OrphaDRGL: A Novel Graph Deep Learning-based Drug Repositioning Approach for Orphan Diseases

Remington Kim
Bergen County Academies
New Jersey, USA

Mentor: Mr. Matthew Wang
Bergen County Academies
New Jersey, USA

OrphaDRGL: A Novel Graph Deep Learning-based Drug Repositioning Approach for Orphan Diseases

Remington Kim

Bergen County Academies, New Jersey, USA

Abstract. More than 95% of the approximately 7000 orphan diseases in the world do not have effective treatments, and since orphan diseases are conditions that affect a limited population, there is little incentive to develop novel treatments for them. Thus, drug repositioning, or finding new uses for existing drugs, has become a viable option as it is faster and more economical than traditional *de novo* drug discovery. To this end, OrphaDRGL was created: a novel computational orphan disease drug repositioning approach utilizing graph deep learning. OrphaDRGL uses open-source disease phenotype, drug side effect, drug chemical structure, and drug-indication data to form a heterogeneous network with drug and medical condition nodes. Edges were created using the Tanimoto coefficient of drug side effects and chemical structures, the Resnik phenotypic similarity of medical conditions and their phenotypes, and existing drug-indication pairs. Morgan fingerprint bit vectors were assigned as explicit node features for drug nodes. A link concealing algorithm was then applied to emulate orphan disease conditions, and a graph convolutional neural network-based link prediction framework was trained on the network in order to identify potential drug repositioning candidates for orphan diseases. After 10-fold cross-validation, OrphaDRGL achieved an average AUC-ROC score of 0.953 and was able to identify both literature-supported and previously unreported drug repositioning candidates for three different orphan diseases. Furthermore, ablation analysis conducted on OrphaDRGL demonstrated that its design was computationally suited for orphan disease drug repositioning. OrphaDRGL is the first of its kind in the scientific literature, and its promising performance helps address the pressing issue of *in silico* identifications of potential drug repositioning candidates for orphan diseases.

Keywords: Graph Neural Networks · Heterogeneous Networks · Drug Repositioning · Orphan Diseases

Table of Contents

1 Introduction

1.1 Orphan Diseases

According to the U.S. Food and Drug Administration, orphan diseases are rare conditions that affect fewer than 200,000 people nationwide [5]. Even though individual orphan diseases are rare, approximately 30 million people suffer from them collectively in the U.S. alone [10]. Still, more than 95% of the approximately 7000 diseases classified as "orphan diseases" in the world do not have effective treatments [10]. One reason for the lack of effective treatments may be that pharmaceutical companies are not incentivized to develop novel treatments for orphan diseases. According to Nosengo [20], it takes 13-15 years and $2-3 billion U.S. to get a *de novo* drug approved so that it can be marketed. This market is relatively small for drugs developed specifically to treat orphan diseases, so pharmaceutical companies have less incentive to devote time and money to those endeavors. Although the FDA Orphan Drug Act in 1983 led to approximately 325 new treatments for orphan diseases, there is still a very pressing need for drugs and treatments for orphan diseases since many are lethal and few have approved treatments [26,10].

1.2 Drug Repositioning

Drug repositioning is an alternative to *de novo* drug discovery for developing potential therapies for orphan diseases [26]. Drug repositioning (also called drug repurposing) is the act of taking existing drugs and finding new uses for them that are outside of its original approved medical indication [2]. The benefits to this strategy include an expedited time to market and potentially less development costs [28]. This is because it is likely that a drug used in a repositioning scenario has already been shown to be safe in humans, so there is a major decrease in costs for the preclinical, phase I, and phase II trials [23]. Also, existing manufacturing and testing infrastructure for a previously approved drug can potentially save a large amount of money and resources. The drug may even skip the phase I trials entirely due to the existence of data from previous safety studies, though parameters such as the dosage and route of administration may differ for the repositioned indication [21]. Additionally, phase IV studies assessing side effects and safety have likely been conducted for potential repositioning candidate drugs on the market [28].

There have been several cases where drug repositioning has been successful in treating both orphan diseases and non-orphan diseases. For example, Rituximab was originally approved to treat several cancers but was then repositioned as a treatment for rheumatoid arthritis [23]; Atomoxetine was originally developed to treat major depression but was then repositioned as an FDA approved treatment for attention deficit hyperactivity disorder (ADHD) [15]; and Tretinoin, a treatment for acne vulgaris was repositioned as an FDA approved treatment for the orphan disease acute promyelocytic leukemia [26].

These potential expense and time-reducing factors as well as the success stories make drug repositioning a favorable option for developing treatments for orphan diseases.

1.3 Computational Network-based Drug Repositioning

In the past, drug repositioning has been a mostly fortuitous process [23], but now, computational drug repositioning has become a viable option. One specific computational approach to drug repositioning is a network-based one. This is where the problem of finding drug repositioning candidates is modelled as a network in which the nodes are drugs and medical conditions, and the edges are the various chemical and biological relation-

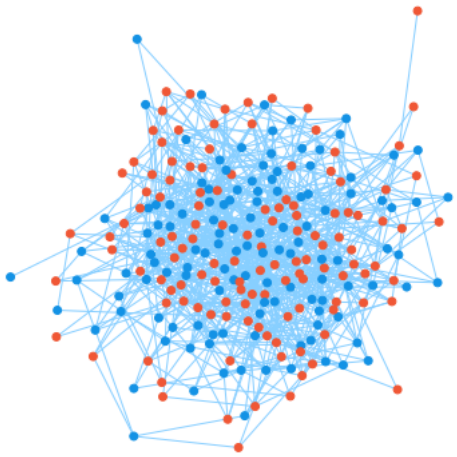


Fig. 1: **Visualization of a Network.**

ships that drugs and medical conditions possess. Deep learning can then be applied to these networks to detect indirect and unapparent patterns and relationships and identify potential drug repositioning candidates.

2 Related Works

2.1 Prior Literature

Several variations of deep learning network-based approaches for drug repositioning, as well as a similar but not identical problem of drug-target interaction prediction, are reported in the scientific literature.

Manoochehri and Nourani [17] formed a drug-target network using interactions, drug similarity, and target similarity and trained a multilayer perceptron (MLP) on truncated adjacency matrices from extracted subgraphs for predicting drug-target interactions. Zhao *et al.* [36] used a graph convolutional neural network (GCN) and an MLP to perform node classification on a graph whose nodes were drug-protein target pairs. Wan *et al.* [31] used a graph neural network and matrix factorization on a network with drug-protein, drug-drug, and protein-protein interactions; drug-disease and protein-disease associations; and drug-drug and protein-protein similarity to predict drug-target interactions. Zeng *et al.* [34] used deep learning to featurize nodes in 15 chemical, genomic, phenotypic, and cellular networks and then utilized matrix factorization to predict new drug-target interactions. Zong *et al.* [37] used deep neural networks to calculate similarities in a drug-target network to predict new drug-target interactions. Ioannidis et al. [9] developed a novel graph neural network model that is suited for predicting rare links to reposition drugs for Covid-19. Lastly, Zeng et al. [1] used one drug-disease, one drug-side-effect, one drug-target, and seven drug-drug networks along with a variational autoencoder to predict new drug-disease associations, or new drug repositioning candidates.

2.2 Commentary and Shortcomings

In the prior literature, there are two key areas which are not addressed.

Firstly, it does not appear that any of the previously reported approaches integrating deep learning and networks/graphs for drug repositioning utilized graph neural networks for subgraph extraction-based link prediction, which is implemented in the SEAL framework [35]. This framework involves using subgraphs, or smaller networks taken from a larger network, to train a GCN for link prediction. It has been shown to have state-of-the-art performance on networks such as airline and academic collaboration networks [35].

In addition, there have been no reported deep learning-based approaches using graph neural networks specifically for orphan disease drug repositioning, which is an inherently different computational problem due to the lack of existing, approved treatments or drugs for orphan diseases. However, there have been non-deep learning drug repositioning approaches such as virtual screening for orphan diseases [7]. Manoochehri and Nourani [17] and Zong et al. [37] did not explicitly design their drug repositioning approaches for diseases or targets missing one key feature: a known drug. They did not remove a disease's known drug links when training to predict if another drug could also act as a possible treatment for it and only found new drug repositioning candidates or new drug-target interactions within a network of known drug-disease associations or drug-target interactions. These approaches are not optimized or currently suitable for orphan diseases because more than 95% of these diseases do not have treatments, meaning there would be no drug-disease associations to use as links [10]. Ioannidis et al. [9] built their model to predict rare link types in the entire network but did not design their model specifically for predicting links from nodes with *zero* edges of certain type, which would be the case with many orphan disease nodes and drug-medical condition edges due to orphan diseases' lack of existing treatments. Zhao et al. [36] found new drug-target interactions for targets that do not have any known drugs, but they did not explicitly use orphan diseases or diseases in general, and they did not use any form of link prediction. Wan et al. [31] evaluated their approach on "unique" drug target interactions – interactions in which the drug or the target has no other interactions – but did not design their approach specifically for targets or diseases with no associations. Zeng et al. [34] evaluated their approach on a subset of targets with less than five known drugs but did not explicitly evaluate diseases that have no known drugs. Finally, Zeng et al. [1], Zeng et al. [34], Manoochehri and Nourani [17], and Zong et al. [37] all did not use graph neural networks.

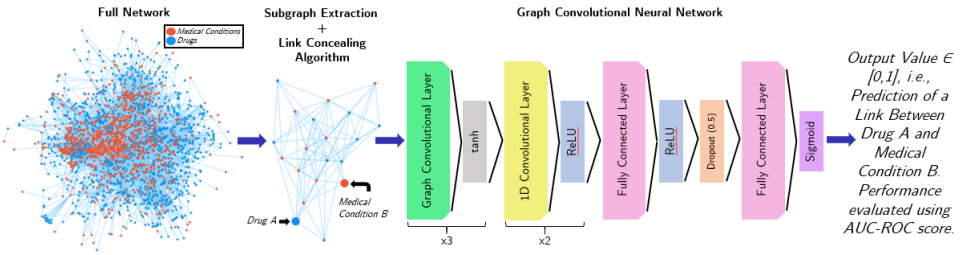


Fig. 2: **High-level Overview of OrphaDRGL's Pipeline.**

3 OrphaDRGL

OrphaDRGL (**Orpha**n disease **D**rug **R**epositioning using **G**raph Deep **L**earning) is a novel, deep learning network-based drug repositioning approach specifically designed for orphan diseases. It uses drug side effect data, which has been shown to be a valid feature in computational drug repositioning [33], drug chemical structure data, disease phenotype data, and existing drug-indication data from four different open-source databases to form a heterogeneous network of drugs and medical conditions. The four different types of edges within the network are drug-medical condition edges, drug-drug side effect similarity edges, drug-drug chemical structure similarity edges, and medical condition-medical condition edges. Subgraphs are extracted from this network and double-radius node labeling is applied as per SEAL [35]. A link concealing is also applied to these subgraphs to emulate orphan disease conditions. Finally, a three-layer GCN is trained on these subgraphs to predict links between drug and medical condition nodes. In other words, it predicts if certain drugs are viable drug repositioning candidates for a given medical condition. A high-level overview of this entire process is displayed in Figure 2.

Notably, OrphaDRGL makes two potentially novel contributions:

- OrphaDRGL is the first of its kind in the scientific literature to use graph neural networks to perform subgraph extraction-based link prediction (SEAL) with explicit node features, drug side effects, and disease phenotypes for drug repositioning.
- OrphaDRGL is the first application of graph neural networks in a drug repositioning approach specifically designed for orphan diseases.

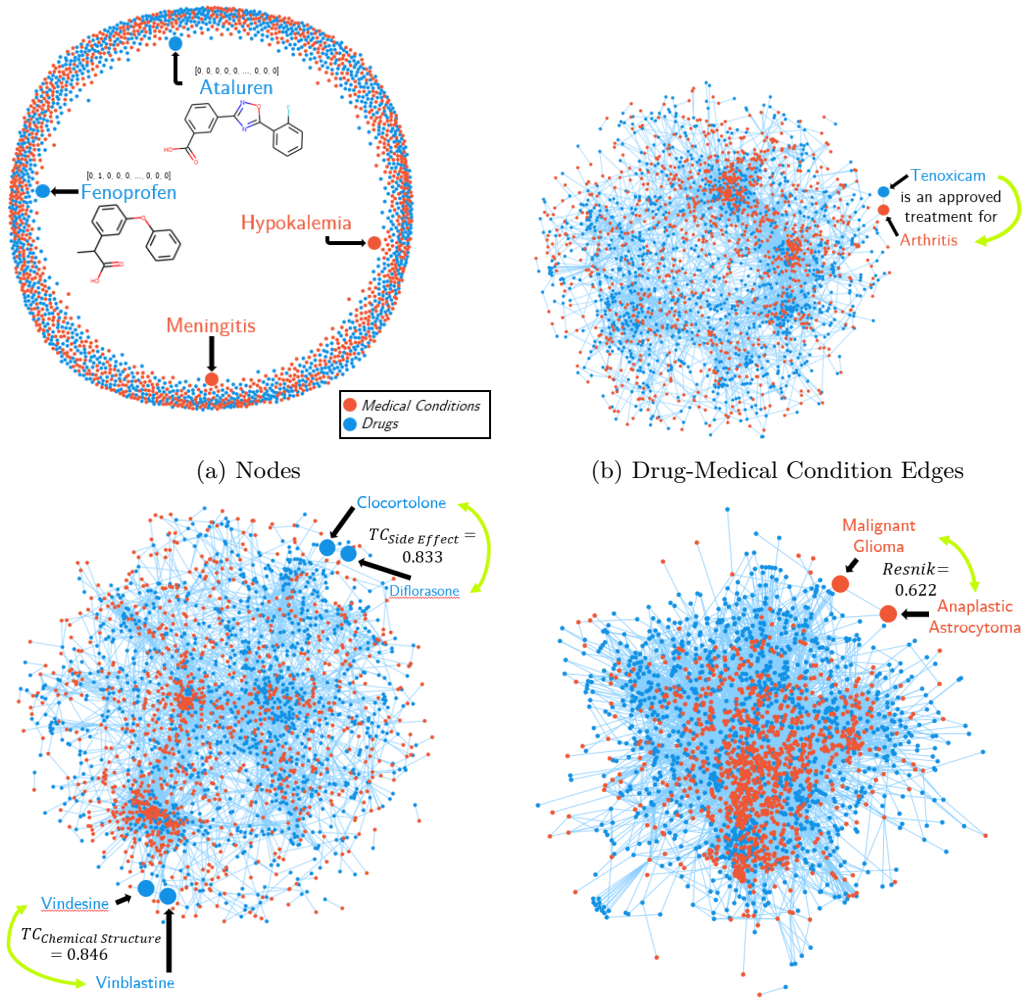
4 Procedure

4.1 Network Formation

As shown in Figure 2, the first step in OrphaDRGL's pipeline is to construct a heterogeneous network of drug and medical conditions nodes with drug-medical condition edges, drug-drug side effect similarity edges, drug-drug chemical structure similarity edges, and medical condition-medical condition edges.

Nodes Every node in OrphaDRGL's network is either a drug or a medical condition taken from the repoDB [4]. All small molecule compound drugs from repoDB [4] along with any associated medical conditions were extracted from the database, and each one became a node. Additionally, each drug node was assigned its 1024-bit Morgan Fingerprint vector, which is a vector-based representation of a compound's molecular

features [25], as its explicit node feature. For medical condition nodes, a 1024-bit filler vector of 0s was assigned as the explicit node feature. There are 1199 medical condition nodes and 1013 drug nodes, and Figure 3a depicts them OrphaDRGL's network.



(a) Nodes

(b) Drug-Medical Condition Edges

(c) Drug-Medical Condition and Drug-Drug Edges

(d) Drug-Medical Condition, Drug-Drug, and Medical Condition-Medical Condition Edges

Fig. 3: **OrphaDRGL Network Formation**

Drug-Medical Condition Edges Drug-medical condition edges in OrphaDRGL's network connect medical conditions to drugs that are approved treatments for it. Approved drug-indication pairs were taken from the repoDB database [4] and used to construct

these edges. There are 5686 drug-medical condition edges, and Figure 3b depicts them in OrphaDRGL's network.

Drug-Drug Edges Drug-drug edges in OrphaDRGL's network connect drug nodes to other drug nodes based on chemical structure similarity and side effect similarity.

Chemical structures of the drugs in the network were extracted from the DrugBank database [32], and the Morgan Fingerprint bit vector of each drug was constructed using the extracted chemical structures. The chemical structure similarity of two drugs was determined by the Tanimoto Coefficient (1), where m_a is the Morgan Fingerprint for drug a. Two drugs with a Tanimoto Coefficient of their Morgan Fingerprints ≥ 0.45 were considered similar, and a drug-drug edge was constructed between them. A similarity threshold of 0.45 was chosen because experimentation by Maggiora *et al.* [16] demonstrated that a 0.45 threshold for Tanimoto Coefficents of Morgan Fingerprints had a significance level of $p <\sim 10^{-4}$.

Then, side effects of the drugs in the network were extracted from the SIDER database [14], and the side effect similarity of two drugs was also determined by the Tanimoto Coefficient (1). However, m_a was replaced by s_a, which is the one-hot encoded side effect vector for drug a. The same threshold as the chemical structure similarity was used.

$$TC(m_1, m_2) = \frac{|m_1 \cap m_2|}{|m_1| + |m_2| - |m_1 \cap m_2|} \tag{1}$$

There are 1221 drug-drug edges, and Figure 3c depicts them in OrphaDRGL's network.

Medical Condition-Medical Condition Edges Medical condition-medical condition edges in OrphaDRGL's network connect medical condition nodes to other medical conditions nodes based on phenotypic similarity.

The phenotypes of medical conditions in the network were first extracted from the Human Phenotype Ontology (HPO) database [13], and the Resnik Information Content Similarity Score (RICSS) (2) [24] on the Human Phenotype Ontology was used to determine the phenotypic similarity of two medical conditions. The Information Content (IC) score of a term, t, represents how specific it is. Thus, terms with lower probabilities of occurring have higher IC scores. The RICSS of two terms is the IC score of their Most Informative Common Ancestor (MICA). For OrphaDRGL, all RICSSs were calculated using the PhenoSimWeb application [22].

$$\begin{aligned} IC(t) &= -log(p(t)) \\ Resnik(t_1, t_2) &= IC(t_{MICA}) \end{aligned} \tag{2}$$

The similarity threshold for the RICSS of two medical conditions classified by HPO as diseases was 0.5, and the similarity threshold for the RICSS of two medical conditions classified by HPO as phenotypes was 0.25. A medical condition-medical condition edge was constructed between any pair of medical condition nodes that met these similarity thresholds, and all diseases were connected to their phenotypes.

There are 7986 medical condition-medical condition edges, and Figure 3d depicts them in OrphaDRGL's network. The summary statistics of this complete network containing all the nodes and edges are given in Table 1.

Table 1: **Summary Statistics of OrphaDRGL's Drug and Medical Condition Network**

2212	Nodes
14893	Edges
13.47	Average Degree
0.33	Average Clustering Coefficient

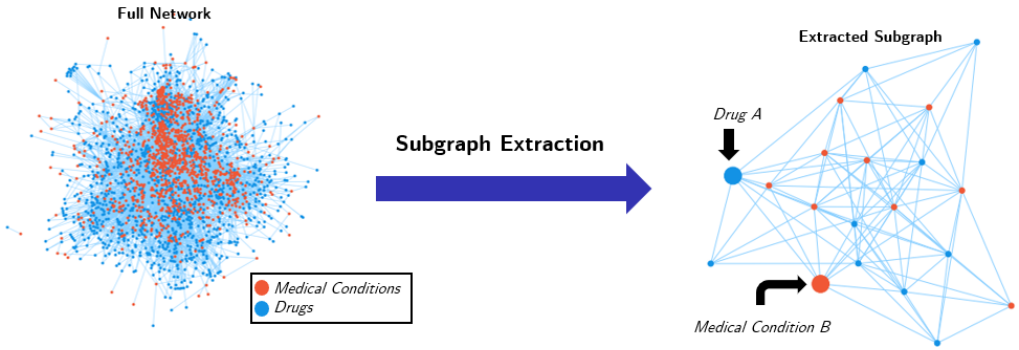


Fig. 4: **A Visualization of SEAL's Subgraph Extraction on OrphaDRGL's Network.** The link being predicted in the subgraph is between Drug A and Medical Condition B. A two-hop subgraph is depicted for visualization purposes.

4.2 Subgraph Extraction for the SEAL Framework

After OrphaDRGL's network was formed, the SEAL framework [35] was applied to it.

SEAL The SEAL framework performs link prediction by extracting k-hop subgraphs surrounding two target nodes, x and y, from a larger network and applying a double-radius node labeling to the subgraphs for marking nodes' roles and retaining structural information; these subgraphs are then used to train a GCN to predict the existence of an edge between x and y [35].

OrphaDRGL's Subgraphs When applying SEAL to OrphaDRGL, the target nodes, x and y were always a medical condition node and drug node, respectively, as the edges OrphaDRGL is trying to predict are drug-medical condition edges. Also, the k parameter in k-hop subgraph extraction was set to 1. A visualization of this process is illustrated in Figure 4.

In total, 11372 subgraphs were extracted from OrphaDRGL's network to act as training and evaluation data. There was a perfect balance between positive and negative samples (5686 each), and available true negative drug-medical condition pairs from the repoDB database, such as those involved in failed clinical trials, were used. The remaining negative samples, however, were randomly sampled from the network.

4.3 Link Concealing Algorithm

When left as is, the medical conditions in OrphaDRGL's network have an average of 4.7 (Std. Dev. of 9.1) and at least one drug-medical condition edges associated with them.

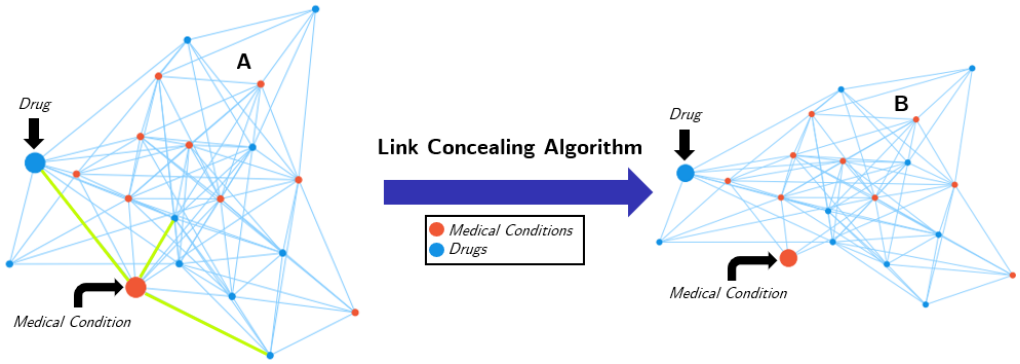


Fig. 5: **A Visualization of the Link Concealing Algorithm Employed on OrphaDRGL's Subgraphs.** The highlighted drug-medical condition edges in subgraph A are the ones that will be concealed to emulate orphan disease conditions. In subgraph B, the edges highlighted in subgraph A are concealed and no longer part of the subgraph. Subgraph B is the final subgraph that will be used to train/evaluate the graph convolutional neural network.

Certain medical conditions have a plethora of drug-medical conditions associated with them. For example, non-insulin-dependent (type II) diabetes mellitus has 23. On the other hand, most orphan diseases will have zero drug-medical condition edges associated with them because they do not have existing, approved treatments [26]. Ergo, there is a significant difference between the distributions of drug-medical condition edges for medical conditions in OrphaDRGL's network and OrphaDRGL's intended application: orphan diseases ($p < 0.0001$; unpaired two-tailed t-test). This is suboptimal because the feature distributions of training, evaluation, and application data being different can cause link prediction algorithms to be futile [8].

To this end, a link concealing algorithm was applied to the subgraphs extracted from OrphaDRGL's network. The Link Concealing Algorithm hides all drug-medical condition edges stemming from the subgraph's target medical condition in order to emulate orphan disease conditions in the training and evaluation data so that OrphaDRGL can be applied to orphan diseases in the real world.

Pseudocode for this algorithm is given in Algorithm 1, and a visualization is depicted Figure 5.

Algorithm 1: OrphaDRGL's Link Concealing Algorithm

Data: S is the adjacency matrix for subgraph with target medical condition node x and $\Gamma(x)$ is the set of all neighbors of x.

begin
  for $i \in \Gamma(x)$ **do**
    if *type(i) = "drug"* **then**
      $S_{xi} = S_{ix} = 0$

4.4 Deep Learning Model

Finally, after network formation, subgraph extraction, and the application of the link concealing algorithm, 10-fold cross validation training and evaluation was performed on a GCN, which is the deep learning model that OrphaDRGL utilizes to predict links between drug and medical condition nodes.

Graph Convolutional Neural Networks A graph convolutional neural network (GCN) is a specialized neural network that contains graph convolutional layers. These layers perform node feature propagations in which each node's feature becomes the linear transformation of the normalized aggregation of its neighbors' features and its own original feature [12]. This process is visually depicted in Figure 6 and represented by (3) [12].

$$Z = \widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2}XW \tag{3}$$

Here, Z is the output of the graph convolutional layer, X is the input, A is the adjacency matrix for the inputted graph and $\widetilde{A} = A + I$, $\widetilde{D}$ is a diagonal degree matrix where $\widetilde{D}_{ii} = \Sigma_j \widetilde{A}_{ij}$, and W is the weight matrix of trainable parameters. $\widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2}X$ performs the node feature propagation and normalization, and multiplying by W performs the linear transformation [12].

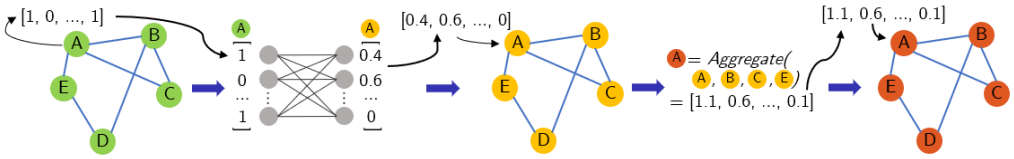


Fig. 6: **A Visualization of a Graph Convolutional Layer Operation.**

10-fold Cross Validation 10-fold cross validation is the process of training a model 10 different times on 10 different training sets containing 90% of the data and evaluating on 10 different testing sets containing 10% of the data and averaging the performance metrics. An advantage of 10-fold cross validation over a single train-test split is that every single sample is part of the testing set once. Thus, 10-fold cross validation gives a better representation of the model's true generalization abilities and performance.

For OrphaDRGL, there were 10234 or 10236 subgraphs, which composed 90% of the samples, in each iteration's training set. These subgraphs were used to train a three-layer GCN with the architecture depicted in Figure 2 for 50 epochs. The learning rate was set to 0.001, the batch size was 16, and the Adam optimizer was used along with a Binary Cross Entropy loss function (4). The remaining 1138 or 1136 subgraphs, which composed 10% of the samples, were in the testing set for evaluation.

$$L = y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y}) \tag{4}$$

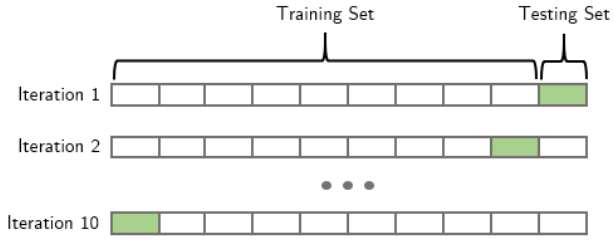


Fig. 7: **The Visualization of OrphaDRGL's 10-fold Cross-validation.**

5 Findings

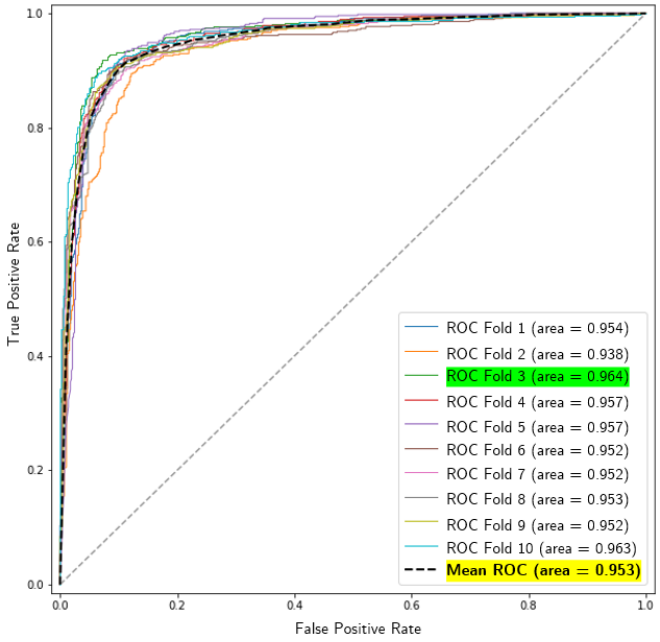


Fig. 8: **The Area Under the Curve Receiver Operating Characteristics (AUC-ROC) for OrphaDRGL's 10-fold cross-validation.** The average AUC-ROC score and the highest AUC-ROC fold are highlighted

5.1 Performance Results

The performance of OrphaDRGL was evaluated using the Area Under the Curve Receiver Operating Characteristics (AUC-ROC) score, which ranges from 0 to 1. Figure 8 displays all 11 ROC curves: one curve for each iteration in the 10-fold cross validation and the average curve.

OrphaDRGL achieved an average AUC-ROC score of 0.953. The best performing model achieved an AUC-ROC score of 0.964 and had an optimal classification threshold of 0.481. These scores demonstrate that OrphaDRGL possesses good generalization abilities.

5.2 Novel and Literature-supported Drug Repositioning Candidate Identifications

Following 10-fold cross validation training and evaluation, OrphaDRGL was applied to three different orphan diseases in order to identify potential drug repostioning candidates. The orphan diseases were Steinert's Disease, a myotonic dystrophy characterized by skeletal muscle weakness and myotonia [27]; Menkes Disease, a copper metabolism disorder [30]; and Agammaglobulinemia, a immunodeficiency disorder [18]. Across these three orphan diseases, OrphaDRGL was able to identify both literature-supported and novel, previously unreported drug repositioning candidates in its top 20 positive link predictions, which are displayed in Table 2.

The literature-supported OrphaDRGL-identified drug repositioning candidates are Testosterone and Prednisone for Steinert's Disease, which are supported by Kingston & Moxley [11] and Trip *et al.* [29], respectively; Penicillamine for Menkes Disease, which is supported by Nadal & Baerlocher [19]; and Ciprofloxacin and Benzylpenicillin for Agammaglobulinemia, which are supported by Ganier [6] and Autenrieth *et al.* [3], respectively. The remaining drugs are novel, previously unreported potential drug repositioning candidate identifications made by OrphaDRGL.

Table 2: **The Top 20 Positive Links or Identified Drugs by OrphaDRGL for Steinert's Disease, Menkes Disease, and Agammaglobulinemia.** The therapeutic effect of highlighted drugs have been supported or suggested in the scientific literature.

Rank	Steinert's Disease	Menkes Disease	Agammaglobulinemia
1	Testosterone	Chlorphenesin	Ibandronate
2	Ataluren	Zoledronic Acid	Lactic Acid
3	Lomustine	Anisotropine Methylbromide	Metaxalone
4	Cholic Acid	Ataluren	Edetic Acid
5	Prednisone	Metaxalone	Pyridostigmine
6	Testosterone Propionate	Ibandronate	Etidronic Acid
7	Amiodarone	Dexpanthenol	Gramicidin D
8	Zinc Sulfate	Prednisone	Ciprofloxacin
9	Carmustine	Oxacillin	Neostigmine
10	Bicalutamide	Baclofen	Lumiracoxib
11	Fluoxymesterone	Lactic Acid	Sodium Ferric Gluconate Complex
12	Ibandronate	Riluzole	Levofloxacin
13	Sucralfate	Dimethyl Fumarate	Zoledronic Acid
14	Dipyridamole	Fosphenytoin	Moclobemide
15	Pyrimethamine	Clobazam	Vortioxetine
16	Tinidazole	Dantrolene	Mannitol
17	Bretylium	Methoxsalen	Dipyridamole
18	Chlorpromazine	Diazepam	Mitoxantrone
19	Temozolomide	Penicillamine	Cromoglicic Acid
20	Duloxetine	Phylloquinone	Benzylpenicillin

5.3 Ablation Analysis

An ablation analysis on OrphaDRGL's Link Concealing Algorithm was performed. When the Link Concealing Algorithm was not applied to the same training samples in 10-fold cross validation, the average AUC-ROC score was 0.778, and the highest AUC-ROC score was 0.846. Indeed, the usage of OrphaDRGL's Link Concealing Algorithm led to a statistically significant performance increase in drug repositioning in orphan disease conditions ($p < 10^{-8}$; unpaired two-tailed t-test). This demonstrates that OrphaDRGL is computationally designed for and applicable to orphan disease drug repositioning, something that the prior literature does not do.

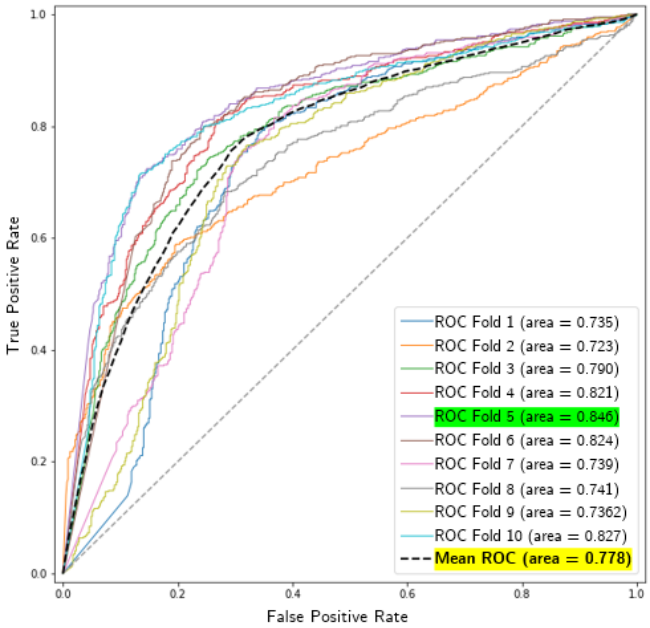


Fig. 9: **The Area Under the Curve Receiver Operating Characteristics (AUC-ROC) Without OrphaDRGL's Link Concealing.** The average AUC-ROC score and the highest AUC-ROC fold are highlighted.

6 Conclusions

6.1 Discussion

OrphaDRGL is a novel deep learning network-based drug repositioning approach for orphan diseases that uses drug side effect data, drug chemical structure data, disease phenotype data, and existing drug-indication data from four different open-source databases to form a drug and medical condition network. Then, OrphaDRGL employs subgraph extraction-based link prediction (SEAL) [35], a link concealing algorithm, and a GCN on its network in order to predict drug repositioning candidates.

OrphaDRGL is the first reported deep learning network-based drug repositioning approach in the scientific literature that utilizes a GCN for subgraph extraction-based link prediction (SEAL) and is specifically designed for orphan diseases. It achieved a high AUC-ROC score when trained and evaluated over 10-fold cross validation. Moreover, it was able to identify both literature-supported as well as novel, previously unreported drug repositioning candidates for three different orphan diseases. This is promising because OrphaDRGL, an *in silico* model, was able to correspond with certain clinical studies that have been carried out.

Ultimately, OrphaDRGL greatly improves upon the fortuitous nature of drug repositioning by providing a comprehensive deep learning-based approach by which potential treatments for orphan diseases can be identified.

6.2 Future Work and Limitations

Currently, OrphaDRGL's network is limited to small molecule drugs. This is because only small molecule compounds have Morgan Fingerprints to use as explicit node features and calculate chemical structure similarity with. Therefore, OrphaDRGL is limited to identifying small molecule drug repositioning candidates and cannot reposition other types of drugs such as biologics. Thus, future work involves developing an equivalent representation for all drugs so that both small molecule drugs and biologics can be incorporated into OrphaDRGL's network, which would broaden the scope and increase the variety of drugs that OrphaDRGL can identify as drug repositioning candidates.

References

1. **35**(24) (May 2019). https://doi.org/10.1093/bioinformatics/btz418, https://doi.org/10.1093/bioinformatics/btz418
2. Ashburn, T.T., Thor, K.B.: Drug repositioning: identifying and developing new uses for existing drugs. Nature Reviews Drug Discovery **3**(8), 673–683 (Aug 2004). https://doi.org/10.1038/nrd1468, https://doi.org/10.1038/nrd1468
3. Autenrieth, I.B., Schuster, V., Ewald, J., Harmsen, D., Kreth, H.W.: An unusual case of refractory campylobacter jejuni infection in a patient with x-linked agammaglobulinemia: Successful combined therapy with maternal plasma and ciproftoxacin. Clinical Infectious Diseases **23**(3), 526–531 (Sep 1996). https://doi.org/10.1093/clinids/23.2.526, https://doi.org/10.1093/clinids/23.2.526
4. Brown, A.S., Patel, C.J.: A standard database for drug repositioning. Scientific Data **4**(1) (Mar 2017). https://doi.org/10.1038/sdata.2017.29, https://doi.org/10.1038/sdata.2017.29
5. Center for Drug Evaluation and Research: Orphan Products: Hope for People With Rare Diseases. FDA (Feb nd), https://www.fda.gov/drugs/information-consumers-and-patients-drugs/orphan-products-hope-people-rare-diseases
6. Ganier, M.: Infantile agammaglobulinemia and immediate hypersensitivity to penicillin g. JAMA: The Journal of the American Medical Association **237**(17), 1852 (Apr 1977). https://doi.org/10.1001/jama.1977.03270440042019, https://doi.org/10.1001/jama.1977.03270440042019
7. Govindaraj, R.G., Naderi, M., Singha, M., Lemoine, J., Brylinski, M.: Large-scale computational drug repositioning to find treatments for rare diseases. npj Systems Biology and Applications **4**(1) (Mar 2018). https://doi.org/10.1038/s41540-018-0050-7, https://doi.org/10.1038/s41540-018-0050-7
8. Hasan, M.A., Chaoji, V., Salem, S., Zaki, M.: Link prediction using supervised learning. In: In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security (2006)
9. Ioannidis, V.N., Zheng, D., Karypis, G.: Few-shot link prediction via graph neural networks for covid-19 drug-repurposing (2020)
10. Kaufmann, P., Pariser, A.R., Austin, C.: From scientific discovery to treatments for rare diseases – the view from the national center for advancing translational sciences – office of rare diseases research. Orphanet Journal of Rare Diseases **13**(1) (Nov 2018). https://doi.org/10.1186/s13023-018-0936-x, https://doi.org/10.1186/s13023-018-0936-x
11. Kingston, W.J., Moxley, R.T.: Treatment of muscular dystrophies and inflammatory myopathies. Clinical Neuropharmacology **9**(4), 361–372 (Aug 1986). https://doi.org/10.1097/00002826-198608000-00003, https://doi.org/10.1097/00002826-198608000-00003
12. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks (2017)
13. Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.P., Gargano, M., Harris, N.L., Matentzoglu, N., McMurry, J.A., Osumi-Sutherland, D., Cipriani, V., Balhoff, J.P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., Segal, M., Jansen, A.C., Muaz, A., Chang, W.H., Bergerson, J., Lauledkind, S.J.F., Yüksel, Z., Beltran, S., Freeman, A.F., Sergouniotis, P.I., Durkin, D., Storm, A.L., Hanauer, M., Brudno, M., Bello, S.M., Sincan, M., Rageth, K., Wheeler, M.T., Oegema, R., Lourghi, H., Rocca, M.G.D., Thompson, R., Castellanos, F., Priest, J., Cunningham-Rundles, C., Hegde, A., Lovering, R.C., Hajek, C., Olry, A., Notarangelo, L., Similuk, M., Zhang, X.A., Gómez-Andrés, D., Lochmüller, H., Dollfus, H., Rosenzweig, S., Marwaha, S., Rath, A., Sullivan, K., Smith, C., Milner, J.D., Leroux, D., Boerkoel, C.F., Klion, A., Carter,

M.C., Groza, T., Smedley, D., Haendel, M.A., Mungall, C., Robinson, P.N.: Expansion of the human phenotype ontology (HPO) knowledge base and resources. Nucleic Acids Research **47**(D1), D1018–D1027 (Nov 2018). https://doi.org/10.1093/nar/gky1105, https://doi.org/10.1093/nar/gky1105
14. Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The SIDER database of drugs and side effects. Nucleic Acids Research **44**(D1), D1075–D1079 (Oct 2015). https://doi.org/10.1093/nar/gkv1075, https://doi.org/10.1093/nar/gkv1075
15. Ledbetter, M.: Atomoxetine: a novel treatment for child and adult ADHD. Neuropsychiatric Disease and Treatment **2**(4), 455–466 (Dec 2006). https://doi.org/10.2147/nedt.2006.2.4.455, https://doi.org/10.2147/nedt.2006.2.4.455
16. Maggiora, G., Vogt, M., Stumpfe, D., Bajorath, J.: Molecular similarity in medicinal chemistry. Journal of Medicinal Chemistry **57**(8), 3186–3204 (Nov 2013). https://doi.org/10.1021/jm401411z, https://doi.org/10.1021/jm401411z
17. Manoochehri, H.E., Nourani, M.: Drug-target interaction prediction using semi-bipartite graph model and deep learning. BMC Bioinformatics **21**(S4) (Jul 2020). https://doi.org/10.1186/s12859-020-3518-6, https://doi.org/10.1186/s12859-020-3518-6
18. Mazhar, M., Waseem, M.: Agammaglobulinemia. In: StatPearls. StatPearls Publishing, Treasure Island (FL) (2021), http://www.ncbi.nlm.nih.gov/books/NBK555941/
19. Nadal, D., Baerlocher, K.: Menkes' disease: long-term treatment with copper and d-penicillamine. European Journal of Pediatrics **147**(6), 621–625 (Aug 1988). https://doi.org/10.1007/bf00442477, https://doi.org/10.1007/bf00442477
20. Nosengo, N.: Can you teach old drugs new tricks? Nature **534**(7607), 314–316 (Jun 2016). https://doi.org/10.1038/534314a, https://doi.org/10.1038/534314a
21. Park, K.: A review of computational drug repurposing. Translational and Clinical Pharmacology **27**(2), 59 (2019). https://doi.org/10.12793/tcp.2019.27.2.59, https://doi.org/10.12793/tcp.2019.27.2.59
22. Peng, J., Xue, H., Hui, W., Lu, J., Chen, B., Jiang, Q., Shang, X., Wang, Y.: An online tool for measuring and visualizing phenotype similarities using HPO. BMC Genomics **19**(S6) (Aug 2018). https://doi.org/10.1186/s12864-018-4927-z, https://doi.org/10.1186/s12864-018-4927-z
23. Pushpakom, S., Iorio, F., Eyers, P.A., Escott, K.J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., Norris, A., Sanseau, P., Cavalla, D., Pirmohamed, M.: Drug repurposing: progress, challenges and recommendations. Nature Reviews Drug Discovery **18**(1), 41–58 (Oct 2018). https://doi.org/10.1038/nrd.2018.168, https://doi.org/10.1038/nrd.2018.168
24. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1. p. 448–453. IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1995)
25. Rogers, D., Hahn, M.: Extended-connectivity fingerprints. Journal of Chemical Information and Modeling **50**(5), 742–754 (Apr 2010). https://doi.org/10.1021/ci100050t, https://doi.org/10.1021/ci100050t
26. Sardana, D., Zhu, C., Zhang, M., Gudivada, R.C., Yang, L., Jegga, A.G.: Drug repositioning for orphan diseases. Briefings in Bioinformatics **12**(4), 346–356 (Apr 2011). https://doi.org/10.1093/bib/bbr021, https://doi.org/10.1093/bib/bbr021
27. Shimizu, T., Nozaki, H., Tokuda, Y.: Steinert's disease. Case Reports **2013**(nov11 1), bcr2013201846–bcr2013201846 (Nov 2013). https://doi.org/10.1136/bcr-2013-201846, https://doi.org/10.1136/bcr-2013-201846
28. Tobinick, E.: The value of drug repositioning in the current pharmaceutical market. Drug News & Perspectives **22**(2), 119 (2009). https://doi.org/10.1358/dnp.2009.22.2.1303818, https://doi.org/10.1358/dnp.2009.22.2.1303818

29. Trip, J., Drost, G.G., van Engelen, B.G., Faber, C.G.: Drug treatment for myotonia. Cochrane Database of Systematic Reviews (Jan 2006). https://doi.org/10.1002/14651858.cd004762.pub2, https://doi.org/10.1002/14651858.cd004762.pub2
30. Tümer, Z., Møller, L.B.: Menkes disease. European Journal of Human Genetics **18**(5), 511–518 (Nov 2009). https://doi.org/10.1038/ejhg.2009.187, https://doi.org/10.1038/ejhg.2009.187
31. Wan, F., Hong, L., Xiao, A., Jiang, T., Zeng, J.: NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. Bioinformatics **35**(1), 104–111 (Jul 2018). https://doi.org/10.1093/bioinformatics/bty543, https://doi.org/10.1093/bioinformatics/bty543
32. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., 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 Research **46**(D1), D1074–D1082 (Nov 2017). https://doi.org/10.1093/nar/gkx1037, https://doi.org/10.1093/nar/gkx1037
33. Yang, L., Agarwal, P.: Systematic drug repositioning based on clinical side-effects. PLoS ONE **6**(12), e28025 (Dec 2011). https://doi.org/10.1371/journal.pone.0028025, https://doi.org/10.1371/journal.pone.0028025
34. Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J., Zhou, Y., Fang, J., Huang, Y., Guo, H., Li, L., Trapp, B.D., Nussinov, R., Eng, C., Loscalzo, J., Cheng, F.: Target identification among known drugs by deep learning from heterogeneous networks. Chemical Science **11**(7), 1775–1797 (2020). https://doi.org/10.1039/c9sc04336e, https://doi.org/10.1039/c9sc04336e
35. Zhang, M., Chen, Y.: Link prediction based on graph neural networks (2018)
36. Zhao, T., Hu, Y., Valsdottir, L.R., Zang, T., Peng, J.: Identifying drug–target interactions based on graph convolutional network and deep neural network. Briefings in Bioinformatics **22**(2), 2141–2150 (May 2020). https://doi.org/10.1093/bib/bbaa044, https://doi.org/10.1093/bib/bbaa044
37. Zong, N., Kim, H., Ngo, V., Harismendy, O.: Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. Bioinformatics **33**(15), 2337–2344 (Apr 2017). https://doi.org/10.1093/bioinformatics/btx160, https://doi.org/10.1093/bioinformatics/btx160

Declaration of Academic Integrity

The participating team declares that the paper submitted is comprised of original research and results obtained under the guidance of the instructor. To the team's best knowledge, the paper does not contain research results, published or not, from a person who is not a team member, except for the content listed in the references and the acknowledgment. If there is any misinformation, we are willing to take all the related responsibilities.

Participant: Remington Kim
Mentor: Mr. Matthew Wang
Date: June 14th, 2021

Acknowledgements

All research was conducted by myself, Remington Kim. I would like to thank my mentor, Mr. Matthew Wang, for giving me valuable guidance and advice throughout my research process.