Graph Neural Networks for Drug Repurposing

Kushagra Agarwal[1] Shreeya Pahune[1] Sumeet Agarwal[1,2]

[1]International Institute of Information Technology, Hyderabad
[2]Microsoft India, Hyderabad

May 4, 2022

Abstract

The COVID-19 pandemic has brought attention to the significance of prioritizing licenced medications for the treatment of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). In this paper, we propose an end-to-end integrated pipeline to screen the plethora of approved drugs to be repurposed for COVID-19. We leverage a network learning paradigm implemented through a Graph Autoencoder performing link prediction. The information is extracted jointly using the Protein-Protein Interaction Graph, Drug-Target Protein Graph, Disease-Protein Graph and the Drug-Disease graph. A novel, step-by-step learning approach was proposed which obtained superior results (AUCROC: 0.92) compared to the node type agnostic approaches mentioned in literature. The model ranked 6158 drugs based on their efficacy in treating COVID-19 and the top ranking hits were verified. Beyond COVID-19, our integrated framework can allow us to uncover drug-repurposing prospects for any other novel/neglected diseases with minimal changes to input protein list.

1 Introduction

1.1 Motivation

Although investment in biomedical and pharmaceutical research and development has increased significantly over the past two decades, the annual number of new treatments approved by the U.S. Food and Drug Administration (FDA) has remained relatively constant and limited. A recent study estimated that pharmaceutical companies spent $2.6 billion in 2015, up from $802 million in 2003, in the development of an FDA-approved new chemical entity drug.

1.2 Problem Definition

Drug re-purposing, represented as an effective drug discovery strategy from existing drugs, could significantly shorten the time and reduce the cost compared to de novo drug discovery and randomized clinical trials. Our problem statement is, "Designing a drug re-purposing networking framework for COVID-19". We aim to create a GNN model which can identify already approved/available drugs and re-purpose them for COVID-19 Fig. 1. Such a robust framework could also be applied to other novel/neglected diseases to gain insights into their treatment. The code is available at `https://github.com/kushagragarwal2443/Drug_Repurposing_GNN`.

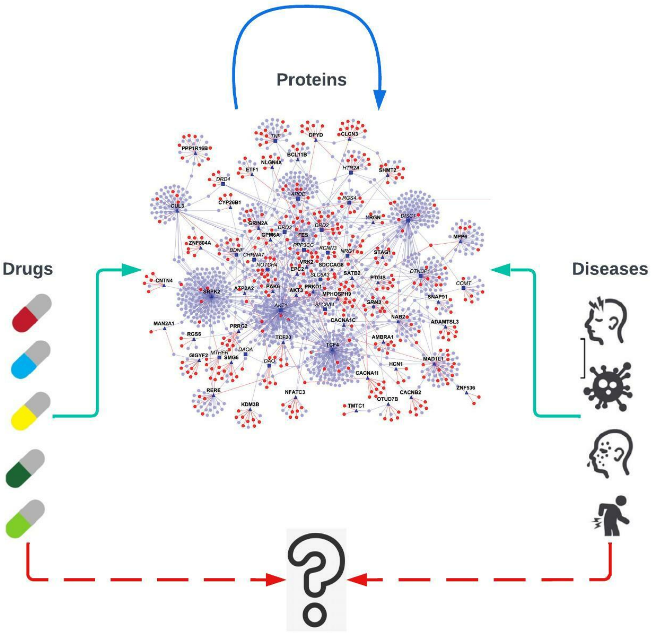


Figure 1: Problem Definition: Predicting Drug-Disease interactions using Protein-Protein, Drug-Protein and Disease-Protein interactions.

1.3 Challenges

- The requisite data is multi-modal, including Protein-Protein interactions, Drug-Protein Targets, Disease-Protein networks and known Drug-Disease pairs. Hence, integrating all this information to extract knowledge under the framework of Graph Neural Network is challenging.
- Non-availability of code and lack of proper evaluation metrics to compare our results. Our main reference work [1] predicted top ranking drugs which were then experimentally and clinically validated. We do not have access to the same, which is why we can only compare our results to theirs.

1.4 Existing Work

Very little work has been done in the area till now, with Gysi et al (2021) [1], paper using a modified Decagon [2] framework being the state-of-the-art. Their graph neural network is an end-to-end trainable model for link prediction on the multimodal graph with (1) an encoder: a graph convolutional network operating on G and producing embeddings for nodes in G; and (2) a decoder: a model optimizing embeddings such that they are predictive of known drug-disease indications.

$$h_i^{(k+1)} = \phi(\Sigma_r \Sigma_{j\epsilon N_r^i} \alpha_r^{(k)} W_r^{(k)} h_j^{(k)} + \alpha_r^i h_i^{(k)}) \tag{1}$$

Their approach removes the distinction between different types of nodes and performs graph convolutions on this union graph. They applied various graph theoretical approaches as well to finally get the consensus ranking of different pipelines.

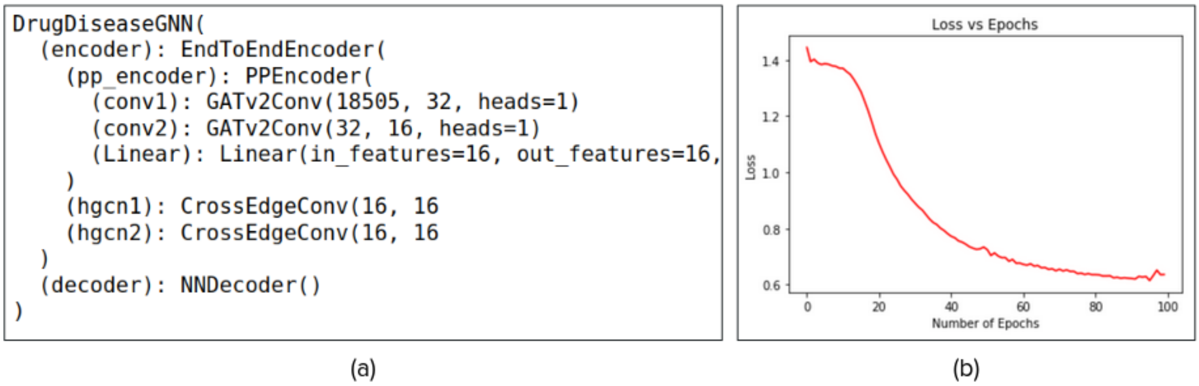
2 Team Contribution

Name	Tasks
Kushagra	1. Problem Formulation 2. Data Preprocessing (Minor Contribution) 3. Graph Neural Network code
Shreeya	1. Data Preprocessing (Major Contribution) 2. Validation/Verification of Predicted Graph
Sumeet	1. Tried alternative Heterograph approach

3 Experiments

3.1 Experiment 1

We integrated PPI, Drug-Protein data, Disease-Protein data and Drug-Disease interactions to perform Graph Attention. End-to-end drug-disease link prediction pipeline was trained using the Encoder-Decoder architecture shown in Fig. 2. Our trained model performed very well on the hold-out test set, achieving an AUCROC of 0.9282 and AUPRC of 0.9053.



```
DrugDiseaseGNN(
  (encoder): EndToEndEncoder(
    (pp_encoder): PPEncoder(
      (conv1): GATv2Conv(18505, 32, heads=1)
      (conv2): GATv2Conv(32, 16, heads=1)
      (Linear): Linear(in_features=16, out_features=16,
    )
    (hgcn1): CrossEdgeConv(16, 16
    (hgcn2): CrossEdgeConv(16, 16
  )
  (decoder): NNDecoder()
)
```

(a)

(b)

Figure 2: (a) Encoder-Decoder architecture (b) Train Loss over 100 epochs

3.2 Experiment 2

Using the trained model, we predicted links between the COVID-19 Disease node (indexed 1446 in our graph) and all the drugs in the database. The top hits were compared with the predictions of Gysi et al., (2021) Table 1:

- Among our top 20 hits, 7 of their top 10 are present.
- Among our top 10, 5 of their top 10 ar present.

4 Novel Ideas

The Gysi et al., (2021), paper did not consider node types and modelled the whole network in one large graph. We thought that this node agnostic approach could lead to inferior results in the presence of pre-defined initial embeddings for different node types. To enable different

Our Rank	Drug Name	CRank in Reference
1	Rifaximin	9
2	Cilostazol	4
3	Cortisone acetate	15
4	Flutamide	7
5	Ritonavir	1
6	Folic acid	16
7	Isoniazid	2
8	Urea	12
9	Deamido-Nad+	NA
10	Betamethasone	18
11	Troleandomycin	3
12	Dimethyl sulfoxide	14
13	Methionine	NA
14	Vindesine	75
15	Gefitinib	34
16	Buspirone	NA
17	Hydroxychloroquine	44
18	Ticlopidine	29
19	Rifabutin	6
20	Quinine	28

Table 1: Rank predictions for drug repurposing. Drugs in red correspond to the top 3 and those in blue to the top 4-10 hits in the reference paper.

node types to be present in the graph, we defined our graph G = (V,E), where V consists of 3 types of nodes: Proteins, Drugs, Diseases and E consists of 4 type of edges: Protein-Protein, Drug-Protein, Disease-Protein, Drug-Disease.

We trained the heterogeneous graph in a step-wise fashion Fig. 3. We first trained our Protein embeddings using the PPI network (2 GATConv layers and 1 Linear layer). Then we combined the initial Drug embeddings with the trained Protein Embeddings and passed them through a custom convolution layer which aggregates incoming Protein embeddings to produce the final Drug Embedding. The same was performed to train the Disease embeddings. Once both the embeddings were obtained, we passed them both through 2 layer fully connected neural networks to get a sigmoid score between 0 and 1 indicating edge probability. The end-to-end pipeline was trained using known drug-disease interactions in our graph.

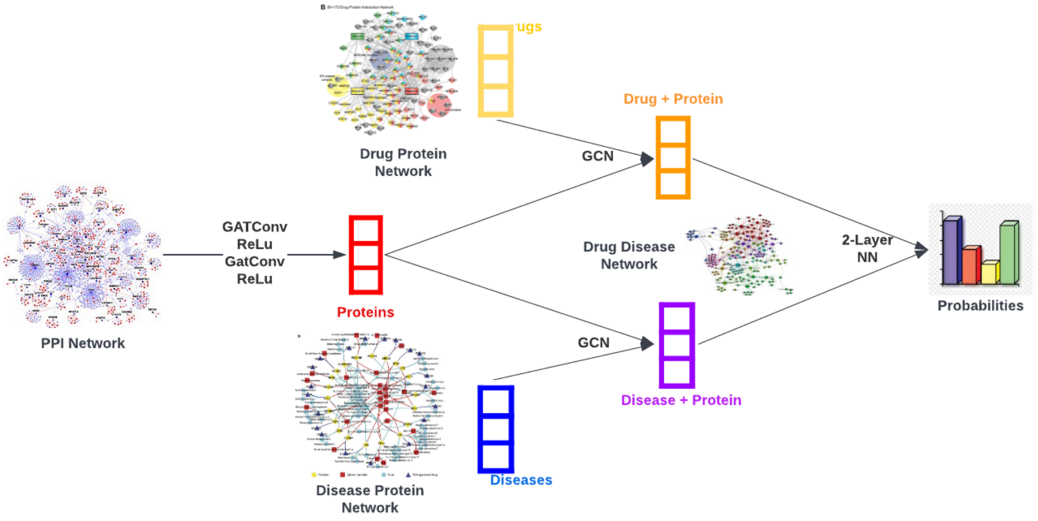


Figure 3: Proposed Pipeline: Protein Embeddings were trained using the PPI network. Next, Drug and Disease Embeddings were trained using custom defined CrossEdge layers. Finally, 2 layer FC NN was used as a decoder to predict edge probabilities.

This step-by-step approach of training our model not only helps us gain superior results (Decagon: AUROC: 0.87, DrugDiseaseGNN: 0.92), but also enables inclusion of pretrained protein, drug and disease embeddings even before training starts to better optimize the loss.

5 Ideas that did not work

- We wanted to add pretrained initial embeddings for proteins, diseases and drugs. We faced the following problems with the same:
  - prot2vec requires amino acid sequences as input. We only had the Entrez IDs of the proteins using which there was no automated way of getting the amino acid sequences (Possible only by using RefSeqID).
  - The Drug2Vec pretrained model only contained embeddings for 540 drugs, only 252 of which intersected with our drug set.
- The Heterograph(Deep Graph) approach: We proposed an end-to-end trainable model with a DGL Multi modal Heterogeneous graph. The Graph again contains 3 types of Nodes and 4 types of Edges. We applied Graph Attention Convolutions on all the 4 edges with mean aggregation function. The embeddings obtained were passed through the Decoder to compute the loss using simple inner-product. We were not able to run the entire Heterograph code due to an error in the loss computation for Positive and Negative class samples separately.

6 References

1. Gysi, Deisy Morselli, et al. "Network medicine framework for identifying drug-repurposing opportunities for COVID-19." Proceedings of the National Academy of Sciences 118.19 (2021).

2. Zitnik, Marinka, Monica Agrawal, and Jure Leskovec. "Modeling polypharmacy side effects with graph convolutional networks." Bioinformatics 34.13 (2018): i457-i466.