on the highest MCC score. This metric was chosen for the optimization because it is an unbiased and balanced measure that works on both balanced and unbalanced data sets, thus it is a more stable measure of performance.

Table 4 presents the parameters and evaluation metrics for the best trial of each method. All models, with the exception of SDNE, performed fairly well, with AUC-ROC and AUC-PR of above 0.9, and MCC of above 0.8. However, the model that performed the best was node2vec with AUC-ROC of 0.977, AUC-PR of 0.981, and MCC of 0.877, which is why it was chosen to be the model used for prediction. The node2vec model worked best for this data set because of the hyperparameters selected. A window size of 4 and walk length of 8 were big enough to include nearby nodes and paths to learn and predict nodes that are not directly related (target - phenotype associations), but not too big as to include unnecessary relations. The number of walks made sure to learn the paths between nodes without overfitting the network so new predictions can be made. Furthermore, high p and q parameters ensured that the model would avoid revisting nodes yet it would be biased toward nodes that are closer to the previous node, which helped in capturing the local and global structures of the network.

To check the randomness and robustness of the models, the training and evaluation of each model was repeated ten times. Figure 6 shows the MCC results of each of the models. These results, with the exception of SDNE, indicate that the models performances are not random and are quite robust.

5.3 Interpretation of Model Predictions

The node2vec embeddings and logistic regression model were trained using the complete graph and exported to be used for predictions. The predictions were created by enquiring the name or identifier of a certain entity. Furthermore, the type of entities to be predicted could also be specified. Three types of relations could be predicted using the model: drug-phenotype, drug-target, and target-phenotype. For each relation predicted, the p-value and MLP value were calculated and used for ranking. Since methods for drug-target associations, or drug target identification, are well developed, those association results were omitted from the interpretations. To validate the new predictions of the model, positive controls were also presented for drug-phenotype associations.

Method	Parameters	Value	AUC-ROC	AUC-PR	MCC
node2vec	Dimensions	300	**0.977**	**0.981**	**0.877**
	Walk length	8			
	Number of walks	8			
	Window size	4			
	Return parameter (p)	2.3			
	In/out parameter (q)	1.9			
DeepWalk	Dimensions	300	0.969	0.974	0.846
	Walk length	8			
	Number of walks	8			
	Window size	2			
HOPE	Dimensions	300	0.937	0.962	0.842
GraRep	Dimensions	300	0.977	0.981	0.866
	k-step	3			
LINE	Dimensions	300	0.979	0.983	0.869
	Proximity order	3			
	Epochs	5			
SDNE	Proximity balance (α)	0.128	0.927	0.949	0.648
	Reconstruction weight (β)	14			
	Epochs	25			

Table 4: Evaluation results for the best model in each NRL method. The model with the best evaluation result is node2vec with AUC-ROC of 0.977, AUC-PR of 0.981, and MCC of 0.877

5.3.1 Predicting the Phenotypes for a Drug

Predicting indications and side effects of a chemical is one of the most common types of tasks used with side effects networks. Table 5 presents the top ten predicted phenotypes for the antipsychotic drug, olanzapine.

Olanzapine is an atypical antipsychotic drug that primarily acts on dopamine and serotonin receptors, and is used to treat schizophrenia and bipolar disorders [79]. Several case reports have presented cardiovascular problems that are associated with olanzapine treatment, one report mentioned that a patient treated with olanzapine experienced bundle branch block, which is a blockage or delay in electrical impulses of the heart [80]. Another case report presented a patient that suffered from cardiomyopathy after being treated with olanzapine [81]. Both phenotypes have been predicted with the model, and Figure 7 shows a subgraph representing some of the paths that are present in the network between olanzapine and three different predicted phenotypes (bundle branch block, cardiomyopathy,

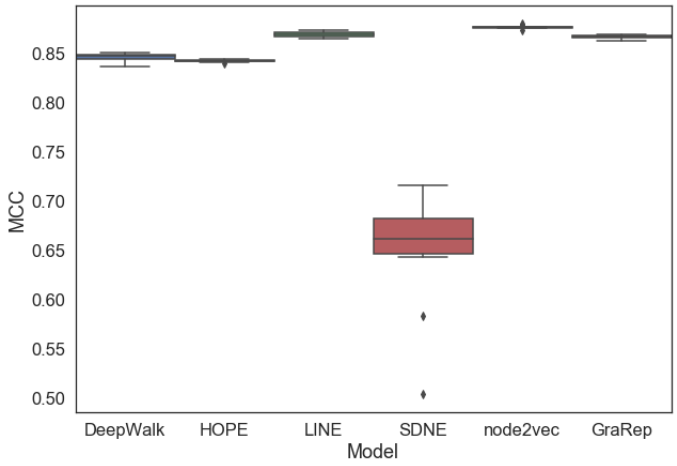


Figure 6: A box plot of the MCC distribution of ten trainings using the best hyperparameters for each NRL model shows their robustness to random sampling.

Namespace	Identifier	Name	p-value	MLP
umls	C0006384	Bundle branch block	0.000	3.475
umls	C0575090	Balance disorder	0.000	3.539
umls	C0878544	Cardiomyopathy	0.001	2.918
umls	C0233794	Memory impairment	0.001	2.870
umls	C0004239	Atrial flutter	0.001	3.297
umls	C0160390	Liver injury	0.001	3.175
umls	C0020676	Hypothyroidism	0.001	2.911
umls	C0002884	Hypochromic anaemia	0.001	3.059
umls	C0034069	Pulmonary fibrosis	0.001	2.901
umls	C0233477	Dysphoria	0.001	3.066

Table 5: Top phenotypic predictions for olanzapine

and balance disorder), which indicate that the predictions could have been made from side-effect similarities.

Studies have been done to investigate the effect of olanzapine, among other atypical antipsychotics, on cognitive function, including attention, memory, and verbal learning and they have confirmed that olanzapine improves those cognitive functions in schizophrenia patients [82–86]. The effect of antipsychotic drugs on the liver has also been studied and it was found that olanzapine can induce hepatic damage [87]. Both association with memory and liver injury were also

Namespace	Identifier	Name	p-value	MLP
umls	C1320474	Nuchal rigidity	0.0	3.318
umls	C0002453	Amenorrhoea	0.0	3.619
umls	C0011168	Dysphagia	0.0	3.408
umls	C0033117	Priapism	0.0	3.515

Table 6: Positive control for olanzapine phenotypic predictions

predicted using the model. Positive controls of olnazapine-phenotype associations are presented in Table 6, those four phenotypes (nuchal rigidity, amenorrhoea, dysphagia, and priapism) are all present in SIDER as side effects of olanzapine.

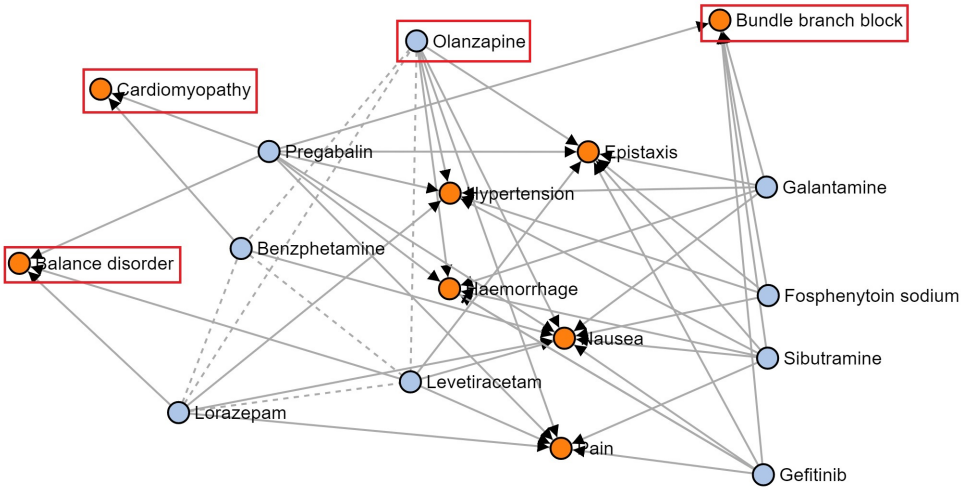


Figure 7: A subgraph showing some of the shortest relation paths between olanzapine and the top three phenotypic predictions

5.3.2 Predicting the Drugs for a Phenotype

The model can also be used to predict chemicals that can be associated with phenotypes. Here, it used to predict chemicals that might best affect Parkinson's disease (PD), a progressive neurodegenerative disorder that affects motor and non-motor functions in variable degrees [88]. Common motor features of PD are tremors, rigidity, slowness (bradykinesia), and impaired balance. Other PD symptoms include cognitive impairment and abnormal neurological behaviors [88]. The top ten predictions of chemicals associated with PD are shown in Table 7.

Namespace	Identifier	Name	p-value	MLP
pubchem.compound	146570	Escitalopram	0.000	3.315
pubchem.compound	5002	Quetiapine	0.001	2.883
pubchem.compound	5486971	Pregabalin	0.001	3.111
pubchem.compound	68617	Sertraline	0.002	2.763
pubchem.compound	5719	Zaleplon	0.002	2.807
pubchem.compound	60853	Ziprasidone HCL	0.002	2.613
pubchem.compound	3345	Fentanyl	0.002	2.630
pubchem.compound	5210	Sibutramine	0.002	2.613
pubchem.compound	44602	Arbaclofen	0.003	2.554
pubchem.compound	154101	Dexmethylphenidate	0.003	2.475

Table 7: Top chemicals predictions for Parkinson's disease

The prediction with the highest significance is escitalopram, a selective serotonin reuptake inhibitor (SSRI) and an *S*-enantiomer of citalopram that is used to treat major depression [89]. Depression is a common complication of PD; approximately 20% to 40% of PD patients have depression and usually the antidepressant treatment of choice is an SSRI [89]. Though the efficacy of SSRIs on depression in PD has not been proven, some studies have found that they could be beneficial [90–93]. One of those studies investigated sertraline, which was also predicted by the model, and found it to be useful in treating depression in PD [91]. Another study evaluated citalopram and found that it can improve the depression symptoms of PD patients [93]. Escitalopram was also investigated and it was concluded that it may be a viable treatment for depression in PD, however more research needs to be conducted to confirm [89, 94]. Escitalopram had many paths leading to its association with PD, these paths were mostly created from side-effect similarities and target similarities between escitalopram and chemicals that are directly associated with PD. Figure 8 shows some of the paths that were found in the network.

Quetiapine is an antipsychotic drug that has been used to treat schizophrenia and bipolar disorders. It is also used in off-label cases such as post-traumatic stress disorder, anxiety disorders, insomnia, and depression in PD [95], since SIDER is curated from drug labels, this kind of off-label use would not be in the training set. Many researches have investigated quetiapine for treating psychotic symptoms in PD, which include delirium, hallucinations, depression, and insomnia among other psychiatric manifestations [96]. Even though those studies have not been able to prove the efficacy of quetiapine on PD, they suggest that the high dropout rate might have influenced the results and follow-up studies with larger sample size are required. Figure 9 shows some of the paths in the network that were

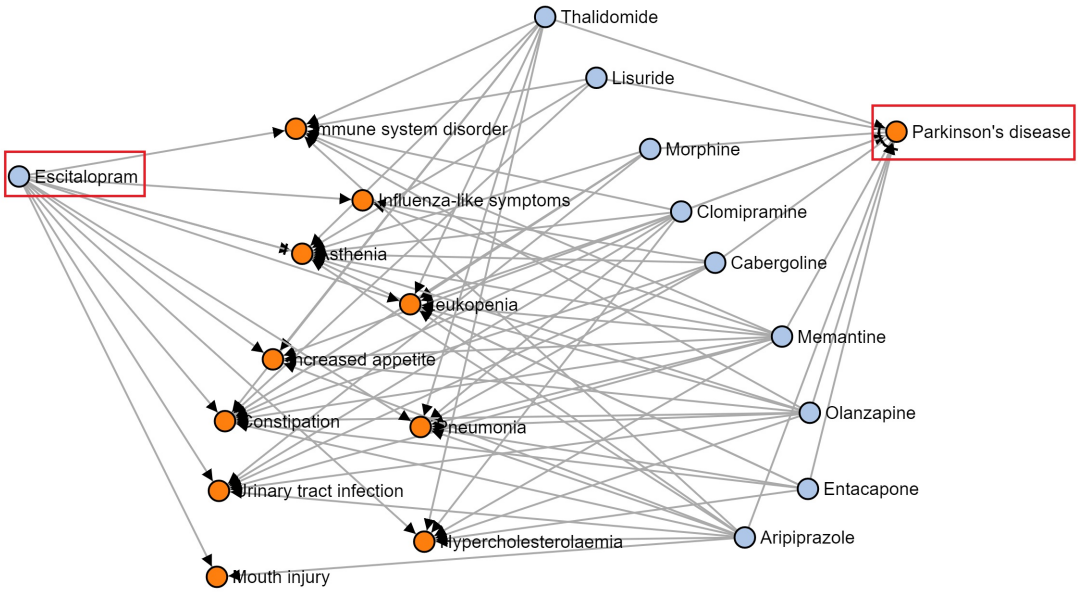


Figure 8: A subgraph showing some of the shortest relation paths between escitalopram and PD

used to predict the association of quetiapine with PD. It is evident from the figure that quetiapine share targets and chemical similarities with two chemicals, bromocriptine and memantine, that are directly associated with PD. It is worth mentioning that the prediction of association between quetiapine and PD was quite consistent, appearing in the top 30 predictions with different predictive models (not shown).

Two high-scoring chemicals chemicals, pregabalin and ziprasidone, have been shown to cause or worsen symptoms of PD [97, 98], while the rest of the predicted chemicals do not have any studies that prove their association with PD. The positive controls for this association are shown in Table 8. These four drugs (i.e., selegiline, aripiprazole, ropinirole, and clomipramine) are already used for treating PD and are indicated as such in DrugBank, which is why their association with the disease exists in the network and they are expected to be predicted by the model.

5.3.3 Predicting the Phenotypes for a Target

Another way to use the model is to predict the association of targets with phenotypes. Since there are no direct relations between targets and phenotypes in

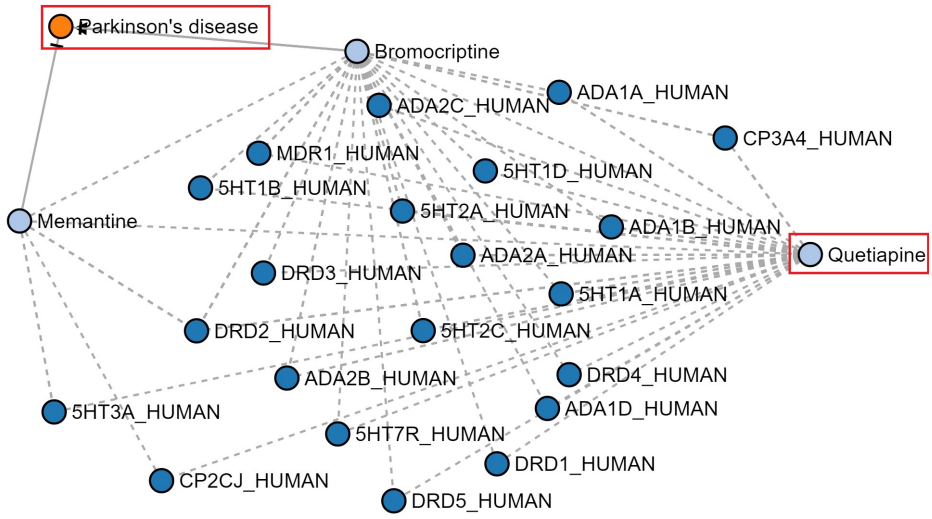


Figure 9: A subgraph showing the shortest relation paths between quetiapine and PD.

Namespace	Identifier	Name	p-value	MLP
pubchem.compound	26757	Selegiline	0.001	2.831
pubchem.compound	60795	Aripiprazole	0.001	2.856
pubchem.compound	5095	Ropinirole	0.001	2.997
pubchem.compound	2801	Clomipramine	0.002	2.823

Table 8: Positive control for chemicals predicted for Parkinson's disease.

the network, the model depends on the indirect chemical-chemical, chemical-phenotype, and chemical-target relations to predict target-phenotype relations. Table 9 presents predicted phenotypes that are associated with muscarinic 2 cholinergic receptor M2 (M2R).

M2R (uniprot entry name: ACM2_HUMAN), encoded by CHRM2 gene, is a receptor belonging to the muscarinic receptors subclass, which contains 5 subtypes (M1-M5) [99]. These receptors are responsible for recognizing the neurotransmitter acetylcholine and are involved in the cholinergic transduction in the central nervous system, basal ganglia, smooth muscles, and other parasympathetic end organs [100]. A recent study has investigated the role of muscarinic receptors in tardive dyskinesia (TD) and concluded that there is an association between variations of CHRM2 and TD [101]. CHRM2 has also been associated with psychiatric and mood disorders such as schizophrenia and depression [102–104]. A common symptom of mood disorders is agitation, which is one of the phenotypes predicted by the model, furthermore, it has been shown that the inhibition of acetylcholine

Namespace	Identifier	Name	*p*-value	MLP
umls	C0013384	Dyskinesia	0.009	2.070
umls	C0015371	Extrapyramidal disorder	0.013	1.889
umls	C0026837	Muscle rigidity	0.015	1.811
umls	C0234133	Extrapyramidal symptoms	0.022	1.649
umls	C0026961	Mydriasis	0.023	1.639
umls	C0013144	Drowsiness	0.025	1.605
umls	C0085631	Agitation	0.025	1.597
umls	C0686347	Tardive dyskinesia	0.038	1.419
umls	C0242422	Parkinsonism	0.044	1.356
umls	C0235063	Respiratory depression	0.053	1.273

Table 9: Top phenotypic predictions for M2R

could cause agitation [105]. Figure 10 shows a subgraph containing the shortest paths from three different phenotypes (agitation, drowsiness, and TD) to M2R, the paths mostly depend on target and phenotype similarities between drugs, under the assumption that if drugs with the same target have the same phenotype, the target could be the cause of the phenotype.

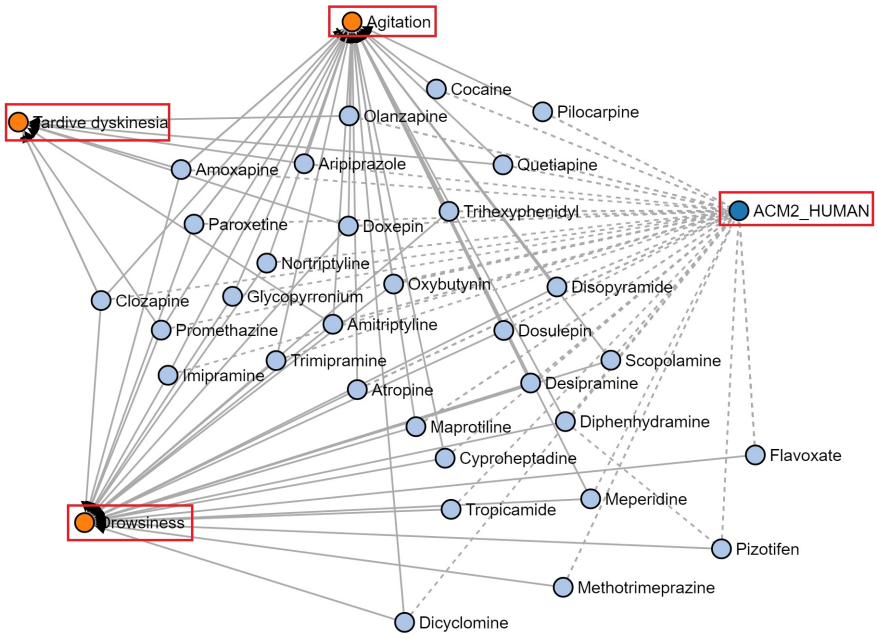


Figure 10: A subgraph of the shortest paths between M2R (ACM2_HUMAN) and agitation, drowsiness, and TD.

5.4 Reproducibility and Software Implementation

The scripts and workflows developed in this thesis are available through the seffnet Python package through GitHub at `https://github.com/seffnet/seffnet`. Each of its components have been wrapped in a command line interface (CLI) such that the results presented in each section of this work (construction of the network, hyper-parameter optimization, prediction) can be generated with a corresponding command following the guidelines described by Grüning *et al.* [106]. The seffnet Python package has a tool chain consisting of flake8 (`https://github.com/PyCQA/flake8`) to enforce code and documentation quality, setuptools (`https://github.com/pypa/setuptools`) to build distributions, pyroma (`https://github.com/regebro/pyroma`) to enforce package metadata standards and tox (`https://github.com/tox-dev/tox`) as a build tool to facilitate the usage of each of these tools in a reproducible way. It leverages community and open source resources to improve its usability by using Travis-CI (`https://travis-ci.com`) as a continuous integration service.

5.5 SEffNet: A Web Application for Link Prediction

The best machine learning model from the workflow was wrapped with a web application using the Flask Python package (`https://github.com/pallets/flask`). It allows users to enter the entity of interest and the types of predictions they want to see (Figure 11) then lists the top results (Figure 12).

Because the web application relies on a logistic regression model, predictions are nearly instantaneous. Additionally, the web application also includes an application programming interface (API) that can be used programmatically and incorporated as a microservice in other workflows.

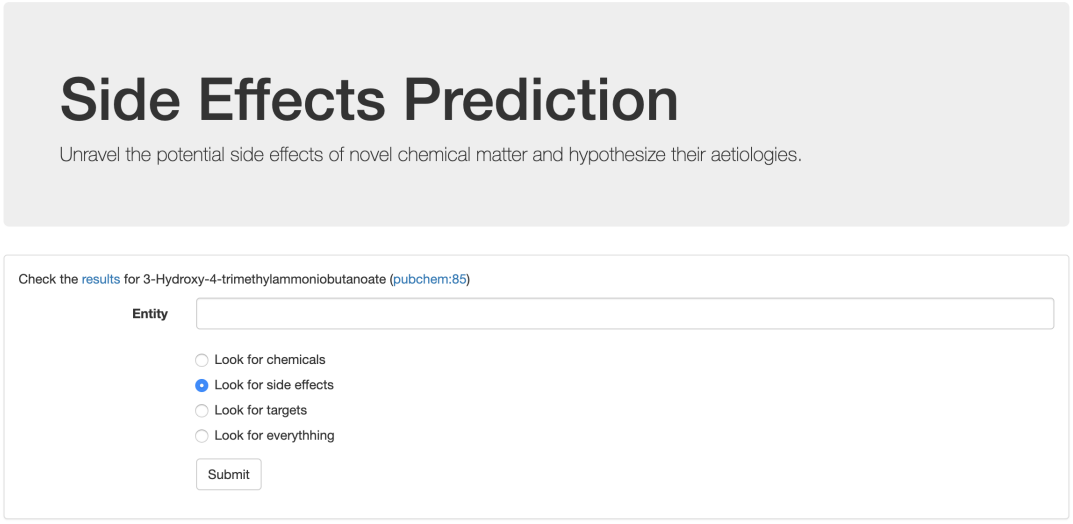
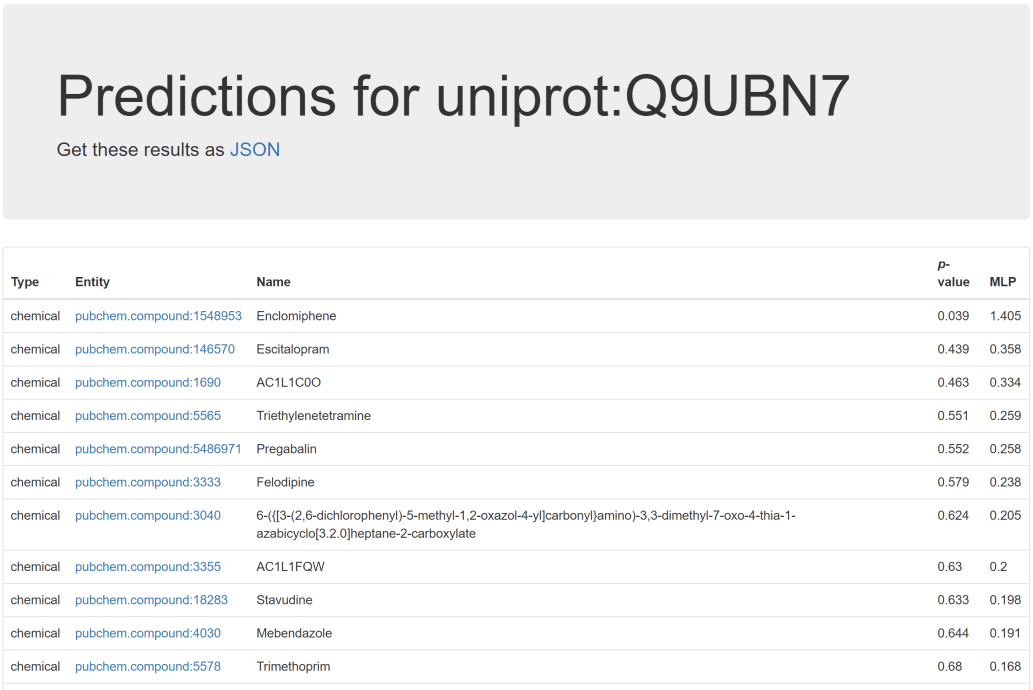


Figure 11: The landing page for the Side Effects Prediction web application.



Type	Entity	Name	*p*-value	MLP
chemical	pubchem.compound:1548953	Enclomiphene	0.039	1.405
chemical	pubchem.compound:146570	Escitalopram	0.439	0.358
chemical	pubchem.compound:1690	AC1L1C0O	0.463	0.334
chemical	pubchem.compound:5565	Triethylenetetramine	0.551	0.259
chemical	pubchem.compound:5486971	Pregabalin	0.552	0.258
chemical	pubchem.compound:3333	Felodipine	0.579	0.238
chemical	pubchem.compound:3040	6-({[3-(2,6-dichlorophenyl)-5-methyl-1,2-oxazol-4-yl]carbonyl}amino)-3,3-dimethyl-7-oxo-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylate	0.624	0.205
chemical	pubchem.compound:3355	AC1L1FQW	0.63	0.2
chemical	pubchem.compound:18283	Stavudine	0.633	0.198
chemical	pubchem.compound:4030	Mebendazole	0.644	0.191
chemical	pubchem.compound:5578	Trimethoprim	0.68	0.168

Figure 12: The predictions page for chemicals that might interact with HDAC6 (uniprot:Q9UBN7), a target of interest in the Human Brain Pharmacome project[1].

6 Conclusion and Future Work

This thesis has demonstrated how a chemical-target-phenotype network can be used in the prediction of indications, side effects, and in the analysis of drugs' mechanisms of action.

6.1 Reflections

Constructing the network needed several preprocessing steps since some of PubChem compounds have different identifiers, which resulted in having duplicate nodes in the network. Because NRL models were already implemented in the BioNEV Python package, using them did not pose any difficulty. However, some modifications were necessary such as the inclusion of reports on additional evaluation metrics (e.g. AUC-PR and MCC). Though the logistic regression classifier performed well in predictions, it would have been interesting to use other well-used classifiers like support vector machines as well as to compare their performances.

6.2 Limitations

Although the resulting predictive model has presented valuable associations between different entities, it has several limitations. One is that there is no directionality or polarity to the edges, which means that the model cannot differentiate

between causality, association, positive correlations, or negative correlations. Moreover, it cannot differentiate between indications and side effects, which means it is not able to tell if the association between a chemical and phenotype is a treatment or an effect of the chemical. Another challenge the machine learning algorithms presented is that they are not formulated as online algorithms, and therefore cannot be easily updated without re-training - if the underlying data set is updated, the whole training process needs to be repeated.

6.3 Future Work

Though the prediction model was able to perform well, there is room for improvement. First, the predictive model can be further enriched by incorporating new data modalities, such as target-target interactions. Second, the implementation of the models embeddings could be enhanced by only training parts of the network at a time, this could be used as a way to create parallel implementation, or as a way to learn newly incorporated data modalities without re-training the whole network. Third, weighted edges could be incorporated to assign importance to edges depending on their significance, this could be especially useful in chemical-chemical similarity association, where an edge between two chemicals that share 60% similarity will have more significance that an edge between two chemicals with 50% similarity, for example. Another example is adding weighted edges between chemical-phenotype, such as more frequent phenotype would be more significant than infrequent or rare phenotypes. Finally, OpenTargets, a database that provides evidence for target-disease associations, could be used to validate target-phenotype relations that are predicted by the model. A further filtering step could be done after the prediction, where the literature co-occurrence of a given pair in predicted relation is calculated and the top predictions are ranked based on the highest co-occurrence frequency.

This thesis has taken the first steps towards a reproducible workflow for the application of network representation in biomedical networks that might be useful for downstream tasks such as prediction, classification and clustering. The thesis's original goal was to create and train a network that can be analyzed to understand drugs' mechanism of action and be able to use that knowledge in predicting new relations, which was achieved with this workflow.

Bibliography

[1] John Arrowsmith. "Phase II failures: 2008–2010". en. In: *Nature Reviews Drug Discovery* 10.5 (May 2011), pp. 328–329.

[2] John Arrowsmith and Philip Miller. "Phase II and Phase III attrition rates 2011–2012". en. In: *Nature Reviews Drug Discovery* 12.8 (Aug. 2013), pp. 569–569.

[3] Fei Wang, Ping Zhang, Nan Cao, Jianying Hu, and Robert Sorrentino. "Exploring the associations between drug side-effects and therapeutic indications". In: *Journal of Biomedical Informatics* 51 (2014), pp. 15–23.

[4] Ted T. Ashburn and Karl B. Thor. "Drug repositioning: identifying and developing new uses for existing drugs". en. In: *Nature Reviews Drug Discovery* 3.8 (Aug. 2004), pp. 673–683.

[5] Humphrey P. Rang, James M. Ritter, Rod J. Flower, and Graeme Henderson. *Rang & Dale's Pharmacology E-Book: with STUDENT CONSULT Online Access*. en. Google-Books-ID: iOLTBQAAQBAJ. Elsevier Health Sciences, Dec. 2014.

[6] R. S. Satoskar and S. D. Bhandarkar &nirmala N. Rege. *Pharmacology and Pharmacotherapeutics*. en. Google-Books-ID: 7d493VOD4P8C. Popular Prakashan, 1973.

[7] Isabella Gashaw, Peter Ellinghaus, Anette Sommer, and Khusru Asadullah. "What makes a good drug target?" en. In: *Drug Discovery Today* 16.23-24 (Dec. 2011), pp. 1037–1043.

[8] Zahra Pourpak, Mohammad R. Fazlollahi, and Fatemeh Fattahi. "Understanding adverse drug reactions and drug allergies: principles, diagnosis

and treatment aspects". eng. In: *Recent Patents on Inflammation & Allergy Drug Discovery* 2.1 (Jan. 2008), pp. 24–46.

[9] E. Baralis and A. Fiori. "Exploring Heterogeneous Biological Data Sources". In: *2008 19th International Workshop on Database and Expert Systems Applications*. Sept. 2008, pp. 647–651.

[10] "UniProt: a worldwide hub of protein knowledge". In: *Nucleic Acids Research* 47.Database issue (Jan. 2019), pp. D506–D515.

[11] Richard J. Roberts. "PubMed Central: The GenBank of the published literature". In: *Proceedings of the National Academy of Sciences of the United States of America* 98.2 (Jan. 2001), pp. 381–382.

[12] "The Gene Ontology project in 2008". In: *Nucleic Acids Research* 36.Database issue (Jan. 2008), pp. D440–D444.

[13] David S Wishart et al. "DrugBank 5.0: a major update to the DrugBank database for 2018". en. In: *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D1074–D1082.

[14] Sunghwan Kim et al. "PubChem Substance and Compound databases". eng. In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D1202–1213.

[15] Sunghwan Kim et al. "PubChem 2019 update: improved access to chemical data". eng. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D1102–D1109.

[16] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. "A side effect resource to capture phenotypic effects of drugs". In: *Molecular Systems Biology* 6 (Jan. 2010).

[17] Olivier Bodenreider. "The Unified Medical Language System (UMLS): integrating biomedical terminology". In: *Nucleic Acids Research* 32.Database issue (Jan. 2004), pp. D267–D270.

[18] Minlie Huang, Aurélie Névéol, and Zhiyong Lu. "Recommending MeSH terms for annotating biomedical articles". In: *Journal of the American Medical Informatics Association : JAMIA* 18.5 (2011), pp. 660–667.

[19] *Introduction to MeSH*. eng. Technical Documentation.

[20] Joanna S. Amberger, Carol A. Bocchini, François Schiettecatte, Alan F. Scott, and Ada Hamosh. "OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders". eng. In: *Nucleic Acids Research* 43.Database issue (Jan. 2015), pp. D789–798.

[21] C. Rosse et al. "Motivation and organizational principles for anatomical knowledge representation: the digital anatomist symbolic knowledge base". eng. In: *Journal of the American Medical Informatics Association: JAMIA* 5.1 (Feb. 1998), pp. 17–40.

[22] Josef Scheiber et al. "Mapping Adverse Drug Reactions in Chemical Space". en. In: *Journal of Medicinal Chemistry* 52.9 (May 2009), pp. 3103–3107.

[23] Giovanna Maria Dimitri and Pietro Lió. "DrugClust: A machine learning approach for drugs side effects prediction". en. In: *Computational Biology and Chemistry* 68 (June 2017), pp. 204–210.

[24] S. Mizutani, E. Pauwels, V. Stoven, S. Goto, and Y. Yamanishi. "Relating drug-protein interaction network with drug side effects". en. In: *Bioinformatics* 28.18 (Sept. 2012), pp. i522–i528.

[25] Nir Atias and Roded Sharan. "An Algorithmic Framework for Predicting Side Effects of Drugs". en. In: *Journal of Computational Biology* 18.3 (Mar. 2011), pp. 207–218.

[26] Edouard Pauwels, Véronique Stoven, and Yoshihiro Yamanishi. "Predicting drug side-effect profiles: a chemical fragment-based approach". en. In: *BMC Bioinformatics* 12.1 (2011), p. 169.

[27] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. "Drug target identification using side-effect similarity". eng. In: *Science (New York, N.Y.)* 321.5886 (July 2008), pp. 263–266.

[28] Maryam Lotfi Shahreza, Nasser Ghadiri, Sayed Rasoul Mousavi, Jaleh Varshosaz, and James R. Green. "A review of network-based approaches to drug repositioning". eng. In: *Briefings in Bioinformatics* 19.5 (2018), pp. 878–892.

[29] Neil Vargesson. "Thalidomide-induced teratogenesis: History and mechanisms". In: *Birth Defects Research* 105.2 (June 2015), pp. 140–156.

[30] M. Kuhn et al. “Systematic identification of proteins that elicit drug side effects”. en. In: *Molecular Systems Biology* 9.1 (Apr. 2014), pp. 663–663.

[31] Lun Yang and Pankaj Agarwal. “Systematic Drug Repositioning Based on Clinical Side-Effects”. en. In: *PLoS ONE* 6.12 (Dec. 2011). Ed. by Peter Csermely, e28025.

[32] Hao Ye, Qi Liu, and Jia Wei. “Construction of Drug Network Based on Side Effects and Its Application for Drug Repositioning”. en. In: *PLoS ONE* 9.2 (Feb. 2014). Ed. by Ozlem Keskin, e87864.

[33] Hongyun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang. “A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications”. In: *arXiv:1709.07604 [cs]* (Sept. 2017). arXiv: 1709.07604.

[34] Palash Goyal and Emilio Ferrara. “Graph Embedding Techniques, Applications, and Performance: A Survey”. In: *Knowledge-Based Systems* 151 (July 2018). arXiv: 1705.02801, pp. 78–94.

[35] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. “Network Representation Learning: A Survey”. In: *arXiv:1801.05852 [cs, stat]* (Dec. 2017). arXiv: 1801.05852.

[36] Nasrullah Sheikh, Zekarias Kefato, and Alberto Montresor. “gat2vec: representation learning for attributed graphs”. en. In: *Computing* (Apr. 2018).

[37] Vlado Dančík, Amrita Basu, and Paul Clemons. “Properties of Biological Networks”. en. In: *Systems Biology: Integrative Biology and Simulation Tools*. Ed. by Aleš Prokop and Béla Csukás. Dordrecht: Springer Netherlands, 2013, pp. 129–178.

[38] Xiaowei Zhu, Mark Gerstein, and Michael Snyder. “Getting connected: analysis and principles of biological networks”. eng. In: *Genes & Development* 21.9 (May 2007), pp. 1010–1024.

[39] Anna D. Broido and Aaron Clauset. “Scale-free networks are rare”. en. In: *Nature Communications* 10.1 (Mar. 2019), pp. 1–10.

[40] Mark E. J. Newman. “Random Graphs as Models of Networks”. In: 2002.

[41] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. “DeepWalk: online learning of social representations”. en. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’14*. New York, New York, USA: ACM Press, 2014, pp. 701–710.

[42] Chris McCormick. *Word2Vec Tutorial - The Skip-Gram Model*.

[43] Jeffrey Pennington, Richard Socher, and Christopher Manning. “Glove: Global Vectors for Word Representation”. en. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543.

[44] Aditya Grover and Jure Leskovec. “node2vec: Scalable Feature Learning for Networks”. en. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16*. San Francisco, California, USA: ACM Press, 2016, pp. 855–864.

[45] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. “Knowledge Graph Embedding: A Survey of Approaches and Applications”. In: *IEEE Transactions on Knowledge and Data Engineering* 29.12 (Dec. 2017), pp. 2724–2743.

[46] Ibrahim Abdelaziz, Achille Fokoue, Oktie Hassanzadeh, Ping Zhang, and Mohammad Sadoghi. “Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions”. en. In: *Journal of Web Semantics* 44 (May 2017), pp. 104–117.

[47] Meng Wang. “Predicting Rich Drug-Drug Interactions via Biomedical Knowledge Graphs and Text Jointly Embedding”. In: *arXiv:1712.08875 [cs]* (Dec. 2017). arXiv: 1712.08875.

[48] Stan Zhao et al. “ContextCare: Incorporating Contextual Information Networks to Representation Learning on Medical Forum Data”. en. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. Melbourne, Australia: International Joint Conferences on Artificial Intelligence Organization, Aug. 2017, pp. 3497–3503.

[49] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. "Translating Embeddings for Modeling Multi-relational Data". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. event-place: Lake Tahoe, Nevada. USA: Curran Associates Inc., 2013, pp. 2787–2795.

[50] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zhigang Chen. "Knowledge Graph Embedding by Translating on Hyperplanes". In: *AAAI*. 2014.

[51] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. "Learning Entity and Relation Embeddings for Knowledge Graph Completion". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI'15. event-place: Austin, Texas. AAAI Press, 2015, pp. 2181–2187.

[52] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. "Knowledge Graph Embedding via Dynamic Mapping Matrix". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 687–696.

[53] Ali Ezzat, Peilin Zhao, Min Wu, Xiao-Li Li, and Chee-Keong Kwoh. "Drug-Target Interaction Prediction with Graph Regularized Matrix Factorization". eng. In: *IEEE/ACM transactions on computational biology and bioinformatics* 14.3 (June 2017), pp. 646–656.

[54] Yoshihiro Yamanishi et al. "DINIES: drug–target interaction network inference engine based on supervised analysis". In: *Nucleic Acids Research* 42.Web Server issue (July 2014), W39–W45.

[55] "Probability-based collaborative filtering model for predicting gene–disease associations". In: 10 ().

[56] Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. "Asymmetric Transitivity Preserving Graph Embedding". In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. event-place: San Francisco, California, USA. New York, NY, USA: ACM, 2016, pp. 1105–1114.

[57] Shaosheng Cao, Wei Lu, and Qiongkai Xu. “GraRep: Learning Graph Representations with Global Structural Information”. In: *CIKM*. 2015.

[58] Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. “A Survey on Network Embedding”. In: *arXiv:1711.08752 [cs]* (Nov. 2017). arXiv: 1711.08752.

[59] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. “Risk Prediction with Electronic Health Records: A Deep Learning Approach”. en. In: *Proceedings of the 2016 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, June 2016, pp. 432–440.

[60] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. “Modeling polypharmacy side effects with graph convolutional networks”. en. In: *Bioinformatics* 34.13 (July 2018), pp. i457–i466.

[61] Jian Tang et al. “LINE: Large-scale Information Network Embedding”. en. In: *Proceedings of the 24th International Conference on World Wide Web - WWW '15*. Florence, Italy: ACM Press, 2015, pp. 1067–1077.

[62] Daixin Wang, Peng Cui, and Wenwu Zhu. “Structural Deep Network Embedding”. en. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. San Francisco, California, USA: ACM Press, 2016, pp. 1225–1234.

[63] Özlem Muslu, Charles Tapley Hoyt, Martin Hofmann-Apitius, and Holger Fröhlich. “GuiltyTargets: Prioritization of Novel Therapeutic Targets with Deep Network Representation Learning”. en. In: *bioRxiv* (Jan. 2019), p. 521161.

[64] William L Hamilton, Rex Ying, and Jure Leskovec. “Representation Learning on Graphs: Methods and Applications”. en. In: (), p. 23.

[65] Charles Tapley Hoyt et al. “Integration of Structured Biological Data Sources using Biological Expression Language”. en. In: *bioRxiv* (May 2019), p. 631812.

[66] Charles Tapley Hoyt. *bio2bel/sider v0.0.1*. Dec. 2018.

[67] Charles Tapley Hoyt. *bio2bel/drugbank v0.0.1*. May 2018.

[68] Darko Butina. “Unsupervised Data Base Clustering Based on Daylight’s Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets”. en. In: *Journal of Chemical Information and Computer Sciences* 39.4 (July 1999), pp. 747–750.

[69] Xiang Yue et al. “Graph Embedding on Biomedical Networks: Methods, Applications, and Evaluations”. en. In: *arXiv:1906.05017 [cs]* (June 2019). arXiv: 1906.05017.

[70] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, and Emmanuel Müller. “VERSE: Versatile Graph Embeddings from Similarity Measures”. en. In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW ’18*. Lyon, France: ACM Press, 2018, pp. 539–548.

[71] Shuhan Yuan, Xintao Wu, and Yang Xiang. “SNE: Signed Network Embedding”. In: *arXiv:1703.04837 [cs]* (Mar. 2017). arXiv: 1703.04837.

[72] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. “Optuna: A Next-generation Hyperparameter Optimization Framework”. In: *arXiv:1907.10902 [cs, stat]* (July 2019). arXiv: 1907.10902.

[73] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. “Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric”. en. In: *PLOS ONE* 12.6 (June 2017). Ed. by Quan Zou, e0177678.

[74] Jorge M. Lobo, Alberto Jiménez-Valverde, and Raimundo Real. *AUC: a misleading measure of the performance of predictive distribution models*. en. Mar. 2008.

[75] Xiaomin Liang et al. “Predicting biomedical relationships using the knowledge and graph embedding cascade model”. In: *PLoS ONE* 14.6 (June 2019).

[76] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. “A Review of Relational Machine Learning for Knowledge Graphs”. en. In: *Proceedings of the IEEE* 104.1 (Jan. 2016). arXiv: 1503.00759, pp. 11–33.

[77] Bing Wang et al. “Unsupervised Learning from Noisy Networks with Applications to Hi-C Data”. In: *NIPS*. 2016.

[78] Chang Su, Jie Tong, Yongjun Zhu, Peng Cui, and Fei Wang. “Network embedding in biomedical data science”. en. In: *Briefings in Bioinformatics* (Dec. 2018).

[79] Kristina Thomas and Abdolreza Saadabadi. “Olanzapine”. eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2019.

[80] Sheeba Ninan, David Hinchcliffe, and Samuel Amo-Korankye. “A case report of olanzapine and related cardiac conduction changes”. en. In: *Progress in Neurology and Psychiatry* 21.4 (Oct. 2017), pp. 13–16.

[81] Beeresha Puttegowda, Joseph Theodore, Ramesh Basappa, and Manjunath Cholenally Nanjappa. “Olanzapine Induced Dilated Cardiomyopathy”. In: *The Malaysian Journal of Medical Sciences : MJMS* 23.2 (Mar. 2016), pp. 82–84.

[82] Susan R McGurk, M.A Lee, K Jayathilake, and Herbert Y Meltzer. “Cognitive Effects of Olanzapine Treatment in Schizophrenia”. In: *Medscape General Medicine* 6.2 (May 2004).

[83] R. C. Smith, M. Infante, A. Singh, and A. Khandat. “The effects of olanzapine on neurocognitive functioning in medication-refractory schizophrenia”. eng. In: *The International Journal of Neuropsychopharmacology* 4.3 (Sept. 2001), pp. 239–250.

[84] M. J. Cuesta, V. Peralta, and A. Zarzuela. “Effects of olanzapine and other antipsychotics on cognitive function in chronic schizophrenia: a longitudinal study”. eng. In: *Schizophrenia Research* 48.1 (Mar. 2001), pp. 17–28.

[85] S. E. Purdon et al. “Neuropsychological change in early phase schizophrenia during 12 months of treatment with olanzapine, risperidone, or haloperidol. The Canadian Collaborative Group for research in schizophrenia”. eng. In: *Archives of General Psychiatry* 57.3 (Mar. 2000), pp. 249–258.

[86] Robert M. Bilder et al. “Neurocognitive effects of clozapine, olanzapine, risperidone, and haloperidol in patients with chronic schizophrenia or schizoaffective disorder”. eng. In: *The American Journal of Psychiatry* 159.6 (June 2002), pp. 1018–1028.

[87] Qinyu Lv and Zhenghui Yi. “Antipsychotic Drugs and Liver Injury”. eng. In: *Shanghai Archives of Psychiatry* 30.1 (Feb. 2018), pp. 47–51.

[88] J Jankovic. “Parkinson’s disease: clinical features and diagnosis”. en. In: *Journal of Neurology, Neurosurgery & Psychiatry* 79.4 (Apr. 2008), pp. 368–376.

[89] Daniel Weintraub et al. “Escitalopram for Major Depression in Parkinson’s Disease: An Open-Label, Flexible-Dosage Study”. In: *The Journal of neuropsychiatry and clinical neurosciences* 18.3 (2006), pp. 377–383.

[90] Liborio Rampello, Santina Chiechio, Rocco Raffaele, Ignazio Vecchio, and Francesco Nicoletti. “The SSRI, citalopram, improves bradykinesia in patients with Parkinson’s disease treated with L-dopa”. eng. In: *Clinical Neuropharmacology* 25.1 (Feb. 2002), pp. 21–24.

[91] R. A. Hauser and T. A. Zesiewicz. “Sertraline for the treatment of depression in Parkinson’s disease”. eng. In: *Movement Disorders: Official Journal of the Movement Disorder Society* 12.5 (Sept. 1997), pp. 756–759.

[92] R. Ceravolo et al. “Paroxetine in Parkinson’s disease: effects on motor and depressive symptoms”. eng. In: *Neurology* 55.8 (Oct. 2000), pp. 1216–1218.

[93] Matthew Menza, Humberto Marin, Kenneth Kaufman, Margery Mark, and Marc Lauritano. “Citalopram treatment of depression in Parkinson’s disease: the impact on anxiety, disability, and cognition”. eng. In: *The Journal of Neuropsychiatry and Clinical Neurosciences* 16.3 (2004), pp. 315–319.

[94] Rohit Verma and Kuljeet Singh Anand. “Efficacy and Tolerability of Escitalopram for Treating Depression in Parkinson’s Disease.” In: 2012.

[95] N. El-Saifi, W. Moyle, C. Jones, and H. Tuffaha. “Quetiapine safety in older adults: a systematic literature review”. en. In: *Journal of Clinical Pharmacy and Therapeutics* 41.1 (Feb. 2016), pp. 7–18.

[96] Philippe Desmarais, Fadi Massoud, Josée Filion, Quoc Dinh Nguyen, and Paulina Bajsarowicz. “Quetiapine for Psychosis in Parkinson Disease and Neurodegenerative Parkinsonian Disorders: A Systematic Review”. eng. In: *Journal of Geriatric Psychiatry and Neurology* 29.4 (July 2016), pp. 227–236.

[97] Santiago Perez Lloret, Mariela Amaya, and Marcelo Merello. “Pregabalin-induced parkinsonism: a case report”. eng. In: *Clinical Neuropharmacology* 32.6 (Dec. 2009), pp. 353–354.

[98] John R. Younce, Albert A. Davis, and Kevin J. Black. “A Systematic Review and Case Series of Ziprasidone for Psychosis in Parkinson’s Disease”. eng. In: *Journal of Parkinson’s Disease* 9.1 (2019), pp. 63–71.

[99] Matthias J Kleinz and Ian Spence. “Chapter 4 - The pharmacology of the autonomic nervous system”. In: *Small Animal Clinical Pharmacology (Second Edition)*. Ed. by JILL E Maddison, STEPHEN W Page, and DAVID B Church. Edinburgh: W.B. Saunders, Jan. 2008, pp. 59–82.

[100] R. S. Aronstam and P. Patil. “Muscarinic Receptors: Autonomic Neurons”. In: *Encyclopedia of Neuroscience*. Ed. by Larry R. Squire. Oxford: Academic Press, Jan. 2009, pp. 1141–1149.

[101] Anastasiia S. Boiko et al. “Pharmacogenetics of tardive dyskinesia in schizophrenia: The role of CHRM1 and CHRM2 muscarinic receptors”. eng. In: *The World Journal of Biological Psychiatry: The Official Journal of the World Federation of Societies of Biological Psychiatry* (Jan. 2019), pp. 1–6.

[102] Wayne C. Drevets, Carlos A. Zarate, and Maura L. Furey. “Antidepressant Effects of the Muscarinic Cholinergic Receptor Antagonist Scopolamine: A Review”. In: *Biological psychiatry* 73.12 (June 2013), pp. 1156–1163.

[103] Won Je Jeon, Brian Dean, Elizabeth Scarr, and Andrew Gibbons. “The Role of Muscarinic Receptors in the Pathophysiology of Mood Disorders: A Potential Novel Treatment?” In: *Current Neuropharmacology* 13.6 (Dec. 2015), pp. 739–749.

[104] B. Dean and E. Scarr. “Possible involvement of muscarinic receptors in psychiatric disorders: a focus on schizophrenia and mood disorders.” eng. In: *Current molecular medicine* 15.3 (2015), pp. 253–264.

[105] Adlei B. Carlson and Gregory P. Kraus. “Physiology, Cholinergic Receptors”. eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2019.

[106] Björn A Grüning, Samuel Lampa, Marc Vaudel, and Daniel Blankenberg. “Software engineering for scientific big data analysis”. en. In: *GigaScience* 8.5 (May 2019).