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

Master's Thesis

A Comparison of Learned and Engineered Features in Network-Based Drug Repositioning

Submitted by
Lingling Xu

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)

November 22, 2019

Abstract

Drug repositioning is an time-efficient and less costly way for drug design. Computational methods especially network-based approaches are applied to integrate biological knowledge and then extract feature embeddings from the network for building a binary classification model to differentiate and predict the relationship between a drug and a disease.

This thesis introduces a workflow that leverages network representation learning methods such as node2vec and edge2vec to generate learned features for Hetionet, an integrated heterogeneous network with comparatively rich biological data (47k nodes and 2000k edges). After optimizing the parameters of node2vec and edge2vec, compare the performances of models trained by learned features and engineered features to evaluate the quality of feature embeddings. At the end, well trained models are used to predict new edges between drug nodes and disease nodes. Case study of HDAC-6 inhibitor, vorinostat, is provided. Diseases that can be potentially treated by vorinostat are predicted and some of them are validated by reserches done by other scientists. Another case study is to predict novel drug candidates for Alzheimer's Disease (AD), also, some of the predictions have been reported. In summary, learned features from node2vec and edge2vec can catch structure and semantic characteristics of the network. And high-quality learned feature embeddings can replace engineered features for drug repositioning tasks.

Acknowledgements

First and foremost, I would like to thank Prof. Dr. Martin Hofmann-Apitius for giving me the opportunity to work on this project and being very supportive and encouraging when I met difficulties.

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 all the discussions, encouragements.

To my colleagues at Fraunhofer SCAI and fellow students, thank you for being there as my accompany and support.

To my lovely friends, to the peaceful but never-stop Rhine, to the silently cold winter, reviving spring, vibrant summer and finally fruitful autumn. All my thanks to this place and time I spent here.

Lastly, to my family, thank you for giving me your unconditional love and support across the whole Europe-Asia continent.

Contents

List of Figures

List of Tables

1 Biological Background

Drug design is a complex, costly, time-consuming process with a low average success rate of 2.1% [1]. While companies usually require 10-15 years to develop a new drug, during which they often spend upwards of 12 billion dollars, the number of drugs approved by the United States' Food and Drug Administration (FDA) has been decreasing since 1995. The probability of a compound entering a phase I trial approved by the regulatory authority is less than 10% and the ultimate number of compounds investigated that make it to market is approximately $1/30,000$. Further, as the need personalized medicine increases, so does the difficult to develop new drugs, and new strategies to develop drugs are necessary [1] [2].

Drug repositioning is the process of finding new uses outside the scope of the original medical indication for existing drugs [2].It is more efficient, less-costly, and less-risky than the conventional drug development (Figure 1). It takes 1-2 years to find targets and 8 years to develop a repositioned drug with 1.6 billion dollars on average. The major advantage of drug-repurposing approaches is that, for an existing drug, both the preclinical information and the clinical profiles (pharmacokinetic, pharmacodynamic, and toxicity) are already available, thereby reducing the development risk. Accordingly, a drug compound can rapidly enter late-stage clinical trials, reducing development cost and time [3].Figure 1 displays the comparison between *de novo* drug development and drug repurposing processes.

Drug repositioning can be performed experimentally or computationally. The latter technique has been widely used to identify new indications for an existing drug (drug-centric) and for identifying effective drugs for treating a disease (disease-centric) by using a strategy of similarity assessment between drugs and/or diseases [4].

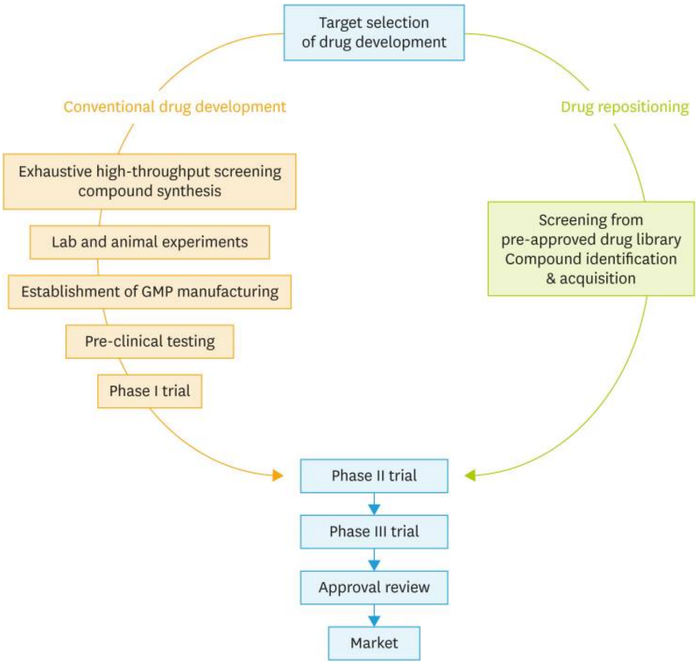
Target selection of drug development
Conventional drug development
Drug repositioning
Exhaustive high-throughput screening compound synthesis
Lab and animal experiments
Establishment of GMP manufacturing
Pre-clinical testing
Phase I trial
Screening from pre-approved drug library Compound identification & acquisition
Phase II trial
Phase III trial
Approval review
Market

Figure 1: A comparison of conventional drug development and drug repositioning. Figure adapted from [5]

There are two trends of computational drug repositioning. While the first trend is to make use of high-throughput data including genomics, proteomics, chemoproteomic, and phenomics [4]. which is flawed by noisy data and limitations of computational methods. For example, Zhang *et al.* (2019) warned that when using genomics data, it is unreasonable to treat each single nucleotide polymorphism (SNP) equally when using the elastic net algorithm because of their different importance in epigenetic studies [6]. The second trend is the development of repositioning algorithms based on retrospective analysis and database maintenance for experimental data, for which the challenge is to apply appropriate methods to integrate and exploit heterogeneous data.

Even computational drug repositioning has its limitations, but it's still a promising method. Several examples of successful repurposed drugs exist in fields such as oncology, diabetes, leprosy, and inflammatory bowel disease. Alfedi *et al.* (2019) performed a high-throughput screening of a library containing 853 FDA approved drugs by using a cell-based reporter assay to monitor variation of frataxin amount to accelerate the development of a new therapy for Friedreich's ataxia [7].They found that *etravirine* represented a promising candidate as a therapeutic for Friedreich's ataxia. Hassan *et al.* (2019) designed a computational and enzyme inhibitory mechanistic approach to fetch promising drugs from the pool of FDA approved drugs against AD. They found that *cinitapride* can be used as a possible therapeutic agent in the treatment of AD [8]. Besides the methods of these two examples, in recent years, network-based drug repositioning strategy is also widely used for predicting novel drug candidates of disease.

1.1 Network-based Drug Repositioning

Network-based drug repositioning has been recently widely used for finding new links between drugs and diseases. Biological networks $G = (V, E)$ consist of nodes $v \subset V$ that can be genes, proteins, anatomies, biological processes, compounds, diseases, etc., and their relationships (i.e., edges) $e \subset E$. Network is a useful data structure from which algorithms can extract structure and semantic information for drug repositioning. Different types of network-based drug repositioning strategies are reviewed below based on the content of network [9].

1.1.1 Gene Regulatory Networks

Molecular perturbations that occurred due to drug administration or a disease can be represented and exploited by building gene regulatory networks, and then the network can be used to prioritize nodes in existing networks to select candidate genes for drug repositioning.

Chen *et al.* (2016) [10] made use of multiple information sources such as gene mutations, gene expression data, functional connectivity, and proximity of module genes based on mining a human functional linkage network for inversely correlated modules of drug and disease gene targets to identify candidates for treating breast and prostate cancer. Two subnetworks were extracted for each drug: one contained upregulated disease genes which were down regulated by the drug and the other one contained downregulated disease genes which were upregulated by the drug. The correlation relationship between the drug and the disease can be exploited from these two sub-networks. For case study, 5 FDA-approved drugs were identified as drug candidates that can serve as potential therapeutic treatments for breast cancer based on therapeutic index tests (e.g., toxic dose, dose for therapeutic response).

While this kind of method can be effective, the signature for a disease or a drug is not always clearly defined, and drug-target gene perturbations may not be detectable. Other network with more solid data needs to be generated and exploited for drug repositioning [9].

1.1.2 Protein-Protein Interaction Networks

Protein-protein interaction Protein-protein interaction (PPI) networks are often used to identify similar proteins for drug targeted proteins based on the assumption that proteins targeted by similar drugs are functionally related and are 'close' in the PPI network [6]. Some example PPI network data resources include UniHi [7], DrugBank [8], GeneCards [11], Gene Expression OmnibusGene Expression Omnibus (GEO) [12], Genotator [13], STRING [14], the Kyoto Encyclopedia of Genes and GenomesKyoto Encyclopedia of Genes and Genomes (KEGG) [15], Online Mendelian Inheritance in ManOnline Mendelian Inheritance in Man (OMIM) [16], and the iRef Index database [17], Keane *et al.* (2015) [18] represented mitochondrial dysfunction and autophagic dysregulation in the cellular MPP[1] model, a neuronal in vitro disease model containing dopaminergic neurons allow the testing of PD drug candidates [19],for parkinson's disease (PD) by generating a PPI networks for respective model. Next, for each node, they extracted topological features and calculated the betweenness centrality, a measure of the number of shortest paths across the network that include the node. The differences in betweenness centrality between each network were calculated to indicate the importance of the node in connecting autophagy and mitochondrial dysfunction in the MPP$^+$ system. The reseachers hypothesized that manipulating nodes with high change in betweenness centrality would regulate MPP$^+$ toxicity. At last, they prioritized P62, GABARAP, GBRL1, and GBRL2 as novel targets associated for PD.

The limitation of this method is that the PPI data has a high degree of variance and it is incomplete due to the diverse range of experimental sources. An integrated and detailed network is required to build a more informative PPI network by standardizing and filtering the PPI data.

[1] $+$

1.1.3 Drug-Target Interaction Networks

Drug-target interactionDrug-Target Interaction (DTI) networks represent the physical interactions of drugs with their targets as swell as their mechanisms of action (e.g., inhibition, agonism, antagonism, reverse agonism). Several data sources include DrugBank [8], KEGG LIGAND, KEGG GENES, KEGG DRUG, KEGG BRITE [15], Drug Combination DatabaseDrug Combination Database (DCDB) [20], Matador [21], BRENDA [22], SuperTarget [21], MalaCards online system [23], and search tool for interactions of chemicals (STITCH) [24].

Yamanishi *et al.* [25] used DTI networks to demonstrate that if two drugs have similar structure, they have a higher chance to interact with similar target proteins. Likewise, two target proteins with high sequence similarity are more likely to interact with similar drugs. Yan *et al.* (2016) [26] built a heterogeneous network G consisting of drugs and their targets, a drug-drug network U and a target-target network V extracted from G based on similarities between drugs and targets respectively. Between the two sub-networks, drugs and targets were connected with edges E to form a bipartite network G = (U, V, E). They developed and applied a method, LPMIHN, to the network to rank targets if the query is a drug and vice versa.

The main limitation of this method is that few drug-target data sets are available. The prediction of models is highly dependent on the training data, so such models cannot predict drugs without target information. Another limitation is that negative samples are difficult to find since there is a scarcity of experiments that were conducted for verifying negative drug-target samples [27]. The solution for the first problem is to make use of highly similar drugs to predict targets (or highly similar targets to predict drugs) for those without relevant information [28]. In machine learning scenarios, the second problem can be addressed with positive-unlabeled learning, in which negative samples can be sampled randomly from unlabeled drug-target pairs. However, this does not address the overarching issue the biocuration community has with structuring negative knowledge. Positive unlabeled learning is discussed in more detail in the methods section.

1.1.4 Drug Similarity Networks

Drug similarity networks support inference methods that rely on the assumption that drugs with similar structures have similar biological effects. The similarities of drugs can be measured by using two-dimensional topological fingerprints, 3D conformations, or biological effects (side effects or gene expression patterns).

Drug similarities are usually measured by structure similarities with fingerprints [29] or chemical graph, which is based on the assumption that compounds with similar structures have similar activity. But this method suffers from that drugs similar in structures often have very different biological effects [30]. To compensate for such a limitation,Similarity-based Inference of drug-TARgets (SITAR) [31] used 5 similarity measurements for drug-drug pairs, which are chemical similarities, side effect similarities [24], gene expression profiles similarities, and Anatomical, Therapeutic and Chemical (ATC) classification. However, this method is still flawed by the similarity measuring and lack of structure information of compounds.

1.1.5 Disease-Disease Networks

Disease-disease networks consist of disease nodes and disease-disease similarity edges. They can be generated from resources like DrugBank [8], OMIM [16], and Disease Ontology (DO) [32]. This network strategy is to find novel drug-disease relationship by measuring the similarity between diseases based on the assumption that similar diseases are similar to be treated by similar drugs. How to measure the similarity between diseases is critical to this disease-disease network strategy. Van Driel *et al.* (2006) [33] created MimMiner to calculate disease similarity based on phenotypes. Huimin Luo *et al.* (2016) [34] improved MimMiner by exploiting the correlative relationships between diseases by comparing the existence of their common drugs. They provided a case study where Levodopa is predicted to treat AD, which had already been tested in clinical trial [35].

1.1.6 Disease-Side effects Networks

Many drugs induce some unintended effects on the living organisms in addition to the desired effects by interacting with off-targets [36] [37]. Those side effects provide a human phenotypic profile for the drug, which suggest additional disease indications. The rationale for drug repositioning based on side effect profiles is that side effects and indications are both measurable behavioral or physiological changes in response to treatment, and if drugs treating a disease share the same side effect, there might be some underlying mechanism-of-action Mechanism-of-

Action (MoA) linking this disease and the side effects. The side effect may thus serve as a phenotypic “biomarker” for this disease.

Yang *et al.* (2011) hypothesized that if a given drug is associated with common side effects as drugs that treat a given disease, then it should also be a candidate for treating that disease. They generated a network with diseases, drugs and side effects. Disease-drug associations were extracted from PharmGKB [38] and drug-side effect associations are extracted from the Side effect resource (SIDER) [39]. A quantitative structure-activity relationship model was built to compensate for the absence of side effect profiles for many drugs. Then they built naive Bayes models to predict indications for 145 diseases using the side effects as features. The area under the receiver operating characteristic (ROC) curve (AUC-ROC) was above 0.8 in 92% of these models, which indicated that closer attention should be paid to the side effects observed in trials not just to evaluate the harmful effects, but also to rationally explore the repositioning potential based on this “clinical phenotypic assay”. Also, they provided a case study that porphyria was suggested to act as antidiabetics and delusions might help with depression [40].

Not many drug repositioning methods are based on this strategy due to some limitations. The relationships between phenotypes and MoAs are unclear because outcomes of drugs’ MoAs are highly dependent on the organism’s genetic map, medication history, and other traits. Therefore, similar phenotypes are not always caused by similar targets. Even the assumption is valid, many drugs don’t have comprehensive and accurate side effects profiles. [6] [36] [9].

1.1.7 Integrated Networks

Networks discussed in previous sections usually contain only two or three types of nodes, such as drugs, targets, diseases, gene signatures, etc. If an integrative network including drugs, targets, phenotypes, diseases, and pathways, can be built to represent the MoA (Figure 2) of a drug, the predictions would be more accurate based on rich information in the network.

Luo *et al.* (2017) [42] built a pipeline, DTINet, which applied learned features to drug-target prediction for drug repositioning. Random Walk with RestartRandom Walk with Restart (RWR) and Diffusion Component AnalysisDiffusion Component Analysis (DCA) were used to generate learned features. DTINet focused on learning a low-dimensional vector representation of features, which accurately illustrated the topological properties of individual nodes in the heterogeneous

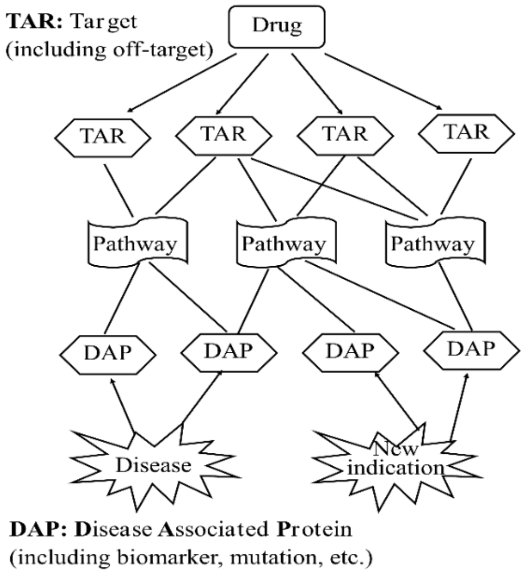

Figure 2: Drug Therapy Model. Figure adapted from [41]

network, and then made prediction based on these representations via a vector space projection scheme. They predicted novel interactions between three drugs and cyclooxygenase proteins and demonstrated potential applications of these identified cyclooxygenase inhibitors as a role of preventing inflammatory diseases.

There are four types of nodes and six types of relationships in DTINet. They extracted the drug nodes from the DrugBank database (Version 3.0) [43] and the protein nodes from the Human Protein Reference Database (HPRD) (Release 9) [44]. The disease nodes were obtained from the Comparative Toxicogenomics Database [45]. The side-effect nodes were collected from the SIDER database (Version 2) [46]. They imported the known DTIs as well as drug–drug interactions from DrugBank (Version 3.0) [43]. The protein–protein interactions were downloaded from the HPRD database (Release 9) [44]. The drug–disease and protein–disease associations were extracted from the Comparative Toxicogenomics Database [45]. They also included the drug–side-effect associations from the SIDER database (Version 2) [46].

Integrative networks contain more comprehensive information for MoA but the limitation is that only a few network representation learning methods are designed for heterogeneous network. Also, it is more difficult to build a high quality network with diverse sources and the standard of adding an edge between two nodes needs to be normalized, for example, the threshold of similarity between nodes and the standard of differentially expressed genes in an anatomy.

1.2 Methods used in Computational Drug Repositioning

Recently, repositioned drugs account for 30% of newly marketed drugs in the US [47]. In the meantime, the explosive and large-scale growth of molecular, genomic and phenotypic data of pharmacological compounds is enabling the development of new area of drug repurposing called computational drug repurposing. There are three main categories, machine learning methods, network-based methods and text-mining methods [4].

1.2.1 Machine Learning Methods

Machine learning techniques that have been applied for drug repositioning include logistic regression,support vector machine (SVM), random forest, neural networks, and deep learning [4]. In this thesis, binary classification is applied to distinguish drug-disease pairs, so only logistic regression andSVM are discussed here.

Logistic regression is a widely used statistic model. It is often applied for binary classification problem, in which samples are usually labeled as two classes, positive and negative. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables and predict labels with given input.

In biological science, Ayers *et al.* applied penalized regression methods to select pertinent predictors for a phenotype [48]. Himmelstein *et al.* [49] used logistic regression for predicting new drug-disease link in Hetionet. Muslu *et al.* trained a logistic regression model by learned features of disease-targets pairs from GAT2VEC to predict targets of a disease [28]. Liu *et al.* [50] generated a web service SPACE (Similarity-based Predictor of ATC CodE), which for each submitted compound, will give candidate ATC codes which is ranked according to the decreasing probability_score. Behind this web service, is a logistic regression model trained by drug similarity features including chemical structures, target proteins, gene expression, side-effects and chemical–chemical associations.

SVM is another popular binary classification model. The objective of the SVM is to find a hyperplane in an N-dimensional space (N is the number of features) that distinctly partition the data points into two parts by maximizing the margin from both sides to the hyperplane.Napolitano *et al.* [51] combined molecular target, drug chemical structure, and gene expression similarity into one feature matrix. Then they predicted drug therapeutic class by an SVM approach based on feature matrices. Shameer *et al.* (2018) [52] use RepurposeDB [53] and DrugBank [54] as datasets and ChemVec as feature engineering strategy to train supervised classification algorithms (naïve Bayes, random forest, SVMs, and an ensemble model combining the three algorithms).

In this thesis, the elastic net, a penalized logistic regression, is used to predict new drug-disease edges because the logistic regression gives not only class predictions but also probability predictions, which is used for ranking the drug candidates in drug repositioning.

1.2.2 Network-based Methods

Based on the graph topological properties, the entities in the same cluster are similar to each other. For biological network, proteins, genes, drugs should cluster with similar nodes. This approach is used to identity drug-disease, drug-target relationships [55]. Lu *et al.* used chemical-chemical interactions and chemical-protein interactions to select candidate compounds that had close associations with approved lung cancer drugs and lung cancer-related genes. K-means clustering and a permutation test were applied to remove compounds with low possibilities to treat lung cancer. The final results suggest that the remaining compounds in cluster has structural similarities with approved drugs for lung cancer and they have potential anti-lung cancer activities [56].

The workflow of network-based propagation approach is that prior information of source nodes propagate to other nodes in network. This approach contains two types: local approaches and global approaches. Local approaches take limited information into account [55]. Global approaches are more accurate by exploring more information. Martinez *et al.* created DrugNet, a new methodology for drug-disease and disease-drug prioritization by integrating drugs, proteins, and diseases into network. They rank the nodes in these networks according to their distance to a query set of nodes. DrugNet achieved a mean AUC-ROC value of 0.9552 ± 0.0015 in 5-fold cross validation tests, which suggests DrugNet could be very useful for discovering new drug use [57].

1.2.3 Semantic and Texting-Mining

The structured information in databases is only a fraction of all knowledge, the biomedical and pharmaceutical literature contains a huge amount of unstructured information available for drugs and diseases, from which new indications of existing drugs can be found through text-mining method. Su *et al.* (2017) [58] used the text mining tools I2E (Linguamatics) and PolyAnalyst (Megaputer) to extract Serious Adverse Events (SAE) data from randomized trials in ClinicalTrials.gov. Through a statistical algorithm, a PolyAnalyst workflow ranked the drugs where the treatment arm had fewer predefined SAEs than the control arm, indicating that potentially the drug is reducing the level of SAE. The results indicated that *telmisartan* may be viable as a repurposed prevention for colon cancer and this was reported by few researches that Telmisartan exerted anti-tumor effects by activating peroxisome proliferator-activated receptor-γ [59] [60] [61].

Ngo *et al.* (2017) made use of continuous-bag-of-words model from word2vec [62] to get word embeddings for cancer related drugs and diseases [63]. The raw corpus is a subset of PubMed abstracts downloaded in October 2013, filtered by the keyword "cancer". After getting embeddings from word2vec model, they performed hierarchical clustering for testing the quality of word embeddings and SVM classification to learn and predict possible relations between drugs and diseases. The clustering showed anti-cancer drugs condensed in one cluster which indicated that the word embedding catched the characteristic of words and the SVM classification model achieved 87.6% accuracy of drug-disease relations in cancer treatment. Also the model succeeded in discovering novel drug-disease relations that were actually reported in recent studies. For example, repositioning of Clarithromycin to anticancer agent was reported in 2015 [64]. Based on the fact that the corpus were downloaded in 2013, the prediction was valid and indicated that the classification of concatenated word vector was a promising approach to in-silico screening of drug-disease relations for drug repositioning.

1.2.4 Validation of Computational Drug Repositioning

The evaluation of computational drug repositioning is usually case studies or unbiased validation methods such as AUC-ROC, which summarise the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds, area under precision-recall (AU-PR), which summarise the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds. Some researchers experimentally validated their results with *in vivo* experiments. The flaw of case studies is that finding evidences in other research paper is inefficient and difficult. Another method, unbiased validation method, is also flawed by the assumption that all drug-disease indications not in databases (e.g., DrugCentral) are false. To address these concerns, Adam *et al.* [53] present repoDB, a database of approved and failed drugs and their indications. repoDB approved indications were drawn from DrugCentral, which contains United Medical Language System Unified Medical Language System (UMLS) indications mapped from free-text mentions in drug labels. In this thesis, due to the purpose of comparing this work to previous work, the case studies, AUC-ROC and AU-PR are used for validating. Applying repoDB will be one of the future work.

1.3 Rephetio Approach

All the networks discussed above in the section 1.1 contain at most 4 types of nodes and 6 types of relationships. But the MoA (Figure 2) behind drug-disease relationships are very complicated and involve many other entities such as anatomy, pathways, molecular functions, etc., so an integrative network is necessary to represent comprehensive biological knowledge accumulated from previous studies. For methods of computational drug repositioning, machine learning methods are very powerful to classify and predict relationships between entities, but one must for machine learning is that feature vectors are indispensable.

Himmelstein *et al.* (2017) [49] generated a heterogeneous network, Hetionet, including biological knowledge from 29 different databases, which contains 11 types of nodes and 24 types of edges, in order to fulfill the absence of a comparatively comprehensive network in network-based drug repositioning field. Furthermore, they generated engineered feature (e.g., DWPC) vectors and the engineered features representing prevalence of metapaths were interpreted by extracting weights of features from logistic regression model. After prediction of novel drug-disease indication, 10 most supportive paths in the network could be found by their algorithm, which not only gives a decent explanation for predictions but also inspiring clues for scientists to verify the predicted indications.

1.3.1 Hetionet

Nodes consist of 1552 small molecules and 137 complex diseases, also genes, anatomies, pathways, biological processes, molecular functions, cellular components, perturbations, pharmacologic classes, drug side effects, and disease symptoms.

Gene nodes are from Entrez Gene, which belongs to the National Center for Biotechnology Information National Center for Biotechnology Information (NCBI) database for gene-specific information. Entrez Gene maintains records from genomes that have been completely sequenced, assigns unique, stable and tracked integers as identifiers to each gene sequence. The content of Entrez Gene is integrated and curated in an automatic way through NCBI's Reference Sequence project [65].

Entrez Gene was chosen but not Hugo Gene Nomenclature Committee (HGNC) identifier due to few advantages of Entrez Gene compared to HGNC. Entrez Gene is more specific to humans, geneIDs are less prone than ambiguous gene symbols in HGNC, and integration of Entrez Gene with many other NCBI services such as HomoloGene can relate orthologous genes across species [66].

Disease nodes are from DO [32], which is a database provides standard hierarchical disease terms. However, the method of Rephetio requires distinct and non-redundant nodes, which means that no terms should be an ancestor or descendant of any other term. Therefore, disease terms should be chosen from an appropriate level, specific enough to be clinically relevant and general enough to be well annotated. To create this DO slim, 108 diseases from GWAS catalog in hetio and 63 cancer terms from TOPNodes_DOcancerslim are combined. Then some overlapped diseases are removed. At last, 137 disease terms are selected for Hetionet.

Compound nodes are from DrugBank with drugs approved, small molecules containing InChI structures [8]. Interacting proteins (targets, enzymes, transporters, carriers) for each drug are extracted for Hetionet. There are two criterias for proteins to be included. One is that the interaction is between a drug and single protein. A target which is a "protein group" (e.g., GABA-A receptor (anion channel) and NMDA receptor) is excluded. Also, a target which is not a protein but a DNA or phosphate is excluded too. Another one is that The protein should be mapped to an entrez gene. At last, 19,906 interactions for 5,878 drugs and 3,757 genes are extracted [67]. The connectivity fingerprints are calculated with Morgan/circular [63] method in order to get the similarities between compounds. RDkit module was used to calculate similarities of compounds with Dice Coefficient (Equation 1, a is the number of "1" in fingerprint of A, b is the number of "1" in fingerprint of B, c is the intersection of fingerprints of A and B) [64]. The similarity threshold is 0.5. Then Compounds from DrugBank are mapped to 30 other compound resources using UniChem. The mapping is based on atomic connectivity and ignores differences in small molecular details, but the mapping with PubChem is based on exact InChi string matches.

$$S(A,B) = 2 * c/a + b \tag{1}$$

Anatomy nodes are from Uber-Anatomy Ontology (UBERON) [68], to represent anatomical structures. UBERON integrates cross-species ontology consisting of over 6,500 classes representing a variety of anatomical entities, organized according to traditional anatomical classification criteria.

Symptom nodes are from Medical Subject Headings (MeSH), a controlled vocabulary produced by the National Library of Medicine (NLM) for cataloging biomedical information, used to annotate PubMed and MEDLINE. Two record types are processed in Hetionet: Descriptors and Supplementary Concept Records [69] [70].

Pathway nodes are from WikiPathways, Reactome, and the Pathway Interaction Database (PID) [71]. Only protein-coding human genes (as Entrez Gene IDs) are included [72] [73]. Pathways files are downloaded as TSV (Tab Separated Values) file and overlapped pathways are removed. At last, 1,862 human pathways (1,341 from Reactome, 298 from WikiPathways, and 223 from PID) are generated.

Disease-gene edges are from DisGeNET. Disease node is from Disease Ontology, and the gene is represented as Entrez Gene. DisGeNET integrates data from expert curated repositories, Gnome-Wide Association Study (GWAS) catalogues, animal models and the scientific literature, and the data are homogeneously annotated with controlled vocabularies and community-driven ontologies [73].

Diseases-differentially expressed genes edges are from STARGEO, which aims to provide disease-specific expression signatures on a broad scale with crowdsources GEO annotation and performs case-control analyses based on user queries. In summary, STARGEO defines case-control queries for 66 diseases. Of these diseases, 49 contained sufficient data which means multiple studies with at least 3 samples per class. Hetionet used STARGEO's random effects meta-analysis and applied a false discovery rate (FDR) p-value threshold of 0.05 to identify differentially expressed genes for each disease. 48,688 Disease–downregulates–Gene and 50,287 Disease–upregulates–Gene relationships were identified [74].

Disease-symptom occurrence edges, disease-disease occurrence edges, disease-localization edges and anatomy-disease co-occurrence edges are calculated between MeSH terms. 23 million journal articles are provided in the NLM. Skilled subject analysts at the NLM typically assign 10–12 MeSH terms per article and denote a subset of these terms as major topics. They infer relationships between nodes in Hetionet based on MEDLINE co-occurrence [70].

Gene-anatomy differential expression edges are from Bgee [75], a database that integrates gene expression data from both microarray and RNA-seq experiments. The Bgee curators annotate samples by their species, anatomical structure, and developmental stage. Bgee leverages anatomical and developmental ontologies to call whether a gene is present or absent and under/over-expressed in a given condition. Gene expression profiles for human anatomical structures (tissues) are generated from Bgee for Hetionet, including whether a gene is expressed in an anatomy, whether a gene is overexpressed or underexpressed in an anatomy.

Protein-protein interaction edges are from three sources, Human Interactome Database (HID) [76], Incomplete Interactome [77], hetio-dag [78]. The edge exists when two proteins physically interact. Besides, for combining an orthogonal resource to the protein interactions, Evolutionary Rate Covariation (ERC) between genes [49], which assesses whether two genes have a similar evolutionary history are calculated. An edge is added when ERC of two genes are bigger than 0.75.

Drug-disease edges are from PharmacotherapyDB [79], an open catalog of medical indications between small molecule compounds and complex human diseases. Containing two parts, Disease Modifying and Symptomatic, both are reviewed by multiple physicians. Diseases and drugs are coded with Disease Ontology and DrugBank identifiers for integrative analysis. Four resources are integrated, LabledIn [80], MEDI-HPS [81] , EHRLink [82], PREDICT [83]. Disease modifying is defined as "a drug that therapeutically changes the underlying or downstream biology of the disease" and Symptomatic as "a drug that treats a significant symptom of the disease." When one of them is satisfied, one edge between the drug and the disease is built.

Compound-gene edges are from Library of Integrated Cellular Signatures (LINCS) [84]. A database provides perturbational profiles across multiple cell and perturbation types, as well as read-outs, at a massive scale. One compound in Hetionet could be matched to multiple LINCS compounds. A single consensus transcriptional profile across multiple signatures is calculated to reduce redundancy [85].

Drug-side effects edges are from SIDER 4 [39], a project to combine data on drugs, targets and side effects into a more complete picture of the therapeutic mechanism of actions of drugs and the ways in which they cause adverse reactions. It contains data on 1430 drugs, 5,880 adverse drug reaction (ADR)s and 140,064 drug-ADR pairs, which is an increase of 40% compared to the previous version. For more precise analyses, they extracted the frequency with which side effects occur for 39% of drug-ADR pairs, and 19% of those pairs can be compared to the frequency under placebo treatment. SIDER furthermore contains a data set of drug indications, extracted from the package inserts using Natural Language Processing. These drug indications are used to reduce the rate of false positives by identifying medical terms that do not correspond to ADRs.There are some other nodes and edges are contained in Hetionet, as shown in Table 1

Nodes\ Edges	**Source**	**Annotation**
Pharmacologic class	DrugBank [8]	It gives the information that which pharmacologic class one drug is belonged to.
Gene Ontology	Gene Ontology	Entrez GeneIDs related to GO terms are provided and set as symbols
Pathway-protein edges	WikiPathways [86], Reactome [87], and Pathway Interaction Database (PID) [71]	As discussed in the section pathway node
Pharmacologic class-drugs edges	Drug Central	If the drug is belonged to a certain pharmacologic class, an edge is created between them.
GO terms-Entrez Gene edges	Gene Ontology [88] and Entrez Gene [65]	If the gene is related to the GO term, an edge is added between them.
Compound-target edges	DrugBank [8]	As discussed in the section compounds nodes
Chemical-chemical similarity edges	DrugBank [8]	The similarity is calculated with finger prints and the threshold is 0.5. More information is in the section compound nodes.

Table 1: Short introduction for some nodes and edges in Hetionet

1.3.2 Rephetio Methods

Himmelstein *et al.* adapted an algorithm originally developed for social network analysis and applied it to Hetionet to identify patterns of efficacy and predict new uses for drugs. The algorithm performs edge prediction through a machine learning framework that accommodates the breadth and depth of information contained in Hetionet [78].

Engineered Features

For capturing the connections between compound nodes and disease nodes, Himmelstein *et al.* [49] engineered topological features to train logistic regression model for drug repositioning. The degree weighted path count (DWPC), prior probability, and node degrees for 14 meta-edges are included in features.

DWPC represents the prevalence of a specific metapath. In Figure 3, on the left is a hypothetical graph, the source node is IRF1, the target node is Multiple Sclerosis. There are several paths from IRF1 to Multiple Sclerosis, one kind of path is one metapath like 'G-e-T-l-D'. Only one path fits this kind of metapath, which is IRF1-expression-Leukocyte-localization-Multiple-Sclerosis. In this path, node IRF1, connecting with two nodes (node type is 'Target') with edge type of 'expression', has a metaedge-specific degree of 2. Then get all metaedge-specific degrees of one path to count path count products (PDP). If there are more than one paths of the same metapath type, add 0PDP of all paths together to get DWPC.

They evaluated all 1206 metapaths that traverse from compound to disease and have a length of 2–4 with DWPC, and compared the performance of each metapath to a baseline computed from permuted networks. The permutation preserved node degree while eliminating edge specificity, in order to isolate the portion of un-permuted metapath performance resulting from actual network paths. They refer to the permutation-adjusted performance measure for a metapath as ΔAUC-ROC. A positive ΔAUC-ROC indicates that paths belonging to the type of metapath tended to occur more frequently between treatments than non-treatments, after accounting for different levels of connectivity (node degrees) in the network. In general terms, ΔAUC-ROC assesses whether paths of a given type were informative of drug efficacy [46]. The DWPCDWPC features were chosen based on the evaluation of ΔAUC-ROC and other standards. At last, 142 DWPC features left for representing drug-disease pairs [89].

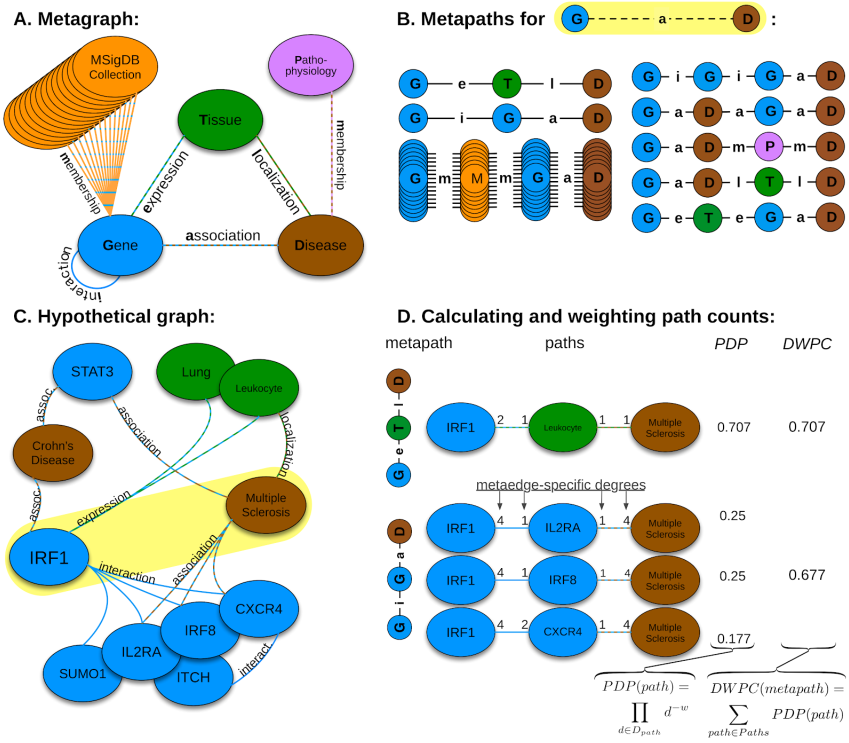

Figure 3: DWPC calculation workflow. Figure adapted from [49]

ince Himmelstein *et al.* engineered features and chose features manually, after prediction, it is possible to interpret the result according to coefficient of each feature. They extracted metapaths of which the coefficient given by the logistic regression model were positive based on the assumption that metapath features (e.g., DWPC) with positive coefficients and positive logistic regression term are considered to contribute positively. Then they find the most supportive paths according to the metrics of DWPC.

Figure 4 shows the ten most supportive paths for treating nicotine dependence with bupropion. These supportive paths possibly explain why bupropion could be used to treat nicotine dependence.

1.3.3 Shortcomings of Rephetio Method and Learned Features

Usually scientists engineered features from similarities based on structures or phenotypes [31], but in network-based drug repositioning, engineering feature vectors becomes more difficult because of a lack of powerful algorithm to represent

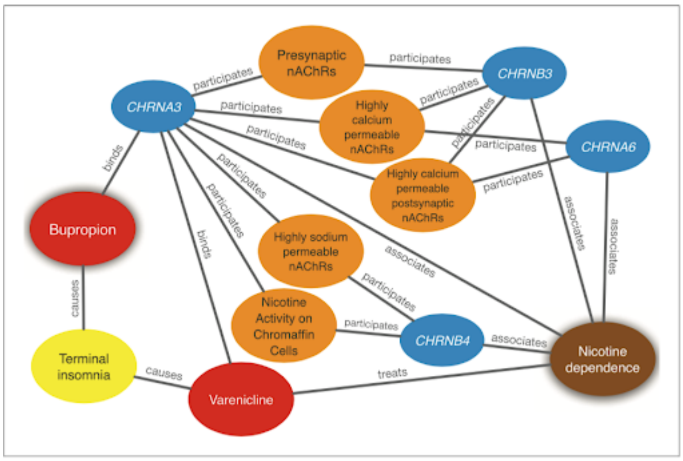

Figure 4: Evidence supporting the repurposing of bupropion for smoking cessation. Figure adapted from [49]

biological entities especially in heterogeneous network. Furthermore, engineering features such as DWPC can only represent limited characteristic of the network. While DWPC for one metapath represents the prevalence of this type of metapath, many other topological information such as neighbours and vicinity are missed. Also, engineered features are are not generalizable. If drug-target features are wanted, DWPCs have to be evaluated and chosen again, which makes this methodology less extensible.

With the advancement of network representation learning methods, it is possible to generate feature vectors with other methods (e.g., node2vec). Muslu *et al.* (2019) [90] used a genome-wide protein-protein interaction network annotated with disease-specific differential gene expression to predict targets for drugs with logistic regression model trained by learned features from DeepWalk model. They evaluated the approach on six diseases of different types (cancer, metabolic, neurodegenerative) within a 10 times repeated 5-fold stratified cross-validation and achieved AUC-ROC values between 0.92-0.94, significantly outperforming a previous approach from Emig *et al.* [91], which relies on manually engineered topological features.

Luo *et al.* (2017) [42] applied another method, DCA, [92] to generate learned features for predicting novel drug-target associations in DTINet. DCA is a method to capture the inherent topological features of a network and represent it as low-dimension feature vectors, which uses diffusion methods for network to get the distribution of nodes with restart random walk. Then approximate each of these distributions by constructing a multinomial logistic model, parameterized by low-dimensional feature vector(s), for each node. Feature vectors of all nodes are jointly learned by minimizing the Kullback-Leibler (KL) divergence (relative entropy) between the diffusion and parameterized-multinomial logistic distributions. Also, their results outperformed the state-of-art methods for drug-target predictions, which indicated that DTINet with learned features can provide a practically useful tool for integrating heterogeneous information to predict new drug–target interactions and repurpose existing drugs.

The shortcomings of engineered features can be overcomed by learned features, which is easier to be generated and keep comparatively comprehensive topological and semantic features of the network. DCA and DeepWalk are both random walk based methods, but the way of them to catch the co-occurrence properties of nodes are different. DCA simply makes use of multinomial logistic model to catch the co-occurrence probabilities between nodes calculated before training, but DeepWalk uses neural network to keep not only the co-occurrence probabilities but also how nodes pairs co-occur with each other by feeding node pairs to the neural network. Furthermore, Ngo *et al.* [61] applied language model (continuous bag-of-word, similar to skip-gram) for drug repositioning as a successful try. Based on these studies, random-walk and language model based network representation learning methods (e.g., node2vec) are promising to represent topological features of a heterogeneous network and for further drug repositioning task. The Computational Background section discusses the methods used in this thesis.

2 Computer Science Background

Random-walk based Network Representation Learning (NRL) methods have been chosen for use in this thesis to address the shortcomings with engineered features. Learned feature vectors of nodes and edges would be generated from these methods of which the core is language model skip-gram (one type of word2vec model). Perozzi et al. [93] developed DeepWalk to apply random walk to network for generating walks analog to sentences for training skip-gram model. In DeepWalk, the walks (sentences) are parsed into a one-layer neural network to learn the occurrence relationships between nodes (words). Then, the weights of the neural network would be kept for generating node (word) vectors. At last, learned feature vectors of nodes can be used for downstream machine learning methods. Node2vec [94] improves DeepWalk by adding in-out (p) and return (q) parameters to generate biased random walk to perform a flexible neighbourhood sampling strategy. Edge2vec [95] was improved further by applying an expectation-maximization (EM) algorithm to calculate a edge-type specific transition probability matrix (TPM) with edge semantics information.

2.1 Language Models

Word2vec is a group of related models to produce word embeddings created by Mikolov et al. [62]. It gives a two-layer neural network to capture the co-occurrence of words in sentences, including two main models, continuous bag-of-word and skip-gram. Continuous bag-of-word predicts the current word from a window of surrounding context words, which is adapted for drug repositioning by Ngo et al. [63], and the skip-gram model conversely uses the word to predict its surrounding words.

When training the neural network, every word in the training set is encoded as one-hot vector. After feeding the pairs of words to a two-layer neural network, backpropagation is applied to update the weights in the hidden layer. The training objective is to minimize the summed prediction error across all context words in the output layer. Each word is then represented by a distribution of weights across the hidden layer.

2.1.1 Skip-Gram Model

In this thesis, the skip-gram model is implemented in node2vec and edge2vec. According to Rong's explanation to word2vec [96]. As illustrated in Figure 5, this model takes one word as input to a log-linear classifier with continuous projection layer in order to predict the context of the word in a certain range. The training complexity is presented in Equation 2

$$Q = C \times (D + D \times log_2 V) \tag{1}$$

Where C is the maximum distance of the words, R in range <1,C> could be selected, R words before and after the target words are used as labels to train the neural network. In this model, there exist two vector representations for each word in the vocabulary: the input vector Vw, and the output vector Vw'. Learning the input vectors is cheap; but learning the output vectors is very expensive. Then hierarchical softmax and negative sampling are used for optimizing computational efficiency [96].

Intuitively, the skip-gram model give the word vector according to its surrounding contexts, which means if two words have similar meanings or tied together often, they should have similar vectors, for example, the vector of "france" is similar to that of "canada".

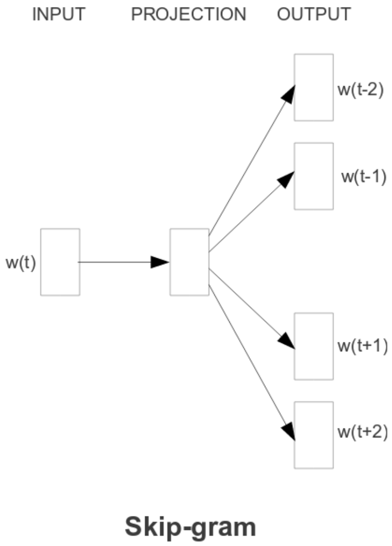

Figure 5: The structure of skip-gram model. Figure adapted from [62]

2.2 Random Walk Based Network Representation Learning Models

Generating random walks is a stochastic method that was first introduced by Karl Pearson in 1905 [97]. A node w is selected from a graph as a starting point, then randomly select one of its neighbors n as next node to go. At last, the sequence of nodes selected in such way is a walk. The probability for the walk to move from w to the n is called transition probability. According to transition probability, the walk is still random but could be more or less possible to one specific node than another.

The transition probabilities contain information about the edge. If the edge e is very important, intuitively, the transition probability to go through e should be higher than many other edges around the starting node w.

2.2.1 DeepWalk

DeepWalk is from Perozzi et al. [93], it uses random walk to make 'sentences' from network, then train the skip-gram model to get the feature vectors of nodes. The algorithm has two parts, the first one is to apply random walk and the second one is to train a skip-gram model. In the first step, the random walk generator takes network G as homogeneous and uniform, starting from vertex Vi to its neighbors until getting to the walk length t. All the walks are 'sentences' to skip-gram model to capture the occurrence probability of 'words' (nodes) in the network.

DeepWalk outperformed other previous methods like SpectralClustering [98], EdgeCluster [99], Modularity [100], wvRN [101], and Majority (This naive method simply chooses the most frequent labels in the training set) with evaluation as micro-F1 and macro-F1. Also, the parameter sensitivity was tested. For dimensionality, the level of changes of results highly depends on the dataset. The performance of the model is sensitive to the number of random walk. When the number of random walks is above 10, the results have no outstanding improvements. It suggests that a small number of random walk could learn a meaningful latent representation.

2.2.2 Node2vec

As illustrated in Figure 6, given a node v, random walk would go to the neighbors x. there are three types of neighbor: one is the previous node t, one is neighbor x_2 connecting with the previous node t, another one is x_2 (or x_3) only connecting with v, but not with the previous node t.So the shortest path distances (d_{tx}) for t, x_1, x_2, x_3 to t are 0, 1, 2, 2. Transition probability π is calculated as Equation 2, w_{vx} is the weight of edge between node v and x. In Equation 2, p, q and weights are integrated in transition probability for generating walks (sentences) for skip-gram model.

$$\pi_{vx} = \begin{cases} \frac{1}{p}\times \mathrm{w}_{vx} & \text{if } d_{tx} = 0 \\ 1\times \mathrm{w}_{vx} & \text{if } d_{tx} = 1 \\ \frac{1}{q}\times \mathrm{w}_{vx} & \text{if } d_{tx} = 2 \end{cases} \tag{2}$$

Node2vec [94] is improved from DeepWalk, adding a return parameter p and an in-out parameter q during calculation of the node TPM. p is the return parameter. When p is smaller than 1 and q, the probability of returning $1/p$ is bigger than 1 and $1/q$. Random walk would be more likely to go back to t (Figure 6), so the walk is more local. q is the in-out parameter. When q is bigger than 1, the probability of going out ($1/q$) is smaller than 1, then still the walk would be more local. If q is smaller than 1, the walk would be more global.

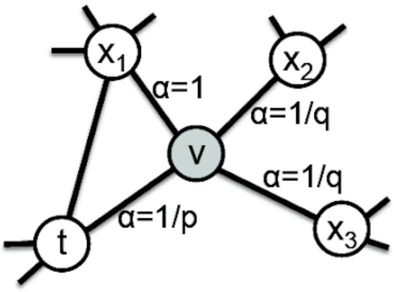

Figure 6: An illustration of random walk procedure in node2vec. Figure adapted from [94]

2.2.3 Edge2vec

In edge2vec algorithm, Zheng et al. [95] improved node2vec by making use of EM algorithm to calculate an additional TPM based on edge types. Edge2vec is based on the assumption that the distribution of edges in different samples are valid estimator of transition correlation in the graph. If two edge types are highly correlated then the transition probability between them is high, and vice versa. With such method, those edge types which are less distributed in the network, but actually important would be more likely to be passed through during random walk. And then the model can capture more comprehensive information of the network.

The edge2vec algorithm contains two parts, generating TPM with EM algorithm and generating walks according to the TPM. The first part starts with a transition probability matrix initiated as all the same number $1/n^2$ with a matrix size $n * n$ (n is the number of edge types in network). Expectation step is to generate walks based on the current TPM. Usually a random walk starts from every node in the network, but in edge2vec, there is a parameter 'max_count' constraining the number of starting nodes. Originally, the biggest 'max_count' is 1,000. It means sampling at most 1,000 starting nodes from all nodes. And then in the Maximization step, the frequency of every edge type is calculated to update the

TPM through evaluation tests as illustrated in Table 2. All these statistic methods are aimed to measure the correlation relationship between two edge types. After m (defined by *em_iter* parameter in edge2vec) times EM, an edge-type transition matrix is generated with edge semantics information.The second part is to operate random walk on the network based on the TPM generated from the first part.

Test	**Explanation**
Wilcoxon signed-rank	A nonparametric test that can be used to determine whether two dependent samples were selected from populations having the same distribution [102].
Entropy	Compute the Cosine distance between two 1-D vectors
Pearson	Evaluates how likely it is that any observed difference between the sets arose by chance [103].
Spearman	A nonparametric measure of rank correlation, assess how well the relationship between two variables can be described using a monotonic function [104].

Table 2: Explanations for test methods in edge2vec

The *em_iter* parameter is designed to constrain how many times EM algorithms to be implemented. If em_iter is set to 5, that means 6 edge-type TPMs would be generated (including the first one). The last matrix would be parsed to the next step to generate random walks.

The random walk in edge2vec is similar to that of node2vec. When the random walk is at node v, it randomly choose a neighbor x to go based on the transition probability between v and x. Also, edge2vec keeps the return parameter p and in-out parameter q in node2vec, which makes edge2vec can perform flexible neighborhood sampling as node2vec doses.

2.2.4 Other Methods

There are some other random walk based methods, which use different strategies to generate transition probability matrices in order to contain more structure and semantic information in network.

Gat2vec

Gat2vec is a random walk based method [105]. It not only extract structural information but also content information. Two sub-networks, structural network and bipartite network, are generated based on one original network. Nodes are connected through attributes in bipartite network based on the assumption that two nodes are similar when they share one attribute. In structural network, edges between nodes only indicate the structural connections in the original network. Then, a shallow neural network is trained by the walks generated from two sub-networks. At last, node vectors are learned through the weights of neural network. This method is more suitable for simple homogeneous networks with attributes, and is therefore not appropriate for Hetionet.

Metapath2vec

Metapath2vec constrain random walk only walk through predefined metapaths by giving zeroes to transition probability to nodes which are not contained in the metapaths in TPM [106]. Implementing metapath2vec needs to pre-define metapaths for a network, which is also time-consuming and inefficient. Also, metapath strategy is applied in Rephetio [46] for generating engineered features and perform drug repositioning, which is similar to metapath2vec methodology. The purpose of this thesis is to compare the learned features and engineered features, so using methods highly different from Rephetio is more reasonable and convincing for accomplishing such a purpose.

3 Motivation & Outline

Drug repositioning is time-efficient and less costly compared to traditional drug design processes. Computational methods are applied for drug repositioning to generate hypotheses for new indications for a drug candidate. With powerful algorithms and models, the process of drug design could be even shorter. As introduced in Biological Background, Himmelstein extract engineered feature vectors form Hetionet then train a logistic regression model to differentiate and predict whether drug-disease edge should exist in network.

Engineered features are more interpretable, but it takes lots of time to engineer and choose them, also lose much information like the neighbors and topological structure. If learned features with network representation learning are applied, it would be much easier for generating embeddings, tuning hyperparameters (e.g., dimensions) and keeping information. Muslu et al. [90] created protein-protein interaction network with differential gene expression as attribute to prioritize targets for diseases (Figure 7). They used learned features generated from DeepWalk to train a logistic regression model and got AUC-ROC values between 0.92 - 0.94.

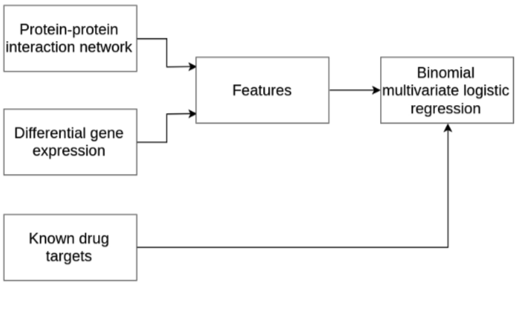
Protein-protein interaction network
Differential gene expression
Known drug targets
Features
Binomial multivariate logistic regression

Figure 7: Workflow of GuiltyTargets. Figure adapted from [90]

These two work inspired the idea that replacing engineered features with learned features, then compare evaluation results between two embedding methods (Figure 8). At last, predict new drug-disease pairs with well tuned models trained by learned features.

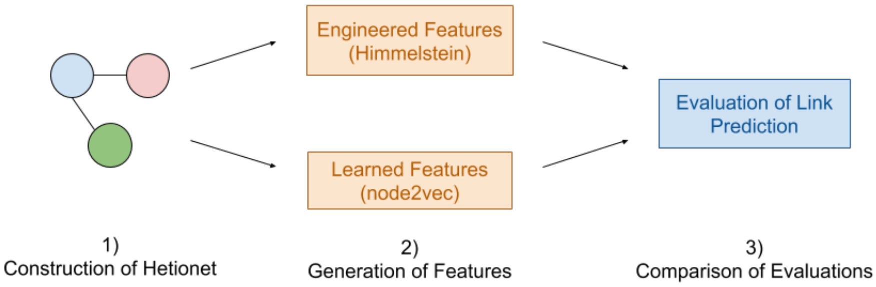
Engineered Features (Himmelstein)
Learned Features (node2vec)
Evaluation of Link Prediction
1) Construction of Hetionet
2) Generation of Features
3) Comparison of Evaluations

Figure 8: Comparison workflow between engineered features and learned features [107]

4 Materials and Methods

This thesis starts with reproducing Rephetio since the data set used in this thesis is from Rephetio project and a solid comparison between engineered features in Rephetio and learned features can only be achieved when Rephetio is fully understood. Then, the parameters of node2vec and edge2vec were optimized to generate the best performing model for comparison and novel drug-disease predictions. A Python package was built for reproducibility of this work. The network Hetionet applied in this thesis is downloaded from GitHub repositories.

4.1 Reproduction of Rephetio

Himmelstein et al. [49] generated a biological knowledge graph with more than 47k nodes and 2000k edges. It includes genes, diseases, compounds, biological processes and so on. The edges are annotated as increase, decrease, target and so on. Himmelstein uses metapath to analyze the graph, create a metric named DWPC, to represent the prevalence of each metapath. With engineered features, to train logistic regression model with known indication [78].

Daniel Himmelstein et al. released all data and note books about his work. The ‘integrate’ repository is to integrate all database and curation to graph, also generate 5 permutation graphs. The ‘learn’ repository is to prepare features and evaluate them, then train and validate the logistic regression model. During the process of reproducing, the most difficulties are come from two parts, environment and the size of the graph is too big. Because he released notebooks, not a python package, even though he gives a configuration file, still many errors come out due to environmental problems. The environmental problem could be solved by changing the version of a package or upgrade something in python.

Due to the size of Hetionet, parsing the whole graph for testing functions is unrealistic. But with a random graph, it is meaningless to analyze it. A meaningful subgraph needs to be generated from Hetionet. With drug-disease pairs and their labels, n pairs of positive samples and m negative samples are randomly chosen. Then all simple paths are extracted between them with a cutoff 3, which means the length of paths. In such a way, the drug-disease paris are kept and metapaths are kept too. With those characteristics, this subgraph is qualified to represent the Hetionet

4.2 Optimize Parameters of node2vec

Parameters were greedily optimized in the following order, from dimensions to q according to the paper of node2vec and Palumbo's work [108]:

dimensions: Dimensions of feature vectors, a parameter from word2vec model, represents the dimensions of word (node) vectors. The most common empirical for word2vec model is 100, also 100 is the default value of word2vec python library. The bigger the dimension is, the more information it contains, easier to overfit for the model. For data in this thesis, less dimensions are better for prediction. 16, 32, 48, 64, 159 are tried to generate node embeddings.

walk_length: Walk_length gives the number of nodes for one walk to go through. It is the length of the 'sentence' parsed into skip-gram model. Node2vec is sensitive to walk length. [89] 30, 50, 100 are tried in this thesis.

window size: Window size is a parameter from word2vec model, illustrate the distance between words which affect each other in the model [91]. For example, in the sentence "Lingling is doing her master thesis in fraunhofer scai", when the window size is three, the word 'her' would be affected by words 'Lingling', 'is', 'doing' on the left and words 'master', 'thesis', 'in' on the right. 2, 3, 4, 10 are tried in this thesis.

number_walk: Number_walk is the number of walks started from every node in the graph. Also the times of random walk to iterate the graph. Node2vec is sensitive to number_walk too [89], [100]. Bigger the number_walk is, more different walks would be generated from the graph, better representative the skip-gram model should be. 10, 20, 30 are tried in this thesis.

p: Return parameter, controls the possibility for a walk to visit the previous node. (more description in method), if set p a high value (greater than max (q, 1)), it is less likely to visit the previous node. The random walk is encouraged to explore further nodes from the visited nodes. If set p a low value (smaller than min (q, 1)), it is more likely to visit the previous node and the walk is more likely to explore the 'local' area around the starting node. 1.0, 2.0, and 0.5 are tried in this thesis.

q: In-out parameter, controls the walk to go to 'inward' node or 'outward' node. If $q > 1$, the walk is biased towards nodes closer to the previous visited node. If $q < 1$, the walk is more likely to go to further node of the previously visited node. 1.0, 2.0, and 0.5 are tried in this thesis.

4.3 Optimize Parameters of Edge2vec

The edge2vec [95] algorithm is based on Node2vec [94]. It uses EM to exploit the correlation level between different edge types in order to capture edge semantics information for biological heterogeneous network. The EM would make transition probability matrix converge after a certain time of updating. It is unknown how many times it would take for getting a converged transition probability matrix. The number of all edges are above 2 million with 24 types. According to the algorithm of edge2vec, it calculates the frequency of every edge types in every walk. So 100 is a reasonable number for one walk to cover all edge types theoretically. Also 100 is a good choice from node2vec to make results more stable. 10,000 gives 10,000 pairs of data to calculate the correlation between two edge types. In consideration of time consuming problem, 10,000 is a suitable number to try. The edge2vec hyperparameters were optimized greedily in the order below:

em_iteration: A number to control how many times for EM to update TPM, is important for the quality of embeddings. Bigger the number is, closer for Transition Probability to converge, more time to run. 5, 10 are tried in this thesis.

max_count: A number to control how many edges to sample for one epoch in EM, is very significant for the quality of TPM. Larger the number is, more edges sampled, the more accurate the transition probability is. 1,000 and 10,000 are tried in this thesis.

e_step: A parameter to choose which test function to be used for evaluate the correlation level of two edge types. 4 test functions are tried in this thesis. More details in Table 2.

4.4 Positive Unlabeled Learning

The way in which Himmelstein *et al.* [49] trained logistic regression model is suitable for proving that engineered features (e.g., DWPC) can represent Hetionet well with an optimistic AUC-ROC value. In this thesis, the goal is to get a prediction model. With a heavily imbalanced training dataset (drug-disease pairs with edges : drug-disease pairs with no edges = 1 : 30), the logistic regression model would not perform well for prediction. The training set was created in the way that positive samples are drug-disease pairs with indication relationships (have an edge between the drug and disease), negative samples are pairs with unknow relationships (have no edge between the drug and the disease). This strategy would cause a problem that some drug-disease pairs are in training data set, then the prediction is inaccurate.

To solve this positive unlabeled (PU) learning problem, the negative samples are generated by sampling the unlabeled drug-disease pairs with the size equal to that of positive samples when the pairs needed to be predicted are out of the unlabeled data set. Also a golden data set repoDB [53] for drug repositioning can be used for training and evaluation.

4.5 Reproducibility Statement

Building a python package makes this work reproducible. Everything is wrapped in package to work automatically after running it through command line. Because data used in this thesis is released by other scientists, so download.py is written to download data from GitHub repositories. The values of parameters are parsed through a configuration file, because the number of parameters is beyond 10, if parsing them through command line interface, it would be confusing and inconvenient. Also, the configuration file is stored in the output directory, which is easier for operating statistics and analysis about them. The package for this thesis is available in GitHub.

5 Results and Discussion

In this thesis, parameters of node2vec and edge2vec are optimized to generate best performance classification models. Then comparison between learned features and engineered features are conducted by replacing engineered feature vectors with learned feature vectors generated by node2vec and edge2vec. Furthermore, the chosen models are used to predict novel drug-disease edges. The below paragraphs demonstrate results of comparisons and predictions.

5.1 A Comparison Between Engineered and Learned Features

For comparing the performances of classification models trained by learned features and engineered features, the first step is to optimize the parameters of node2vec and edge2vec to generate high-quality feature vectors. After the parameter optimization, comparisons are conducted between the chosen models and Rephetio's.

5.1.1 Parameter Optimization of Node2vec and Comparisons

As discussed in section 4.2 and 4.3, dimensions, num_walks, *walk_length*, *p*, *q* and operator are optimized sequentially in this thesis. First of all, the AUC-ROC is chosen to evaluate the models trained by learned features and then parameters are selected based on evaluations.

The fisrt optimized parameter is *dimensions*, As illustrated in Figure 9, *dimensions* is changed from 16, 32, 48, 64, 159. The AUC-ROCs drop to 0.5 when $dimensions = 64$, but return to similar values with $dimensions = 32, 48$ when $dimensions = 159$, which doesn't mean 64 dimensions is not suitable for this method because it may be caused by the randomness of node2vec and edge2vec during random walk. Even though, this result is meaningful in which it shows that node2vec and hetionet is not sensitive to *dimensions*. Since smaller *dimensions* has comparable performance with larger *dimensions*, the smaller one 64 is chosen for avoiding overfitting.

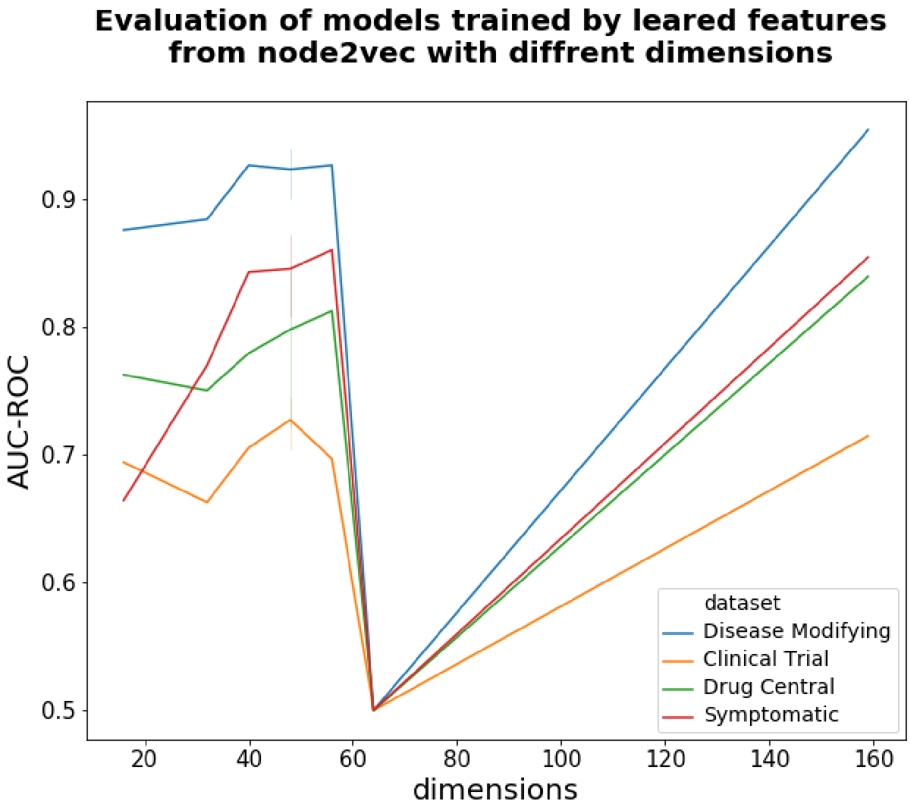

Figure 9: Performance of logistic regression models with parameters: $p = 1$, $q = 1.0$. $repeat = 1$, $walk\_length = 30$, $window\_size = 10$, $number\_walk = 10$, $dimensions = 16, 32, 48, 64, 159$ Star signs are values of Rephetio.

The randomness challenges the process of parameter optimization when the results fluctuate with same parameters. The next step is to control the effects of randomness by tuning relative parameters. In node2vec and edge2vec, the core model is skip-gram, whose actual training data is 'walks' generated by random walk. If the random walk is operated for more times, the training data should be more comprehensive and representative for the network. Then *num_walks* is tested with 10, 20, 30. Surprisingly, among all repeated experiments with same other parameters except *num_walks*, $num\_walks = 10$ gives 90% valid models (logistic regression models trained by learned features from node2vec whose evaluation results are not 0.5), $num\_walks = 20$ giving 40% ones, and $num\_walks = 30$ giving 40% ones too. The reason of such results might be that the hetionet is a complexed network, when *num_walks* gets larger, the 'walks' becomes more diverse and sparse for the skip-gram model. Then the logistic regression models fail with low-quality learned features. Also, it was reported that node2vec can catch useful structure information of network with small *num_walk*. At last, $num\_walks = 10$ is chosen.

Another important parameter is *window size*, which controls distance of nodes affecting the center node in walks in node2vec scenario. For evaluating the results properly, 10 repeated experiments are operated with same parameters to reduce the confusion from randomness. 2, 6, 10 are tried in this thesis. As illustrated in Figure 10, the average of AUC-ROC values get better when *window size* gets larger for Drug Central and Symptomatic. For Clinical Trial, the results of $windowsize = 2, 6, 10$ are similar but $windowsize = 10$ makes the performance of models more stable. $windowsize = 10$ is chosen at last.

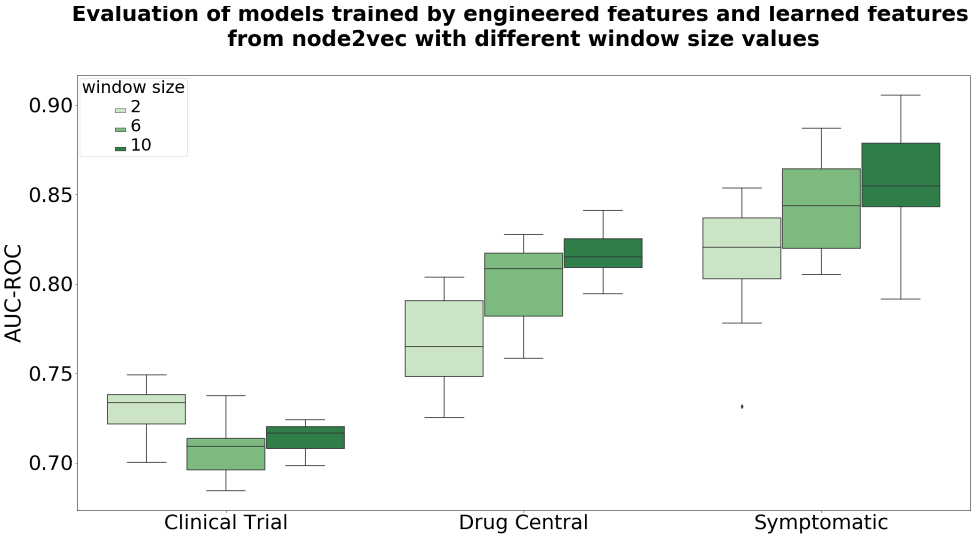

Figure 10: Performance of logistic regression models with parameters: $p = 1$, $q = 1.0$, $repeat = 10$, $walk\_length = 30$, $windowsize = 2, 6, 10$, $number\_walk = 10$.

The next step is to optimize p and q, As shown Figure 11, experiments are repeated ten times with same hyperparameters to test the robustness. Comparing performances of models to star signs (AUC-ROC values of Rephetio), two data sets, Clinical Trial and Symptomatic give better evaluation results than Rephetio. Disease Modifying, the data set including training data, gives a smaller AUC-ROC result with which shows that the models are less prone to be overfitting. Drug Central demonstrates comparable evaluation result to Rephetio. The comparison suggests that learned features can replace engineered features for building classification models with a better and more stable performance.

The performance of models improves with q getting larger, suggesting that models with more local neighborhood information of the network are better than models with more global neighborhood information. This indicates that for a biological network, the local information like neighbors in small distance is more meaningful than that in large distance. In Rephetio work, the length of metapath is 2, 3 or 4, which is another indication that local information is comparatively more valuable than global information.

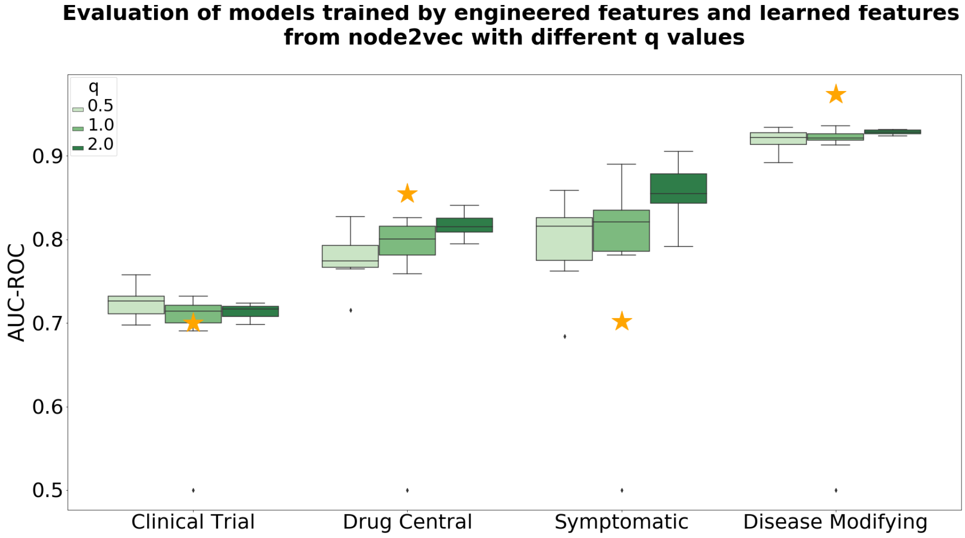

Figure 11: Performance of logistic regression models with parameters: $p = 1$, $repeat = 10$, $walk\_length = 30$, $window\_size = 10$, $number\_walk = 10$, $q = 0.5, 1.0, 2.0$. Star signs are values of Rephetio.

walk_length is the parameter to affect the length of walks. Larger *walk_length* makes random walk go through more nodes, then more neighbourhood information should be catched. As shown in Figure 12, it is obvious that the Symptomatic data set are getting narrower when walk length gets larger. It shows that larger walk_length makes the performance more stable, and the AUC-ROC values of $walk\_length = 100$ are comparable with those of $walk\_length = 30, 50$. In this case, $p = 1$ and $q = 2$, make the random walk more local, and with walk_length = 100, the random walk could explore over the local vicinity of the starting node. As a consequence, the performances of models tends to be similar because of the random walk catches similar and comprehensive information. For the other three data sets, changing of walk_length doesn't make much difference to the evaluation results.

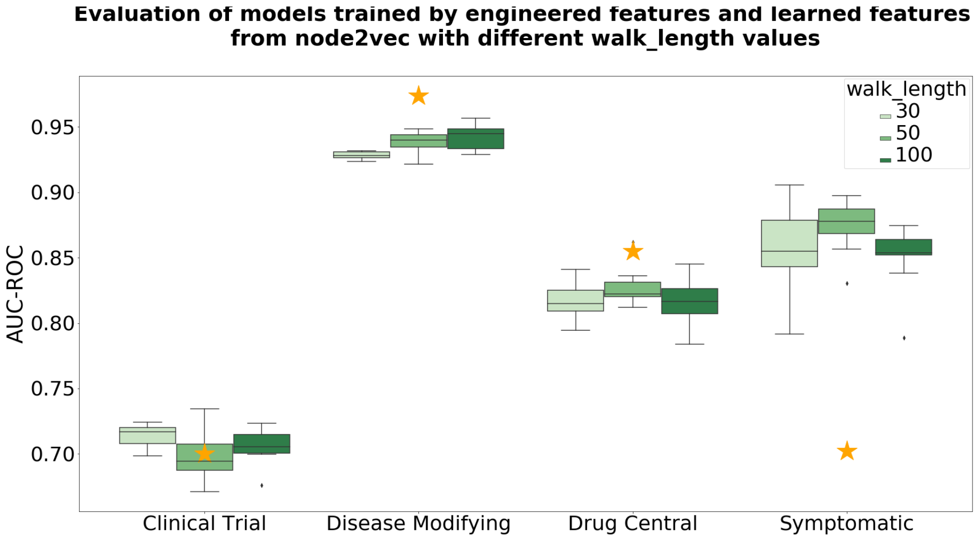

Figure 12: Performance Logistic regression models with parameters : $p = 1$, $q = 2$, $repeat = 10$, $window\_size = 10$, $number\_walk = 10$, $walk\_length = 30, 50, 100$. Star signs are values of Rephetio.

AU-PR is more suitable for evaluating imbalanced data sets. As illustrated in Figure 13, Two evaluation data sets (Clinical Trial and Symptomatic) show better results than Rephetio's. In summary, learned features from node2vec can replace enginnered features in Rephetio, and learned features can perform better with an easy implementation.

5.1.2 Parameter Optimization of Edge2vec and Comparisons

Edge2vec is improved from node2vec. The parameters of edge2vec are optimized based on the results of optimizing parameters of node2vec. But there are some unique parameters in edge2vec such as *em_iter* and *max_count* (discussed in section 4.3). *em_iter* controls the epochs of sampling the network to update

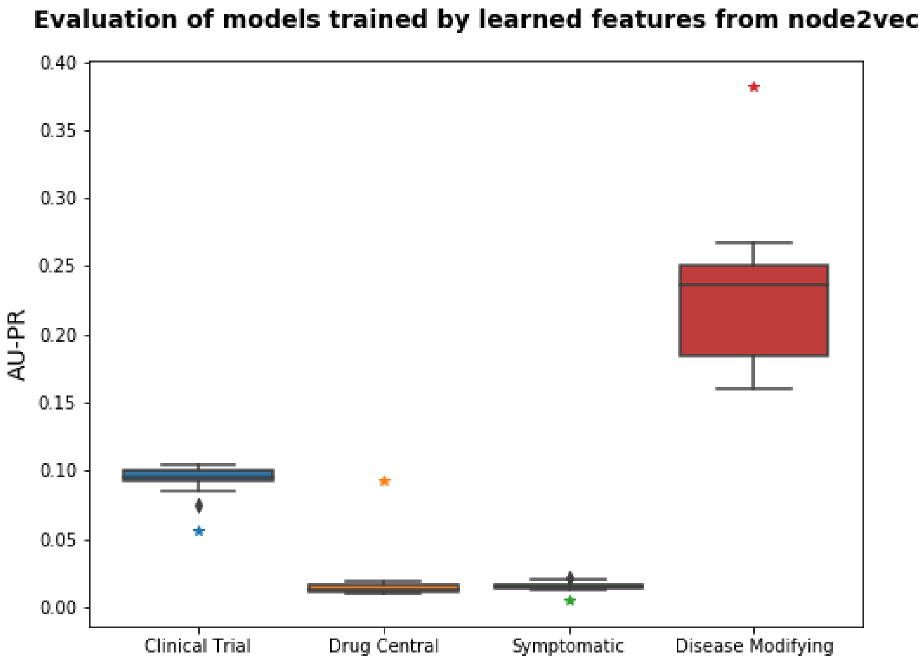

Figure 13: Performance of logistic regression models with parameters : $p = 1$, $q = 2$, $repeat = 10$, $window\_size = 10$, $number\_walk = 10$, $walk\_length = 100$. Star signs are AU-PR values of Rephetio.

TPM, $em\_iter = 10$ is chosen. max_count is important for the quality of TPM, $max\_count = 10000$ is chosen.

Another important parameter is the e_step, which controls the test function to evaluate correlation level between two edge types (explained in Table 2). As in node2vec, 10 times repeated experiments are operated to test the robustness of randomness with same parameters. 4 different tests, wilcoxon signed-rank ($e\_step = 1$), entropy ($e\_step = 2$), pearson ($e\_step = 3$), spearman ($e\_step = 4$) are all tried in this thesis as shown in Figure 14.

The evaluation results of 4 tests all have high variance. The first test, wilcoxon signed-rank performs comparatively better than other three test functions. Further optimization should be done to control the high variance. Then *window size* is changed to 3, in Figure 15, the evaluation results show that $window\_size = 3$ is more suitable for edge2vec applied in hetio, so $e\_step = 1$ and $window\_size = 3$ are chosen.

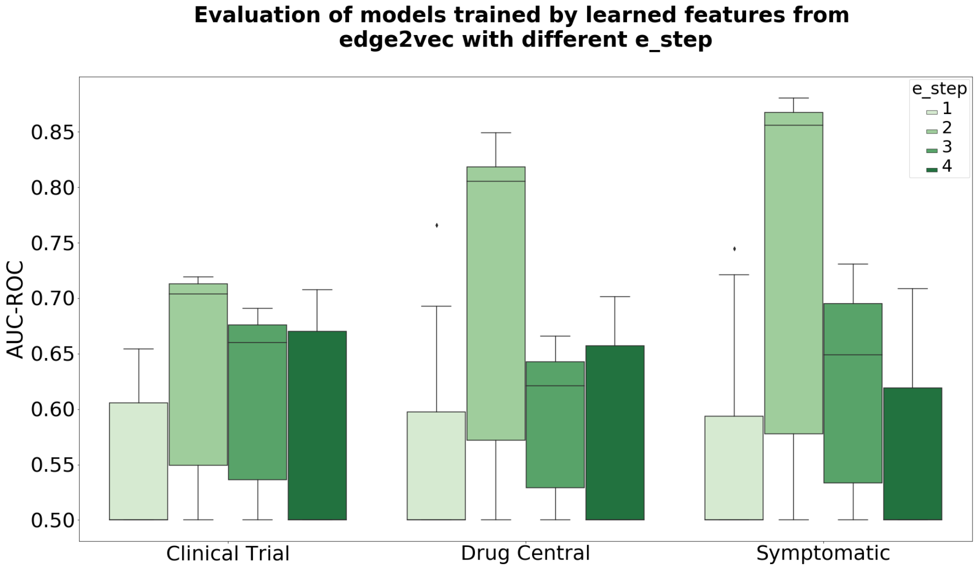

Figure 14: Performance Logistic regression models with parameters : $p = 1, q = 1, repeat = 10, window\_size = 10, number\_walk = 10, walk\_length = 100, em\_iter = 10, max\_count = 10000, e\_step = 1, 2, 3, 4$

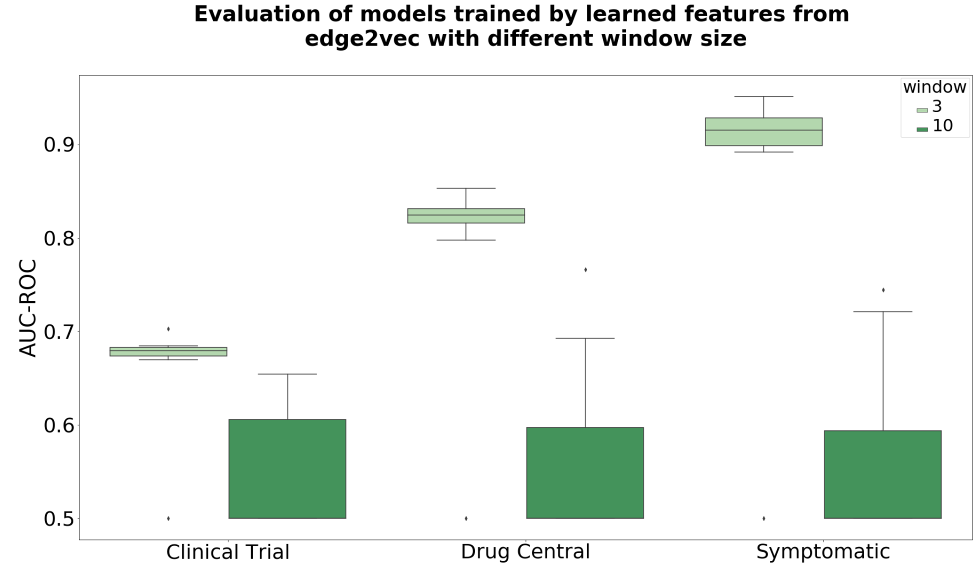

Figure 15: Performance Logistic regression models with parameters : $p = 1, q = 1, repeat = 10, window\_size = 3, number\_walk = 10, walk\_length = 100, em\_iter = 10, max\_count = 10000$. Star signs are values of Rephetio.

As illustrated in Figure 16, comparing the performance of models to star signs (AUC-ROC values of models from Rephetio), the evaluation results of Symptomatic data set surpass Rephetio's with a gap around 0.2. For the other three data sets, performances are comparable to Rephetio's. In Figure 17, AU-PR values of Clinical Trial and Symptomatic of edge2vec outperform that of Rephetio.

Symptomatic data set is from the same source of Disease Modifying (training data included). The outstanding performance of Symptomatic data set suggests that the learned feature vectors from edge2vec represent the network better than engineered features. Furthermore, the performances of 4 data sets of edge2vec are comparatively more stable than that of node2vec , which indicates that edge2vec performs better because it's able to capture the meaningful structure of the network with a more representative TPM for the network.

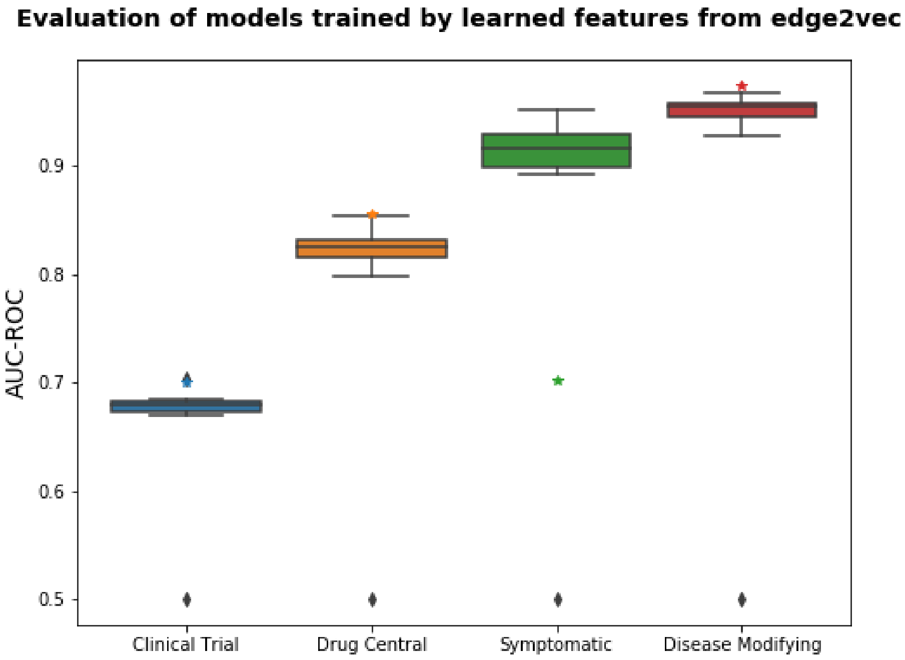

Figure 16: Performance of logistic regression models with parameters : $p = 1$, $q = 1$, $repeat = 10$, $window_size = 3$, $number_walk = 10$, $walk_length = 100$, $em\_iter = 10$, $max\_count = 10000$. Star signs are values of Rephetio.

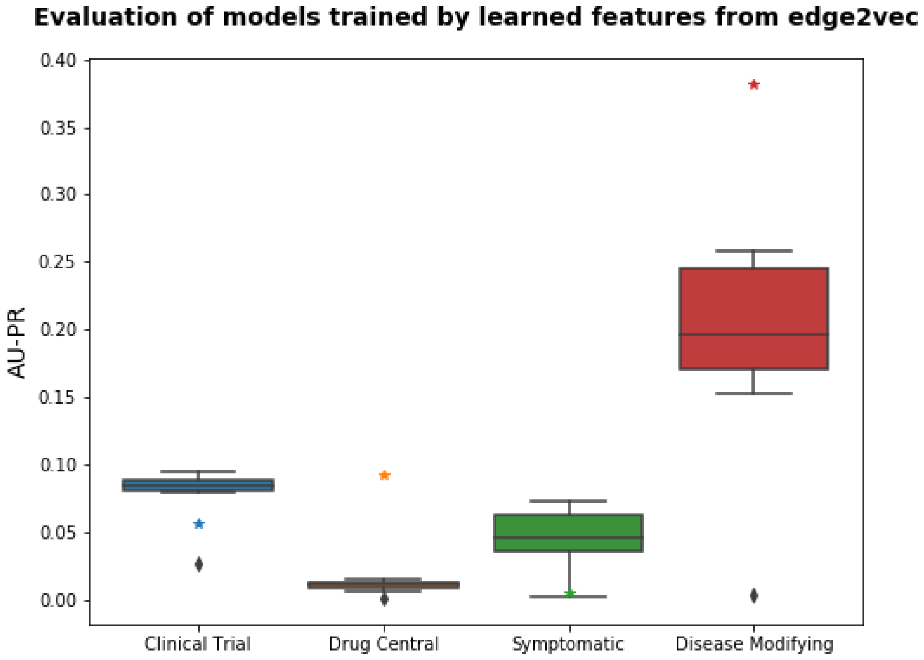

Figure 17: Performance of logistic regression models with parameters : $p = 1$, $q = 2$, $repeat = 10$, $window\_size = 10$, $number\_walk = 10$, $walk\_length = 100$. Star signs are AU-PR values of Rephetio.

5.1.3 Information of Classification Models with Best Performances

To summary, models of best performance from Rephetio, node2vec, and edge2vec and their AUC-ROC values are presented in Table 3. In Table 4, parameters of chosen best performance models are presented.

	Clinical Trial	Drug Central	Symptomatic	Disease Modifying
Rephetio	0.700000	0.855000	0.702000	0.974000
Node2vec	0.714927	0.844998	0.852863	0.956823
Edge2vec	0.702855	0.829167	0.927183	0.967201

Table 3: AUC-ROC values of chosen models

	num_walk	walk_length	window_size	p	q	dimensions	operator	em_iter	max_count	e_step
Node2vec	10	100	10	1	2	48	Hadamard			
Edge2vec	10	100	3	1	1	48	Hadamard	10	10000	1

Table 4: Parameters of best performance models for node2vec and edge2vec

5.2 Predictions of Diseases for Vorinostat

Vorinostat is a histone deacetylase (HDAC) inhibitor, affecting the HDAC enzymes regulated the phenotypic and genotypic expression in cells in order to homeostasis and neoplastic growth. Apart from cutaneous T cell lymphoma the role of vorinostat for other types of cancers is being investigated [109]. Because the affection of HDACs in neurodevelopment, memory formation, and cognitive processes, HDAC inhibitors are expected to be novel agents neurodegenerative disorders such as AD.

As shown in Table 5, "probability" represents the probability of the prediction being true. Table 5 is sorted by the descending value of probability. 18 is generated from Drug Repurposing project in Hetionet, finding supportive paths between vorinostat and AD.

Disease Ontology Identifier	**Disease Name**	**Probability**
10652	alzheimer's disease	0.566710
1312	focal segmental glomerulosclerosis	0.551427
9744	type 1 diabetes mellitus	0.547426
4159	skin cancer	0.543505
10608	celiac disease	0.537503
14330	Parkinson's disease	0.532105
10976	membranous glomerulonephritis	0.528259
332	focal segmental glomerulosclerosis	0.522691
4989	pancreatitis	0.514514

Table 5: Part of repositioning prediction results of HDAC6 inhibitor, vorinostat

As shown in Figure 5, the ten most supportive path for vorinostat-AD pair. HDAC enzymes are keys to connect vorinostat and AD [110]. Rustenhoven et al. also suggest vorinostat could be a new drug for AD due to its utility of limiting microglial-mediated inflammatory responses [111]. Figure 5 gives paths that vorinostat and memantine share the same side effects Thromboembolism and Squamous cell carcinoma, which could be another underlying mechanism that vorinostat could work in a similar way as memantine.

The highest probability of prediction for vorinostat is colon cancer. And the probability of vorinostat - AD pair is not high, which is reasonable because HDAC-6 is most relevant to cancers based on studies and researches until now, AD

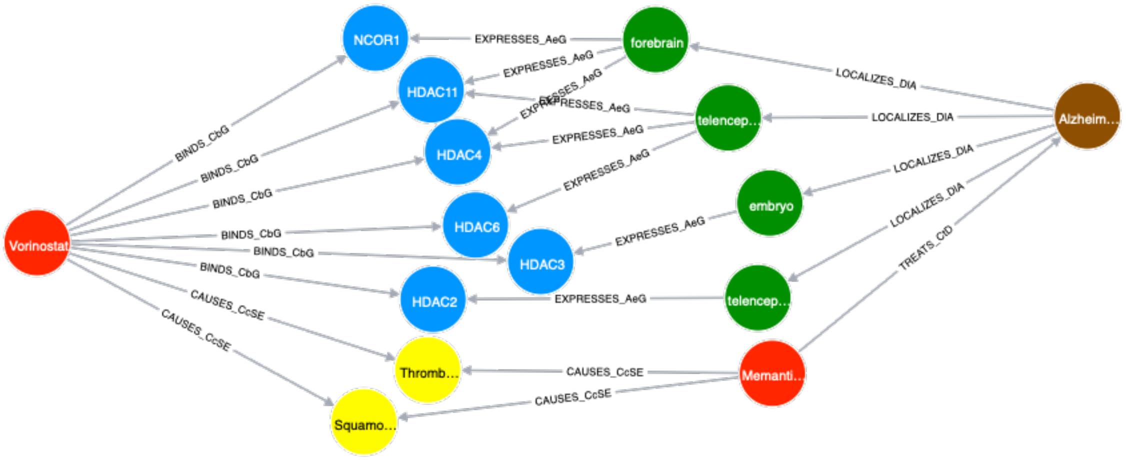

Figure 18: Ten most supportive paths connecting vorinostat and Alzheimer's disease

and Parkinson's disease are predicted to have an edge with vorinostat with a comparatively low but still positive probability, which suggests that the model is able to catch information of graph and give convincing predictions.

5.3 Predictions of Drugs for Alzheimer's Disease

Table 6 shows ten of the drug candidate predictions for AD.

DrugBank Compound ID	Compound Name	Probability
DB06287	Temsirolimus	0.90302
DB00193	Tramadol	0.89986
DB00715	Paroxetine	0.89033
DB00877	Sirolimus	0.88865
DB00724	Imiquimod	0.85494
DB00285	Venlafaxine	0.85105
DB08828	Vismodegib	0.84252
DB00494	Entacapone	0.82564
DB00289	Atomoxetine	0.82018
DB00996	Gabapentin	0.81912

Table 6: Part of predictions for Alzheimer's disease

Temsirolimus is an inhibitor of mammalian target of rapamycin (mTOR) kinase, a component of intracellular signaling pathways involved in the growth and proliferation of cells [112]. It was approved by Food and Drug Administration (FDA) in 2007 to be used for metastatic renal-cell carcinoma [113]. Jiang et al. found that temsirolimus promotes autophagic clearance of amyloid-β and provides protective effects in cellular and animal models of AD [114]. Christelle et al. also suggested that temsirolimus was able to alleviate tau pathology in mutant tau transgenic mice [115]. As illustrated in Figure 19, the underlying mechanism might be that temsirolimus affects PPP3CA, and PPP3CA can alter the sensitivity of animals to electrophysiological and cognitive impairments caused by amyloid-β exposure [116]. The third ranking drug is paroxetine, which was researched that it lowers intracellular APP while maintaining APLP-1 levels [117], which could be the underlying mechanism of paroxetine connecting to AD.

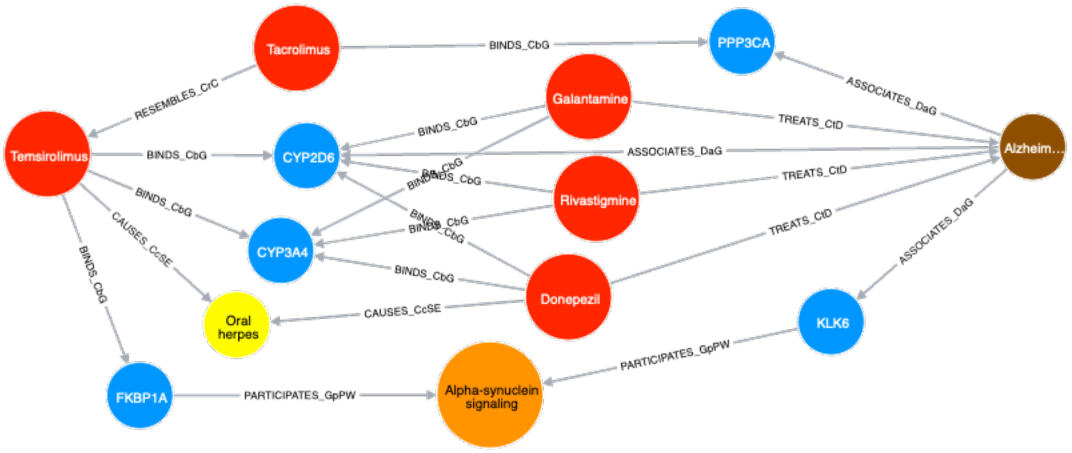

Figure 19: Ten most supportive metapaths connecting temsirolimus and Alzheimer's Disease. Temsirolimus-resembles-Tacrolimus-binds-PPP3CA-associates-Alzheimer's Disease is the path having supportive researching evidence.

In conclusion, network is a data structure good for applying computational algorithms and visualization. The work of this thesis indicates that node2vec and edge2vec are powerful NRL methods for generating high-quality feature vectors from networks to train downstream machine learning models in order to conduct drug repositioning tasks. And compared with engineered features, the advantages of learned features are simple implementation and making machine learning models more generalizable.Therefore, learned features can replace engineered features for drug repositioning tasks with a better performance with one set of optimized parameters.

6 Conclusion and Future Work

6.1 Reflections

This thesis truly benefited from open resource data. Hetionet is open resource on GitHub, all the original data and the pre-processing are available, organized and easy to get access to. Also, reading all thinklab notes about Rephetio project, gives me a teaching that how other scientists think and work, which is very inspiring and makes his work more convincing and reliable. Besides, it is also easier to understand and evaluate the work with all data and codes. Edge2vec python package is released in GitHub, but the codes were unorganized and gave errors when running with Hetionet. Modifying edge2vec package and implementing multiprocessing method into it consumed almost one month in this thesis. Even open-resource data is common in recent years, quality of the data is the further issue to be addressed.

The logistic regression model is powerful to differentiate classes. Other classification methods can be implemented too such as support vector machine. And with embeddings, not only drug-disease pairs could be classified, drug-target, target-disease and many other models could be built, which haven't been completed in this thesis due to time limitations.

6.2 Limitations

Hetionet was released 3 years ago. Some compounds are not in the network and diseases are limited to 136 kinds of diseases, way smaller than disease types in the real world. If more information contained in Hetionet, such as phosphorylation and degradation, the network would be more informative and the models would perform better.

Then another problem comes after having a large network, which is to improve the efficiency and reduce training time. Even with the Hetionet, it was a problem for optimizing parameters with 200k edges. When running a experiment takes 2-4 hours and it was necessary to repeat experiments for reduce the confusion of randomness, optimizing one set of parameters took for almost one day, which limited the options of parameters. Therefore, a more efficient algorithm is in highly demanding.

Another serious limitation is that node2vec and edge2vec basically catch structure information of the network. Node2vec only catch topological structures and treat a heterogeneous network as homogeneous, at last, lots of semantic information is lost. For edge2vec, the correlation information between edge types is considered, but semantic information of nodes are abandoned too.

For training a binary classification model, positive samples are important. Negative samples are significant to the quality of models too. But in the real world, there is not much reliable source for negative samples of drug-disease relationships, which means the drug can't treat the disease. In such a situation, negative samples are randomly chosen from unlabeled drug-disease pairs.

6.3 Future Work

The models built in this thesis perform well in predicting new drug-disease edges. But there are some space for improvement. The first one is to update the network with new data and enrich the network with more specific biological information such as phosphorylation, degradation and quantitative assay data (e.g., IC50). The quantitative data can be used as weights for edges in the network to further constrain random walk. Then the following issue is to normalize the quantitative assay data.

The second one is to improve the computational speed of algorithms. Because the logistic regression models in this thesis were trained by drug-disease edge vectors, so they can only predict novel drug-disease relationships. But in reality, drug-target, drug-disease predictions are also important for drug repositioning, so the third future work is to build other models for predicting drug-target, target-disease with respective vectors generated from node2vec or edge2vec model.

The third one is to apply repoDB [53] to train and evaluate classification models. RepoDB contains drugs failed in clinical trials, which can be negative samples to train models to make classification models more realistic. Then the predictions would be more accurate. Also repoDB can be used for evaluate the current logistic models. The evaluation results would be more solid and convincing.

Bibliography

[1] Michael Hay, David W. Thomas, John L. Craighead, Celia Economides, and Jesse Rosenthal. “Clinical development success rates for investigational drugs”. eng. In: *Nature Biotechnology* 32.1 (Jan. 2014), pp. 40–51.

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

[3] Natalia Novac. “Challenges and opportunities of drug repositioning”. en. In: *Trends in Pharmacological Sciences* 34.5 (May 2013), pp. 267–272.

[4] Kyungsoo Park. “A review of computational drug repurposing”. en. In: *Translational and Clinical Pharmacology* 27.2 (2019), p. 59.

[5] Yusuke Kobayashi, Kouji Banno, Haruko Kunitomi, Eiichiro Tominaga, and Daisuke Aoki. “Current state and outlook for drug repositioning anticipated in the field of ovarian cancer”. en. In: *Journal of Gynecologic Oncology* 30.1 (2019), e10.

[6] Yan-Fen Dai and Xing-Ming Zhao. “A Survey on the Computational Approaches to Identify Drug Targets in the Postgenomic Era”. en. In: *BioMed Research International* 2015 (2015), pp. 1–9.

[7] Ravi Kiran Reddy Kalathur et al. “UniHI 7: an enhanced database for retrieval and interactive analysis of human molecular interaction networks”. eng. In: *Nucleic Acids Research* 42.Database issue (Jan. 2014), pp. D408–414.

[8] Vivian Law et al. “DrugBank 4.0: shedding new light on drug metabolism”. en. In: *Nucleic Acids Research* 42.D1 (Jan. 2014), pp. D1091–D1097.

[9] 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.

[10] Hsiao-Rong Chen, David H. Sherr, Zhenjun Hu, and Charles DeLisi. "A network based approach to drug repositioning identifies plausible candidates for breast cancer and prostate cancer". eng. In: *BMC medical genomics* 9.1 (2016), p. 51.

[11] Marilyn Safran et al. "GeneCards Version 3: the human gene integrator". eng. In: *Database: The Journal of Biological Databases and Curation* 2010 (Aug. 2010), baq020.

[12] Emily Clough and Tanya Barrett. "The Gene Expression Omnibus Database". eng. In: *Methods in Molecular Biology (Clifton, N.J.)* 1418 (2016), pp. 93–110.

[13] Dennis P. Wall et al. "Genotator: a disease-agnostic tool for genetic annotation of disease". eng. In: *BMC medical genomics* 3 (Oct. 2010), p. 50.

[14] Damian Szklarczyk et al. "STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets". eng. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D607–D613.

[15] Minoru Kanehisa and Yoko Sato. "KEGG Mapper for inferring cellular functions from protein sequences". en. In: *Protein Science* (Aug. 2019), pro.3711.

[16] Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders". eng. In: *Nucleic Acids Research* 33.Database issue (Jan. 2005), pp. D514–517.

[17] Sabry Razick, George Magklaras, and Ian M. Donaldson. "iRefIndex: a consolidated protein interaction database with provenance". eng. In: *BMC bioinformatics* 9 (Sept. 2008), p. 405.

[18] Harriet Keane, Brent J. Ryan, Brendan Jackson, Alan Whitmore, and Richard Wade-Martins. "Protein-protein interaction networks identify targets which rescue the MPP+ cellular model of Parkinson's disease". eng. In: *Scientific Reports* 5 (Nov. 2015), p. 17004.

[19] Evelyn Braungart, Manfred Gerlach, Peter Riederer, Ralf Baumeister, and Marius C. Hoener. "Caenorhabditis elegans MPP+ model of Parkinson's disease for high-throughput drug screenings". eng. In: *Neuro-Degenerative Diseases* 1.4-5 (2004), pp. 175–183.

[20] Yanbin Liu et al. "DCDB 2.0: a major update of the drug combination database". eng. In: *Database: The Journal of Biological Databases and Curation* 2014 (2014), bau124.

[21] Stefan Günther et al. "SuperTarget and Matador: resources for exploring drug-target relationships". eng. In: *Nucleic Acids Research* 36.Database issue (Jan. 2008), pp. D919–922.

[22] Ida Schomburg, Antje Chang, and Dietmar Schomburg. "BRENDA, enzyme data and metabolic information". eng. In: *Nucleic Acids Research* 30.1 (Jan. 2002), pp. 47–49.

[23] Noa Rappaport et al. "MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search". eng. In: *Nucleic Acids Research* 45.D1 (2017), pp. D877–D887.

[24] Michael Kuhn, Christian von Mering, Monica Campillos, Lars Juhl Jensen, and Peer Bork. "STITCH: interaction networks of chemicals and proteins". eng. In: *Nucleic Acids Research* 36.Database issue (Jan. 2008), pp. D684–688.

[25] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. "Prediction of drug-target interaction networks from the integration of chemical and genomic spaces". eng. In: *Bioinformatics (Oxford, England)* 24.13 (July 2008), pp. i232–240.

[26] Xiao-Ying Yan, Shao-Wu Zhang, and Song-Yao Zhang. "Prediction of drug-target interaction by label propagation with mutual interaction information derived from heterogeneous network". eng. In: *Molecular bioSystems* 12.2 (Feb. 2016), pp. 520–531.

[27] Hailin Chen and Zuping Zhang. "A Semi-Supervised Method for Drug-Target Interaction Prediction with Consistency in Networks". en. In: *PLoS ONE* 8.5 (May 2013). Ed. by Ozlem Keskin, e62975.

[28] Salvatore Alaimo, Rosalba Giugno, and Alfredo Pulvirenti. “Recommendation Techniques for Drug-Target Interaction Prediction and Drug Repositioning”. eng. In: *Methods in Molecular Biology (Clifton, N.J.)* 1415 (2016), pp. 441–462.

[29] Peter Willett. “Similarity-based virtual screening using 2D fingerprints”. en. In: *Drug Discovery Today* 11.23-24 (Dec. 2006), pp. 1046–1053.

[30] Dagmar Stumpfe and Jürgen Bajorath. “Exploring Activity Cliffs in Medicinal Chemistry: Miniperspective”. en. In: *Journal of Medicinal Chemistry* 55.7 (Apr. 2012), pp. 2932–2942.

[31] Liat Perlman, Assaf Gottlieb, Nir Atias, Eytan Ruppin, and Roded Sharan. “Combining drug and gene similarity measures for drug-target elucidation”. eng. In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 18.2 (Feb. 2011), pp. 133–145.

[32] L. M. Schriml et al. “Disease Ontology: a backbone for disease semantic integration”. en. In: *Nucleic Acids Research* 40.D1 (Jan. 2012), pp. D940–D946.

[33] Marc A. van Driel, Jorn Bruggeman, Gert Vriend, Han G. Brunner, and Jack A. M. Leunissen. “A text-mining analysis of the human phenome”. eng. In: *European journal of human genetics: EJHG* 14.5 (May 2006), pp. 535–542.

[34] Huimin Luo et al. “Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm”. en. In: *Bioinformatics* 32.17 (Sept. 2016), pp. 2664–2671.

[35] *ClinicalTrials.gov (2006a) Docetaxel, Doxorubicin, and Prednisone in Treating Patients With Advanced Prostate Cancer That Has Not Responded to Hormone Therapy.* Tech. rep.

[36] Zikai Wu, Yong Wang, and Luonan Chen. “Network-based drug repositioning”. eng. In: *Molecular bioSystems* 9.6 (June 2013), pp. 1268–1281.

[37] Kun Yang, Hongjun Bai, Qi Ouyang, Luhua Lai, and Chao Tang. “Finding multiple target optimal intervention in disease-related molecular network”. en. In: *Molecular Systems Biology* 4.1 (Jan. 2008), p. 228.

[38] Caroline F. Thorn, Teri E. Klein, and Russ B. Altman. “PharmGKB: the Pharmacogenomics Knowledge Base”. eng. In: *Methods in Molecular Biology (Clifton, N.J.)* 1015 (2013), pp. 311–320.

[39] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. "The SIDER database of drugs and side effects". eng. In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D1075–1079.

[40] 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.

[41] Hao Ye, Jia Wei, Kailin Tang, Ritchie Feuers, and Huixiao Hong. "Drug Repositioning Through Network Pharmacology". en. In: *Current Topics in Medicinal Chemistry* 16.30 (Oct. 2016), pp. 3646–3656.

[42] Yunan Luo et al. "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information". en. In: *Nature Communications* 8.1 (Dec. 2017), p. 573.

[43] Craig Knox et al. "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs". eng. In: *Nucleic Acids Research* 39.Database issue (Jan. 2011), pp. D1035–1041.

[44] T. S. Keshava Prasad et al. "Human Protein Reference Database–2009 update". eng. In: *Nucleic Acids Research* 37.Database issue (Jan. 2009), pp. D767–772.

[45] Allan Peter Davis et al. "The Comparative Toxicogenomics Database: update 2013". eng. In: *Nucleic Acids Research* 41.Database issue (Jan. 2013), pp. D1104–1114.

[46] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. "A side effect resource to capture phenotypic effects of drugs". en. In: *Molecular Systems Biology* 6.1 (Jan. 2010), p. 343.

[47] Robert M. Plenge, Edward M. Scolnick, and David Altshuler. "Validating therapeutic targets through human genetics". en. In: *Nature Reviews Drug Discovery* 12.8 (Aug. 2013), pp. 581–594.

[48] Kristin L. Ayers and Heather J. Cordell. "SNP Selection in genome-wide and candidate gene studies via penalized logistic regression". en. In: *Genetic Epidemiology* 34.8 (Dec. 2010), pp. 879–891.

[49] Daniel Scott Himmelstein et al. "Systematic integration of biomedical knowledge prioritizes drugs for repurposing". en. In: *eLife* 6 (Sept. 2017), e26726.

[50] Zhongyang Liu et al. "Similarity-based prediction for Anatomical Therapeutic Chemical classification of drugs by integrating multiple data sources". en. In: *Bioinformatics* 31.11 (June 2015), pp. 1788–1795.

[51] Francesco Napolitano et al. "Drug repositioning: a machine-learning approach through data integration". en. In: *Journal of Cheminformatics* 5.1 (Dec. 2013), p. 30.

[52] Khader Shameer et al. *Prioritizing Small Molecule as Candidates for Drug Repositioning using Machine Learning*. en. preprint. Bioinformatics, May 2018.

[53] Khader Shameer et al. "Systematic analyses of drugs and disease indications in RepurposeDB reveal pharmacological, biological and epidemiological factors influencing drug repositioning". en. In: *Briefings in Bioinformatics* 19.4 (July 2018), pp. 656–678.

[54] 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.

[55] Hanqing Xue, Jie Li, Haozhe Xie, and Yadong Wang. "Review of Drug Repositioning Approaches and Resources". en. In: *International Journal of Biological Sciences* 14.10 (2018), pp. 1232–1244.

[56] Jing Lu et al. "Identification of new candidate drugs for lung cancer using chemical–chemical interactions, chemical–protein interactions and a K-means clustering algorithm". en. In: *Journal of Biomolecular Structure and Dynamics* 34.4 (Apr. 2016), pp. 906–917.

[57] Víctor Martínez, Carmen Navarro, Carlos Cano, Waldo Fajardo, and Armando Blanco. "DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data". en. In: *Artificial Intelligence in Medicine* 63.1 (Jan. 2015), pp. 41–49.

[58] Eric Wen Su and Todd M. Sanger. “Systematic drug repositioning through mining adverse event data in ClinicalTrials.gov”. eng. In: *PeerJ* 5 (2017), e3154.

[59] Juan Li et al. “Telmisartan exerts anti-tumor effects by activating peroxisome proliferator-activated receptor-γ in human lung adenocarcinoma A549 cells”. eng. In: *Molecules (Basel, Switzerland)* 19.3 (Mar. 2014), pp. 2862–2876.

[60] Zhichen Pu, Min Zhu, and Fandou Kong. “Telmisartan prevents proliferation and promotes apoptosis of human ovarian cancer cells through upregulating PPARγ and downregulating MMP-9 expression”. eng. In: *Molecular Medicine Reports* 13.1 (Jan. 2016), pp. 555–559.

[61] Tony Tong-Lin Wu, Ho-Shan Niu, Li-Jen Chen, Juei-Tang Cheng, and Yat-Ching Tong. “Increase of human prostate cancer cell (DU145) apoptosis by telmisartan through PPAR-delta pathway”. eng. In: *European Journal of Pharmacology* 775 (Mar. 2016), pp. 35–42.

[62] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient Estimation of Word Representations in Vector Space”. In: *arXiv:1301.3781 [cs]* (Jan. 2013). arXiv: 1301.3781.

[63] Duc Luu Ngo et al. “Application of Word Embedding to Drug Repositioning”. In: *Journal of Biomedical Science and Engineering* 09.01 (2016), pp. 7–16.

[64] Pan Pantziarka, Vidula Sukhatme, Gauthier Bouche, Lydie Meheus, and Vikas P. Sukhatme. “Repurposing Drugs in Oncology (ReDO)-itraconazole as an anti-cancer agent”. eng. In: *Ecancermedicalscience* 9 (2015), p. 521.

[65] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. “Entrez Gene: gene-centered information at NCBI”. eng. In: *Nucleic Acids Research* 39.Database issue (Jan. 2011), pp. D52–57.

[66] *Daniel Himmelstein, Casey Greene, Alexander Pico (2015) Using Entrez Gene as our gene vocabulary. Thinklab. doi:10.15363/thinklab.d34.*

[67] *Daniel Himmelstein, Sabrina Chen (2015) Protein (target, carrier, transporter, and enzyme) interactions in DrugBank. Thinklab. doi:10.15363/thinklab.d65.*

[68] Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa A Haendel. "Uberon, an integrative multi-species anatomy ontology". en. In: *Genome Biology* 13.1 (2012), R5.

[69] Daniel S. Himmelstein. *User-Friendly Extensions To Mesh V1.0*. Feb. 2016.

[70] *Daniel Himmelstein, Alex Pankov (2015) Mining knowledge from MEDLINE articles and their indexed MeSH terms. Thinklab. doi:10.15363/thinklab.d67*.

[71] Carl F. Schaefer et al. "PID: the Pathway Interaction Database". eng. In: *Nucleic Acids Research* 37.Database issue (Jan. 2009), pp. D674–679.

[72] Daniel S. Himmelstein and Alexander R. Pico. *Dhimmel/Pathways V2.0: Compiling Human Pathway Gene Sets*. Apr. 2016.

[73] Janet Piñero et al. "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes". eng. In: *Database: The Journal of Biological Databases and Curation* 2015 (2015), bav028.

[74] *Daniel Himmelstein, Frederic Bastian, Dexter Hadley, Casey Greene (2015) STARGEO: expression signatures for disease using crowdsourced GEO annotation. Thinklab. doi:10.15363/thinklab.d96*.

[75] Frederic Bastian et al. "Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species". en. In: *Data Integration in the Life Sciences*. Ed. by Amos Bairoch, Sarah Cohen-Boulakia, and Christine Froidevaux. Vol. 5109. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 124–131.

[76] Katja Luck et al. *A reference map of the human protein interactome*. en. preprint. Systems Biology, Apr. 2019.

[77] J. Menche et al. "Uncovering disease-disease relationships through the incomplete interactome". en. In: *Science* 347.6224 (Feb. 2015), pp. 1257601–1257601.

[78] Daniel S. Himmelstein and Sergio E. Baranzini. "Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes". en. In: *PLOS Computational Biology* 11.7 (July 2015). Ed. by Hua Tang, e1004259.

[79] *Daniel Himmelstein (2016) Announcing PharmacotherapyDB: the Open Catalog of Drug Therapies for Disease. Thinklab. doi:10.15363/thinklab.d182.*

[80] Ritu Khare, Jiao Li, and Zhiyong Lu. "LabeledIn: Cataloging labeled indications for human drugs". en. In: *Journal of Biomedical Informatics* 52 (Dec. 2014), pp. 448–456.

[81] Wei-Qi Wei et al. "Development and evaluation of an ensemble resource linking medications to their indications". en. In: *Journal of the American Medical Informatics Association* 20.5 (Sept. 2013), pp. 954–961.

[82] Allison B McCoy et al. "Development and evaluation of a crowdsourcing methodology for knowledge base construction: identifying relationships between clinical problems and medications". en. In: *Journal of the American Medical Informatics Association* 19.5 (Sept. 2012), pp. 713–718.

[83] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharan. "PREDICT: a method for inferring novel drug indications with application to personalized medicine". en. In: *Molecular Systems Biology* 7.1 (Jan. 2011), p. 496.

[84] Amar Koleti et al. "Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data". en. In: *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D558–D566.

[85] *Daniel Himmelstein, Caty Chung (2015) Computing consensus transcriptional profiles for LINCS L1000 perturbations. Thinklab. doi:10.15363/thinklab.d43.*

[86] Martina Kutmon et al. "WikiPathways: capturing the full diversity of pathway knowledge". eng. In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D488–494.

[87] David Croft et al. "Reactome: a database of reactions, pathways and biological processes". eng. In: *Nucleic Acids Research* 39.Database issue (Jan. 2011), pp. D691–697.

[88] The Gene Ontology Consortium. "Gene Ontology Consortium: going forward". en. In: *Nucleic Acids Research* 43.D1 (Jan. 2015), pp. D1049–D1056.

[89] *Daniel Himmelstein (2016) Our hetnet edge prediction methodology: the modeling framework for Project Rephetio. Thinklab. doi:10.15363/thinklab.d210.*

[90] Oezlem Muslu, Charles Tapley Hoyt, Martin Hofmann-Apitius, and Holger Froehlich. *GuiltyTargets: Prioritization of Novel Therapeutic Targets with Deep Network Representation Learning*. en. preprint. Bioinformatics, Jan. 2019.

[91] Dorothea Emig et al. "Drug Target Prediction and Repositioning Using an Integrated Network-Based Approach". en. In: *PLoS ONE* 8.4 (Apr. 2013). Ed. by Patrick Aloy, e60618.

[92] Hyunghoon Cho, Bonnie Berger, and Jian Peng. "Diffusion Component Analysis: Unraveling Functional Topology in Biological Networks". eng. In: *Research in computational molecular biology: ... Annual International Conference, RECOMB ...: proceedings. RECOMB (Conference: 2005-)* 9029 (Apr. 2015), pp. 62–64.

[93] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. "DeepWalk: Online Learning of Social Representations". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14* (2014). arXiv: 1403.6652, pp. 701–710.

[94] Aditya Grover and Jure Leskovec. "node2vec: Scalable Feature Learning for Networks". eng. In: *KDD: proceedings. International Conference on Knowledge Discovery & Data Mining* 2016 (Aug. 2016), pp. 855–864.

[95] Zheng Gao et al. "edge2vec: Representation learning using edge semantics for biomedical knowledge discovery". In: *arXiv:1809.02269 [cs]* (Sept. 2018). arXiv: 1809.02269.

[96] Xin Rong. "word2vec Parameter Learning Explained". In: *arXiv:1411.2738 [cs]* (Nov. 2014). arXiv: 1411.2738.

[97] Karl Pearson. "The Problem of the Random Walk". en. In: *Nature* 72.1865 (July 1905), pp. 294–294.

[98] Ulrike von Luxburg. "A tutorial on spectral clustering". en. In: *Statistics and Computing* 17.4 (Dec. 2007), pp. 395–416.

[99] Lei Tang and Huan Liu. "Scalable learning of collective behavior based on sparse social dimensions". In: *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*. Proceeding of the 18th ACM conference. Hong Kong, China: ACM Press, 2009, p. 1107.

[100] Lei Tang and Huan Liu. “Relational learning via latent social dimensions”. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’09*. the 15th ACM SIGKDD international conference. Paris, France: ACM Press, 2009, p. 817.

[101] Sofus A. Macskassy and Foster Provost. “A Simple Relational Classifier”. In: ().

[102] R. F. Woolson. “Wilcoxon Signed-Rank Test”. en. In: *Wiley Encyclopedia of Clinical Trials*. Ed. by Ralph B. D’Agostino, Lisa Sullivan, and Joseph Massaro. Hoboken, NJ, USA: John Wiley & Sons, Inc., Sept. 2008, eoct979.

[103] Carl R. Rogers. “Toward a Science of the Person”. en. In: *Journal of Humanistic Psychology* 3.2 (Apr. 1963), pp. 72–92.

[104] Leann Myers and Maria J. Sirois. “Spearman Correlation Coefficients, Differences between”. en. In: *Encyclopedia of Statistical Sciences*. Ed. by Samuel Kotz, Campbell B. Read, N. Balakrishnan, Brani Vidakovic, and Norman L. Johnson. Hoboken, NJ, USA: John Wiley & Sons, Inc., Aug. 2006, ess5050.pub2.

[105] Nasrullah Sheikh, Zekarias Kefato, and Alberto Montresor. “gat2vec: representation learning for attributed graphs”. en. In: *Computing* 101.3 (Mar. 2019), pp. 187–209.

[106] Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. “metapath2vec: Scalable Representation Learning for Heterogeneous Networks”. en. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’17*. Halifax, NS, Canada: ACM Press, 2017, pp. 135–144.

[107] Lingling Xu. *comaprison_workflow.png*. 2019.

[108] Enrico Palumbo et al. “Knowledge Graph Embeddings with node2vec for Item Recommendation”. In: *The Semantic Web: ESWC 2018 Satellite Events*. Ed. by Aldo Gangemi et al. Vol. 11155. Cham: Springer International Publishing, 2018, pp. 117–120.

[109] Aditya Kumar Bubna. “Vorinostat-An Overview”. eng. In: *Indian Journal of Dermatology* 60.4 (Aug. 2015), p. 419.

[110] Angela De Simone and Andrea Milelli. "Histone Deacetylase Inhibitors as Multitarget Ligands: New Players in Alzheimer's Disease Drug Discovery?" en. In: *ChemMedChem* 14.11 (June 2019), pp. 1067–1073.

[111] Justin Rustenhoven et al. "PU.1 regulates Alzheimer's disease-associated genes in primary human microglia". en. In: *Molecular Neurodegeneration* 13.1 (Dec. 2018), p. 44.

[112] T. Schmelzle and M. N. Hall. "TOR, a central controller of cell growth". eng. In: *Cell* 103.2 (Oct. 2000), pp. 253–262.

[113] Gary Hudes et al. "Temsirolimus, Interferon Alfa, or Both for Advanced Renal-Cell Carcinoma". en. In: *New England Journal of Medicine* 356.22 (May 2007), pp. 2271–2281.

[114] Teng Jiang et al. "Temsirolimus promotes autophagic clearance of amyloid-β and provides protective effects in cellular and animal models of Alzheimer's disease". en. In: *Pharmacological Research* 81 (Mar. 2014), pp. 54–63.

[115] Christelle Frederick et al. "Rapamycin Ester Analog CCI-779/Temsirolimus Alleviates Tau Pathology and Improves Motor Deficit in Mutant Tau Transgenic Mice". In: *Journal of Alzheimer's Disease* 44.4 (Feb. 2015), pp. 1145–1156.

[116] Russell E. Nicholls et al. "PP2A methylation controls sensitivity and resistance to β-amyloid-induced cognitive and electrophysiological impairments". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 113.12 (Mar. 2016), pp. 3347–3352.

[117] Sandra Payton, Catherine M. Cahill, Jeffrey D. Randall, Steven R. Gullans, and Jack T. Rogers. "Drug Discovery Targeted to the Alzheimer's APP mRNA 5'-Untranslated Region: The Action of Paroxetine and Dimercaptopropanol". en. In: *Journal of Molecular Neuroscience* 20.3 (2003), pp. 267–276.

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, November 22, 2019

....

Lingling Xu