Link Prediction in Disease-Disease Interactions Network Using a Hybrid Deep Learning Model

Ashwag Altayyar and Li Liao

March 10, 2025

Link Prediction in Disease-Disease Interactions Network Using a Hybrid Deep Learning Model

Ashwag Altayyar[1[0000-0003-3517-4214]] and Li Liao[1[0000-0002-1197-1879]]

[1] University of Delaware, Newark DE 19716, USA
`{ashwag, liliao}@udel.edu`

Abstract. Discovering disease-disease association based on the underlying biological mechanisms is an essential biomedical task in modern biology as understanding these relationships will assist biologists in discovering the pathogenesis, diagnosis, and intervention of human diseases. Recently, deep learning on graph and graph neural networks have achieved promising performance in modeling complex biological structures and learning compact representations of interconnected data. Inspired by the success of graph neural networks in learning subgraph representations, we propose a novel framework, SNN-VGA, designed to predict potential disease comorbid pairs. We first model disease-associated genes as subgraphs in the protein-protein interactions network and learn disentangled disease module representations using a subgraph neural network model. The learned embeddings are leveraged by the variational graph auto-encoder to predict disease comorbidity in the disease-disease interactions network. Empirical results from a benchmark dataset demonstrate that our method performs competitively compared with the state-of-the-art model, with an AUROC of 0.96.

Keywords: Association Prediction, Comorbidity, Disease, Graph Convolution Network, Subgraph Neural Networks, Variational Graph Auto-Encoder.

1 Introduction

In cells, the majority of cellular components exert their functions through the interactions with other cellular components [1]. Cellular functions are regulated by a complex network of molecular interactions, known as the interactome, which involves physical and functional interactions between various biological macromolecules such as proteins [2]. Since protein–protein interactions (PPIs) are intrinsic to most of the complex biological processes, any disruption of these interactions may cause malfunction and potentially lead to diseases. It has been shown that, the analysis of PPIs is important for understanding the molecular mechanisms of diseases, which can improve the prognostics and treatment for human disorders [3]. Often, the interconnectivity of the PPIs network allows genetic abnormalities to propagate through the network connections and indirectly influence the activity of other gene products [1]. Therefore, perturbations in the PPIs network can lead to the simultaneous presence of two or more diseases in the same individual, a phenomenon referred to as comorbidity [4]. The etiology of disease comorbidities involves several mechanisms. Previous studies have identified the comorbidity patterns through shared associated genes between diseases [5, 6]. Beyond genetic overlap, network-based structure has made substantial contributions to the advancement of biological systems [7], which allows for the exploration of cellular-level connections encoded by PPIs to reveal the underling mechanism of comorbidity. Therefore, direct interactions between causative proteins of two diseases were analyzed to uncover the molecular mechanisms driving disease comorbidity [8]. Other studies have suggested that diseases may co-occur because they are co-regulated by high-level biological factors, such as shared cellular processes and biological pathways [9, 10]. The random walk algorithm was proposed to explore unexplained disease similarity by analyzing the connections between disease-related genes in the PPIs network [8, 11].

Most of the approaches described above are based on analyzing biological factors and local network structures underlying the development of comorbid diseases. However, the PPIs network is large and complex, which requires more advanced methods to reveal intricate relationships to explain or predict disease comorbidity. Indeed, studies have been developed to consider the disease module theory to quantify the associations between diseases [6, 12]. In recent years embedding representation has been applied to disease biology. LINE [13] was used to map each gene in the PPIs network into a low-dimensional vector space to capture the intricate similarities between diseases [14]. CoGO is a model that used graph convolutional network (GCN) to measure disease similarity according to the structure of gene ontology and the gene interaction network [15]. Another work employed isometric feature mapping (Isomap), an extension of multi-dimensional scaling (MDS) that applied geodesic distance on the PPIs network for identifying disease comorbidities. In this approach, the nodes' coordinates were derived by preserving the shortest path distances between node pairs through eigenvalue decomposition and double centering of the distance matrix [16]. Despite these advances, the mentioned studies have certain limitations when inferring disease associations including:

1. Many comorbid disease pairs remain undiscovered in the medical literature. As a result, negative samples, which represent disease pairs that do not co-occur more frequently than expected by random chance, are sparse leading to imbalanced training data.
2. Disease modules contain rich higher-order connectivity patterns, both internally among member genes and externally through interactions with the rest of the network. Most of the previous work elucidates disease associations depending mainly on learning the representation of each gene associated with each disease separately without considering the interconnections of genes related to each disease module.
3. Some of the afore-mentioned methods rely on the location of disease modules within the PPIs network to predict disease relationships. They assumed that gene products associated with a disease segregate in the same neighborhood. In reality, many of the disease modules can be localized in one region of the network or distributed across multiple local neighborhoods, each with non-trivial internal topology.

By analyzing a benchmark PPIs dataset used for comorbidity prediction with disease module separation [6] and PPIs network Isomap embedding [16], we made the following observations:

- All disease modules form multiple disjoint components in the PPIs network. The minimum number of connected components observed across all disease modules is three. However, the maximum number of connected components observed across all disease modules is 276, which represents a high degree of fragmentation within a particular disease subgraph.
- The largest connected component (LCC) across all disease modules contains 85% of the proteins belonging to a specific subgraph. Despite the presence of large and connected components, the majority of the largest connected components, approximately 93%, include less than half of the proteins within the disease module.

These observations indicate the fragmented nature of disease-related genes throughout PPIs network, as they are not organized into cohesive clusters but rather existed in isolated groups. Motivated by the above analyses, we propose a deep learning framework, Subgraph Neural Networks-Variational Graph Auto-encoder (SNN-VGA), as depicted in Fig. 1. It consists of two models Subgraph Neural Networks (SUBGNN) [17] and Variational Graph Auto-Encoder (VGAE) [18]. SUBGNN is used to generate meaningful representations for disease subgraphs on the PPIs network while considering the fragmented topology of each disease subgraph. These learned representations by SUBGNN are further leveraged during the construction of a disease-disease interactions network (DDIs) to denote the features associated with diseases in the network. Then, we formulate disease comorbidity prediction using the constructed disease graph as a link prediction problem and exploit the advancement of VGAE to determine whether there is a missing link between two diseases in the DDIs network.

2 Materials and Methods

2.1 Biological Data

Protein-Protein Interactions (PPIs). The PPIs interactome describes the interactions between proteins within the cell. Our PPIs data is derived from [6, 16], contains 13,460 proteins and 141,296 interactions, including regulatory, binary, literature-curated, metabolic enzyme-coupled, protein complexes, kinase-substrate pairs, and signaling interactions. We model the PPIs interactome as a graph $G_{PPIs} = (V, E)$ that contains two main elements $V = \{1, \ldots, n\}$ is the set of nodes representing proteins, and $E \subseteq V \times V$ is the set of undirected edges that indicate the interactions between the proteins. The largest connected component in this graph includes 13,329 nodes and 141,150 edges, covering more than 99% of the nodes and edges in the dataset used for this study. We focus on the LCC because it represents the most biologically relevant interactions, where the involved proteins frequently participate in significant cellular processes. It is generally believed that small connected components (many of them are singletons) in the current incomplete PPIs network are a result of missing edges, which correspond to interactions yet discovered, conceivably due to their minor/obscure roles, and that those small connected components, with missing edges once detected, will be connected to form into a larger component or merge to the LCC. Therefore, it has been a common practice adopted in similar and related studies to focus on LCC for PPIs networks [16, 17].

Disease Data. The disease-gene associations dataset is obtained from [6, 16]. The dataset contains a list of 299 diseases, and each disease has a set of genes that are known to be associated with the disease.

2.2 Disease Modules in the PPIs Network

Disease-gene associations and the interactions between them can be modeled in a PPIs network as subgraphs consisting of both known human diseases and disease-related genes. Each subgraph represents a disease module that contains a set of proteins which collectively contribute to a cellular function within the PPIs network and are implicated in causing the disease. In this work, we have constructed 299 disease modules as subgraphs, each consisting of gene products related to a specific disease.

2.3 Disease Comorbidity

To validate our proposed method, we utilize a Medicare dataset of disease history that includes 10,743 disease pairs [6, 16]. In order to quantify the comorbidity for each disease pair, the relative risk (RR) of observing a pair of diseases d_i and d_j, affecting the same patient, is computed using the following equation:

$$\mathrm{RR}_{ij} = \frac{C_{ij}N}{P_iP_j} \tag{1}$$

where C_{ij} is the number of patients affected by both diseases, N is the total number of patients in the population, and P_i and P_j are the prevalence of diseases i and j respectively. The prevalence of a disease refers to the proportion of the total population that is affected by a given disease. When the RR exceeds a specific threshold, it indicates that the co-occurrence of two diseases is more frequent than would be expected by a random chance. In this study, we set the threshold for the RR at two different values: 0 and 1 to investigate how it may affect the learning and performance of the model. When the threshold on RR is set at 1, the data contains 6,269 comorbid disease pairs, whereas setting the RR value to zero gives rise to 8,874 disease pairs, which are used to construct DDIs network, as described in the following sections.

2.4 Disease Network Representation

Given the dataset of disease-associated genes and the PPIs network, we adopt a subnetwork embedding model called SUBGNN that captures the topology of disease subgraphs. It creates representations for all disease modules, which have varying sizes and multiple distributed connected components throughout the graph, as shown in Fig. 1(a).

Subgraph Representations. Given a PPIs network as a graph $G_{PPIs} = (V, E)$, where $V = \{1, \ldots, n\}$ consists of a set of nodes denote the proteins, and edges $E \subseteq V \times V$ represent the interactions between them. $S = (V', E')$ is a disease subgraph of G_{PPIs} if $V' \subseteq V$ and $E' \subseteq E$ where nodes in each disease subgraph denote the product of genes associated with the disease, and the edges indicate the interactions between them. Each subgraph has a unique label y_S defines distinct disease and may include multiple connected components $S^{(c)}$. Given disease subgraphs $\mathcal{S} = \{S_1 , S_2, \ldots, S_n \}$, SUBGNN is designed to identify the unique structure of subgraphs via three property-aware channels, each designated to explore a different aspect of subgraph topology which are position, neighborhood, and structure described in Table 1. SUBGNN specifies a mechanism that propagates neural messages at the subgraph level, between the subgraph components and randomly sampled anchor patches. Anchor patches $\mathcal{A}_\mathrm{x} = \{A_\mathrm{x}^{(1)}, \ldots, A_\mathrm{x}^{(n\,A)}\}$ are subgraphs that are randomly sampled from the underlying graph G_{PPIs} in a channel-specific manner, where each anchor patch corresponds to one of the SUBGNN's channels, defined as $\mathcal{A}_\mathrm{P}$, $\mathcal{A}_\mathrm{N}$ and $\mathcal{A}_\mathrm{S}$. Each propagated message conveys information about the relationship between a specific anchor patch and a subgraph component as follows:

$$\mathrm{MSG}_\mathrm{X}^{A\to S} = \gamma_\mathrm{x}(S^{(c)},\ A_\mathrm{x}) \,.\, \mathbf{a}_\mathrm{x} \tag{2}$$

where X is the channel, $\gamma_\mathrm{x}$is a similarity function between the component $S^{(c)}$ and the anchor patch A_x, and $\mathbf{a}_\mathrm{x}$ is the learned embedding of A_x . There are three types of similarity functions that determine the relative weighting of each anchor patch in building the subgraph component representations. For the position channel, the similarity function is defined as follow:

$$\gamma_\mathrm{P}\left(S^{(c)}, A_\mathrm{P}\right) = \frac{1}{\left(d_{sp}\left(S^{(c)}, A_\mathrm{P}\right) + 1\right)} \tag{3}$$

where d_{sp} represents the average shortest path (SP) on the graph between the connected component $S^{(c)}$ and the anchor patch A_P specified for position channel. In contrast, for the neighborhood channel, the similarity function is $\gamma_\mathrm{N}\left(S^{(c)}, A_\mathrm{N}\right) = 1$ in the case of an internal neighborhood and $\gamma_\mathrm{N}\left(S^{(c)}, A_\mathrm{N}\right) \leq K$ for a border neighborhood that includes the subset of neighbor nodes within a k-hop distance from the connected component nodes. For the structure channel, the similarity function is given by:

$$\gamma_\mathrm{S}\left(S^{(c)}, A_\mathrm{S}\right) = \frac{1}{\left(\mathrm{DTW}\left(d_{S^{(c)}}, d_{A_\mathrm{s}}\right) + 1\right)} \tag{4}$$

here, $d_{S^{(c)}}$ and $d_{A_\mathrm{s}}$are the ordered degree sequences for the subgraph component and anchor patch, respectively, which are compared by the normalized dynamic time warping (DTW) measure [19], a similarity measure that calculates the optimal alignment between two sequences by minimizing the cumulative distance. The messages are then transformed into an order-invariant hidden representation $\mathbf{h}_{\mathrm{x},c}$ for the subgraph component $S^{(c)}$, as follows:

$$\mathbf{g}_{\mathrm{x},c} = \mathrm{AGG}_M\left(\left\{\mathrm{MSG}\,_\mathrm{X}^{A_\mathrm{x}\to S^{(c)}}\ \forall\, A_\mathrm{x} \in \mathcal{A}_\mathrm{x}\right\}\right) \tag{5}$$

$$\mathbf{h}_{\mathrm{x},c} \longleftarrow \sigma\left(\mathbf{W}_\mathrm{x} \,.\, [\mathbf{g}_{\mathrm{x},c};\, \mathbf{h}_{\mathrm{x},c}]\right) \tag{6}$$

The outcome of applying these equations is a channel specific hidden representation $\mathbf{h}_{\mathrm{x},c}$ for each connected component $S^{(c)}$ of subgraph S and channel X, where $\mathbf{W}_{\mathrm{x}}$ is a layer-wise learnable weight matrix for channel X, σ is a non-linear activation function, AGG_{M} is a function that aggregates messages received from anchor patches, and $\mathbf{h}_{\mathrm{x},c}$ is the representation of the connected component at the previous layer, which gets updated and passed to the next layer of the model. The model is designed as such to learn a d_s-dimensional subgraph representation $\mathbf{z}_S \in \mathbb{R}^{d_s}$ for each disease subgraph $S \in \mathcal{S}$. This representation encapsulates the collective properties of all subgraph components using three channels across all layers, which can be then used for comorbidity prediction.

2.5 Disease-Disease Interaction Prediction

We address disease comorbidity as a task of predicting potential edges between diseases in a network as shown in Fig. 1(b). We consider a Graph $G_{disease} = (V, E)$, where $V = \{1, \ldots, n\}$ represents the set of nodes each denoting a disease, and $E \subseteq V \times V$ is a set of edges that capture the interactions between diseases. The adjacency matrix of $G_{disease}$ denoted by $\mathbf{A} \in \mathbb{R}^{n\times n}$ satisfies $A_{ij} \neq 0$ if and only if $(v_i, v_j) \in E$ suggesting the existence of a relationship between disease pairs. Specifically, with the RR threshold set to 1, only disease pairs with an RR value of 1 or higher are considered connected by positive edges in the disease graph, indicating their comorbidities. Conversely, assigning a more relaxed threshold at RR = 0 allows disease pairs with an RR value of 0 or higher to be connected by positive edges. Additionally, each node in the graph is associated with a d-dimensional feature vector generated by SUBGNN model. All disease feature vectors are stored in the disease feature matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$.

Variational Graph Auto-Encoder (VGAE). VGAE is a framework for unsupervised learning specifically designed for graph-structured data. It combines the power of GCN with probabilistic modeling to learn low-dimensional latent representations of nodes in a graph. In particular, the latent representations for an undirected graph are learned by leveraging the graph structure represented by an adjacency matrix **A** and observed node attributes **X** to encode the graph structure and produce a posterior approximation $q_{\phi}(\mathbf{Z} \mid \mathbf{X}, \mathbf{A})$ over the latent variables **Z**. Subsequently, the decoder reconstructs the original graph structure from these latent variables that consist of a compressed representation of the graph's structure and features. We introduce a component for disease comorbidity prediction based on the VGAE model in our designed formwork, as illustrated in Fig. 1(b). To the best of our knowledge, our model is the first attempt to implement VGAE for comorbidity prediction.

Inference Model. The inference model aims to compute latent representations **Z** via multiple graph convolution layers to capture the structural similarities between diseases. We initially adopt two convolutional layers of a GCN to learn more informative representations of diseases. We then embed these representations into a low-dimensional latent space. The encoder model is defined as:

$$q(\mathbf{Z} \mid \mathbf{X}, \mathbf{A}) = \prod_{i=1}^{N} q(\mathbf{z}_i \mid \mathbf{X}, \mathbf{A}) \tag{7}$$

$$q(\mathbf{z}_i \mid \mathbf{X}, \mathbf{A}) = \mathcal{N}(\mathbf{z}_i \mid \boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i^2)) \tag{8}$$

where $\boldsymbol{\mu} = \mathrm{GCN}_{\boldsymbol{\mu}}(\mathbf{X}, \mathbf{A})$ and $\log \boldsymbol{\sigma} = \mathrm{GCN}_{\boldsymbol{\sigma}}(\mathbf{X}, \mathbf{A})$ are the matrices of $\boldsymbol{\mu}_i$ and $\log \boldsymbol{\sigma}_i$ representing the parameters of the learned distribution that describes the latent variables $\mathbf{Z}$. $\mathrm{GCN}_{\boldsymbol{\mu}}(\mathbf{X}, \mathbf{A})$ and $\mathrm{GCN}_{\boldsymbol{\sigma}}(\mathbf{X}, \mathbf{A})$ denote a two-layer GCN defined as $\mathrm{GCN}(\mathbf{X}, \mathbf{A}) = \tilde{\mathbf{A}}\ \mathrm{ReLU}(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}_0)\mathbf{W}_1$, where $\mathbf{W}_i$ are the weight matrices. The symmetrically normalized adjacency matrix $\tilde{\mathbf{A}}$ is given by $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\ \mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is the degree matrix. The Rectified Linear Unit function is defined as ReLU $(\cdot) = \max(0, \cdot)$.

Generative Model. The Generative model maps disease feature vectors from the latent space generated by the encoder into the original disease graph. The structure of the decoder component influences the model's flexibility and ability to capture the expressiveness of the learned features. Therefore, to enhance these aspects, a multilayer perceptron (MLP) neural network is employed to predict the probability of links between diseases in the network, as illustrated in Fig. 1(b). The latent representations corresponding to each disease pair are concatenated and fed into a MLP neural network to predict the likelihood of edges in the disease network. We propose the following decoder network to reconstruct the original disease graph $\mathbf{A}$:

$$p(\mathbf{A} \mid \mathbf{Z}) = \prod_{i=1}^{N} \prod_{j=1}^{N} p(A_{ij} \mid \mathbf{z}_i, \mathbf{z}_j) \tag{9}$$

For each pair of nodes i and j in the disease network, the probability of the existence of an edge between them is calculated using MLP as the following expression:

$$p(A_{ij} = 1 \mid \mathbf{z}_i, \mathbf{z}_j) = \sigma\ (\mathbf{W}_2(\mathrm{ReLU}(\mathbf{W}_1\mathbf{Z}_{ij} + \mathbf{b}_1)) + \mathbf{b}_2) \tag{10}$$

where $\mathbf{Z}_{ij} = [\ \mathbf{z}_i, \mathbf{z}_j]$ represents the concatenated latent representations corresponding to diseases i and j, and the parameters $\mathbf{W}_i$ and $\mathbf{b}_i$ are the decoder weight matrix and bias vectors, respectively. $\sigma\ (\cdot)$ is defined as the logistic sigmoid function. The final output determines the predicted probability of a link between diseases i and j.

Training Objective. We optimize the variational lower bound $\mathcal{L}$ w.r.t. the variational parameters $\mathbf{W}_i$, given by:

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{Z} \mid \mathbf{X}, \mathbf{A})}\,[\log p(\mathbf{A} \mid \mathbf{Z})] - \mathrm{KL}[q(\mathbf{Z} \mid \mathbf{X}, \mathbf{A}) \,\|\, p(\mathbf{Z})] \tag{11}$$

here, $\mathrm{KL}[q(\cdot) \,\|\, p(\cdot)]$ denotes the Kullback-Leibler divergence between $q(\cdot)$ and $p(\cdot)$. We assume a Gaussian prior for $p(\mathbf{Z})$, expressed as $p(\mathbf{Z}) = \prod_i p(\mathbf{z}_i) = \prod_i \mathcal{N}(\mathbf{z}_i \mid 0, \mathbf{I})$.

3 Experiments

3.1 Datasets

Our experiment is conducted on benchmark datasets for PPIs and disease-associated genes [6, 16], which form an underlying base graph including subgraphs with their associated labels as known diseases. Because of the use of LCC in our method, genes that are not in the LCC will be dropped from our experiments, which may cause some loss of information and make the prediction task more challenging. On the other hand, use of LCC allows us to focus on genes that are on the LCC and hence more informative in terms of the degree of interconnectivity and interactions with other genes. The statistics of the datasets are summarized in Table 2.

3.2 Experimental Setup and Evaluation Methods

We build our implementation of SNN-VGA by leveraging two distinct platforms: the Facebook machine learning library “PyTorch” [20-22], and the scikit-learn machine learning library [23]. We detail the experimental setups and evaluation methods for disease module representations and comorbidity prediction, respectively.

3.3 Disease Module Representations

We use the experimental setups proposed by SUBGNN. Initially the model is trained using Graph Isomorphism Network (GIN) [24] on link prediction to generate node and meta node embeddings for each node within the subgraph of the PPIs network. Subsequently, these trainable nodes embeddings are utilized to implement SUBGNN model, which generates feature vectors for each disease module.

3.4 Predicting Comorbidities Between Disease Modules

For the experimental settings of VGAE, we employ a transductive link prediction split in which the same graph structure is partitioned into the training, validation, and test sets. From the entire graph, 70% of the edges are designated as positive samples for the training set. Additionally, we sample 20% of the edges for validation and 10% for testing, which serve as positive samples, i.e., node pairs that are connected with an edge. Concurrently, for the training, validation, and test sets, we also randomly sample an equal number of negative samples, i.e., node pairs that are unconnected.

Parameters Selection for VGAE Architecture. The architecture of VGAE significantly influences the prediction performance of the model. Accordingly, we empirically set the dimensions of both the hidden layer and latent variables to 128 and 64, respectively. These values were selected based on validation set performance to balance model complexity and generalization. Additionally, we initialize the weights as described in reference [25]. We train the model for 50 epochs using Adam optimizer [26] with a learning rate of 0.001.

Evaluation Measures. We apply a nested cross-validation procedure [27], for model assessment and selection. Our model is trained using a 10-fold-within-5-fold nested-CV procedure to obtain an unbiased estimate of model performance while simultaneously optimizing the parameters. We calculate the reconstruction probability of the test edges to evaluate the ability of the model to classify comorbid versus non-comorbid disease pairs. We employ common evaluation metrics to measure the prediction performance of the SNN-VGA model, which include accuracy, precision, recall, F-measure (F1), average precision (AP), and the receiver operating characteristic (ROC) curve score.

4 Results

The averaged model performance for the comorbidity prediction task is reported in Table 3. We evaluate our method's performance by setting the comorbidity RR threshold values at 0 and 1. A threshold of 1 emphasizes stronger disease associations, while a threshold of 0 incorporates a wider range of associations, thus increasing edges between diseases and enhancing both graph connectivity and model training. Moreover, it enables the model to capture more complex relationships between the diseases and learn meaningful representations. As illustrated in Table 3, with an RR threshold of 0, SNN-VGA achieves remarkably high scores with an area under the ROC curve (AUROC) of 0.96 and an AP of 0.95. At the stricter RR threshold of 1, although there is a slight decline in the performance, SNN-VGA still yields strong results with an AUROC of 0.94 and an AP of 0.92. The superior performance of our method, particularly at the RR = 0 threshold can be attributed to its ability to effectively leverage the increased connectivity within the disease network, which in turn leads to more comprehensive analysis of potential disease associations. Fig. 2 represents the ROC curves and their related areas under the curves that exhibit the performance of SNN-VGA across distinct test sets. For each test set, we run the model with different random initialization, and we then obtain the mean result and standard error derived from 10 runs that further emphasize the model consistency and statistical reliability under different conditions.

In our comparison, we include the recent state-of-the-art method “Weighted Geometric Embedding” [16] for predicting comorbid diseases. This method mapped the PPIs network into a low-dimensional geometric space using the MDS technique. Each disease module was characterized by features derived from its projection in the geometric space, which were subsequently used to train support vector machine and random forest classifiers for comorbidity classification. Table 3 presents the results of this performance comparison. It can be observed that our SNN-VGA model achieves outstanding performance in the disease comorbidity prediction task as compared with the “Weighted Geometric Embedding" method. In particular, our approach enhances the AUROC score and accuracy by 6% and 2%, respectively.

5 Conclusion

In this study, we introduce SNN-VGA, a novel computational approach that integrates biological data into a single network and employs graph deep learning paradigms to predict disease comorbidity. We develop two distinct models. Initially, a SUBGNN is adopted to produce a set of feature vectors, each representing a specific disease module. Then, a model based on VGAE is applied to reconstruct the DDIs network for predicting disease comorbidities. By addressing shortcomings observed in related work, our method significantly outperforms a state-of-the-art method in cross-validation experiments on a benchmark dataset, as measured by common metrics. It demonstrates that our approach, by integrating a network comprised of diseases and a network of PPIs, cross linked via known disease-gene associations, offers a powerful platform for analyzing disease similarities with a unified graph-theoretic framework.

References

1. Barabási, A.-L., Gulbahce, N., Loscalzo, J.: Network medicine: A network-based approach to human disease. Nat. Rev. Genet. 12(1), 56–68 (2011). doi: 10.1038/nrg2918
2. Luck, K., Sheynkman, G.M., Zhang, I., Vidal, M.: Proteome-scale human interactomics. Trends Biochem. Sci. 42(5), 342–354 (2017). doi: 10.1016/j.tibs.2017.02.006
3. Gonzalez, M.W., Kann, M.G.: Chapter 4: Protein interactions and disease. PLoS Comput. Biol. 8(12), e1002819 (2012). doi: 10.1371/journal.pcbi.1002819
4. Feinstein, A.R.: The pre-therapeutic classification of comorbidity in chronic disease. J. Chronic Dis. 23(7), 455–468 (1970). doi: 10.1016/0021-9681(70)90054-8
5. Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., Barabási, A.-L.: The human disease network. Proc. Natl. Acad. Sci. U.S.A. 104(21), 8685–8690 (2007). doi: 10.1073/pnas.0701361104
6. Menche, J., et al.: Uncovering disease-disease relationships through the incomplete interactome. Science 347(6224), 1257601 (2015). doi: 10.1126/science.1257601
7. Barabási, A.-L., Oltvai, Z.N.: Network biology: Understanding the cell's functional organization. Nat. Rev. Genet. 5(2), 101–113 (2004). doi: 10.1038/nrg1272
8. Ko, Y., Cho, M., Lee, J.-S., Kim, J.: Identification of disease comorbidity through hidden molecular mechanisms. Sci. Rep. 6(1), 39433 (2016). doi: 10.1038/srep39433
9. Rual, J.-F., et al.: Towards a proteome-scale map of the human protein–protein interaction network. Nature 437(7062), 1173–1178 (2005). doi: 10.1038/nature04209
10. Rubio-Perez, C., et al.: Genetic and functional characterization of disease associations explains comorbidity. Sci. Rep. 7(1), 6207 (2017). doi: 10.1038/s41598-017-04939-4
11. Hamaneh, M.B., Yu, Y.-K.: DeCoaD: Determining correlations among diseases using protein interaction networks. BMC Res. Notes 8(1), 226 (2015). doi: 10.1186/s13104-015-1211-z
12. Ni, P., Wang, J., Zhong, P., Li, Y., Wu, F.-X., Pan, Y.: Constructing disease similarity networks based on disease module theory. IEEE/ACM Trans. Comput. Biol. Bioinf. 17(3), 906–915 (2020). doi: 10.1109/TCBB.2018.2817624
13. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: LINE: Large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web, pp. 1067–1077. Florence, Italy (2015). doi: 10.1145/2736277.2741093
14. Li, Y., Keqi, W., Wang, G.: Evaluating disease similarity based on gene network reconstruction and representation. Bioinformatics 37(20), 3579–3587 (2021). doi: 10.1093/bioinformatics/btab252
15. Chen, Y., Hu, Y., Hu, X., Feng, C., Chen, M.: CoGO: A contrastive learning framework to predict disease similarity based on gene network and ontology structure. Bioinformatics 38(18), 4380–4386 (2022). doi: 10.1093/bioinformatics/btac520
16. Akram, P., Liao, L.: Prediction of comorbid diseases using weighted geometric embedding of human interactome. BMC Med. Genomics 12(S7), 161 (2019). doi: 10.1186/s12920-019-0605-5
17. Alsentzer, E., Finlayson, S.G., Li, M.M., Zitnik, M.: Subgraph neural networks. In: Advances in Neural Information Processing Systems 33, pp. 8017–8029. Vancouver, Canada (2020). doi: 10.48550/arXiv.2006.10538
18. Kipf, T.N., Welling, M.: Variational graph auto-encoders. In: NIPS Workshop on Bayesian Deep Learning, pp. 1–3. Barcelona, Spain (2016). doi: 10.48550/arXiv.1611.07308
19. Mueen, A., Keogh, E.: Extracting optimal performance from dynamic time warping. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2129–2130. San Francisco, USA (2016). doi: 10.1145/2939672.2945383
20. Fey, M., Lenssen, J.E.: Fast graph representation learning with PyTorch Geometric. In: International Conference on Learning Representations Workshop on Representation Learning on Graphs and Manifolds, New Orleans, USA (2019). doi: 10.48550/arXiv.1903.02428

21. Paszke, A., et al.: PyTorch: An imperative style high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8026–8037. Vancouver, Canada (2019). doi: 10.48550/arXiv.1912.01703

22. Falcon, W., et al.: PyTorchLightning/pytorch-lightning: 0.7.6 release. Zenodo, (2020). doi: 10.5281/zenodo.3828935

23. Pedregosa, F., et al.: Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011). doi: 10.48550/arXiv.1201.0490

24. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? In: International Conference on Learning Representations, New Orleans, USA (2019). doi: 10.48550/arXiv.1810.00826

25. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, pp. 249–256. Sardinia, Italy (2010)

26. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations, San Diego, USA (2015). doi: 10.48550/arXiv.1412.6980

27. Stone, M.: Cross-validatory choice and assessment of statistical predictions. J. Roy. Stat. Soc. Ser. B Methodol. 36(2), 111–133 (1974). doi: 10.1111/j.2517-6161.1974.tb00994.x

Table 1. Six properties of subgraph topology in Subgraph Neural Network.

Position	Internal	The distances between S_i's components
	Border	The distances between S_i and the rest of G_{PPIs}
Neighborhood	Internal	Defines a set of internal nodes of S_i
	Border	Defines a set of border nodes of S_i
Structure	Internal	The internal connectivity of $S^{(c)}$ within S_i
	Border	The border connectivity of $S^{(c)}$ within S_i

Table 2. Statistics of the benchmark datasets.

Dataset	**#Nodes**	**#Edges**
Protein–Protein Interactions	13,460	141,296
Disease–Disease Interactions (RR = 0)	299	8,874
Disease–Disease Interactions (RR = 1)	299	6,269

Table 3. Comparison of averaged model performance using our method and state-of-the-art method for thresholds RR = 0 and RR = 1.

Model		**Indicators**					
		AUROC	Accuracy	Precision	Recall	F1	AP
SNN-VGA (Ours)	RR=0	0.96 ± 0.01	0.92 ± 0.01	0.91 ± 0.01	0.94 ± 0.01	0.92 ± 0.01	0.95 ± 0.00
	RR=1	0.94 ± 0.00	0.89 ± 0.01	0.87 ± 0.00	0.92 ± 0.01	0.89 ± 0.01	0.92 ± 0.00
Weighted Geometric Embedding [16]	RR=0	0.90	0.90	0.90	0.90	0.89	-
	RR=1	0.76	0.70	0.70	0.70	0.69	-

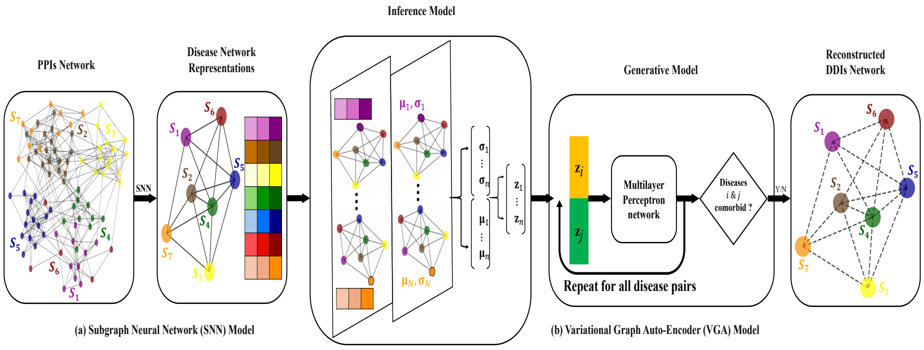
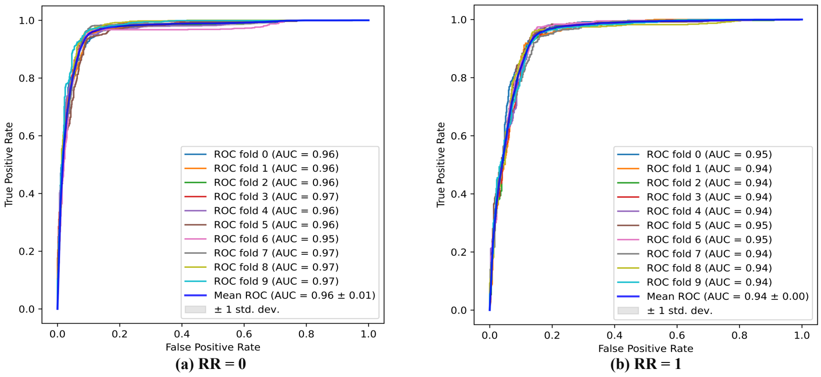

Inference Model
PPIs Network
Disease Network Representations
Generative Model
Reconstructed DDIs Network
S_7
S_2
S_3
S_5
S_4
S_6
S_1
SNN
μ_1, σ_1
μ_N, σ_N
σ_1
σ_n
μ_1
μ_n
z_1
z_n
z_i
z_j
Multilayer Perceptron network
Diseases i & j comorbid ?
Y/N
Repeat for all disease pairs
(a) Subgraph Neural Network (SNN) Model
(b) Variational Graph Auto-Encoder (VGA) Model

Fig. 1. The Hybrid Deep Learning Architecture of SNN-VGA Combining Subgraph Neural Network and Graph Variational Auto-Encoder. **(a) SNN Model for Disease Representations:** This step takes as input a protein-protein interactions (PPIs) network, in which we identify disease modules via known disease-gene associations. Each disease module corresponds to a subgraph comprised of nodes representing genes associated with the disease. The output of this step is a disease network (much like a condensation graph of these disease modules), along with a feature vector for each disease module, characterizing its positional and structural relationships with other disease modules learned by SNN from the underlying PPIs network. **(b) VGA Model for Predicting Disease Comorbidities: Inference Model:** It consists of a two-layer Graph Convolutional Network that processes the disease feature vectors and the structure of the disease network to infer the latent disease embedding used to predict interactions between diseases. The mean (**μ**) and variance (**σ**) vectors represent the parameters of the learned probabilistic distribution for each node in the latent space. These vectors are used to sample the latent space to create an embedding vector **Z** for each disease node, capturing the essential features of each disease node in a lower-dimensional space. **Generative Model:** For each pair of diseases *i* and *j* (colored yellow and green respectively), their **Z** vectors are concatenated into a single representation, which is then processed by a Multilayer Perceptron network to determine the comorbidity between these two diseases: if yes, an edge is added to connect these two disease nodes; otherwise, they remain unconnected. This process is repeated for all disease pairs. **Reconstructed Disease-Disease Interactions Network:** The output network generated by the model aims to predict disease comorbidities by leveraging the latent embedding and incorporating DDIs ground truth to enhance prediction accuracy.



(a) RR = 0 **(b) RR = 1**

Fig. 2. ROC curves for each fold of the 10-fold cross-validation, along with the mean result and standard error for the test sets across different RR values.