Dissertation
On

Evidence-based Drug Repurposing using Graph Neural Networks

Thesis Submitted in the partial fulfillment of the requirements for the degree
of
Master of Engineering
in
Computer Science and Engineering

Submitted By
Somtirtha Mukhopadhyay

Examination Roll number:
M4CSE22006
Registration number:
154130

Under Guidance of
Prof. (Dr.) Ujjwal Maulik
Jadavpur University

Dept. of Computer Science & Engineering
Faculty Council of Engineering and Technology
JADAVPUR UNIVERSITY
KOLKATA – 700032
2021 – 2022

Department of Computer Science & Engineering
Faculty Council of Engineering and Technology
JADAVPUR UNIVERSITY, KOLKATA – 700032

Certificate of Recommendation

This is to certify that Somtirtha Mukhopadhyay (Roll number: 002010502006) has completed his dissertation entitled "Evidence-based Drug Repurposing Using Graph Neural Network", under the supervision and guidance of Prof. (Dr.) Ujjwal Maulik, Jadavpur University, Kolkata. We are satisfied with his work, which is being presented for the partial fulfillment of the degree of Master of Engineering in Computer Science & Engineering, Jadavpur University, Kolkata – 700032.

Prof. (Dr.) Ujjwal Maulik
Teacher in Charge of Thesis
Professor, Dept. of Computer Science & Engineering
Jadavpur University, Kolkata – 700 032

Prof. (Dr.) Anupam Sinha
HOD, Dept. of Computer Science & Engineering
Jadavpur University, Kolkata – 700 032

Prof. (Dr.) Chandan Mazumdar
Dean, Faculty Council of Engineering and Technology
Jadavpur University, Kolkata – 700 032

Faculty Council of Engineering and Technology
JADAVPUR UNIVERSITY, KOLKATA – 700032

Certificate of Approval*

The foregoing thesis entitled "Evidence-based Drug Repurposing Using Graph Neural Network", is hereby approved as a creditable study for Master of Engineering in Computer Science & Engineering and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned does not necessarily endorse or approve any statement made, opinion expressed, or conclusion therein but approves this thesis only for the purpose for which it is submitted.

Final Examination for Evaluation of the Thesis

Signature of Examiners

*only in case the thesis is approved

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that this thesis contains a literature survey and original research work by the undersigned candidate, as part of her Master of Engineering in Computer Science & Engineering.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

Name	Somtirtha Mukhopadhyay
Exam Roll Number	M4CSE22006
Roll Number	002010502006
Registration Number	154130
Thesis Title	Evidence-based Drug Repurposing Using Graph Neural Network

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Signature with date:

Signature with date:

Acknowledgments

This dissertation would not have been possible without the support of many people. First and foremost, I would like to thank my advisor, Dr. Ujjwal Maulik, Professor, Department of Computer Science and Engineering, Jadavpur University, Kolkata. He has provided me with the perfect balance of guidance and freedom. He is the first person who introduced me to the world of deep learning and provided me with the guidance that was essential in making this dissertation. I also want to thank him for his perspective and for helping me pursue and define projects with more impact.

I would also like to thank Mr. Rangan Das, Research Scholar, Department of Computer Science and Engineering, Jadavpur University for his immense contribution to my knowledge of deep learning and bioinformatics. A lot of this dissertation would not have happened without the long, brainstorming sessions and the technical discussions we had during the coffee breaks at the university. He has questioned me at every step of the way, allowing me to make numerous improvements on my project. In many ways, he has shown me the path for how to apply, develop and understand deep learning.

This dissertation would be incomplete without the support of Mrs. Ashmita Dey, Research Scholar, Department of Computer Science and Engineering, Jadavpur University. I would not be able to complete this dissertation without her support and encouragement. She helped me a lot by proofreading and how to write the validation and test results in my papers and suggesting necessary changes. I'm also thankful to Mr. Rajarshee Banerjee, B.E., Department of Computer Science and Engineering, Jadavpur University for his invaluable suggestions regarding the coding of the model described in this dissertation.

Finally, I would like to thank Jadavpur University for providing me with the opportunity to work in such a productive environment. I would also like to thank all the other professors and research scholars of the department who have extended their helping hands whenever I needed help.

Date:

Place:

Somtirtha Mukhopadhyay, M.E in CSE
Roll Number: 002010502006

Contents

Acknowledgments 5

Abstract 7

Introduction 8

- Background 8
- Motivation 10
- Implementation Framework 11
- Organization of Thesis Work 12

Literature survey 13

- Graph Neural Network 13
- Previous Works on Drug Repurposing 14

Dataset 20

- Disease-Gene Interaction: 20
- Gene-Gene Interaction: 21
- Drug-Drug Interaction: 22
- Drug-Gene Interaction: 23

Implementation 25

- Relational Graph Convolution Networks (RGCN) 25
- Use of Jumping-Knowledge Neural Networks 29

Validation and Test Results 34

- Model Training Results 34
- Validation Data Description 37
- Validation of Results 38
- Validation of the predicted drug-gene associations 39

Conclusion 41

Future Scope 42

References 44

Abstract

Graph Neural Network is the class of Deep Learning algorithms that can work directly on Graphs and take advantage of their structural information. In real life, we all know that there are many such datasets that can be easily and efficiently represented in a network so we can take advantage of them using GCN here we are mainly focusing on using Graph Convolution Network (GCN) and Jumping Knowledge Network (JK-Net).

Here we see that the dataset of ours contains genes & drugs is a network of how genes are affected by drugs, how the drugs interact among them, and how genes interact among themselves can be very easily represented by using Hetero-graph and DGL Graph Library so that any new relations can be easily learned from them using an algorithm that takes advantage of that structural information in those Graphs. Moreover what is the basis of this work is to merge those different Graphs into a single Heterogeneous-Graph and use the GCN algorithm to predict new relations between them.

The goal of this work is to take advantage of the Graph Learning algorithm to predict links that will eventually lead to the Re-purposing of the Drugs which is comparatively an easier process than developing a new drug for a new variant of the disease which is a long process and will take not less than 10-15 years and a lot of resources and wealth but if the existing drugs can be repurposed successfully, will not only save time and effort but also lives.

Introduction

Background

The last few years saw a surge in the deep learning research domain and extended CNN over irregularly structured data. While performing such research it has been found that the CNN may sometime fail to correctly produce the desired results over irregular data. Further studies show that in comprehending the irregularly structured data the Graph structure can be easily and efficiently used in representing the complex domain of irregular data like molecular structure, to computer and social and traffic networks. The extension of CNNs using the graph leads to a new domain of research the GCNNs which improve the inferential powers and the efficiency of the graph-based convolution neural networks. The majority of GCNNs are based on operations and certain properties. There are many challenges in Graph-deep learning among them lies the problem of learning with the edge signals and information.

The CNN's basic limitation is that it works well with regular structures and data. In general, there are many such scenarios where the data is irregular and thus CNN will not work well in such cases. Moreover, if we forcefully try to convert irregular data to a regular structure we may lose some insight and features of the data which may completely destroy the data and our whole model may fail so in such a case we have to use irregular structures and models that support those structures like GCNNs.

Over many irregular geometric shapes, Graphs have significant roles to play in the machine learning domain. It has been found that we can represent many complex structures by using graphs like the irregular relations between different entities. For example data from the network of computers or users, the molecular bonds in a chemical compound, etc. Thus in the last half-decade, there has been a lot of research on extending the CNN to a graph.

Previous methods that were used for drug repurposing can be classified as signature-based and network-based. In the case of those network-based approaches, the networks were constructed on drug-drug links information from the interactions and similarities taken from drug, disease, and targets. Some studies use descriptors for each drug-disease feature and their similarities to finally construct a new logistic regression model or some statistical models to predict new links as was [1,2]. Other models as [3] (GPSnet) Genome-wide

Positioning System Network where he used networks to predict drug responses in cancer cell lines accurately by integrating them with transcriptome profiles, and whole-exome sequencing, the drug-target network and drug-induced microarray data into human protein-protein interactome [4]. Several studies were made that further inferred new drug indications by information flow or random walks on the networks which were built on those relationships [5, 6]. Integrated diverse prior knowledge of drugs and diseases through non-negative matrix factorization and then made the predictions according to their projections in low-dimensional feature space [5]. Individualized Network-based Co-Mutation is another approach for quantifying the acknowledged genetic interactions in cancers. It can promote the significant identification of candidate therapeutic pathways. Cheng et al. [5] developed a network-based infrastructure to identify targets or new interactions for existing drugs by directly targeting the important mutated genes or their neighbors in the protein interaction network [6]. Further using the adjacency matrix of the disease-drug relationships. Several methods have been proposed by constructing the negative graph i,e to build drug-disease networks based on known drug-disease relationships and then complement the adjacency matrix of the network with some different algorithms given by [7].

In the case of signature-based methods which had been successfully applied in the field of drug discovery, [8]. With advancements in microarray and the next-generation sequencing techniques, massive amounts of genomics data are accumulated using big-data techniques. The Connectivity Map (CMap) contains many gene expression signatures, which can be used to explore functional connections between diseases, genes, and therapeutics [8]. In paper [9] Donertas identified repurposing for the long-term purpose of drugs by comparing changes in gene expression with drug-affecting expression profiles in the Connectivity Map [8]. As the successor of CMap, the Library of Integrated Network-Based Cellular Signatures (LINCS) where the project consists of assay results from primary human cells treated with bioactive small molecules, ligands, or genetic perturbations [10]. As we know drugs and their indications often share commonly related genes on which drugs execute their functions. the idea was simple more common genes, the drugs and diseases share among them, the more likely the drug is to be associated with the disease. Some studies have also been proposed that finds the associations of drugs and diseases based on their related genes or gene expressions [11]. Comparatively, some methods have also been proposed using the protein complexes shared by the drugs and the diseases [12] and they're common concerning genes [13].

The signature-based methods cannot be applied to drugs and diseases without finding something common between the related genes or proteins. In general, both these two kinds of methods have their advantages and disadvantages. As in the case of Network-based methods integrates the relationships between the drugs but ignores the prior knowledge gained in that process. In the case of signature-based methods grasps the characteristics of drugs or diseases themselves but cannot utilize the potential knowledge from the drug-drug links information.

Another approach is using both of this network-based as well as the signature-based methods for the prediction of the drug repurposing in case of breast cancer where they tried to predict the results using Graph Neural Network algorithm GraphSAGE model and the used the drug-exposure gene expression from the LINCS project of signature-based method and the drug-drug links from STITCH database. Then they constructed the network of those drug-drug links and ran their algorithm on that network using the network-based method. Furthermore, they benchmarked their GraphRepur algorithm comparing it with the other deep learning models like deepDR [14] and LLE-DML (signature-based) [15], BiFusion (network-based) this work can be found in [16].

In this paper, I would like to discuss an approach mainly using GCN and compare the model with another algorithm of Graph the JK-Net on Heterogeneous Graph basically our model would be a more like the combination of both the network and signature-based approach where we have taken data from sites like DrugBank, DisGeNet and then used our model using the DGL library of python to run the algorithm on our Heterogeneous graph data to obtain the task of link prediction to eventually obtain a bigger goal of Drug-Repurposing.

Motivation

As said earlier GCN is an algorithm that can learn very well from the irregular structural information of a network or a Graph. Previously there it has been used to learn and predict the molecular structures of drugs so we thought why not use the same technique for a network of diseases-genes-drugs if we can predict the links between nodes of genes and drugs then we can use the drug for some gene set and thus it's, in short, the drug-repurposing technique. As we all know that the invention of a new drug may take a span of even up to 10 years so it's a long process overall. So if we can use a developed drug for a different purpose it will be very helpful as we have just seen a few months it was what happened in case of the Covid-19. What scientists did was just repurposed the drugs for Pneumonia to treat Covid-19. While a new drug for Covid-19 is still in progress will take a long time to be developed. So it was the basic motivation of my paper.

Implementation Framework

In this thesis work two different algorithms are used to do the final task of link prediction the task is simple if we can produce the links that can be biologically validated and if it can be produced as a new link that was never there between some genes and some drugs and then if we can link those genes with a disease then a drug is basically repurposed which was predicted by our model. To achieve this task of drug being repurposed one algorithm that is the simple Graph Convolution Network is used and then another method is used to do a comparative study another algorithm known as the Jumping Knowledge Network is used for comparing the results predicted by two different models of the Graph Neural Network.

Those models were fed with the Heterogeneous Graph of drug-drug interactions, drug-gene interactions, and the gene-gene interactions as well during the training procedure. Finally, the link between the two unknown pairs of gene and drug was given as input and their link's probability is produced as the result of the trained model.

Since Python is a language that is having a lot of in-built libraries and frameworks to efficiently work on data science here I have used Python as the programming language. The Deep Graph Library in Python is the library that provides us immense power to make Graphs both Homogeneous as well as Heterogeneous from basic NetworkX Graph which is another network building library in Python. So firstly, the Hetero-Graph was made and it was passed to the GCN model for the training. The Graph is having almost 12,000 nodes and over 6,00,000 edges that were used for the training purpose. Finally, the model was tested by giving the nodes as input and obtaining the probability of them being a valid connection is predicted by the model which was then biologically validated using again a network of disease and genes taken from the DisGeNet database.

Organization of Thesis Work

CHAPTER-I: Introduces the thesis. This section tells about the Graph Neural Network Models used and a very little about the line of approach of others and mine towards the goal of evidence-based Drug-Repurposing using Graph Neural Network.

CHAPTER-II: This section will discuss the Models used and the previous works in this field and how they approached the problem. This chapter is the Literature Survey of this thesis work.

CHAPTER-III: This part consists of the Description of the dataset that has been used in this dissertation, which has been prepared from more than one well-known website their complete description and references are well stated in this part with a line-by-line description.

CHAPTER-IV: This section mainly contains the Implementation of the RGCN and the JK-Net usage and how all these concepts were used in making the whole model for the link prediction task in detail.

CHAPTER V: This section contains a detailed description of the results and the biological proof of the results that are being produced by the model. It is mainly developed keeping in focus the two psychological diseases but could have been done on any disease present in our database.

CHAPTER VI: This section contains the future developments and proceedings that could be done and the model results could be even better (dynamic or changing graph in terms of edges and nodes). The Future Scope and the Conclusion could be found in this chapter.

Literature survey

Graph Neural Network

Graph Neural Network is a semi-supervised algorithm that takes advantage of the irregular structure of a network to achieve the tasks of either node classification, link prediction, or graph classification as a whole. Here basically focusing on the task of link prediction between a drug and a gene pair and providing the probability of them being connected by constructing a negative and a positive graph and using the Graph Convolution Network (GCN), and the Jumping Knowledge Network (JK-Net).

Let's first discuss the GCN which is very much similar to the classical Convolution Neural Network (CNN) the basic difference lies in the fact that GCN is a more generalized version of CNN with the same process of moving around a kernel filter to obtain the features followed in GCN known as weight sharing. GCN is a very powerful tool that can expand CNN on irregular data networks like Graphs. Here we will use the same equation of CNN

$$H^{[i+1]} = \sigma(W^{[i]}.H^{[i]} + b^{[i]})$$

The above equation of CNN just takes the form of GCN as given below and stated by Thomas Kipf and Welling [1] in their paper on GCN. Where the function takes the modified Adjacency matrix of the Graph on which GCN is trained.

$$f\left(H^{(i)}, A\right) = \sigma\left(H^{(i)}.W^{(i)}.A^{*}\right)$$

$$f\left(H^{(i)}, A\right) = \sigma\left(D^{-\frac{1}{2}}.A^{*}.D^{-\frac{1}{2}}.H^{(i)}.W^{(i)}\right)$$

The above equations are pretty basic equations that are being used from the start of using GCN. The whole concept of message passing is based on the above-mentioned equations stated [17].

Previous Works on Drug Repurposing

The previous works were mainly dependent on the sole criteria of finding the new and more evident interactions between the drugs. So the previous works mainly can be classified into two broad groups, the similarity-based and the network-based approaches.

Start with the similarity-based approach is an integrative method using heterogeneous information [18, 19]. The method is quite simple it naturally depends on the notion that the similarly structured drug affects similar diseases. They take into consideration the shared characteristics between drugs, as in drug–targets, chemical structures, and their adverse effects. Then constructs similarity features to build the final prediction models. For example, PREDICT [20] which is a similarity-based framework that sums up drug-drug similarities (based on drug-protein interactions, sequence, and gene-ontology) and then the disease–disease similarities (disease–phenotype and human phenotype ontology), then the authors took them as the key feature set to apply logistic regression and predict similar drugs for similar diseases and achieves the final goal.

Whereas in the case of network-based approaches [21, 22, 23] it models graph-structured information between different biological networks to enhance the performance of Drug Repurposing. Mainly, in these models, the nodes in the networks represent either drugs, diseases, or genes and the edges denote the interactions or relationships among them. For example Cheng identified hundreds of new drug-disease associations by quantifying the network propinquity of disease genes and drug targets in the human protein-protein interactome. The deepDR Zeng learned high-level features of drugs from the heterogeneous networks by a multimodal deep autoencoder and applied a variational autoencoder to infer candidates for approved drugs. However, deepDR only considers information sources in the drug domain only without considering their interactions in the disease domain.

Bipartite graphs are such graphs in which the nodes can be divided easily into two disjoint sets in [21, 22], this feature of a bipartite graph can be very well used in representing the interaction between two biological node types like disease-symptoms, or drug-targets or even protein-protein. For example, in [23] built a bipartite graph to analyze relationships between drug targets and disease–gene products by constructing a network in [24] constructed a bipartite graph to analyze the relationships between human genetic diseases.

In the field of Drug-repurposing, in [25] developed a bipartite drug–target network graph using drug pair similarity and integrating the drug chemical structure similarities, common drug targets, and interactions between them. Zheng also constructed a bipartite graph with a known relationship between drugs and their target proteins. However, most of these methods mainly rely on predefined drug similarity features and ignore the important information sources that can be found in the disease domain. Although the utilization of the PPI information, the relationship between drug-target and disease–gene in the context of biological interaction network has been very recently investigated by another prior work the BiFusion in [15].

Year	Model	Key Features
2020	KGNN: Knowledge Graph Neural Network for Drug-Drug Interaction Prediction in [55]	• The most recent work on Drug-Drug Interaction prediction using the GNN model. • The Drug interactions are drawn from the Knowledge Graph using and the score is predicted. •The Loss function used is the Binary Loss Entropy Function.
2021	SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization	•Local subgraph for identifying useful information. •Subgraph summarization scheme for generating reasoning path. •Multi-channel data and knowledge integration for improved multi-typed DDI predictions. •It almost increases the F1 score percetnage by 5.54% and it excels at a very low resource setting it is a multi type relation network.
2020	Bi-Level Graph Neural Networks for Drug-Drug Interaction Prediction	•Lower Level Representation Graph Embedding. •Follows a multi-scale readout taken from the concept of GIN and GAT •Higher Level Interaction Node Embedding.
2019	Predicting Drug–Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation	•Novel Deep Learning Method is used. •Distance aware Graph Attention Network to differentiate various intermolecular interactions. •Graph Features extracted from intermolecular structure using embedded 3D structure of protein ligand binding pose. •Better performance than Docking and other models achieved AUROC of 0.968 for the DUD-E test set and AUROC of 0.935 for the PDBbind test set.

Year	Model	Key Features
2020	SkipGNN: predicting molecular interactions with skip-graph networks	•Here the molecular interactions are predicted using GNN but with a twist of not only considering the neighbours the authors considered even the two hop neighbours as well. •Here instead of one GNN model two GNN is used on the skip graph to obtain the results desired and then they are merged with each other to get the skip graph model. •Perfoms well incase of small dataset and incomplete ones since the hidden interctions are found to some extent.
2019	Drug-drug Interaction Prediction with Graph Representation Learning	•Model finds the most important local atoms using attention mechanism. •Can be used for large as well as smaller datasets. •Can overcome the two problems of previous models i.e., the scalability and the robustness.
2018	Enhancing Drug-Drug Interaction Extraction from Texts by Molecular Structure Information	•It is a novel neural network approach for extracting drug-drug interaction features using external molecular structures of the drugs. •Using two models one is textual drug pairs with convolutional neural networks and then their molecular pairs with graph convolutional networks (GCNs), then finally concatenating the outputs of the two networks.
2019	Drug-Drug Interaction Predicting by Neural Network Using Integrated Similarity in [56]	•NDD calculates many different informations about the drugs then. •Heuristic selection function at first with a non-linear similarity fusion method to achieve high-level features •Then it uses the neural network to predict the interactions.
2019	Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings in [57]	•First step the RDF knowledge graph is made. •Then the drug feature vector is extracted by using many different graph embedding thechniques like RDF2Vec, TRANSD, TRANSE. •Then the DDI are predicted using 3 different classifiers those are Logistic Regression, Naive Bayes and Random Forest.

Year	Model	Key Features
2019	Drug-Drug Interaction Prediction Based on Knowledge Graph Embeddings and Convolutional-LSTM Network in [58]	•Over 12,000 drug features taken into consideration from the three sources and RDFization is done they are DrugBank, KEGG, PharmGKB and Knowledge graph was formed. •ComplEx embedding method was used and PyTorch-BigGraph (PBG) with a Convolutional-LSTM network and classic machine learning-based prediction models. •Finally the ensemble of 3 best classifier model is used to predict the results.
2020	Drug–drug interaction prediction with Wasserstein Adversarial Autoencoder-based knowledge graph embeddings	•The concept of Adverserial Auto Encoders(AAE) is used for the embedding of KG is proposed. •Wasserstein distances and Gumbel-Softmax relaxation for DDI tasks were used. Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposingThen the used or Wasserstein distance in removing vanishing gradient problem were also used to achieve significant improvements.
2020	Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing in [59]	•The most relevant work to ours is this work where first the approach of getting more DDI was being done on Heterogeneous Graph structures using graph neural network. •The datasets taken from the DrugCentral, DisGeNet, DGI db are the Heterogeneous Bipartite graph is formed. •Finally the Knowledge Graph is passed through a Graph Neural network model for specially designed for the Bipartite graph to get the desired new relations between drugs and thus achieving the broad picture of drug repurposing.
2021	Bi-Level Graph Neural Networks for Drug-Drug Interaction Prediction	•Lower-Level Representation Graph Embedding. •Follows a multi-scale readout taken from the concept of GIN and GAT •Higher Level Interaction Node Embedding.

Table 1: A chronological list of related works.

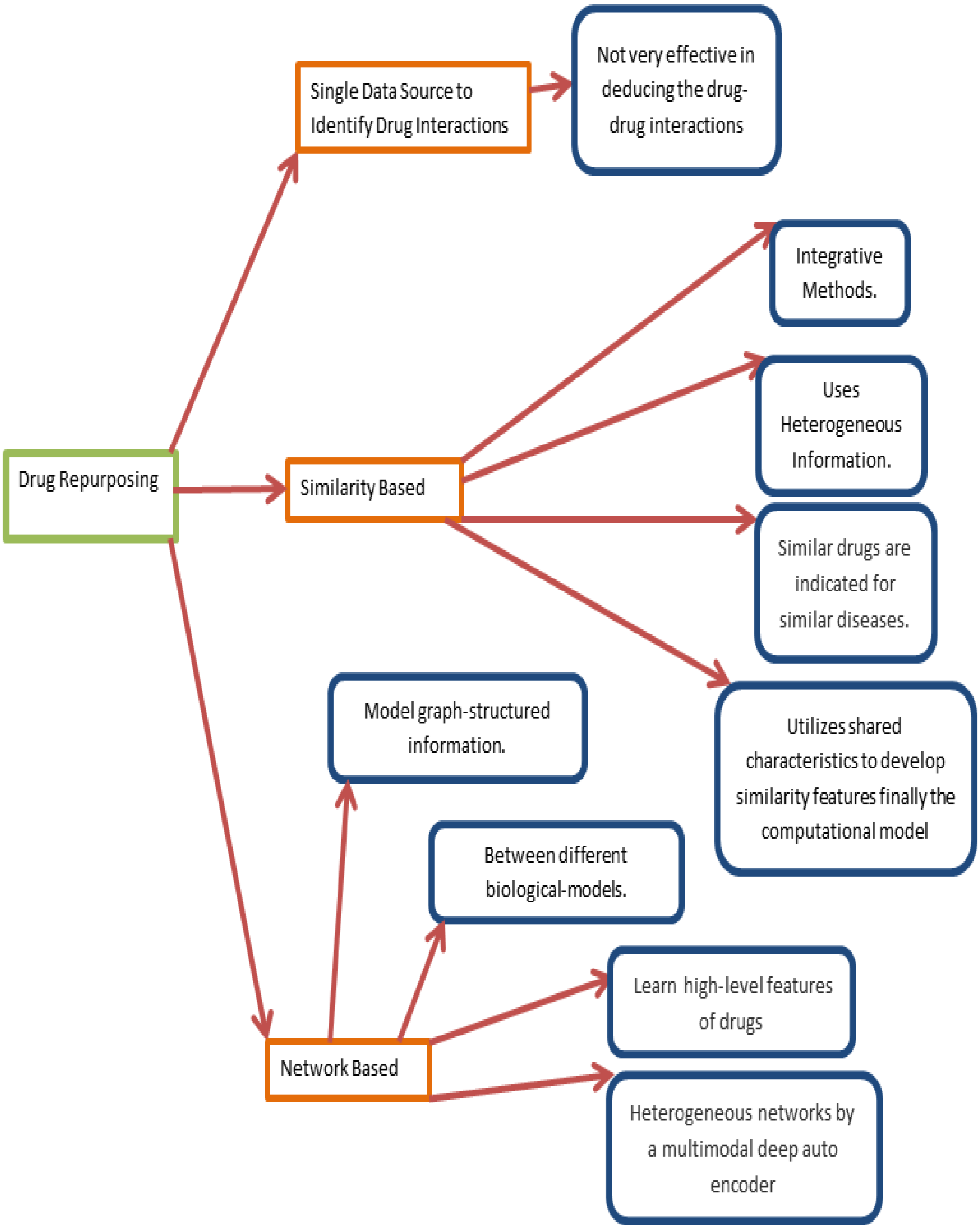

Drug Repurposing
Single Data Source to Identify Drug Interactions
Not very effective in deducing the drug-drug interactions
Similarity Based
Integrative Methods.
Uses Heterogeneous Information.
Similar drugs are indicated for similar diseases.
Utilizes shared characteristics to develop similarity features finally the computational model
Network Based
Model graph-structured information.
Between different biological-models.
Learn high-level features of drugs
Heterogeneous networks by a multimodal deep auto encoder

Fig 1: Overview of Drug-Repurposing prior works

Dataset

Disease-Gene Interaction:

The dataset in this dissertation has been prepared from various sources firstly the disease and the genes relations have been taken from the DisGeNET website where there is a database SQLite relational database file which consists of many tables the structure of these database files is given below:

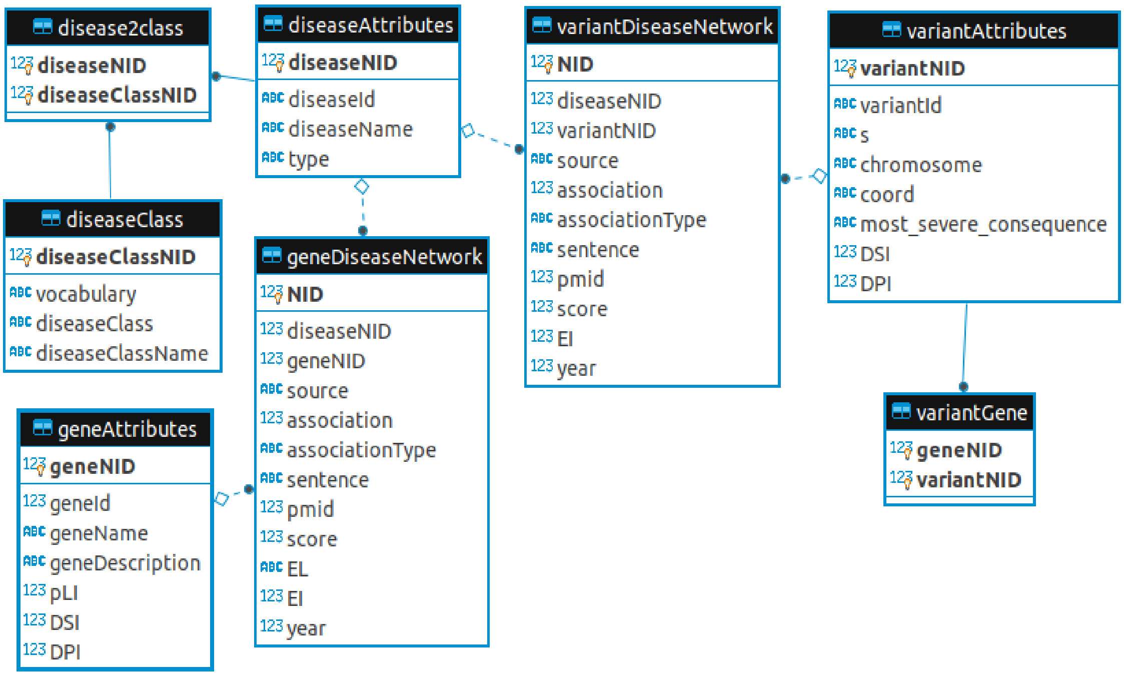


Fig 2: The Database structure of DisGeNet

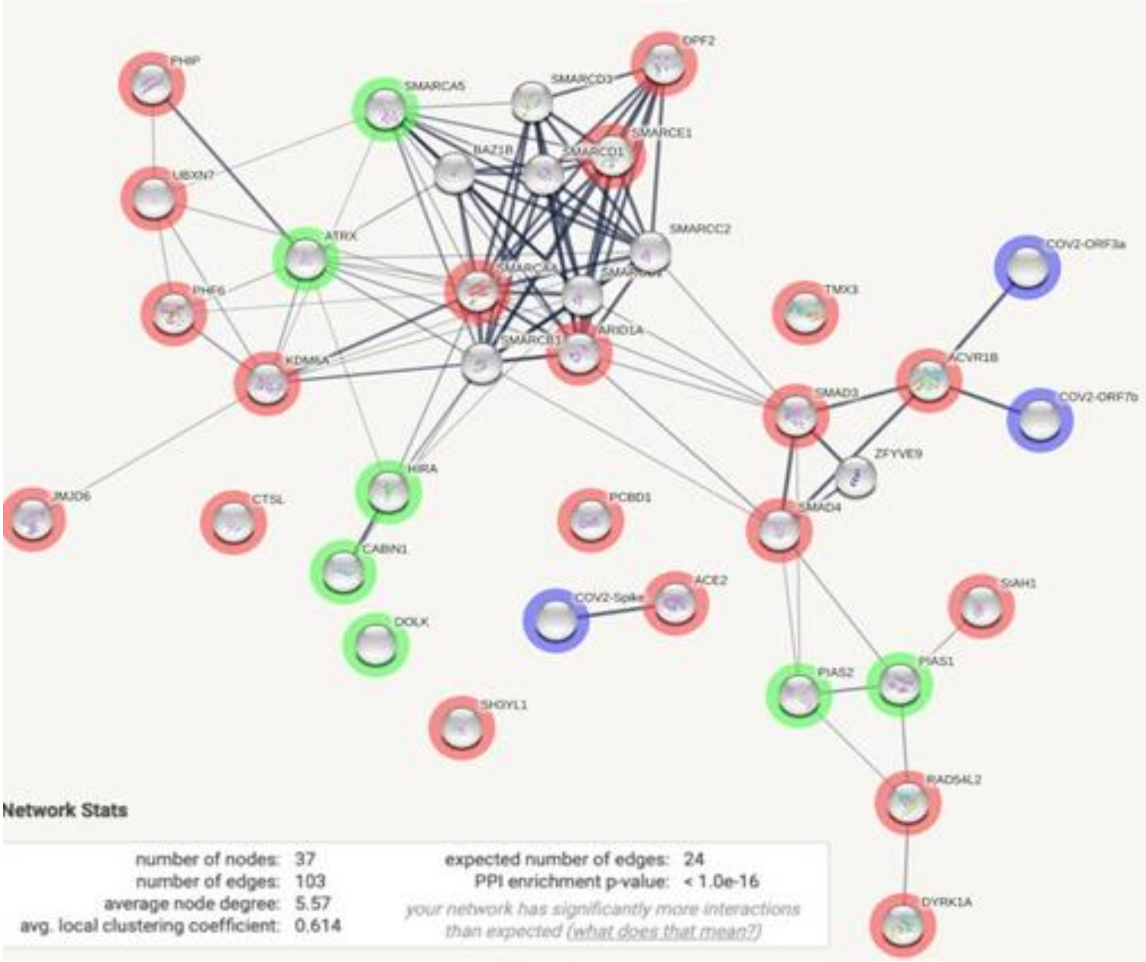
Thus from this .db File, the table considered were diseaseAttributes, geneAttributes, geneDiseaseNetwork, and disease class, and the links between the diseases and the genes are taken out to build the main graph. These Disease-Gene interactions are later used to verify the efficiency of the model. Actually, we will be using these relations as the ground truths between the interactions between the diseases and the genes so after our model predicts the new links between the drugs and the genes we can cross verify them and this is what the term evidence in our title of this thesis is based on.

Moreover, a tool using the flask and python is provided from this dataset taking the diseases common and their symptoms also and the database tool is also provided with this thesis where the graphs can be visualized and can be seen for a better understanding of the tool will show the disease and the genes linked with it as well as its symptoms and their links with the common genes in short a linkage between the disease –genes-symptoms-genes graph network.

Gene-Gene Interaction:

These gene-gene interactions were taken from the https://string-db.org/ website where there are links between 67.6 million Proteins in 14094 organisms. From this huge pool of data, only the common human genes that were in intersection with those genes of the disease-gene set were taken into consideration only and their links were taken out and saved as a sub-graph of gene-gene interactions having 9 attributes in simpler terms the 9 edge features i.e., various scores. They are namely:

1. neighborhood_on_chromosome
2. gene_fusion
3. homology
4. phylogenetic_cooccurrence
5. coexpression
6. combined_score
7. experimentally_determined_interaction
8. database_annotated
9. automated_textmining



Network Stats

number of nodes:	37	expected number of edges:	24
number of edges:	103	PPI enrichment p-value:	< 1.0e-16
average node degree:	5.57		
avg. local clustering coefficient:	0.614		

your network has significantly more interactions than expected (what does that mean?)

Fig 3: A small subset representation of Gene-Gene Interactions

From this huge set here only a common subset of 1473 genes and the gene-gene interaction file contains a set of 39307 gene-gene interaction edges from Stringdb and intersecting it with the DrugBank and DisGeNET.

Drug-Drug Interaction:

The best way to get information about drugs is to search from the well-known site the DrugBank. From where the drug-drug interactions were taken out using the text-mining and are basically done by web-scraping. These drugs are taken by their Id and the attributes that are considered are basically the MACCS which is considering the chemical molecular smiles that is if a certain structure is present the keys positions are switched on (putting 1) otherwise putting 0 at that position that keys are of length 167 chemical compound is considered. From there the drugs are taken out which are 11160 with the SMILES list the details about MACC's keys are here at this link (Paper on MACCS keys: https://www.biorxiv.org/content/10.1101/853762v1.full.pdf).

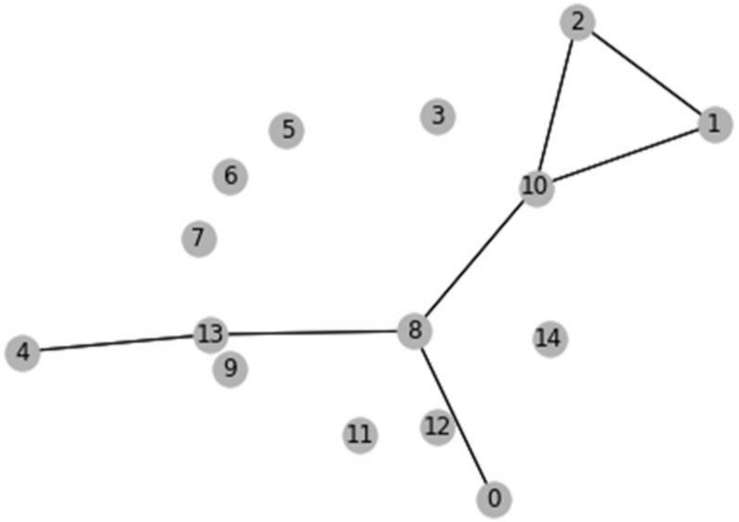


Fig 4: Subset of Drug_Id-Drug_Id Interaction Graph.

This is a small example of the drug-drug interaction graph which is considered from the DrugBank site consisting of 583548 edges and that graph is changed to a bi-directional graph so the number of edges is doubled consisting of 1167096 edges in the drug-drug graph which was further converted into the heterogeneous graph structure.

Drug-Gene Interaction:

The drug-gene interaction links are also taken into consideration which is also from the DrugBank i.e., mentioned by the prior works as the drug-protein links which is also the same intersection set of those 1473 genes or proteins. This is the graph which is having 599570 edges but no edge attributes as such were considered which can be considered but, in this case, no such suitable attributes were seen in any of those online sites. This can be considered in the future scope of this thesis and will be discussed in detail as this thesis proceeds.

Finally, the hetero graph is made having the meta-graphs as

1. The drug-drug interactions.
2. The gene-gene interactions.
3. The drug-gene interactions.

Using the DGL python library which is an exceptionally powerful library in forming heterogeneous graphs and implementing the GCN algorithms on the hetero-graphs to finally predict the new drug-gene links which are nothing but the drugs used for new genes thus achieving the final task of drug-repurposing and validated using the graph of disease-genes.

Implementation

The basic Graph Convolution Network devised in [6] was originally for the homogeneous graphs which are extremely efficient in calculating and understanding the irregular structure of a graph and the tasks of Link classification, Node classification, and graph classification were efficiently powerful. A paper was written on semi-supervised learning of node prediction on the Karate Club graph structure there it was seen that GCN was able to learn the trends by one pass only and the reason was given that it follows the concept of Weisfeiler Lehman Algorithm which just makes the GCN a very powerful method all-over. A recent paper by Perozzi on DeepWalk says that the same results can be carried forward in the case of unsupervised learning embeddings too.

For all nodes $v_i \in G$: Weisfeiler Lehman Algorithm

- Get features $\{h_{vj}\}$of neighboring nodes $\{v_j\}$
- Update node feature $h_{vi} \longleftarrow \text{hash}(\sum_j h_{vj})$, where hash(·) is (ideally) an injective hash function
- Repeat for k steps or until convergence.
- $h_{vi}^{(l+1)} v = \sigma(\sum_j \frac{1}{c_{ij}} h_{vj}^{(l)} W^{(l)})$

Relational Graph Convolution Networks (RGCN)

Going further a step ahead GCN can be stretched to concentrate on a local neighborhood which was introduced in [6] on large-scale relational datasets. This can be also defined as a differentiable framework of Graph neural networks.

$$h_i^{(l+1)} = \sigma(\sum_{m \in \mu_i} g_m(h_i^{(l)}, h_j^{(l)})$$

Where $h_i^{(l)} \in R^{d^{(l)}}$ is the representation of the l-th layer node v_i with dimensionality $d^{(l)}$. The incoming messages coming through the $g_m(*,*)$ function are passed through another pointwise function like ReLU(*) or Tanh(*). μ_i is the messages which typically equal the number of incoming edges to node v_i. $g_m(h_i^{(l)}, h_j^{(l)})$= W($h_j^{(l)}$) which is basically kept simple and message specific as in [35]. This type of transformation was found to be very effective in finding and accumulating local neighborhood features which led to significant improvements in the fields of graph classification and semi-supervised learning.

Taking motivation from the above concept another equation was devised for a multi-graph structure with more than one relation type:

$$\mathrm{h}_i^{(\mathrm{l}+1)} \;=\; \sigma(\sum_{\mathrm{r} \in \mathrm{R}} \sum_{\mathrm{j} \in N_i^r} \frac{1}{c_{i,r}} \mathrm{W}_r^{(l)} \; \mathrm{h}_j^{(l)} \; + \mathrm{W}_0^{(l)} \; \mathrm{h}_i^{(l)})$$

Here N_i^r is the set of a neighbor index of node 'i' in relation to r ∈ R . Whereas the $c_{i,r}$ is a problem-specific normalization constant that can be learned or can be predefined as the cardinality of the N_i^r. The concept is pretty intuitive which is nothing but accumulating the transformed normalized sum of the i nodes in feature vector forms. Contradicting the concept of the regular GCNs the authors have used a relation-specific transformation that is depending on the type of relation the edge is having since the graph is a heterogeneous entity where there are edges of more than one type of relation. To preserve the representations of the self-node layer by layer that is of node 'i' they have further introduced a self-loop concept with a special kind of relationship at the start of the algorithm only. All the nodes at a layer can be modified in parallel here using the sparse matrix multiplication to avoid various computational losses. This model is the whole concept behind the R-GCN (the relational graph convolution network).

Now the most obvious question that comes to mind is the huge increase in the parameters of the whole network. So, what's done generally the weights $\mathrm{W}_r^{(l)}$ of the R-GCN are regularized into 2 parts the (basis and the block diagonal decomposition).

Where the Basis Decomposition for each layer $\mathrm{W}_r^{(l)}$ takes the shape:

$$\mathrm{W}_r^{(l)} = \sum_{b=1}^{B} \mathrm{a}_{rb}^{(l)} \, \mathrm{V}_b^{(l)}$$

Where $\mathrm{V}_b^{(l)} \in R^{d^{(l+1)} * d^{(l)}}$ with coefficient $\mathrm{a}_{rb}^{(l)}$ with one dependency that is r. Then the block-diagonal decomposition of each $\mathrm{W}_r^{(l)}$ is defined as the summation of the lower dimensional matrices i.e.,

$$\mathrm{W}_r^{(l)} = {}_{b=1}^{B}\bigoplus \mathrm{Q}_{br}^{(l)}$$

Thus the $\mathrm{W}_r^{(l)}$ becomes the block-diagonal matrices: diagonal ($\mathrm{Q}_{1b}^{(l)}$, . . ., $\mathrm{Q}_{rb}^{(l)}$) with $\mathrm{Q}_{rb}^{(l)} \in R^{\frac{d^{(l+1)}}{B} * d^{\frac{l}{B}}}$.

What is Link Prediction and how is it used in this dissertation?

The answer to this question is what my project is completely based on in short, the link or the edge set of the whole graph is what is given to the R-GCN network to learn and to work upon but in my case, it is considered to be an incomplete set of that edge set of that particular type of relation which needs to the requirement of finding the score of a valid edge that belongs to that incomplete set of edges that must be calculated to complete the edge set of that relation or type. G = (V, Є, R) where Є is the mentioned incomplete set for which a valid edge's score needs to be found f(s, r, o) to possible edges (s, r, o) that it may belong to Є. The positive and the negative scores are both given for a given pair that is what this dissertation paper is based upon the link between the drug and a gene and the score of it so that it can be used on disease caused by that gene or not.

In R-GCN there is also a use of an auto-encoder for solving the score-related problems. The two parts are the entity encoder and the scoring part(decoder). The edges are reconstructed by the decoder which in turn relies on vertex representations, scores are given on triples of (subject, relation, object) over the function of $R^d \times R \times R^d \to R$. In the case of the existing models, the scores are calculated using the (tensor and neural factorization methods as stated in [34],[36],[37]. But in the case of R-GCN, the dependency is solely on the encoder not on a single, real-valued vector $e_i \; for \; every \; v_i \in V$ only optimized during training. The encoder is using the formula $e_i = h_i^{(l)}$ just like the concept of auto-encoder that was coined in the paper [35] for unlabelled and undirected Graphs. The picture of the R-GCN is given below:

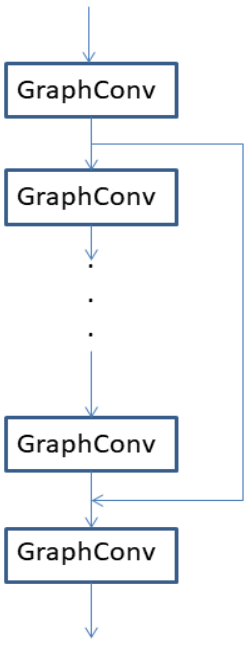


Fig 5: Residual Graph Convolution Neural Network

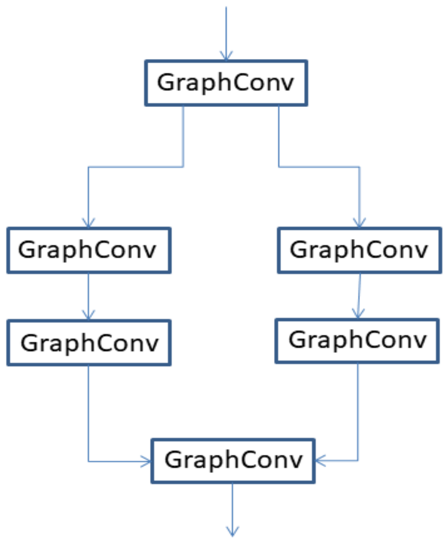


Fig 6: Parallel Graph Convolution Neural Network

The DisMult factorization by [38] is the decoder (scoring function) known to perform well on standard link prediction models on its own. Basically, the driving formula in DisMult is the diagonal matrix of every relation r (i.e., $R_r \in R^{d\times d}$ and the triple score which is described above (s,r,o)

$$f(s,r,o) = e_s^T R_r e_0$$

Further following the research papers of [36] another training was done on the negative graph of the original graph which is nothing but some random corruption either corrupting the subject or the object keeping the relation same sampling ω negative ones for each positive examples. Then the cross-entropy was optimized using the formula

$$L = -\frac{1}{1+\omega|\xi|} \sum_{(s,r,o,y)\in T} y \log l\big(f(s,r,o)\big) + (1-y)\log(1 - l(f(s,r,o))$$

T= Total set of triples both corrupted and valid ones and l is the logistic sigmoid function.

Use of Jumping-Knowledge Neural Networks

But another very important point that crossed our minds during the project was the neighborhood of a particular node can change with the positional as well as the molecular representation of that particular node in the case of biomedical projects this problem, again and again, cropped up in the past too. The solution that we came up with was that the radius of the neighborhood of each node can be dynamically learned on the go in an adaptive way depending on the specific task. This above-mentioned problem is quite persistent in the case of GCN, especially in the social media networks, biomedical, financial networks, molecular networks, etc. Respective pictures will be shown where the same random walk steps will be seen producing different results for the same subset of the graph. The over-crowding of neighbors during the walk will lead to an over-smoothing effect that may sometimes lose the actual flavor of a feature or may average the whole structure too broadly may make us lose the actual feature. This is happening because most of the recent models of deep-learning are considering the message passing neighborhood aggregation scheme that was said in their paper [6] in this paper the model learns the features by iteratively aggregating the hidden features of every node in the graph with each of their neighboring nodes to find the new hidden features and so on. This iterative aggregation and message passing are directly proportional to the number of layers it was found that for a lesser number of layers for example 2 the algorithm works perfectly following the Weisfeiler-Lehman graph isomorphism learning the hidden features well. The same problem in computer vision was solved by the techniques of residual connections but here that trick is not applicable also. It was seen that the smaller layered networks perform well but the deeper networks with access to more information don't perform well. The state-of-the-art algorithm of GCN was found only by using 2- layers only. After using residual techniques also the result was not at all good in-case of the networks like the citation network.

So the above problem was discussed well in the paper in [50], [51] about Jumping Knowledge Networks. Where they have addressed mainly the above-stated problem in two parts:

1. Studying and finding the results and the limitations of the neighborhood aggregation scheme
2. Based on the analysis stated in point number 1, they have devised an adaptive structure-aware representation learning algorithm that works exceptionally well in the case of larger and complex graphs with diverse sub-graph structures.

Furthermore, they have stated that "one size fits all" cannot be true in this case also so the whole process of representation learning may vary as in [54] a lot in case there is a change in the sub-graph structure of a graph and also dependant on the specific task to be performed. In short, the flaw in the concept of the nearest neighborhood of a node was found.

More formally if it can be said would be that random walks performed on a graph from a node may sometimes learn more information in only 2 steps that focus on local neighborhoods only than some higher-order version of it where the feature information may be washed out or shows a fading symptom due to averaging. Another factor also seems to be very important was the changing of locality that can be linked directly with this dissertation topic where in a biological network most of the nodes are having very less connections whereas some nodes have a ton of connections suddenly with other nodes (core node). Such phenomenon can also be seen in the case of the social network structures too where an expander or core part is present in most of the cases.

The subgraph with a variety of structures can also be an important factor besides the node features the change in structure with the time spent expanding as it proceeds will also affect the whole static process of iterative aggregation and finding of features only based on the local neighborhood only mixing of times may cause the sub-graph to change drastically. The same no of iterations can lead to the production of different random walks thus different features for the different localities.

Next let's come to the solution of the problem as suggested in the above-mentioned paper, as it was told previously that the residual connection method failed to some extent like when the aggregation of the neighborhood features was skipped and combined later breaking the formula of aggregation mentioned as eqn.1:

$$\mathrm{h}_{v}^{(l)} = \sigma \left(\mathrm{W}_l \cdot \mathrm{AGGREGATE}\left(\left\{ \mathrm{h}_{u}^{(l-1)}, \forall \mathrm{u} \in \mathrm{N(v)}\right\}\right)\right) \ldots 1$$

To the two small formulas

Eqn 2 and 3:

$$\mathrm{h}_{N_v}^{(l)} = \sigma \left(\mathrm{W}_l \cdot \mathrm{AGGREGATE}_N \left(\left\{\mathrm{h}_{u}^{(l-1)}, \forall \mathrm{u} \in \mathrm{N(v)}\right\}\right)\right) \ldots 2$$

$$\mathrm{h}_{v}^{(l)} = \mathrm{COMBINE}\left(\mathrm{h}_{u}^{(l-1)}, \mathrm{h}_{N_v}^{(l)}\right) \qquad \ldots 3$$

Where the COMBINE function is the main catch to the concept of "skip connection" between different layers. But this strategy also fails since it creates another problem that if we skip a layer $\mathrm{h}_{v}^{(l)}$ and not aggregate it in the short skip it then all subsequent units will also have to skip this in other words we cannot say that if a higher-up representation $\mathrm{h}_{u}^{(l+j)}$ uses the skip operation and another one doesn't. So, this skip-connection also cannot adaptively adjust the neighborhood sizes.

These above observations lead to the fact whether a fixed but structure-dependant radius can be taken but the answer remains no in the case of large radii as well as small radii. So, the authors proposed two powerful architectural changes in the jump connection and the subsequent adaptive aggregation mechanism.

The main idea of this model is that the common neighborhood aggregation techniques in the previously described networks say that each layer will aggregate the influence of its previous layers and the whole thing will generally lead to a fading of the actual feature vector. But in the JK-Net what happens is that each node is taken care of as each intermediate representation of its results is taken into consideration at the last layer as shown in the figure (this “jump” to the last layer) is done independently. Thus, if this is done explicitly for each node then the whole problem of adaptive learning of the radius for each node is solved. Let $h_v^{(1)}, \ldots, h_v^{(k)}$ be the jumping representations of node v (from k layers) that are to be aggregated into one. Then again 3 possibilities were tried out

1. Concatenation: Most simplistic way of combining the $h_v^{(1)}, \ldots, h_v^{(k)}$ layers and then performing a linear transformation. This approach is not node adaptive instead it combines such that it works well for the whole dataset overall. This approach works well with smaller and less complex structures.
2. Max-pooling: This approach selects element-wise max($h_v^{(1)}, \ldots, h_v^{(k)}$) from the set which is the most informative layer of feature coordinates. This max-pooling has the advantage that it is simple and uses no extra parameters to learn. Globally selects and prefers the higher-order layers mostly in general.
3. LSTM-attention: This uses an attention mechanism to get the most useful neighborhood ranges for each node by computing an attention score at each layer which represents the importance of a feature learned on that layer for that particular node.

A simple block diagram of the JK-Net is given below in this dissertation for better understanding:

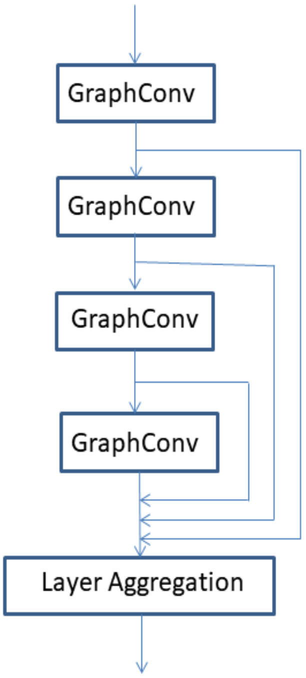


Fig 7: Jumping-Knowledge Graph Convolution Neural Network

The JK-Net in general tries to find the importance of a node's sub-graph features at different ranges after looking at the learned features at all the layers, than just blindly trying to optimize and fix the same predefined weights for all nodes.

Thus, in our case also this can be easily understood that we cannot predict before whether a drug will be linked to a gene-set of the same radii maybe it will differ with the structure of the sub-graph or the number of nodes that particular drug is linked with. Thus, using the JK-Net here will according to us increase the probability that the net features learned using the hidden layers will be more enriched and will produce better end results achieving the final task of link prediction further the main goal of drug repurposing.

Validation and Test Results

Though this can be tried and tested on any of the diseases present in the database and the drugs and genes linked with it is present. But to provide an example and make the explanation simple here 2 diseases are selected and the whole process of validation is done here in this dissertation. The diseases that were chosen are Bipolar Disorder and Dementia but to remind us again it is just done for simplicities sake that these 2 diseases were chosen otherwise any other diseases can be chosen.

To proceed with the validation and implementation and result from calculation let's know in brief about the diseases and how the whole process is structured here.

Though the exact cause of Bipolar Disorder is not known it is found that genetics, environment, and altered brain structure and chemistry play a role in its causing. It is broadly known to have 3 symptoms: Manic episode includes high energy, insomnia, and losing the touch with reality, whereas the Depressive episode may result in lack of energy, lack of interest in daily activities, low motivation, and the Mood episodes may last for a longer time period which may cause suicidal tendencies. The long treatment process comprises both medicinal and psychotherapy.

Dementia: It is a group of symptoms that may affect the social and thinking abilities of a person and may cause trouble in the daily life of a person. The group of symptoms includes forgetfulness and lessening the thinking and social abilities of a person. Medication and therapies may help to manage those symptoms but are irreversible most of the time.

Model Training Results

The Results when only residual GCN above mentioned is used without much tweaking the concept just stacking one layer on another the observation was found and recorded below:

MODEL	LAYERS	TEST AUC	TEST HITS@10	TEST HITS@100
GCN	2	0.991	0.799	0.889
GCN	3	0.990	0.864	0.935
GCN	4	0.991	0.836	0.957

Table 2: Residual Graph Convolution Neural Network Results.

Here the Concept of Parallel RGCN is used and the observed results were recorded various combinations were tried out in this process the observed results were found in the HITS@K where 'K' is any choice of number. It is nothing but a random subset of given size 'K' will be taken out and only the positive results will be considered among them and the ratio of the Hits will be calculated.

MODEL	**LAYERS**	**TEST AUC**	**TEST HITS @10**	**TEST HITS@100**
GCN with Residual Connections	4	0.989	0.818	0.926
GCN with Residual Connections	8	0.992	0.827	0.945
GCN with Residual Connections	12	0.992	0.859	0.919
GCN with Parallel Connections	3,4 and 5	0.993	0.817	0.967
GCN with Parallel Connections	4,4 and 4	0.994	0.833	0.947
GCN with Parallel Connections	2,4 and 6	0.993	0.875	0.907

Table 3: Parallel Graph Convolution Neural Network Results.

Finally, the modified JK-Net that was described in detail above was applied to the given dataset and the results were obtained and recorded as follows:

MODEL	**LAYERS**	**TEST AUC**	**TEST HITS@10**	**TEST HITS@100**
Jumping Knowledge Network (Concatenation)	5	0.993	0.855	0.914
Jumping Knowledge Network (Concatenation)	7	0.993	0.882	0.939
Jumping Knowledge Network (Concatenation)	9	0.991	0.863	0.897
Jumping Knowledge Network (Max -pooling)	5	0.994	0.858	0.952
Jumping Knowledge Network (Max -Pooling)	7	0.993	0.872	0.927
Jumping Knowledge Network (Max -Pooling)	9	0.991	0.830	0.934
Jumping Knowledge Network (LSTM -attention)	5	0.993	0.904	0.912
Jumping Knowledge Network (LSTM -attention)	7	0.994	0.890	0.942
Jumping Knowledge Network (LSTM -attention)	9	0.994	0.888	0.939

Table 4: JK-Net Results.

Validation Data Description

The validation data is mainly comprised of the selected drug-gene pairs which are not in our main dataset and the resulting score that is the presence of a valid link between those two is predicted. The above-mentioned model will predict 2 scores as output those are positive and negative edge scores respectively and their summation will be '1'. Here the table that will follow after this as validation test results will have only the positive edge value column in it which tells us the possibility of an edge being present in between those two nodes. Here in our case, the two nodes were programmatically found out using the Node2Vec algorithm then considering 2-4 hops of the gene-gene interaction graph to find out the gene nodes which did not have any link in the drug-gene interaction graph i.e., in the test set on which the model was trained. A set of drug-gene pair lists was given to the model related to the 2 diseases above mentioned were sent as input and we got the output list containing the positive and negative scores then we sorted out the links according to the highest positive scores and we got the results and only put here those for which viable proof was found there were such results too which were still trial drugs so no validation for such pair was found.

Now in the next part, the result of the model prediction is attached to the table and the validation of those links will be provided as the link maybe some of them are proven some of them are still under trial but the links will be sufficient for the proof in this case of our model this results will provide us, the knowledge about the validity and the power of the above model described in this dissertation.

Validation of Results

DrugBank ID	Drug Name	Predicted off-site target	Prediction score	Known Target	Validation reference
DB00382	Tacrine	APOE	0.816	ACHE	[30],[31]
DB00472	Fluoxetine	ALK	0.741	ALB	[32]
DB00382	Tacrine	MAOA	0.856	BCHE	[33]
DB00472	Fluoxetine		0.684	ALB	[38]
DB00674	Galantamine	SNAP25	0.695	CHRNA4	[34]
DB00458	Imipramine	SNCA	0.807	SLC6A3	[35]
DB00334	Olanzapine	GABBR2	0.675	CHRM3	[36]
DB00334	Olanzapine	MAOA	0.625	HTR1A	[37]

Validation of the predicted drug-gene associations

Tacrine is a well-known drug which is known more commonly as EC 3.1. 1.7 (acetylcholinesterase) inhibitor. There was a lot of research on this drug and the references above mentioned in the table the drug has a clear relation with "ACHE" but is returned to have a hidden connection with the apolipoprotein E (ApoE) phenotype for the treatment of Alzheimer's disease. Tacrine has also been seen to occur in position 3 in the above table for the gene Monoamine Oxidase A which is also a valid proof that it can work well in treating people suffering from acute AD to inhibit the production of acetylcholine but there was also a problem of Tacrine caused toxicity for which many Tacrine analogous compounds were developed.

Fluoxetine which is a drug that can be used to treat vascular dementia has a direct connection with the 'ALB' gene found to be also having a link with Alkaloids (ALK) from Trichilia monadelpha possess antidepressant which can be used as a treatment for the rapid antidepressant and by direct search in Google we can even find that this drug is also can be used in treating over compulsive disorder (OCD).

Galantamine is known to work well with dementia symptoms of the Alzheimer's disease and is known to have a connection with ABCA, ApoE3, CYP2D, CHAT, CHRNA, and ESR but has no direct link with SNAP25-like genes but it has been predicted it out and it was found that SNAP-25 is a promising cerebrospinal fluid for synapse degeneration in Alzheimer' and the valid reference would be this paper mentioned below published in 2014. "SNAP25_is_a_promising_novel_cerebrospinal_fluid_biomarker_for_synapse_degeneration_in_Alzheimer's_disease."

Next on the list comes Imipramine which is a known drug used for the treatment of Bipolar-Disorder is an antidepressant by nature is providing a positive result in the case of "SNCA" gene which can be seen as having a direct relation with the treatment of Parkinson's Disease a detailed study was performed except that which was mentioned as the proof which was an exclusive drug-disease paper known as "Double-blind Placebo" Study which proves the significance of Imipramine in treating Parkinson's disease.

Olanzapine which is known to work very well with diseases Bipolar disorder and Schizophrenia shows linkage through the positive score in the case of GABBR2 and MAOA which in turn is linked with the GABAergic system genes in neuroblastoma IMR-32 cells which are known to be linked with the Alzheimer's disease again but a study shows it must be used as in a combination of "Selank" which exhibits the effects of the Olanzapine. In similar results, it is shown to have a positive linkage with MAOA gene mentioned above to have links with AD patients. Though it needs to be mentioned this drug is not clinically approved for AD treatments in the case of an adult over 60 ages. But it is also seen to be used directly in action with AD patients hospitalized with excessive agitation-related symptoms.

Fluoxetine can also be used as an anti-inflammatory property which is caused by continuous treatment when the patient becomes resistant to a specific treatment this can be used in place of some anti-depressant to work for a while which has been found in research [38] mentioned in the references.

Conclusion

As has been seen in the above results table and the elaborated description of it the positive scores have a real-time significance in the world of biomedicine. The results in most of the cases were having valid proof is given here. Apart from these, there are such results too which are under trial. This will be an evolutionary experiment and this model will be having real-time significance if we can find a subset of the similar diseases and the new disease as it was in the case of Covid-19 all pneumonia medicines were tried out in this case also from the similar subset of the diseases we can get the medicines and we can try out and test various new pair and links linked with the new disease and then get some help in sorting which medicines to be used and valid repurposing can happen. Furthermore in conclusion I would say that there are still many shortcomings of this developed model it can be fine-tuned and the changes will lead to more precise and accurate results it is our belief. The rest future scope and the further developments that could be thought of as of now are listed below in the Future Scope part which will follow the Conclusion.

Future Scope

This dissertation just initiates this research work field since no other paper or research has been published in this field till now except that of the BiFusion paper mentioned above many a time. The scope of proceeding in this field is maximum. So this is almost the end of this dissertation here I will discuss mainly our ideas and hypothesis about the future of this dissertation which cannot be implemented because of the lack of time that was provided and due to the lack of manpower coding ability altogether. First and the foremost thing that I would like to mention is the implementation of the concept of a temporal graph neural network in the backbone in place of the R-GCN model. As it is named temporal GCN works on graphs that change over time and new links can be generated over time-period. As we know the paper on temporal GCN by Yi Liu and others is described and tested on a traffic signaling graph but just like a traffic signaling graph this drug-gene-disease graph can also change over time since new diseases as well as existing drugs under trial can produce active links for some genes. Then the whole model would be again trained this is where the temporal GCN may come in handy.

The second point would be that there is a lack of data, not every drug or gene, or disease can be listed it is a hectic task only some popular drugs and diseases are listed. The data in some cases found to be not up to the mark since it was listed but not by any person who are having specific domain knowledge may lead to false data, even data which are not absolutely correct.

Thirdly there are many different connections a disease further depends upon not only genes those factors can be also linked with this graph to form an overall heterogeneous graph to train a more sophisticated model to achieve better results, we have all seen the use of an existing drug repurposing in recent times during the covid-19 period.

The fourth point would be that there are different factors like a drug can react with another drug, or maybe a drug will react with a gene that suppresses it which is not at all good or may react to form some complications these factors will also play a vital role in making the model more sophisticated.

Very few number of layers are given in our model the increase in layer numbers will make the model predictions less efficient that problem may be solved in some other way using some other model than that of the Jumping Knowledge model it can also be explored further which we couldn't pursue because time was limited.

Then lastly the features used by us were very less may be somehow the features of drugs-genes-diseases can be increased or accumulated to perform better model predictions use of more efficient evaluation metrics for the heterogeneous graphs overall could also help in the process of model training in our case it was only the Hits@K and accuracy we had to validate later using extensive browsing this can be made simpler and more efficient using some other intuitive evaluation metrics.

References

[1] Oerton E, Roberts I, Lewis PSH, Guilliams T, Bender A. Understanding and predicting disease relationships through similarity fusion. Bioinformatics. 2019;35(7):1213-1220. doi:10.1093/bioinformatics/bty754

[2] Gottlieb A. et al. (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol. Syst. Biol., 7, 496.

[3] Cheng, F., Lu, W., Liu, C. et al. A genome-wide positioning systems network algorithm for in silico drug repurposing. Nat Commun 10, 3476 (2019). https://doi.org/10.1038/s41467-019-10744-6.

[4] Cheng L. et al. (2014) SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. PLoS One, 9, e99415.

[5] Cheng F. et al. (2019) Network-based prediction of drug combinations. Nat. Commun., 10, 1–11.

[6] Ma, T. et al. (2018). Drug similarity integration through attentive multi-view graph auto-encoders. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018.

[7] Guney E. et al. (2016) Network-based in silico drug efficacy screening. Nat. Commun., 7, 10331.

[8] Zhang P. et al. (2013) Computational drug repositioning by ranking and integrating multiple data sources. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 579–594. Springer, Berlin, Heidelberg.

[9] Kipf T.N. , Welling M. (2017) Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (ICLR).

[10] Li J. , Lu Z. (2012) A new method for computational drug repositioning using drug pairwise similarity. In: 2012 IEEE International Conference on Bioinformatics and Biomedicine, pp. 1–4. IEEE Press.

[11] “Common genetic associations between age-related diseases", Donertas, Handan Melike and Fabian, Daniel K and Fuentealba, Matias and Partridge, Linda and Thornton, JanetM,

[12] Pavlopoulos G.A. et al. (2018) Bipartite graphs in systems biology and medicine: a survey of methods and applications. Gigascience, 7, giy014.

[13] Kontou P.I. et al. (2016) Network analysis of genes and their association with diseases. Gene, 590,

[14] Zeng X. et al. (2019) deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics, 35, 5191–5198.

[15] Zichen Wang, Mu Zhou, Corey Arnold, Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing, Bioinformatics, Volume 36, Issue Supplement_1, July 2020, Pages i525–i533, https://doi.org/10.1093/bioinformatics/btaa437.

[16] Saberian, Nafiseh et al. “A new computational drug repurposing method using established disease-drug pair knowledge.” Bioinformatics (Oxford, England) vol. 35,19 (2019): 3672-3678. doi:10.1093/bioinformatics/btz156.

[17] Napolitano F. et al. (2013) Drug repositioning: a machine-learning approach through data integration. J. Cheminf., 5, 30.

[18] Luo H. et al. (2016) Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. Bioinformatics, 32, 2664–2671.

[19] He C. et al. (2020) Bipartite graph neural networks for efficient node representation learning. In: Thirty-Fourth AAAI Conference on Artificial Intelligence. AAAI press.

[20] Feng Q. et al. (2018) PADME: a deep learning-based framework for drug-target interaction prediction. arXiv preprint arXiv:1807.09741.

[21] Yildirim M.A. (2007) Drug–target network. Nat. Biotechnol., 25, 1119–1127. 68–78

[22] Fralick M. et al. (2019) Assessment of use of combined dextromethorphan and quinidine in patients with dementia or Parkinson's disease after our food and drug administration approval for pseudobulbar affect. JAMA Internal Med., 179, 224–230.

[23] Drug Repurposing for Cancer: An NLP Approach to Identify Low-Cost Therapies. Shivashankar Subramanian, Ioana Baldini, Sushma Ravichandran, Dmitriy A. Katz-Rogozhnikov, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Kush R. Varshney, Annmarie Wang, Pradeep Mangalath, Laura B. Kleiman arXiv:1911.07819 [cs.CL].

[24] Hamilton W. et al. (2017) Inductive representation learning on large graphs. In: Advances in Neural Information Processing Systems, pp. 1024–1034. Curran Associates Inc., Red Hook, NY, United States.

[25] Scarselli F. et al. (2008) The graph neural network model. IEEE Trans. Neural Netw., 20, 61–80.

[26] Hamosh A. et al. (2004) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res., 33, D514–D517.

[27] Li J. , Lu Z. (2012) A new method for computational drug repositioning using drug pairwise similarity. In: 2012 IEEE International Conference on Bioinformatics and Biomedicine, pp. 1–4. IEEE Press.

[28] Lubecka K. et al. (2018) Novel clofarabine-based combinations with polyphenols epigenetically reactivate retinoic acid receptor beta, inhibit cell growth, and induce apoptosis of breast cancer cells. Int. J. Mol. Sci., 19, 3970.

[29] Luo H. et al. (2016) Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. Bioinformatics, 32, 2664–2671.

[30] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, Max Welling (arXiv:1703.06103 [stat.ML], (or arXiv:1703.06103v4 [stat.ML] for this version), https://doi.org/10.48550/arXiv.1703.06103.

[31] Bao, J.; Duan, N.; Zhou, M.; and Zhao, T. 2014. Knowledgebased question answering as machine translation. In ACL.

[32] Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modelling multi-relational data. In NIPS.

[33] Chang, K.-W.; tau Yih, W.; Yang, B.; and Meek, C. 2014.Typed Tensor Decomposition of Knowledge Bases for Relation Extraction. In EMNLP.

[34] Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. arXiv preprint arXiv:1704.01212.

[35] Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive representation learning on large graphs. arXiv preprint arXiv:1706.02216.

[36] Kipf, T. N., and Welling, M. 2016. Variational graph autoencoders. arXiv preprint arXiv:1611.07308.

[37] Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2016. Gated graph sequence neural networks. In ICLR.

[38] Toutanova, K., and Chen, D. 2015. Observed versus latent features for knowledge base and text inference. In Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality, 57–66.

[39] Toutanova, K.; Lin, V.; Yih, W.-t.; Poon, H.; and Quirk, C.2016. Compositional learning of embeddings for relation paths in knowledge base and text. In ACL.

[40] Sjögren, M., Hesse, C., Basun, H. et al. Tacrine and rate of progression in Alzheimer's disease – relation to ApoE allele genotype. J Neural Transm 108, 451–458 (2001).

[41] The apolipoprotein E ε4 allele and the response to tacrine therapy in Alzheimer’s diseaseA.-S. Rigaud,L. Traykov,L. Caputo,M.-C. Guelfi,F. Latour,R. Couderc,F. Moulin,J. De Rotrou,F. Forette,F. Boller. First published: 09 October 2008. https://doi.org/10.1046/j.1468-1331.2000.00073.x

[42] Glycine/NMDA Receptor Pathway Mediates the Rapid-onset Antidepressant Effect of Alkaloids From Trichilia Monadelpha - Basic and Clinical Neuroscience (iums.ac.ir) BCN 2021, 12(3): 395-408

[43] Bilqees Sameem, Mina Saeedi, Mohammad Mahdavi, Abbas Shafiee, A review on tacrine-based scaffolds as multi-target drugs (MTDLs) for Alzheimer's disease, European Journal of Medicinal Chemistry, Volume 128,2017, Pages 332-345, ISSN 0223-5234, htts://doi.org/10.1016/j.ejmech.2016.10.060.

[44] Sumirtanurdin R, Thalib AY, Cantona K, Abdulah R. Effect of genetic polymorphisms on Alzheimer's disease treatment outcomes: an update. Clin Interv Aging. 2019;14:631-642. Published 2019 Mar 29. doi:10.2147/CIA.S200109.

[45] Edward C. Lauterbach, Psychotropic drug effects on gene transcriptomics relevant to Parkinson's disease, Progress in Neuro-Psychopharmacology and Biological Psychiatry, Volume 38, Issue 2, 2012, Pages 107-115, ISSN 0278-5846, https:/doi.org/10.1016/j.pnpbp.2012.03.011.

[46] Filatova E, Kasian A, Kolomin T, et al. GABA, Selank, and Olanzapine Affect the Expression of Genes Involved in GABAergic Neurotransmission in IMR-32 Cells. Front Pharmacol. 2017;8:89. Published 2017 Feb 28. doi:10.3389/fphar.2017.00089.

[47] Chen ML, Chen CH. Chronic antipsychotics treatment regulates MAOA, MAOB and COMT gene expression in rat frontal cortex. J Psychiatr Res. 2007 Jan-Feb;41(1-2):57-62. doi: 10.1016/j.jpsychires.2005.03.005. Epub 2005 Jun 17. PMID: 15964593.

[48] T116. Anti-Inflammatory Parp Inhibitor Demonstrates Antidepressant Activity in Animal Model of Treatment-Resistant Depression. Gregory Ordway, W. Drew Gill, Joshua B. Coleman, Hui Wang-Heaton, Michelle Chandley, Russell Brown. DOI:https://doi.org/10.1016/j.biopsych.2019.03.439.

[49] Yang, Fei & Zhang, Huyin & Tao, Shiming & Hao, Sheng. (2022). Graph representation learning via simple jumping knowledge networks. Applied Intelligence. 1-19. 10.1007/s10489-021-02889-z.

[50] Representation Learning on Graphs with Jumping Knowledge Networks by Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, Stefanie Jegelka, arXiv:1806.03536v2

[51] Hammond, D. K., Vandergheynst, P., and Gribonval, R. Wavelets on graphs via spectral graph theory. Applied and Computational Harmonic Analysis, 30(2):129–150, 2011.

[52] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016a.

[53] Hochreiter, S. and Schmidhuber, J. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[54] Perozzi, B., Al-Rfou, R., and Skiena, S. Deepwalk: Online learning of social representations. In ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 701–710, 2014.

[55] Lin, Xuan & Quan, Zhe & Wang, Zhi-Jie & Ma, Tengfei & Zeng, Xiangxiang. (2020). KGNN: Knowledge Graph Neural Network for Drug-Drug Interaction Prediction. 2711-2717.

[56] Rohani, N., Eslahchi, C. Drug-Drug Interaction Predicting by Neural Network Using Integrated Similarity. *Sci Rep* **9,** 13645 (2019). https://doi.org/10.1038/s41598-019-50121-3.

[57] Celebi, R., Uyar, H., Yasar, E. et al. Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings. BMC Bioinformatics 20, 726 (2019). https://doi.org/10.1186/s12859-019-3284-5.

[58] Drug-Drug Interaction Prediction Based on Knowledge Graph Embeddings and Convolutional-LSTM Network, Md. Rezaul Karim arXiv:1908.01288**.**

[59] Wang Z, Zhou M, Arnold C. Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing. Bioinformatics. 2020 Jul 1;36(Suppl_1):i525-i533. doi: 10.1093/bioinformatics/btaa437. PMID: 32657387; PMCID: PMC7355266.

[60] Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ, von Mering C. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021 Jan 8;49(D1): D605-D612. doi: 10.1093/nar/gkaa1074. Erratum in: Nucleic Acids Res. 2021 Oct 11;49(18):10800. PMID: 33237311; PMCID: PMC7779004.

[61] STRING is part of the ELIXIR infrastructure: it is one of ELIXIR's https://string-db.org/.

[62] DisGeNet reference link can be found here : https://www.disgenet.org/.

[63] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017 Nov 8. doi: 10.1093/nar/gkx1037.