International Journal Of Public Mental Health And Neurosciences
ISSN: 2394-4668
(An Official publication of Sarvasumana Association)

Network-Driven Drug Re-purposing: Insights from Graph Neural Networks and Recommendation Models

Kushagra Agarwal
School of Computer Science
Carnegie Mellon University
Pennsylvania, USA
kagarwa2@andrew.cmu.edu

Shreeya Pahune
Center for Computational Natural Sciences and Bioinformatics
Institute of Information Technology - Hyderabad
Hyderabad, India
shreeya.pahune@research.iiit.ac.in

Hemant Chandak
Department of Information and Communication Technology
Rajasthan, India
hemant.209303369@muj.manipal.edu

Sumanth Agrawal
Avalara Technologies Private Limited
Maharashtra, India
sumeetagrawal.srm@gmail.com

Abstract*—The rising cost of drug discovery, coupled with a stagnation in the approval of novel treatments, highlights the urgent need for innovative strategies such as drug re-purposing. Pharmaceutical companies invest roughly 10-15 years and $2.6 billion to get a single FDA-approved drug to market. The COVID-19 pandemic further underscored the necessity of quickly identifying existing drugs with potential efficacy against a fast-spreading virus to curtail the pandemic. In this study, we perform a comparative analysis of several Graph Neural Networks (GNNs) and recommendation system models to address drug re-purposing. We construct an integrated graph that combines Protein-Protein Interaction networks, Drug-Target Protein graphs, Disease-Protein associations, and Drug-Disease links. We leverage a network learning paradigm implemented over this complex graph via both node-agnostic and heterogeneous graph techniques for link prediction in drug-disease pairs. We implement a Heterogeneous Graph Transformer (HGT) model that processes three node types (drugs, diseases, proteins) and four edge types. The HGT achieved an AUC-ROC of 0.985 and an F1-score of 0.90, demonstrating its efficacy in predicting drug re-purposing candidates. Additionally, we compared several node-agnostic GNN architectures, including Graph Convolutional Networks, Graph Attention Networks, GraphSAGE, and Graph Isomorphism Networks. All architectures performed comparably, with an AUC-ROC of around 0.98. However, when framing the drug re-purposing task as a recommendation problem using Matrix Factorization with side information, we observed a significant drop in performance, with the AUC-ROC falling to 0.82. This performance degradation highlights the importance of incorporating Protein-Protein Interaction networks in the modelling process, as matrix factorization fails to capture these complex network effects critical for drug repurposing. Our models ranked 6,158 drugs based on their predicted efficacy in treating COVID-19, providing a valuable tool for prioritizing clinical trials and further research. Beyond COVID-19, such an integrated framework can allow us to uncover drug re-purposing prospects for any other novel diseases in a significantly more efficient and cost-effective way.

***Index Terms*— Graph Neural Network, Drug re-purposing, Machine Learning, COVID-19.**

I. Introduction

It is well established that developing safe and effective drugs is a tedious challenge today. According to a study conducted by DiMasi et al. (2016) [1], it can take pharmaceutical companies up to 15 years and approximately $2.6 billion in order to get FDA approval for a single new drug. The current drug development process lacks the required efficiency as it is both costly and time intensive.

Drug repurposing offers a promising solution to aid this time-consuming and expensive process. Rather than developing drugs from scratch, repurposing existing ones allows researchers to leverage the safety and efficacy data already collected for approved drugs, thus significantly reducing the time it takes for new drugs to reach the market [2].

Drug repurposing has already played a key role in finding effective treatments for diseases like COVID-19 [3, 4], Ebola [5], and some forms of cancer [6]. Machine learning-based approaches have previously been employed for this purpose, including both network-based models [7, 8] and deep learning techniques [9, 10]. Network-based models aim to capture the underlying relationships between different entities, while deep learning models analyse various modalities of data, such as transcriptomic or pharmacological profiles, to predict new

therapies for diseases with similar pathways. These advanced computational frameworks can predict potential treatments for novel diseases by learning complex relationships between existing drugs, known diseases and other biological data.

One of the best examples of the urgent need for expedited drug discovery techniques is the COVID-19 pandemic. When such a rapidly spreading disease emerges, developing a novel treatment from start to finish would take too long. Drug repurposing thus offers a necessary shortcut, allowing us to quickly identify potential treatments and proceed with their clinical trials [11].

In our study, we utilise graph neural networks in addition to a recommendation system model to tackle the challenge of drug repurposing. Protein-protein interaction networks, Drug-Protein graphs, Disease-Protein associations, and Drug-Disease links are all combined into an integrated graph. This complex network enables us to identify potential drugs for a specific disease by learning underlying patterns. Our goal is to develop a methodology that can identify existing drugs for the treatment of emerging threats like COVID-19. This approach could reduce the high expenses associated with creating novel drugs and shorten the drug discovery process.

II. RESULTS

We compared Heterogeneous Graph Neural Networks (HGT), Node-agnostic Graph Neural Networks and Matrix Factorization, on the task of link prediction between drug-disease pairs. Table I compares the two graph-based algorithms using Precision, Recall, F-1 score, and AUC-ROC.

We can observe that node-agnostic (homogeneous) graph neural networks gave better results than heterogeneous counterparts even though they have a simpler representation. This can be attributed to the likely lesser number of drug-disease pairs that train the weights. Within the homogeneous methods, the highest performance was achieved by GraphSAGE with an AUC-ROC of 0.9911 and an F1-score of 0.9610. This is followed closely by the heterogeneous approach with an AUC-ROC of 0.9849 and an F1-score of 0.9032.

TABLE I. PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	AUC – ROC	F1-Score	Precision	Recall
GraphSAGE	0.991	0.961	0.944	0.979
GCN	0.990	0.954	0.933	0.976
GAT	0.978	0.920	0.859	0.989
GIN	0.980	0.715	0.561	0.986
HGT	0.985	0.903	0.824	1.000
Matrix Factorization*	0.817	-	-	-

*Matrix Factorisation had Precision@10 and Recall@10 values

A. *Top hits for COVID-19*

We leveraged these methods to get the top 20 ranked drugs which could potentially be effective against the COVID-19 disease (MESH ID: D000086382). Interestingly, there was a significant overlap between the top 20 ranked drugs between the various methods. Seven drugs were highlighted by both the GCN and SAGE methods, and Heparin was common to GCN, GAT and HGT predictions. The results are summarized in Table II.

We identified a set of 31 drugs that could be used as candidates against COVID19 across the six approaches. Of the 31, the majority are antivirals (64.5%)). This is followed by 22.6% anticoagulants, 9.7% immunomodulators, and finally, 3.2% corticosteroids.

The large number of drugs in the antiviral category are in line with the viral origin of COVID-19 (spread by SARS-CoV-2). Additionally, we observed an overlap in the predictions and existing literature and currently used treatments for COVID-19. Anticoagulants, including Heparin, are associated with a reduction in mortality rate [12], and interferons and other immunomodulatory substances may affect the system's reaction in COVID-19 patients [13].

TABLE II. PERFORMANCE COMPARISON OF DIFFERENT MODELS

DrugBank ID	Drug Name	Category	Predicted By
DB00977	Heparin	Anticoagulant	GCN, SAGE, HGT
DB00316	Acyclovir	Antiviral	GCN, SAGE
DB00313	Amantadine	Antiviral	GCN, SAGE
DB00091	Ribavirin	Antiviral	GCN, SAGE
DB00783	Saquinavir	Antiviral	GCN, SAGE
DB01174	Amprenavir	Antiviral	GCN, SAGE
DB00515	Famciclovir	Antiviral	GCN, SAGE
DB00396	Methylprednisolone	Corticosteroid	GCN
DB00550	Prednisone	Corticosteroid	GCN
DB00755	Interferon alfa-2b	Immunomodulator	GCN

III. METHODS

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

A. *Dataset Construction*

For our study, we started by processing data from multiple networks downloaded from the Decagon project hosted on the Stanford Network Analysis Platform (SNAP) at Stanford University (*https: //snap.stanford.edu/decagon/*). These included protein-protein interactions (PPI), drug-protein associations, disease-protein associations, and drug-disease associations and were also used in the study conducted by Zitnik et al. (2018) [14]. Our final network consisted of 18505 proteins (denoted by ENTREZ gene IDs), 6158 drugs (DrugBank IDs), and 1448 diseases (MESH IDs). Further, we retained only those entities that were present across each of the datasets considered, resulting in 327924 protein-protein interactions, 22230 drug-

protein interactions and 195811 validated drug-disease links to train our models and predict the best potential drugs for COVID-19 (Figure 1).

The PPI network was represented by a symmetric adjacency matrix, and the remaining interaction pairs were represented using binary adjacency matrices, where 1 indicated a known interaction, and 0 indicated the absence of an interaction.

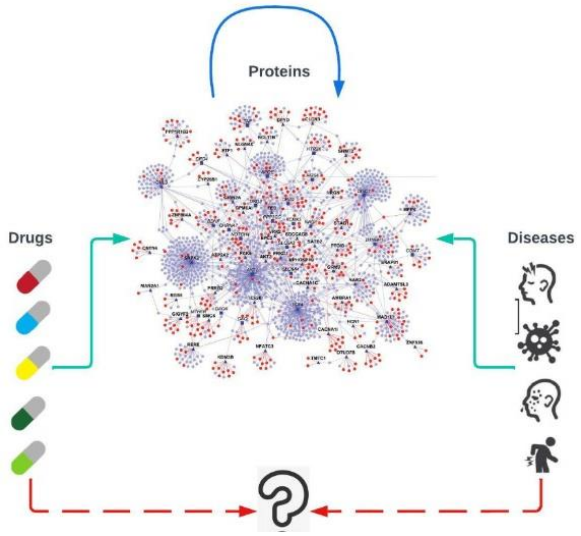

Fig. 1. Problem Definition: Predicting Drug-Disease interactions by leveraging Protein-Protein Interactions (PPIs), Drug-Protein interactions, Disease-Protein interactions and Drug-Disease associations extracted from the integrated network

B. *Homogeneous Graph Neural Networks*

We implemented four GNN architectures as part of our node-agnostic graph approach:

1. Graph Convolutional Networks (GCN) [15] which learns neighbourhood information using graph convolutions.
2. Graph Attention Networks (GAT) [16] which apply attention mechanisms to understand important segments of the graph.
3. GraphSAGE [17] which samples neighbouring nodes and aggregates their information.
4. Graph Isomorphism Networks (GIN) [18] which identifies differences within the structure of graphs.

These GNN architectures ensured that we captured high-level as well as specific representations within the graph. Each of the above architectures consisted of two graph convolutional layers along with ReLU activation functions. The hidden and output layers correspondingly consisted of 64 and 16 dimensions.

We then projected the features into embeddings of 10-dimension feature spaces for each entity (disease, drug, and protein). We performed negative sampling to generate a balanced set of drug-disease interactions by creating a number of negative, that is, non-interacting, drug-disease linkages.

Using a train, validation and test split of 70%, 15% and 15% correspondingly, we trained the models using binary cross-entropy loss with Adam optimizer for 50 epochs. Multiple metrics were used to test the performance of each model, including AUC-ROC, F1 score, precision and recall.

Following this, we predicted the top 10 drugs for COVID-19 (MESH ID: D000086382) for each model. Corresponding to each drug, we computed its score using the inner product of the learned disease embedding and all drug embeddings. Finally, we obtained a set of identified potential drug candidates for COVID-19 highlighted across the four architectures.

C. *Heterogeneous Graph Neural Networks*

We created a heterogeneous graph transformer (HGT) model that utilised all four types of associations as edges: disease-drug, drug-protein, disease-protein, and protein-protein interactions and the corresponding three types of nodes with node feature vectors. The model architecture consisted of 64-dimensional hidden layer channels for each entity and two HGT convolution layers, with four attention heads each.

The data was split using a 75-25 train-test split ratio. Following this, we trained the model over 50 epochs with loss as binary cross-entropy using the Adam optimizer with a learning rate of 0.001. Negative class sampling was used to create negative drug-disease pairs consistent with the homogeneous approach. The model's performance was then evaluated, and drug predictions for COVID-19 were obtained.

D. *Matrix Factorization*

In addition to the GNN-based approaches, we redefined the task as a recommendation problem that utilised Matrix Factorization to suggest the best drug options for a given disease, in this instance, COVID-19. Based on existing research, this method shows promising applications for drug discovery [19].

The initial feature matrices were constructed using the one-hot encoding of each drug and disease's associations with proteins. This resulted in two feature matrices, with 1 denoting a known interaction and 0 denoting its absence. Additionally, the binary adjacency matrix for drug-disease relations was used.

We used the LightFM [20] framework to implement our matrix factorization model. Our model used the drug-protein and disease-protein interaction features, both embedded into a shared 10-dimensional latent space, as input and recommended drugs for diseases. The model was trained using logistic loss across ten epochs with an 80-20 train-split. Metrics such as Precision@k, Recall@k, and AUC-ROC were calculated to assess performance across each epoch. Finally, we obtained predictions for the top 20 drugs that the model suggested for COVID-19.

All of our code, written in Python using the PyTorch framework for GNN models and the LightFM library for matrix

factorization, is available on GitHub at *https://github.com/kushagragarwal2443/Drug Repurposing GNN.*

IV. Discussion

Our findings highlight the superiority of graph-based approaches in drug repurposing. Node-agnostic GNN architectures, namely, GCN, GAT, GraphSAGE, and GIN, showed high performance with AUC-ROC around 0.98. Within these four models, GraphSAGE achieved the highest performance (AUC-ROC: 0.99 and F1 Score: 0.96). The Heterogeneous Graph Transformer (HGT) model showed similar performance with an AUC-ROC of 0.985 and an F1-score of 0.90, proving that the model could effectively capture the complex relationships between each node or entity (drugs, diseases, and proteins) using the diverse types of edges or interactions.

The matrix factorization approach, however, showed a significantly lower performance with an AUC-ROC of 0.82. This can be attributed to the absence of PPI networks in the feature inputs for this approach, which seem to contribute important information that can strengthen our understanding and, hence, the prediction of drug-disease associations.

By ranking all of the 6158 drugs, we obtained a set of 31 candidate drugs for COVID-19 using the top 20 predictions from each model. We noted that many of the top candidates were antiviral drugs against SARS-CoV-2 [21, 22]. These include Acyclovir, Amantadine, and Ribavirin, which are flagged by both GCN and our best-performing GraphSAGE model. Further, anticoagulants like Heparin (GCN, GAT and HGT model) are known to prevent clotting in COVID-19 patients, thus reducing mortality rates [12]. Drugs belonging to the immunomodulatory family, like Interferon alfa-2b (GIN), may help control immune responses for patients affected by COVID-19 [13]. Our findings indicate a strong motivation for applying a focused strategy for the treatment of COVID-19 and the potential for immediate applications in clinical trials. The framework designed in this study can be extended to other emerging diseases. It has the potential to significantly shorten the timeline from drug discovery to clinical testing by identifying reliable candidate drugs for a specific disease.

V. Conclusion

In this study, we demonstrated the effectiveness of network-based approaches for drug repurposing, exploring three model implementations: node-agnostic (homogeneous graph-based), heterogeneous graph-based, and matrix factorization techniques. Homogeneous GNN models achieved the highest results, with AUC-ROC scores between

0.98 and 0.99 across all four models, while the heterogeneous graph transformer also performed well with an AUC-ROC of 0.985. Both graph-based methods outperformed the matrix factorization model (AUC-ROC: 0.82), highlighting the importance of incorporating relevant protein-protein interactions. For COVID-19, several drugs were identified as potential antivirals aligned with known treatments, further validating model performance. Among the 31 top candidates, many have recognized clinical use, reinforcing the practical applications of this approach. Our findings demonstrated that our models could successfully capture complex relationships between proteins, drugs, and diseases, and this framework can be extended to a wide range of diseases, thus enabling a more efficient and cost-effective drug development process.

[1] **Supplementary information**. Supplementary File: List of 31 candidate drugs, including DrugBank ID, name, therapeutic category, and prediction methods identifying each as a potential candidate.

References

[1] DiMasi, J.A., Grabowski, H.G., Hansen, R.W.: Innovation in the pharmaceutical industry: New estimates of R&D costs. Journal of Health Economics 47, 20–33 (2016)

[2] Pushpakom, S., Iorio, F., Eyers, P.A., Escott, K.J., al.: Drug repurposing: progress, challenges and recommendations. Nature Reviews Drug Discovery 18, 41–58 (2019)

[3] Mercorelli, B., Palu`, G., Loregian, A.: Drug repurposing for viral infectious diseases: How far are we? Trends in Microbiology 26(10), 865–876 (2018)

[4] Gordon, D.E., Jang, G.M., Bouhaddou, M., al.: A sars-cov-2 protein interaction map reveals targets for drug repurposing. Nature 583, 459–468 (2020)

[5] Veljkovic, V., Loiseau, P.M., Figadere, B., Glisic, S., Veljkovic, N., Perovic, V.R., Cavanaugh, D.P., Branch, D.R.: Virtual screen for repurposing approved and experimental drugs for candidate inhibitors of ebola virus infection. F1000Research 4, 34 (2015)

[6] Pantziarka, P., Verbaanderd, C., Huys, I., Bouche, G., Meheus, L.: Repurposing drugs in oncology: From candidate selection to clinical adoption. Seminars in Cancer Biology 68, 186–191 (2021)

[7] Cheng, F., Zhou, Y., Li, J., Skol, A.L., Zhao, J., al.: Prediction of drug–target interactions and drug repositioning via network-based inference. PLoS Computational Biology 12(5), 1004975 (2016)

[8] Fiscon, G., Conte, F., Farina, L., Paci, P.: A comparison of network-based methods for drug repurposing along with an application to human complex diseases. International Journal of Molecular Sciences 23(7), 3703 (2022)

[9] Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., Zhavoronkov, A.: Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. Molecular Pharmaceutics 13(7), 2524–2530 (2016)

[10] Amiri Souri, E., Chenoweth, A., Karagiannis, S.N., Tsoka, S.: Drug repurposing and prediction of multiple interaction types via graph embedding. BMC Bioinformatics 24(1) (2023)

[11] Nosengo, N.: Can you teach old drugs new tricks? Nature 534, 314–316 (2016)

[12] Tang, N., Bai, H., Chen, X., Gong, J., Li, D., Sun, Z.: Anticoagulant treatment is associated with decreased mortality in severe coronavirus disease 2019 patients with coagulopathy. Journal of Thrombosis and Haemostasis 18(5), 1094–1099 (2020)

[13] Wang, N., Shang, J., Jiang, S., Du, L.: Subunit vaccines against emerging pathogenic human coronaviruses. Frontiers in Microbiology 11, 298 (2020)

[14] Zitnik, M., Agrawal, M., Leskovec, J.: Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 34(13), 457–466 (2018)

[15] Kipf, T., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2017)

[16] Velickovic, P., Fedus, W., P., R.P., al.: Graph attention networks. Proceedings of the International Conference on Learning Representations (ICLR) (2018)

[17] Hamilton, W.L., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: Neural Information Processing Systems (NeurIPS) (2017)

[18] Xu, W.W., Zhu, J., Jegelka, S.: How powerful are graph neural networks? Proceedings of the International Conference on Learning Representations (ICLR) (2018)

[19] Huang, F., Qiu, Y., Li, Q., Liu, S., Ni, F.: Predicting drug-disease associations via multi-task learning based on collective matrix factorization. Frontiers in Bioengineering and Biotechnology 8 (2020)

[20] Kula, M.: Metadata embeddings for user and item cold-start recommendations. arXiv, ??? (2015)

[21] Dyall, J., Gross, R., Kindrachuk, J., Johnson, R.F., Olinger, G.G., Hensley, L.E., Frieman, M.B., Jahrling, P.B.: Middle east respiratory syndrome and severe acute respiratory syndrome: Current therapeutic options and potential targets for novel therapies. Drugs 77(18), 1935–1966 (2017)

[22] Tay, M.Z., Poh, C.W., Renia, L., al.: The trinity of covid-19: immunity, inflammation and intervention. Nature Reviews Immunology 20, 363–374 (2020) Science, 1989.