Self Supervised Prediction of Genetic Associations in Comorbid Diseases With Masked Autoencoder using Hypergraph Representations

Saikat Biswas*, Vibhanshu Ranjan*, Pabitra Mitra, and Krothapalli Sreenivasa Rao

***Abstract*—Comorbid disease association refers to the simultaneous occurrence of a disease with the coexistence of another primary disease. Due to the complex traits of these co-occurring multi-diseases, it is crucial to know the underlying genetic molecular basis of the prevalent diseases. The inference of common genetic association based on gene co-expression data helps to unveil the pathogenesis of comorbid diseases. There exist a few disease-specific gene co-expression-based analyses to predict the hub genes causing these diseases. However, works lack multi-relational biological data integration. In addition, there still does not exist any unified method to predict the common genetic associations from the co-expression graph across comorbid diseases. Hence, we introduce a generalized and novel approach to predict overlapping genetic associations from disease-specific gene co-expression networks with a self-supervised edge-masking technique catapult with a hypergraph-based pre-embedding learning approach. The advantage of hypergraph learning is that it induces higher-order rich biological information of candidate genes. In addition, we use the self-supervised-based edge masking strategy to attain model training over only a few numbers of edge labels. Our proposed approach outperforms the six baseline models for our case-study datasets and also predicts novel genetic associations across comorbid disease pairs. Our implementation is available at https://github.com/VRa11/HyperSSL.**

***Index Terms*—genetic association, co-morbidity, gene co-expression network, hypergraph, graph neural network, self-supervised learning, link prediction.**

I. Introduction

COMORBID disease condition refers to the occurrence of an adverse medical condition along with a primary (index) disease. A patient suffering from a comorbid disease(s) is significantly susceptible to multiple risk factors and possesses a higher mortality risk than with a disease alone. It has been shown that during the Covid-19 pandemic, around 61.9% of the population in Italy who died from Covid disease were suffering from at least 3 or more fatal comorbidity factors [1]. Comorbid conditions in an individual increase with age and have been seen as the primal factor for declining life expectancy [2]. The presence of comorbid conditions also increases the number of medical consultations, prescriptions, and hospitalizations, which immensely affects the psychological state of patients, and further leads to the risk of postoperative complications. These adverse effects increase the overall healthcare cost of the individuals [3]. The presence of two diseases at the same time also complicates the appropriate choice of optimal treatment due to multiple drug interactions.

It is very much needed to characterize the key hidden drivers of the comorbid conditions to initiate better medications and efficient treatment. Here, these key drivers can be associated with environmental, pharmacological, or genetic factors [4]. Among these, genetic factors can be assumed as one of the key molecular components for comorbid diseases. The hidden molecular information is essential [5] to understand the interrelated connection between comorbid diseases. Recently, Cheng et al. [5] propose that the common target factors of comorbid diseases can be better understood while analyzing the hidden molecular information. Most of the present studies regarding this approach to comorbidity revolve around the identification of genes and their associated mutations. Modern advancement of omics data, namely gene expression profiles (including transcriptomics and proteomics data) plays a crucial role in present-day disease-specific analysis [6], [7] due to their direct participation in disease progression.

Gene co-expression network can be projected as a gene-gene similarity matrix for downstream analysis [8] which also helps to find candidate disease genes, participating in transcriptional regulatory programs. Recently, Zhu et al. [9] explored the hidden molecular mechanism between comorbid diseases namely autism spectrum disorder and inflammatory bowel disease using gene expression data. In another recent study, the potential disease biomarkers in peripheral blood are being identified in depression patients using gene co-expression networks in addition to a few machine learning methods [10]. However, the efficacy of mere gene co-expression networks for knowing disease mechanisms is limited due to the lesser ability to infer causality phenomena [8]. In some of the approaches, the PPI (protein-protein interaction) network is incorporated with gene co-expression network [11], [12] to get more interpretable disease-causing drivers. But these methods also carry a few issues for better comprehension of disease causation. For instance, the general graph proposed in these works may fail to shed light on the true interpreting genes for a specific disease. Moreover, to obtain a better explainability of underlying disease mechanisms it is necessary to reveal the genes and gene modules solely associated with

Saikat Biswas is with Advanced Technology Development Centre, Indian Institute of Technology, Kharagpur, 721302, India.

Vibhanshu Ranjan is with the Department of Electrical Engineering, Indian Institute of Technology, Kharagpur, 721302, India.

E-mail: saikatbiswas17@iitkgp.ac.in, pabitra@cse.iitkgp.ac.in

Pabitra Mitra and Krothapalli Sreenivasa Rao are with the Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, 721302, India.

Manuscript received xxxxx xx, xxxx; revised xxxxx xx, xxxx.

*These authors contributed equally to this work

the specific diseases. A simple PPI network fails to capture significant genetic information in the network, which may play a detrimental role in various complex comorbid diseases. It has been observed that metabolic pathways, phenotypes, miRNA, Transcription factor (TF) proteins, and SNPs (single nucleotide polymorphisms) are a few of the critical role players in causing fatal comorbid conditions [13], [14]. Ma et al. [16], [17] proposed various tensor decomposition-based approaches for human-virus PPI interactions to better interpret disease causation. So infusing molecular component information of candidate genes with disease-specific gene association can better interpret the causality of the complex diseases along with their prevalent diseases.

Graph convolution network (GCN) is a popular known approach of imposing traditional convolution to graphs by propagating messages of neighboring nodes for each central node [18]. However, in recent studies, it has been found that Hypergraph Neural Networks (HGNN) attain better performance than GCN in citation networks, while efficiently integrating latent higher dimensional relational information in the large complex networks [19]–[21]. Among the recent works on hypergraph embedding, a logistic matrix factorization based on hypergraph was proposed to predict potential metabolite–disease interactions [22] while incorporating the higher-order biological relationships. In another study, miRNAs-disease associations are predicted by a hypergraph regularized bipartite local model (HGBLM), based on a hypergraph-embedded Laplacian support vector machine (LapSVM) [23]. In this work, the author incorporated microRNA-disease-associated higher-order complex biological information using hypergraph embedded LapSVM model.

Graph-based deep learning on massive data often requires precise annotations, which are generally very expensive and time-consuming. To address this issue, self-supervised learning (SSL) is emerging as a new technique for extracting informative knowledge through well-designed pretext tasks without relying on manual labels [24]. SSL method has helped to reduce the dependency on annotated labels and enable the training on massive unlabeled data. The primary goal of SSL is to learn transferable knowledge from abundant unlabeled data with well-designed pretext tasks and then generalize the acquired knowledge to downstream tasks with specific supervision signals. Recently, SSL methods have also emerged in bio-medical and genomic data research due to their capability of inferring unbiased results while learning on huge unlabeled data [25]. Our proposed novel **HyperSSL** method introduces the following unique contributions in comparison to the existing methods:

- We employ a tailored hypergraph-based pre-embedding approach to incorporate the latent higher-dimensional genetic features.
- Despite of a large number of edge-dropping in gene co-expression graph, with the usage of hypergraph representation approach our self-supervised-based model outperforms the existing state-of-the-art models.
- The suggested approach also shows that the rich embedding learning method can effectively make up for the lack of topological knowledge in a self-supervised setting.

II. Related Works

We review a few studies based on disease comorbidity analysis using gene co-expression data. In addition, we also cite a few of the recent studies on biological networks using hypergraph-based approaches and SSL techniques. In recent work, Bharadhwaj et al [26] identify the shared patterns among Schizophrenia, Bipolar Disorder, and Type 2 Diabetes using disease-specific gene expression datasets. In their work, they investigate the overlapping pathways and genetic associations among these comorbid diseases while analyzing gene co-expression patterns. It is evident that in comparison with mere disease-associated genes, gene expression data can better reflect the causation and consequence of different complex diseases [27]. Thus, it is highly relevant to study the concurrence of comorbid diseases by analyzing the disease-specific gene expression data [26], [28]. In recent work, it is highlighted that the crucial factor that triggers a specific group of patients in higher susceptibility to a given disease than the other group of patients by analyzing the disease-specific RNA-Seq data [29]. Gaudelet et al. [30] proposed a multi-scale neural network-based framework by integrating the microarray-based gene expression data with the integrated gene-pathway information to explore the comorbid disease association mechanisms. An age-dependent comorbidity analysis regarding Type 2 Diabetes (T2D) is being recently explored using tissue-specific gene expression data analysis [31]. This study sheds light on the primal age variant factors that can initiate T2D-associated multiple condition-affected individuals while analyzing their different age-dependent gene expression profiles.

A recent study by Li et al. [32] showed the importance of a weighted gene co-expression module for identifying the key genes in atrial fibrillation disease. Their proposed work explains that the trait-specific gene co-expression module identification can better reflect the hidden cause of underlying disease mechanisms. This work indicates that altered atrial metabolism can be a crucial factor in the occurrence of atrial fibrillation in patients. Another recent work inferred a plausible link between the C4A gene and Schizophrenia disease using brain tissue gene co-expression network analysis [33]. Wang et al. [10] proposed a method based on a gene co-expression network to identify hub genes responsible for depression patients associated with concurrent comorbid conditions like coronary heart disease and Parkinson's disease.

Ma et al. [22] proposed a novel approach with hypergraph-based logistic matrix factorization to predict the potential interactions between metabolites and diseases by exploring higher-order disease-metabolite relationships and also for SARS-CoV-2 drug repositioning [34]. A weighted hypergraph is used in another study to predict potential microbial-drug associations while employing a generalized matrix factorization-based approach [15]. In recent work, Jain et al. [35] proposed that biological pathway-encoded hyperedges can help to build disease-specific hypergraphs that efficiently explore the novel repurposing drug targets for complex prevalent diseases. A multi-tissue and cell-type gene expression integration approach is developed using hypergraph factorization-based parameter-efficient graph representation learning method [36]. The pro-

posed method can better describe the latent collective biological mechanisms that spawn complex comorbid conditions in a patient. Burke et al. [37] introduced a direct hypergraph generation framework to analyze the multi-morbidity disease progression by capturing multi-dimensional disease-causing relationships.

Due to the recent advancement of self-supervised learning methods, there exists a considerable amount of work in disease diagnosis studies. Li et al. [38] proposed a multi-modal data-based self-supervised feature learning method for retinal diseases. In their method, they leverage the modality-variant feature with the patient-similarity features to diagnose the disease state in an individual. In another recent work, self-supervised-based contrastive learning is used for single-cell RNA-Seq data analysis [39]. In this framework, the gene expression profile is mapped into low-dimensional space by keeping similar functional cell representations together and distinct cells to the distant embedding space.

In our proposed hypergraph-embedding induced self-supervised model, **Hyper-SSL**, we investigate the comorbid disease pair relationship from their core molecular basis perspectives to shed light on the key genetic associations that can better explain the primal functional mechanisms for the concurrent prevalence of these diseases in a patient. As is observed, hypergraph learning can efficiently capture the higher-order relationships in a graph. We integrate the higher-order rich biological information of candidate genes in a disease-specific co-expression graph by introducing a hypergraph learning approach with a gene-auxiliary information bipartite graph. This method can generate the most informative gene representations. Next, the masked-autoencoder-based self-supervised model is used for its proven efficacy in predicting masked edges with only a fewer edge-label learning framework. As a whole, this framework can efficiently learn on less but most informative gene representations in a disease-specific gene co-expression graph to reconstruct missing edges as well as predict new most important common genetic associations across comorbid diseases. In comparison, the other existing comorbid analysis methods are not capable of giving a unified, generalized framework for inferring common genetic associations in comorbid disease pairs. Our **Hyper-SSL** framework, to our best knowledge, is the first disease-specific gene-coexpression data-based computational approach to identify the crucial genetic associations that can give a better explainability of the complex comorbid diseases.

III. Materials and Methods

A. Data description

In this method, we have chosen four comorbid disease cases, namely, Type 2 Diabetes (T2D), Parkinson's disease (PD), Huntington's disease (HD), and Schizophrenia (Sch), as the case studies for our proposed method. We have used the disease-specific microarray-based gene expression data from the GEO Omnibus datasets public repository. The used gene expression data are, namely, GSE64998 [40] for T2D, GSE99039 [41] for PD, GSE3790 [42], [43] for HD, and GSE53987 [44] for Sch.

In our work, we build a disease-specific gene co-expression network for each of the comorbid diseases. In addition, we use other gene-associated auxiliary information regarding pathways, phenotypes, miRNAs, SNPs, and Transcription factor (TF) proteins collected from other biological databases as below:

- Human biological signaling pathway interactions are obtained from the MSigDB c2 (canonical pathway) database.
- Human phenotypes and their genetic associations are obtained from OMIM (Online Mendelian Inheritance in Man) database [45].
- The SNPs associated with the corresponding human genes are obtained from the GWAS (Genome-wide association studies) database.
- The miRNA and interacting gene information is collected from miRTarBase [46] database.
- The TF-proteins and their target gene associations are obtained from hTFtarget [47] database.

In the following sections, we describe the method for constructing co-expression networks and their embeddings.

B. Gene expression analysis and Gene co-expression network construction

Gene expression profiling is a method to determine the pattern of genes expressed. The gene expression is studied at the transcription level, under conditioned specific circumstances, or in a desired cell to get a global view of cellular functions for further molecular-level analysis. In the gene expression analysis, we first extract the feature data from gse series matrix file and normalize the gene expression file. The corresponding gene information and expression file are then merged to obtain the required dataframe to perform the WGCNA (Weighted Gene Co-expression Network Analysis) analysis [48]. WGCNA is performed to calculate the edge matrix to construct a disease trait-specific gene co-expression network. In this procedure, first, the pairwise correlation is calculated for two nodes (i.e. genes), say g_i and g_j. The co-expression similarity matrix c_{ij} is calculated as the absolute value of the correlation coefficient between the profiles of nodes g_i and g_j as below:

$$c_{ij} = |cor(g_i, g_j)| \tag{1}$$

This co-expression similarity matrix is then converted to an adjacency matrix while using the power with soft threshold β computed by *pickSoftThreshold* function in the WGCNA package. Thus, we obtain the adjacency matrix c_{ij} showing large adjacency values between genes with similar expressions.

After binarizing the continuous values in the adjacency matrix c_{ij}, we finally obtain the edge matrix A_{ij} as below:

$$A_{ij} = \begin{cases} 1, \ c_{ij} > c_{thresh} \\ 0, \ otherwise, \end{cases} \tag{2}$$

where, the c_{thresh} is optimized by the WGCNA method itself.

We have selected the most significant gene co-expression network modules which are highly correlated with disease traits among all of the generated gene co-expression network modules. It is being observed that the genetic associations manifesting with high betweenness centrality scores in the gene coexpression network are being identified as the primal candidates participating in complex diseases in humans [49]–[52]. Accordingly, we first identify the edges with high edge betweenness centrality scores and then select the largest connected component among the obtained gene co-expression network modules. Nodes with similar gene expressions are connected through interconnected edges in the generated gene co-expression network. Langfelder and Horvath (2008) proposed in their WGCNA method, that genes with similar expressions are always likely to be associated with similar biological functions. In a few of the recent works [53]–[55], researchers have noticed that, the complex disease mechanism and biological processes can better be described through the correlated co-functional gene modules due to the gene's property of collective participation in regulating biological processes [56]. Hence, the co-functional gene modules connected on the coexpression graph can be used to extract efficient features of biological gene modules which outweighs the task of genetic association prediction across complex comorbid diseases. Compared with earlier methods [12], [57], which construct gene graphs based on general PPI data, the gene co-expression graph in our proposed method based on disease-specific gene expression data evidently can better reflect the disease-specific genetic interactions.

C. ***Hyper-SSL:*** *Hyper embedding infused Self Supervised Learning model for link analysis*

1) Mathematical Background of Hypergraphs: We briefly discuss some of the fundamental concepts of hypergraphs based on the seminal works [21], [58]–[64]. A hypergraph is a generalization of graphs in which its hyperedges can join any number of nodes. Mathematically, an unweighted hypergraph can be represented as, $\mathcal{H} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes, $\mathcal{E} = \{e_1, e_2, \ldots, e_m\}$ is the set of hyperedges with $e_q \subseteq \mathcal{V}$ for $q = 1, 2, \ldots, m$. In the hypergraph, any two nodes are called adjacent, if they are in the same hyperedge. A hypergraph is referred to as connected if there always exists a path between any of the node pairs through a hyperedge. If all the hyperedges in a hypergraph contain the same number of nodes, then $\mathcal{H}$ is called a k-uniform hypergraph. Hence, a graph is just a 2-uniform hypergraph. An incidence matrix of a hypergraph $\mathbf{H}$ can be denoted by $\mathbf{H} \in \mathbb{R}^{n \times m}$, consists of logical values which indicate the relationships between nodes and hyperedges as,

$$\mathbf{H}(\mathbf{v}, \mathbf{e}) = \begin{cases} 1, if v \in e \\ 0, otherwise \end{cases}$$

So, if a node v_i is involved in a hyperedge e_q, then the (i, q)-th entry of $\mathbf{H}$, i.e. $\mathbf{H_{iq}}$ has the value 1 otherwise, it is equal to 0. The degree of a node in the hypergraph is the number of hyperedges containing that node, which can be calculated as, $d(i) = \sum_q H_{iq}$. Similarly, the cardinality of a hyperedge is the number of nodes contained in that hyperedge and can be represented as, $C_q = \sum_i H_{iq}$. The diagonal node degree matrix and the diagonal hyperedge cardinality matrix of a hypergraph can be defined as $\mathbf{D_{v_i}} \in \mathbb{R}^{n \times n}$ and $\mathbf{D_{e_q}} \in \mathbb{R}^{m \times m}$, respectively.

For a given hypergraph, the classification task refers to classifying the nodes on the hypergraph, where the labels on the hypergraph are required to be smoothed, through the hypergraph structure. This task can be formulated as a regularization introduced in [62]:

$$\arg\min_f \{R_{emp}(f) + \Omega(f)\}, \tag{3}$$

where $\Omega(f)$ is a regularizer on hypergraph, $R_{emp}(f)$ refers to the supervised empherical loss and $f(\cdot)$ is a classifcation function. The regularizer can be defined as the following:

$$\Omega(f) = \frac{1}{2} \sum_{e_q \in \mathcal{E}} \sum_{\{u_i, v_i\} \in V} \frac{w(e_q) H(u_i, e_q) H(v_i, e_q)}{C_q} \left(\frac{f(u_i)}{\sqrt{d(u_i)}} - \frac{(v_i)}{\sqrt{d(v_i)}} \right)^2, \tag{4}$$

Here say, $\Theta = \mathbf{D_{v_i}^{\frac{-1}{2}}} \mathbf{H W D_{e_q}^{-1} H^T D_{v_i}^{\frac{-1}{2}}}$ and $\Delta = I - \Theta$, and the normalized $\Omega(f)$ can be rewritten as $\Omega(f) = f^T \Delta f$, where $\mathbf{W}$ is the hyperedge weight matrix, Δ is positive semi-definite, and usually called the hypergraph Laplacian.

2) Problem Formulation: The overall framework of our proposed Hyper-SSL model is illustrated in Figure 1. At a high level, in the proposed method, we formulate two graph structures. The first one, is a gene-gene co-expression graph, say, $\mathcal{G}_g = (\mathcal{V}_g, \mathcal{E}_g)$, which represents the disease-specific gene co-expression network graph, where $\mathcal{V}_g = \{v_{g_1}, v_{g_2}, \ldots, v_{g_n}\}$ denotes the gene nodes of the gene co-expression network, and $\mathcal{E}_g = \{e_{g_1}, e_{g_2}, \ldots, e_{g_m}\}$ denotes the corresponding edges between the gene pairs and $|\mathcal{V}_g| = n$ and $|\mathcal{E}_g| = m$. Here, the subscript "g" symbolizes *gene*. Next, to capture the more detailed biological information of the participating genes in the gene co-expression graph $\mathcal{G}_g$, we build a bipartite graph, which represents the genes and their associated biological information namely, *pathway*, *phenotype*, *miRNA*, *SNP*, and *TF-proteins*. Now, out of this bipartite graph using k-hop neighborhood-based approach [21], we formulate an auxiliary hypergraph say, $\mathcal{H}_a = (\mathcal{V}_a, \mathcal{E}_a)$, where $\mathcal{V}_a = \{v_{a_1}, v_{a_2}, \ldots, v_{a_p}\}$ define the nodes in the graph and $\mathcal{E}_a = \{e_{a_1}, e_{a_2}, \ldots, e_{g_q}\}$ refers to the hyperedges built-up using the 1-hop (here in k-hop, k=1) neighbors around a central node while learning the hypergraph structure and $|\mathcal{V}_a| = p$ and $|\mathcal{E}_a| = q$. Here, the subscript "a" symbolizes *auxiliary*. Since in the graph, the genes are connected with auxiliary information through the 1-degree neighborhood and vice-versa for the auxiliary nodes, we adapt the 1-hop neighborhood approach for hypergraph structure learning in our method.

3) Hypergraph learning-based Pre-embedding Generation: We integrate the rich biological auxiliary information of the candidate genes into the co-expression network using the hypergraph embedding learning approach. For that, we employ the hypergraph convolution approach from a spatial domain perspective [21]. In this method, the graph convolution takes

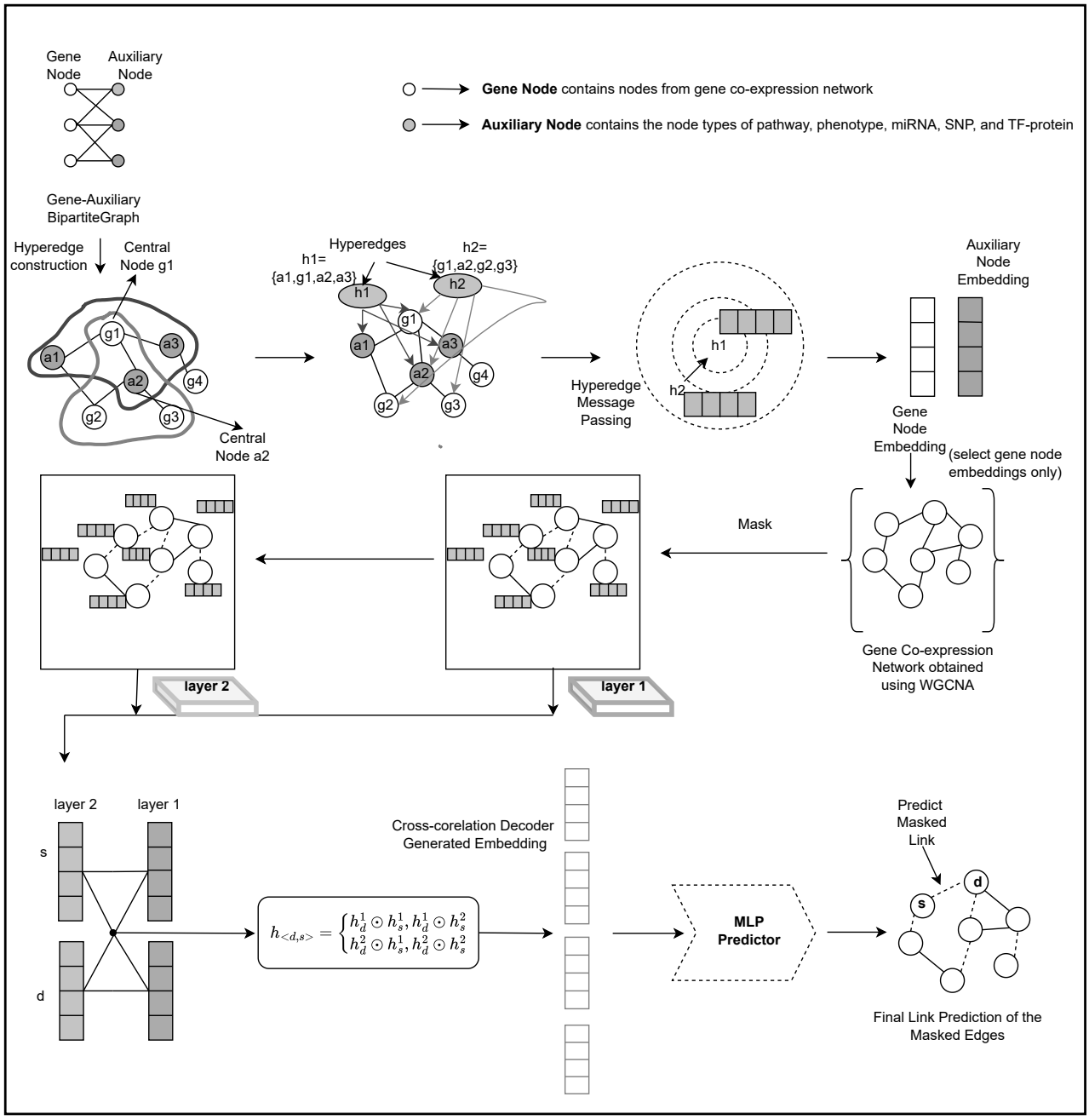

Fig. 1: Workflow of our proposed Hyper-SSL model

the aggregation of the neighboring nodes, to get a new representation of the central node. Here, we define the spatial convolution on the hypergraph $\mathcal{H}_a$. For each node, $\mathcal{V}_a$ in the hypergraph we aggregate its neighboring node messages to update itself according to the path between the central node and each node in its neighborhood (which be either gene or auxiliary node). To propagate the message from the node to hyperedge and that from the hyperedge to node using the hyperpath [21], which can be defined as the *Inter Neighbor Relation* N over node set $\mathcal{V}_a$ and hyperedge set $\mathcal{E}_a$. Say, we denote the node inter-neighbor set $\mathcal{N}_v(e_q)$ of the hyperedge e_q and the hyperedge inter-neighbor set $\mathcal{N}_e(v_p)$ of the node v_p based on the inter-neighbor relation. For a given auxiliary node $v_p \in \mathcal{V}$ of the hypergraph $\mathcal{H}_a$, we aim to aggregate messages from its hyperedge inter-neighbor set $\mathcal{N}_e(v_p)$. To acquire those hyperedge messages for each hyperedge e_q in the hyperedge inter-neighbor set $\mathcal{N}_e(v_p)$, we aggregate the messages from its node inter-neighbor set $\mathcal{N}_v(e_q)$. Finally, the hypergraph convolution makes a closed message passing loop from the node feature set $\mathcal{X}^l$ to $\mathcal{X}^{(l+1)}$. The employed spatial hypergraph convolution in the $l-th$ layer can be defined as:

$$\begin{cases} a^l_{e_q} = \sum_{v_p \in N_v(e_q)} A^l_v(x^l_{v_p}) \\ y^l_{e_q} = U^l_e(w_{e_q}, m^l_{e_q}) \\ a^{l+1}_{v_p} = \sum_{v_p \in N_e(v_p)} A^l_e(x^l_{v_p}, y^l_{e_q}) \\ x^{l+1}_{v_p} = U^l_v(x^l_{v_p}, a^{l+1}_{e_q}), \end{cases} \tag{5}$$

where, $x^l_{v_p} \in X^l$ in the input feature vector of the node $v_p \in V$ in the layer $l = 1, 2, \ldots, L$ and $x^{l+1}_{v_p}$ is the updated feature of the node v_p. $a^l_{e_q}$ is the message of the hyperedge $e_q \in E$, and w_{e_q} is the associated weight to the hyperedge e_q. $a^{l+1}_{v_p}$ refers to the message of the node v_p. $y^l_{e_q}$ is the hyperedge feature of the hyperedge e_q, which is an element of the feature set $Y^l = \{y^l_1, y^l_2, \ldots, y^l_A\}$, where $y^l_i \in \mathbb{R}^{C_l}$. $A^l_v(\cdot), U^l_e(\cdot), A^l_e(\cdot), U^l_v(\cdot)$, are the node message functions, hyperedge update function, hyperedge message function, and node update functions, respectively. These functions can be defined as follows:

$$\begin{cases} A^l_v(x^l_{v_p}) = \frac{x^l_{v_p}}{|N_v(e_q)|} \\ U^l_e(w_{e_q}, a^l_{e_q}) = w_{e_q} \cdot a^l_{e_q} \\ A^l_e(x^l_{v_p}, y^l_{e_q}) = \frac{y^l_{e_q}}{|N_e(v_p)|} \\ U^l_v(x^l_{v_p}, a^{l+1}_{v_p}) = \sigma(a^{l+1}_{e_q} \cdot \Theta^l), \end{cases} \tag{6}$$

where $\Theta^l \in \mathbb{R}^{C^l \times C^{l+1}}$ is a trainable parameter of layer l, which can be learned in the training phase. $\sigma(\cdot)$ is an arbitrary non-linear activation function like $ReLU(\cdot)$. $\frac{x^l_{v_p}}{|N_v(e_q)|}$ and $\frac{y^l_{e_q}}{|N_e(v_p)|}$ denote the normalized node and hyperedge features, respectively. Hence, such a node is guided to aggregate the hyperedge feature set Y^l, which can be formulated as, $Y^l = WD^{-1}_{e_q}H^T X^l$ and similarly, the updating node feature set X^{l+1} from the hyperedge feature set Y^l can be formulated as, $X^{l+1} = \sigma(D^{-1}_{v_p} H Y^l \Theta^l)$. Thus, the final matrix format of the hypergraph convolution can be formulated as below:

$$X^{l+1} = \sigma(D^{-1}_{v_p} H W D^{-1}_{e_q} H^T X^l \Theta^l) \tag{7}$$

Finally, from all the obtained node embeddings (i.e. the genes and auxiliary nodes), we only select the gene node embeddings to be fed into the masked graph autoencoder model.

4) Masked Graph Autoencoder-based model for link analysis: In the next phase of our proposed **Hyper-SSL** model, we use the obtained hypergraph learning-based node embeddings and perform the self-supervised learning-based masked graph autoencoder method [65], [66] on the co-expression graph, $\mathcal{G}_g$. Here, we employ an edge-level self-supervised learning by graph masking strategy. we perturb the input gene co-expression graph $\mathcal{G}_g$ by randomly dropping the edges in the graph. A sample set of edges say $\mathcal{E}_{g_{mask}}$ from the observed edges (i.e. $\mathcal{E}_g$ with some masking ratio w, and then acquire the perturbed input graph as $\mathcal{G}_{g_{perb}}$ as below:

$$\mathcal{G}_{g_{perb}} = (\mathcal{V}_g, \mathcal{E}_{g_{remain}}), \mathcal{E}_{g_{remain}} = \mathcal{E}_g - \mathcal{E}_{g_{mask}} \tag{8}$$

where $\mathcal{E}_{g_{remain}}$ denotes the remaining gene-gene edge set after the graph masking. For the masking strategy, we employ the undirected masking strategy proposed by the authors Tan et al. [65], [66]. Here, as our graph is undirected, consequently after random sampling over $\mathcal{E}_g$, the obtained masked edge set $\mathcal{E}_{g_{mask}}$ and the perturbed graph $\mathcal{G}_{g_{perb}}$ also become undirected. Now, we use the well-established GNN model as the encoder for the perturbed co-expression graph $\mathcal{G}_{g_{perb}}$. The primary objective of GNN models is to update the node representation by imposing the representation from itself and the neighboring nodes as well. This can be defined as below:

$$h^l_d = COM(h^{l-1}_d, AGG(\{h^{l-1}_s : s \in N_v\})), \tag{9}$$

where h^l_d denotes the embedding of the gene node d at the $l-th$ layer, and $N_v = \{s|s|e_{d,s} =< d, s >\in \mathcal{E}_g\}$ is the direct neighbors of the gene node d with $h^{(0)}_d = x_d$. Here, AGG is used to aggregate (by averaging) the features from the neighbors and the function COM is applied to combine (by concatenating) the aggregated neighbor information in addition to its node feature generated from the previous layer. Hence, for a GNN encoder with L layers, there are L node representations as $\{h^{(1)}_v, h^{(2)}_v, \ldots, h^{(L)}_v\}$ being generated, where $h^{(l)}_v$ holds the neighborhood structure within l hops.

The large masking ratio of the edges in the original gene co-expression graph $\mathcal{G}_g$, makes the input graph $\mathcal{G}_{g_{perb}}$ incomplete by nature. This inevitably fails to reconstruct the masked edges by the mere application of the general GNN decoder. Hence, to address this issue we use the cross-correlation decoder [65], [66] (shown in Figure 1.). In specific, for a given L hidden representations of the gene node pair d and s (i.e. $\{h^{(l)}_d, h^{(l)}_s\}^L_{l=1}$), the generated cross-correlations can be defined as follows:

$$h_{<d,s>} = \|^L_{l,j=1} h^{(l)}_d \odot h^{(j)}_s, \tag{10}$$

where, $||$ denotes the concatenation and $\odot$ refers to the element-wise multiplication. In addition, $h_{d,s} \in \mathbb{R}^{dL^2}$ is the final representation between gene nodes d and s, considering their $l-th$ order neighborhood and the $j-th$ order neighborhood, respectively. After obtaining the cross-correlation representation of the edge-connecting d and s. The MLP layer is being used to predict its probability by, $g(d, s) = MLP(h_{<d,s>}) \in \mathbb{R}$. The proposed approach recovers the original gene co-expression graph $\mathcal{G}_g$ while learning the masked edges $\mathcal{E}_{g_{mask}}$. This technique is inspired by the recent self-supervised-based work, proposed by Tan et al. [65], [66].

By employing the proposed **Hyper-SSL** method, the encoder graph can effectively induce the hypergraph information in the gene nodes and get updated efficiently to generate final robust node embeddings and then the cross-correlation decoder finally reconstructs the original gene co-expression graph $\mathcal{G}_g$. The model is being trained by adopting the standard graph-based loss function as below:

$$\mathcal{L}_{Hyper-SSL} = -\frac{1}{|\mathcal{E}_{g_{mask}}|} \sum_{(d,s) \in \mathcal{E}_{g_{mask}}} \log \frac{exp(g(d,s))}{\sum_{r \in \mathcal{V}_g} exp(g(d,r))} \tag{11}$$

In the proposed method, the summation operation of Eq. 11 is approximated by negative sampling to accelerate the training.

IV. Experimental Details

Our proposed *Hyper-SSL* model is built upon on Pytorch and Pytorch-Geometric library. For generating the hyper-embedding we apply the *HGNNP* [21] hypergraph learning approach. In this hypergraph learning setting, we split the dataset into 60, 20, and 20 ratios for training, validation, and testing, respectively. The input dimension is 128 and the number of classes is 2 (i.e. the gene and auxiliary nodes) for our generated dataset. Then in the next phase of our proposed method, we employ a self-supervised masked graph autoencoder technique and predict the masked edges at the final stage. We set the input dimension to 128 (obtained from the hypergraph embedding), the hidden dimension to 128, and the decoder dimension to 256 in the masked autoencoder experimental setting. We split the edges into 85%, 5%, and 10% for training, validation, and testing, respectively. We randomly mask 70% edges and fix the number of encoder layers to 4 and learning rate (lr) to 0.0001. We have randomly selected the negative edges as the edges which are not present in the original co-expression graph and take an equal number of negative samples as the positive ones. For training, We set the number of epochs to 100 with Adam optimizer and early stopping patience of 50 epochs.

Link prediction is a fundamental task in network biology to get an approximated interaction probability between biological node entities [67], [68]. To identify the common functional genes in the comorbid disease pairs, it is important to first predict the disease-specific top-scoring gene-gene pairs, which infer the most biologically significant edges in the co-expression graph. Hence, we choose the link prediction task to address this problem. We perform the link prediction task in disease-specific (i.e. T2D, PD, SCH, and HD for our case study) gene co-expression graphs. In our **Hyper-SSL** model, We first generate the pre-embeddings of the candidate gene nodes in the gene co-expression graph using hypergraph learning (using the auxiliary graph described earlier) and then use those biologically rich embeddings in the self-supervise-based edge mask autoencoder for the link prediction of the masked edges in the co-expression graph. We follow the well-known [69] study to construct the train/valid/test edge splits. We perform extensive experiments and compare our proposed method with the baseline methods as follows:

- **GraphSAGE** is an inductive framework that leverages node representation information to efficiently generate unseen node representation [70].
- **DGI** is self-supervised learning that generates node features using maximal mutual information between patch representations and corresponding high-level summaries of graph [71].
- **GIC** is a self-supervised graph representation learning approach that captures cluster-level information content while applying the K-means method and maximizing the same cluster inter-node mutual information [72].
- **MGAE** is a self-supervised learning approach that randomly masks a large proportion of edges in a graph and tries to reconstruct those missing edges during training [66].

TABLE I: The Disease Dataset Statistics

Disease Data	# Gene Co-expression Graph		# Gene-Auxiliary Bipartite Graph		
	# gene nodes	# edges	# gene nodes	# auxiliary nodes	# edges
PD	**498**	**2000**	**498**	**4324**	**19348**
T2D	**692**	**3919**	**692**	**4500**	**27572**
HD	**963**	**3000**	**963**	**5436**	**35682**
SCH	**741**	**4500**	**741**	**4758**	**30372**

- **JMSSMMA** is joint masking and self-supervised-based model to predict small molecular miRNA associations while employing probability distribution-based masking technique [73].
- **VKBNMF** is non-linear matrix factorization-based approach to predict human-virus PPI interactions introducing a Bayesian kernel method [17].

We follow the same experimental settings for all the methods for fair comparison. To measure the effectiveness and importance of hypergraph embedding learning, we use random embeddings as the pre-embeddings for the baseline methods. We evaluate the model performance using the popular AUC, AUPR, and F1-measure evaluation metrics.

V. Parameter Sensitivity Analysis

It is essential to select the right hyperparameters to ensure optimal model performance. As part of our sensitivity analysis, we have conducted a study in which we varied the learning rate (lr) and the number of layers. The impact of different learning rates is depicted in Figures S5, S6, S7, and S8, while the effects of varying the number of layers on specific disease testing results are presented in Figures S9, S10, S11, and S12. These illustrations can be found in the supplementary file, providing visual insights into the outcomes when different parameter values are employed.

VI. Results

TABLE II: The result comparisons of our proposed method with Baseline methods on the T2D (Type 2 Diabetes) Dataset

T2D (Type 2 Diabetes)			
	AUC (%)	**AUPR (%)**	**F1 Measure**
GraphSAGE	59.18 ± 0.71	54.19 ± 0.60	0.61 ± 0.07
DGI	73.14 ± 1.56	70.87 ± 2.21	0.72 ± 0.01
GIC	72.81 ± 2.28	72.68 ± 2.14	0.71 ± 0.001
MGAE	92.70 ± 1.03	92.18 ± 0.23	0.80 ± 0.04
JMSSMMA	71.62 ± 2.20	80.87 ± 0.71	0.71 ± 0.01
VKBNMF	75.23 ± 4.01	76.31 ± 3.82	0.72 ± 0.01
Hyper-SSL (our proposed)	**93.58 ± 0.63**	**92.86 ± 0.79**	**0.86 ± 0.03**

A. Result Analysis

The link prediction performance on disease-specific gene co-expression graphs obtained using baseline models and the proposed approach are presented in Tables II, III, IV, and V, respectively, for diseases T2D, PD, SCH, and HD. From the obtained results, it is observed that our proposed *Hyper-SSL* approach outperforms all the baseline methods. The high scores in AUC, AUPR, and F1-measure metrics justify our

TABLE III: The result comparisons of our proposed method with Baseline methods on the PD (Parkinson's Disease) Dataset

PD (Parkinson's Disease)			
	AUC (%)	**AUPR (%)**	**F1 Measure**
GraphSAGE	60.00 ± 2.78	57.04 ± 1.71	0.63 ± 0.02
DGI	63.62 ± 0.19	62.93 ± 0.41	0.65 ± 0.008
GIC	77.14 ± 0.73	77.72 ± 0.65	0.72 ± 0.005
MGAE	90.72 ± 1.71	91.12 ± 1.53	0.78 ± 0.05
JMSSMMA	70.45 ± 4.26	79.45 ± 1.20	0.69 ± 0.01
VKBNMF	73.22 ± 6.04	75.29 ± 5.13	0.71 ± 0.03
Hyper-SSL (our proposed)	**91.95 ± 0.89**	**92.24 ± 1.06**	**0.79 ± 0.06**

TABLE IV: The result comparisons of our proposed method with Baseline methods on the SCH (Schizophrenia Disease) Dataset

SCH (Schizophrenia Disease)			
	AUC (%)	**AUPR (%)**	**F1 Measure**
GraphSAGE	59.04 ± 2.14	59.52 ± 1.71	0.58 ± 0.11
DGI	67.76 ± 2.35	64.63 ± 2.65	0.68 ± 0.01
GIC	72.96 ± 3.61	71.62 ± 3.90	0.71 ± 0.01
MGAE	92.18 ± 4.43	92.70 ± 3.41	0.81 ± 0.05
JMSSMMA	74.87 ± 6.18	83.36 ± 2.69	0.75 ± 0.01
VKBNMF	77.79 ± 2.53	80.31 ± 2.31	0.75 ± 0.02
Hyper-SSL (our proposed)	**94.89 ± 0.47**	**94.32 ± 0.77**	**0.88 ± 0.01**

TABLE V: The result comparisons of our proposed method with Baseline methods on the HD (Huntington's Disease) Dataset

HD (Huntington's Disease)			
	AUC (%)	**AUPR (%)**	**F1 Measure**
GraphSAGE	59.51 ± 1.66	50.06 ± 1.95	0.52 ± 0.23
DGI	63.30 ± 4.70	63.07 ± 3.00	0.61 ± 0.06
GIC	73.25 ± 0.77	71.08 ± 0.84	0.70 ± 0.009
MGAE	95.72 ± 2.03	95.16 ± 1.48	0.86 ± 0.03
JMSSMMA	77.67 ± 6.34	85.33 ± 2.81	0.78 ± 0.01
VKBNMF	74.12 ± 3.11	73.57 ± 5.16	0.73 ± 0.03
Hyper-SSL (our proposed)	**97.64 ± 0.56**	**97.21 ± 0.48**	**0.88 ± 0.07**

proposed method's considerable efficiency and robustness. For instance, in T2D, the *Hyper-SSL* method gives high AUC, AUPR, and F1 scores of 93.58 ± 0.63, 92.86 ± 0.79, and 0.86 ± 0.03, respectively. The other well-known self-supervised model *MGAE* gives the second highest results of 92.70 ± 1.03, 92.18 ± 0.23, and 0.80 ± 0.04 for AUC, AUPR, and F1 measure, respectively. Similarly, for other diseases, namely, PD, SCH, and HD, the *Hyper-SSL* method gives the AUC scores of 91.95 ± 0.89, 94.89 ± 0.47, and 97.64 ± 0.56, AUPR scores of 92.24 ± 1.06, 94.32 ± 0.77, and 97.21 ± 0.48. Similarly, our model provides the highest F1 scores of 0.79 ± 0.06, 0.88 ± 0.01, and 0.88 ± 0.07 for PD, SCH, and HD, respectively. All of these results outperform the state-of-the-art models, namely, *GraphSAGE*, *DGI*, *GIC*, *MGAE*, *JMSSMMA*, and *VKBNMF*. With these encouraging results, it is clear that the proposed hypergraph-induced learning approach improves the self-supervised masked method to make the best node representations, in contrast to the high proportion of edge masking ratios. Hence, it is evident that our proposed method can efficiently reconstruct the missing edges and outshines the link prediction task compared to other well-known baseline methods. The graphical representations of the ROC_AUC curves can be seen in supplementary Figures S1, S2, S3, and S4.

B. Novel Gene Associations Predictions of the Comorbid Diseases

In addition to the model comparisons of our proposed method with the baseline methods, we predict some new novel genetic associations for comorbid disease pairs as produced in Table IX with their functional enrichment analysis (on biological processes) and biological validations. To attain these comorbid disease-pair genetic associations, we first choose the link prediction inferred top-scoring gene-gene pairs across all comorbid diseases. Next, among these edges, we identify the common genes (with their associated biological functions) participating in both of the prevalent diseases. As the obtained top links represent the most significant neighborhood genes in each of the diseases, the individual candidate genes of those edges which also participate in both of the comorbid diseases consequently infer the functional modules associated with the sharing genes. We produce the Gene Ontology (GO) enrichment analysis based on *biological process* (GOBPID), *molecular function* (GOMFID), and *cellular component* (GOCCID) for the predicted novel associations using the Enrichr package [74] in the Tables, VI, VII, and VIII, respectively. These findings reflect the most important common functional modules responsible for these prevalent diseases. It observed that there exists a significant association of LXβ (NR1H2 gene) polymorphisms in the causation of T2D [75]. The LXR or Liver X receptors are a subclass of nuclear receptors that cause alterations in the cellular level of the endogenous lipid ligands. It is already well-established that the NR1H2 gene is a significant regulator of glucose homeostasis which leads to T2D disease. On the other hand, the LXβ also causes increased phagocytic activities of microglia [76]. These phagocytic activities initiate the abnormal accumulation of pathogenic proteins that lead to PD in a patient [77]. The other predicted genetic association RAB1B also shows participation in T2D [78]. It is experimentally proven that the Rab family genes initiate GTPases with α-synuclein, Leucine-rich repeat kinase 2, and vacuolar protein sorting 35, 3 key proteins in PD pathogenesis [79]. Impaired lipid metabolism in patients plays a crucial role in the function of T2D [80]. The pathological progression of PD disease is highly dependent on the up-regulation of lipid-protein metabolism [81]. It is observed that the ceramide proteins, which are regulated by OSBP, can cause SCH in a patient [82]. OSBP is also linked to controlling cholesterol efflux activity, which plays a crucial role in insulin secretion in an individual and leads to T2D [83]. Another association of common genetic association predicted from our method can be seen as ATP6AP1 and TRIM23 for the SCH and HD comorbid disease pair. The deficiency of the ATP6AP1 gene results in broad-spectrum neurodegenerative disorders including SCH and HD [84]. The TRIM variant genes play an important regulator in neuropsychiatric disorders, such as SCH [85]. COX11 participates in an abnormal copper binding

TABLE VI: GO enrichment analysis based on the biological process of a few predicted novel gene associations in comorbid disease pairs

Comorbid Disease Pairs	Predicted Genetic Associations	GOBPID	P-Value	Term
T2D & PD	NR1H2 RAB1B PPP1CA OS9	GO:0051247	0.000937	Positive Regulation Of Protein Metabolic Process
		GO:0032373	0.001249	Positive Regulation Of Sterol Transport
		GO:0060331	0.001499	Negative Regulation Of Response To Type II Interferon
		GO:0010288	0.001748	Response To Lead Ion
		GO:0006621		Protein Retention In ER Lumen
T2D & SCH	PPP1R11 RNF26 PSMB1 PSMC2 ARF1 ACTR1A TMED2 ATP9A OSBP GNB1	GO:0006511	0.000219	Ubiquitin-Dependent Protein Catabolic Process
		GO:0016192	0.000338	Vesicle-Mediated Transport
		GO:0019941	0.000536	Modification-Dependent Protein Catabolic Process
		GO:0006686	0.004243	Sphingomyelin Biosynthetic Process
		GO:0007191	0.004243	Adenylate Cyclase-Activating Dopamine Receptor Signaling Pathway
T2D & HD	PSENEN CASP7 PNPO MAGT1 SLC8A2 PLA1A	GO:0051604	0.000501	Protein Maturation
		GO:0019438	0.001499	Aromatic Compound Biosynthetic Process
		GO:0046184	0.001798	Aldehyde Biosynthetic Process
		GO:0050890	0.000149	Cognition
		GO:0036150	0.002797	Phosphatidylserine Acyl-Chain Remodeling
SCH & HD	GNA11 SNAPC5 ATP6AP1 NUP50 TRIM23	GO:0060158	0.003295	Phospholipase C-activating Dopamine Receptor Signaling Pathway
		GO:0042796	0.003844	snRNA Transcription By RNA Polymerase III
		GO:0007603	0.004392	Phototransduction, Visible Light
		GO:0006886	0.000636	Intracellular Protein Transport

TABLE VII: GO enrichment analysis based on the molecular function of a few predicted novel gene associations in comorbid disease pairs

Comorbid Disease Pairs	Predicted Genetic Associations	GOMFID	P-Value	Term
T2D & PD	NR1H2 RAB1B PPP1CA	GO:0034190	0.001249	Apolipoprotein Receptor Binding
		GO:0051117	0.018118	ATPase Binding
		GO:0005525	0.049254	GTP Binding
		GO:0098641	0.004243	Cadherin Binding Involved In Cell-Cell Adhesion
T2D & SCH	OSBP HNRNPDL ARF1 PGK1 TSR1	GO:0008142	0.005089	Oxysterol Binding
		GO:0034046	0.005935	poly(G) Binding
		GO:0035639	0.007105	Purine Ribonucleoside Triphosphate Binding
T2D & HD	CASP7 PNPO MAGT1 SLC8A2	GO:0097200	0.001499	Cysteine-Type Endopeptidase Activity Involved In Execution Phase Of Apoptosis
		GO:0016641	0.002098	Oxidoreductase Activity, Acting On The CH-NH2 Group Of Donors, Oxygen As Acceptor
		GO:0010181	0.002996	FMN Binding
		GO:0046873	5.653e-05	Metal Ion Transmembrane Transporter Activity
SCH & HD	GNA11 GNB5 TRIM23 FEN1	GO:0003924	0.000350	GTPase Activity
		GO:0017111	0.000370	Ribonucleoside Triphosphate Phosphatase Activity
		GO:0005525	0.005207	GTP Binding
		GO:0017108	0.002747	5'-Flap Endonuclease Activity

that leads to SCH and HD [86]. COX11 is a COX assembly factor that is crucially important for copper insertion into the COX subunit I [87]. Disruption of intracellular chloride homeostasis in the central nervous system plays a crucial role in the etiopathogenesis of several neurodegenerative diseases including SCH and HD [88]. Another study showed that many inflammatory mediators cause a disparity in GABAergic signaling by altering the intracellular chloride homeostasis and leading to neurological disorders [89]. The significant participation of the ATP6AP1 gene in the GABAergic process is being recently investigated [90]. A study suggests the involvement of COX11 in mitochondrial regulation with the incorporation of copper plays a vital role in the causation of complex prevalent diseases like SCH and HD [91].

TABLE VIII: GO enrichment analysis based on the cellular component of a few predicted novel gene associations in comorbid disease pairs

Comorbid Disease Pairs	Predicted Genetic Associations	GOCCID	P-Value	Term
T2D & PD	RAB1B OS9	GO:0033116	0.012933	Endoplasmic reticulum-Golgi Intermediate Compartment Membrane
		GO:0000836	0.001748	Hrd1p Ubiquitin Ligase Complex
T2D & SCH	PSMB1 PSMC2 GNS	GO:0060205	0.004212	Cytoplasmic Vesicle Lumen
		GO:1904813	0.004803	Ficolin-1-Rich Granule Lumen
T2D & HD	PPP2R1A DERL2	GO:0000159	0.004790	Protein Phosphatase Type 2A Complex
		GO:0000839	0.001998	Hrd1p Ubiquitin Ligase ERAD-L Complex
SCH & HD	GNA11 GNB5 VCAN GNA11 TRIM23 ATP6AP1 NMNAT2	GO:0031234	0.000786	Extrinsic Component Of Cytoplasmic Side Of Plasma Membrane
		GO:0005764	0.002244	Lysosome
		GO:0098588	0.008821	Bounding Membrane Of Organelle

TABLE IX: Literature validation of a few of the predicted top genetic associations of the comorbid disease pairs with their corresponding biological process terms

Comorbid Disease Pairs	Predicted Genetic Associations	Biological Process Term	Literature Validation
T2D & PD	NR1H2 RAB1B PPP1CA	Positive Regulation Of Protein Metabolic Process	PMID: 34178836
			PMID: 34676987
			PMID: 27502188
		Response To Lead Ion	PMID: 29315801
			PMID: 35767566
			PMID: 32790695
T2D & SCH	PPP1R11 RNF26 PSMB1 PSMC2 OSBP	Ubiquitin-Dependent Protein Catabolic Process	PMID: 12436343
			PMID: 27336055
			PMID: 36085258
			PMID; 27797829
		Sphingomyelin Biosynthetic Process	PMID: 23924720
			PMID: 23408933
			PMID: 31275749
			PMID: 33141894
			PMID: 37455011
			PMID: 33683021
T2D & HD	PSENEN CASP7 MAGT1	Protein Maturation	PMID: 35418220
			PMID: 22155432
			PMID: 23193206
		Cognition	PMID: 19782763
			PMID: 35277073
			PMID: 29412701
SCH & HD	ATP6AP1 COX11	Intracellular Monoatomic Cation Homeostasis	PMID: 34576238
			PMID: 21334439
			PMID: 23554570
			PMID: 33498264
	ATP6AP1 TRIM23	Intracellular Protein Transport	PMID: 28069866
			PMID: 38225560

VII. Discussion and Conclusion

In this paper, we propose a novel computational approach to identify the common genetic associations in comorbid disease pairs that can help explore the hidden mechanisms of these complex prevalence of diseases. Here, we use two different types of graphs, namely, (a) a disease-trait-a specific gene co-expression graph and (b) a biologically enriched auxiliary information graph for the candidate genes in this co-expression graph. Next, our proposed unique method integrates a hypergraph learning approach with the self-supervised learning technique. Using the hypergraph, we intend to capture the rich biological information of the participating genes in the co-expression graph. The pre-embedded genes can take benefit of its external biological interactions regarding pathways, phenotypes, miRNAs, SNPs, and TF-proteins. The purpose of using these biological associations is to fuse a myriad of information that can give a subtle pathophysiological representation of the genes instead of their mere topological

features. In addition, our *Hyper-SSL* approach dynamically constructs these hyperedges while automating the selection of the central node and corresponding 1-hop neighborhoods. So, the higher-order relational messages propagate and update the hypernodes using an efficient spatial convolution technique. We further use these well-learned pre-embeddings of the genes with co-expression graph topology for the final edge prediction task. To accomplish this task, we use the self-supervised based edge masking procedure. This learning approach reflects the advantage of efficient learning with only a few edge labels. Finally, we compare our method with the other well-recognized baseline models for the link prediction task.

Despite masking a large number of edges from the original co-expression graph, our proposed method outperforms the existing other popular self-supervised graph learning models. Thus, through our approach, we also demonstrate that the high-dimensional hypergraph embedding approach can effectively boost the prediction performance of the edge-masking autoencoder model, which can never be attained through random node embedding initialization. Hence, it can be justified to say that external gene-specific biological information is predominantly important for efficient feature learning through a highly masked self-supervised approach. We take an integrated node representation approach to study the overlapping genetic association across comorbid disease pairs, as every genetic association's auxiliary information is highly significant for discovering the sharing trait mechanisms between the prevalent diseases.

In summary, in search of comorbid disease sharing core genetic associations, we take advantage of high-relational hypergraph information node representation accompanied by the least number of edge-label-based strategies by imposing a mask auto-encoder-based self-supervised model to get a robust outcome in considerably accelerated learning criteria on disease-specific gene co-expression graphs.

There still exists scope for further modification of our method. In our method, we only use five types of biological information to build the auxiliary bipartite graph. So, here, we can integrate additional genetic and cellular information with a more guided learning approach. Next, we can modify the hyperedge construction by introducing more than 1-hop to accommodate more than one-degree neighborhood information. Finally, we hope that our generalized method for predicting genetic association on comorbid diseases will be inevitably useful for the in-depth pathogenesis of inter-related complex comorbid disease mechanisms, which can be used to propose better therapeutics and shed light on the recent advent of drug repurposing hypothesis.

REFERENCES

[1] S. Gjerstad and A. Molle, "Comorbidity factors influence covid-19 mortality much more than age," 2020.

[2] K. Barnett, S. W. Mercer, M. Norbury, G. Watt, S. Wyke, and B. Guthrie, "Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study," *The Lancet*, vol. 380, no. 9836, pp. 37–43, 2012.

[3] A. Dugravot, A. Fayosse, J. Dumurgier, K. Bouillon, T. B. Rayana, A. Schnitzler, M. Kivimaki, S. Sabia, and A. Singh-Manoux, "Social inequalities in multimorbidity, frailty, disability, and transitions to mortality: a 24-year follow-up of the whitehall ii cohort study," *The Lancet Public Health*, vol. 5, no. 1, pp. e42–e50, 2020.

[4] J. Sánchez-Valle and A. Valencia, "Molecular bases of comorbidities: present and future perspectives," *Trends in Genetics*, 2023.

[5] L. Cheng, H. Zhao, P. Wang, W. Zhou, M. Luo, T. Li, J. Han, S. Liu, and Q. Jiang, "Computational methods for identifying similar diseases," *Molecular Therapy-Nucleic Acids*, vol. 18, pp. 590–604, 2019.

[6] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome biology*, vol. 18, no. 1, pp. 1–15, 2017.

[7] V. A. Yépez, C. Mertes, M. F. Müller, D. Klaproth-Andrade, L. Wachutka, L. Frésard, M. Gusic, I. F. Scheller, P. F. Goldberg, H. Prokisch *et al.*, "Detection of aberrant gene expression events in rna sequencing data," *Nature Protocols*, vol. 16, no. 2, pp. 1276–1296, 2021.

[8] S. Van Dam, U. Vosa, A. van der Graaf, L. Franke, and J. P. de Magalhaes, "Gene co-expression analysis for functional classification and gene–disease predictions," *Briefings in bioinformatics*, vol. 19, no. 4, pp. 575–592, 2018.

[9] J. Zhu, H. Meng, L. Zhang, and Y. Li, "Exploring the molecular mechanism of comorbidity of autism spectrum disorder and inflammatory bowel disease by combining multiple data sets," *Journal of Translational Medicine*, vol. 21, no. 1, p. 372, 2023.

[10] Z. Wang, Z. Meng, and C. Chen, "Screening of potential biomarkers in peripheral blood of patients with depression based on weighted gene co-expression network analysis and machine learning algorithms," *Frontiers in Psychiatry*, vol. 13, p. 1009911, 2022.

[11] J. Das and H. Yu, "Hint: High-quality protein interactomes and their applications in understanding human disease," *BMC systems biology*, vol. 6, no. 1, pp. 1–12, 2012.

[12] Y. Kong and T. Yu, "A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data," *Bioinformatics*, vol. 34, no. 21, pp. 3727–3737, 2018.

[13] A. S. Cristino, S. M. Williams, Z. Hawi, J. An, M. A. Bellgrove, C. E. Schwartz, L. d. F. Costa, and C. Claudianos, "Neurodevelopmental and neuropsychiatric disorders represent an interconnected molecular system," *Molecular psychiatry*, vol. 19, no. 3, pp. 294–301, 2014.

[14] J. Shen, Y. Feng, M. Lu, J. He, and H. Yang, "Predictive model, mirna-tf network, related subgroup identification and drug prediction of ischemic stroke complicated with mental disorders based on genes related to gut microbiome," *Frontiers in Neurology*, vol. 14, p. 1189746, 2023.

[15] Y. Ma and Q. Liu, "Generalized matrix factorization based on weighted hypergraph learning for microbe-drug association prediction," *Computers in Biology and Medicine*, vol. 145, p. 105503, 2022.

[16] Y. Ma and J. Zhong, "Logistic tensor decomposition with sparse subspace learning for prediction of multiple disease types of human–virus protein–protein interactions," *Briefings in Bioinformatics*, vol. 24, no. 1, p. bbac604, 2023.

[17] Y. Ma, Y. Zhao, and Y. Ma, "Kernel bayesian nonlinear matrix factorization based on variational inference for human–virus protein–protein interaction prediction," *Scientific Reports*, vol. 14, no. 1, p. 5693, 2024.

[18] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[19] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3558–3565.

[20] X. Sun, H. Yin, B. Liu, H. Chen, J. Cao, Y. Shao, and N. Q. Viet Hung, "Heterogeneous hypergraph embedding for graph classification," in *Proceedings of the 14th ACM international conference on web search and data mining*, 2021, pp. 725–733.

[21] Y. Gao, Y. Feng, S. Ji, and R. Ji, "Hgnn+: General hypergraph neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3181–3199, 2022.

[22] Y. Ma and Y. Ma, "Hypergraph-based logistic matrix factorization for metabolite–disease interaction prediction," *Bioinformatics*, vol. 38, no. 2, pp. 435–443, 2022.

[23] Y. Ding, L. Jiang, J. Tang, and F. Guo, "Identification of human microrna-disease association via hypergraph embedded bipartite local model," *Computational Biology and Chemistry*, vol. 89, p. 107369, 2020.

[24] L. Wu, H. Lin, C. Tan, Z. Gao, and S. Z. Li, "Self-supervised learning on graphs: Contrastive, generative, or predictive," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[25] R. Krishnan, P. Rajpurkar, and E. J. Topol, "Self-supervised learning in medicine and healthcare," *Nature Biomedical Engineering*, vol. 6, no. 12, pp. 1346–1352, 2022.

[26] V. S. Bharadhwaj, S. Mubeen, A. Sargsyan, G. M. Jose, S. Geissler, M. Hofmann-Apitius, D. Domingo-Fernández, and A. T. Kodamullil, "Integrative analysis to identify shared mechanisms between schizophrenia and bipolar disorder and their comorbidities," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 122, p. 110688, 2023.

[27] J. Sánchez-Valle, H. Tejero, J. M. Fernández, D. Juan, B. Urda-García, S. Capella-Gutiérrez, F. Al-Shahrour, R. Tabarés-Seisdedos, A. Baudot, V. Pancaldi *et al.*, "Interpreting molecular similarity between patients as a determinant of disease comorbidity relationships," *Nature Communications*, vol. 11, no. 1, p. 2854, 2020.

[28] S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie, and A. J. Butte, "Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets," *PLoS computational biology*, vol. 6, no. 2, p. e1000662, 2010.

[29] B. Urda-García, J. Sánchez-Valle, R. Lepore, and A. Valencia, "Patient stratification reveals the molecular basis of disease comorbidities," *medRxiv*, pp. 2021–07, 2021.

[30] T. Gaudelet, N. Malod-Dognin, J. Sánchez-Valle, V. Pancaldi, A. Valencia, and N. Pržulj, "Unveiling new disease, pathway, and gene associations via multi-scale neural network," *PloS one*, vol. 15, no. 4, p. e0231059, 2020.

[31] P. H. Guzzi, F. Cortese, G. C. Mannino, E. Pedace, E. Succurro, F. Andreozzi, and P. Veltri, "Analysis of age-dependent gene-expression in human tissues for studying diabetes comorbidities," *Scientific Reports*, vol. 13, no. 1, p. 10372, 2023.

[32] W. Li, L. Wang, Y. Wu, Z. Yuan, and J. Zhou, "Weighted gene co-expression network analysis to identify key modules and hub genes associated with atrial fibrillation," *International Journal of Molecular Medicine*, vol. 45, no. 2, pp. 401–416, 2020.

[33] M. Kim, J. R. Haney, P. Zhang, L. M. Hernandez, L.-k. Wang, L. Perez-Cano, L. M. O. Loohuis, L. de la Torre-Ubieta, and M. J. Gandal, "Brain gene co-expression networks link complement signaling with convergent synaptic pathology in schizophrenia," *Nature neuroscience*, vol. 24, no. 6, pp. 799–809, 2021.

[34] Y. Ma, J. Zhong, and N. Zhu, "Weighted hypergraph learning and adaptive inductive matrix completion for sars-cov-2 drug repositioning," *Methods*, vol. 219, pp. 102–110, 2023.

[35] A. Jain, M.-L. Charpignon, I. Y. Chen, A. Philippakis, and A. Alaa, "Generating new drug repurposing hypotheses using disease-specific hypergraphs," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*. World Scientific, 2023, pp. 261–275.

[36] R. Viñas, C. K. Joshi, D. Georgiev, P. Lin, B. Dumitrascu, E. R. Gamazon, and P. Liò, "Hypergraph factorization for multi-tissue gene expression imputation," *Nature machine intelligence*, vol. 5, no. 7, pp. 739–753, 2023.

[37] J. Burke, A. Akbari, R. Bailey, K. Fasusi, R. A. Lyons, J. Pearson, J. Rafferty, and D. Schofield, "Representing multimorbid disease progressions using directed hypergraphs," *medRxiv*, pp. 2023–08, 2023.

[38] X. Li, M. Jia, M. T. Islam, L. Yu, and L. Xing, "Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4023–4033, 2020.

[39] W. Han, Y. Cheng, J. Chen, H. Zhong, Z. Hu, S. Chen, L. Zong, L. Hong, T.-F. Chan, I. King *et al.*, "Self-supervised contrastive learning for integrative single cell rna-seq data analysis," *Briefings in Bioinformatics*, vol. 23, no. 5, p. bbac377, 2022.

[40] H. Kirchner, I. Sinha, H. Gao, M. A. Ruby, M. Schönke, J. M. Lindvall, R. Barrès, A. Krook, E. Näslund, K. Dahlman-Wright *et al.*, "Altered dna methylation of glycolytic and lipogenic genes in liver from obese and type 2 diabetic patients," *Molecular metabolism*, vol. 5, no. 3, pp. 171–183, 2016.

[41] R. Shamir, C. Klein, D. Amar, E.-J. Vollstedt, M. Bonin, M. Usenovic, Y. C. Wong, A. Maver, S. Poths, H. Safer *et al.*, "Analysis of blood-based gene expression in idiopathic parkinson disease," *Neurology*, vol. 89, no. 16, pp. 1676–1683, 2017.

[42] A. Hodges, A. D. Strand, A. K. Aragaki, A. Kuhn, T. Sengstag, G. Hughes, L. A. Elliston, C. Hartog, D. R. Goldstein, D. Thu *et al.*, "Regional and cellular gene expression changes in human huntington's disease brain," *Human molecular genetics*, vol. 15, no. 6, pp. 965–977, 2006.

[43] L. Jones, D. R. Goldstein, G. Hughes, A. D. Strand, F. Collin, S. B. Dunnett, C. Kooperberg, A. Aragaki, J. M. Olson, S. J. Augood *et al.*, "Assessment of the relationship between pre-chip and post-chip quality measures for affymetrix genechip expression data," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–18, 2006.

[44] T. A. Lanz, V. Reinhart, M. J. Sheehan, S. J. S. Rizzo, S. E. Bove, L. C. James, D. Volfson, D. A. Lewis, and R. J. Kleiman, "Postmortem transcriptional profiling reveals widespread increase in inflammation in schizophrenia: a comparison of prefrontal cortex, striatum, and hippocampus among matched tetrads of controls with subjects diagnosed with schizophrenia, bipolar or major depressive disorder," *Translational psychiatry*, vol. 9, no. 1, p. 151, 2019.

[45] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online mendelian inheritance in man (omim), a knowledge-base of human genes and genetic disorders," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D514–D517, 2005.

[46] H.-Y. Huang, Y.-C.-D. Lin, J. Li, K.-Y. Huang, S. Shrestha, H.-C. Hong, Y. Tang, Y.-G. Chen, C.-N. Jin, Y. Yu *et al.*, "mirtarbase 2020: updates to the experimentally validated microrna–target interaction database," *Nucleic acids research*, vol. 48, no. D1, pp. D148–D154, 2020.

[47] Q. Zhang, W. Liu, H.-M. Zhang, G.-Y. Xie, Y.-R. Miao, M. Xia, and A.-Y. Guo, "htftarget: a comprehensive database for regulations of human transcription factors and their targets," *Genomics, proteomics & bioinformatics*, vol. 18, no. 2, pp. 120–128, 2020.

[48] P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, no. 1, pp. 1–13, 2008.

[49] F. J. Azuaje, "Selecting biologically informative genes in co-expression networks with a centrality score," *Biology direct*, vol. 9, no. 1, pp. 1–23, 2014.

[50] F. J. Azuaje, S. Rodius, L. Zhang, Y. Devaux, and D. R. Wagner, "Information encoded in a network of inflammation proteins predicts clinical outcome after myocardial infarction," *BMC medical genomics*, vol. 4, no. 1, pp. 1–10, 2011.

[51] F. E. Dewey, M. V. Perez, M. T. Wheeler, C. Watt, J. Spin, P. Langfelder, S. Horvath, S. Hannenhalli, T. P. Cappola, and E. A. Ashley, "Gene coexpression network topology of cardiac development, hypertrophy, and failure," *Circulation: cardiovascular genetics*, vol. 4, no. 1, pp. 26–35, 2011.

[52] F. Azuaje, L. Zhang, C. Jeanty, S.-L. Puhl, S. Rodius, and D. R. Wagner, "Analysis of a gene co-expression network establishes robust association between col5a2 and ischemic heart disease," *BMC medical genomics*, vol. 6, pp. 1–10, 2013.

[53] B. Fa, C. Luo, Z. Tang, Y. Yan, Y. Zhang, and Z. Yu, "Pathway-based biomarker identification with crosstalk analysis for robust prognosis prediction in hepatocellular carcinoma," *EBioMedicine*, vol. 44, pp. 250–260, 2019.

[54] G. Muzio, L. O'Bray, and K. Borgwardt, "Biological network analysis with deep learning," *Briefings in bioinformatics*, vol. 22, no. 2, pp. 1515–1530, 2021.

[55] H. Wang, P. Sham, T. Tong, and H. Pang, "Pathway-based single-cell rna-seq classification, clustering, and construction of gene-gene interactions networks using random forests," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 6, pp. 1814–1822, 2019.

[56] X. Xing, F. Yang, H. Li, J. Zhang, Y. Zhao, M. Gao, J. Huang, and J. Yao, "Multi-level attention graph neural network based on co-expression gene modules for disease diagnosis and prognosis," *Bioinformatics*, vol. 38, no. 8, pp. 2178–2186, 2022.

[57] S. Rhee, S. Seo, and S. Kim, "Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification."

[58] C. Chen and I. Rajapakse, "Tensor entropy for uniform hypergraphs," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 4, pp. 2889–2900, 2020.

[59] C. Chen, A. Surana, A. M. Bloch, and I. Rajapakse, "Controllability of hypergraphs," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 2, pp. 1646–1657, 2021.

[60] Y. Gao, Z. Zhang, H. Lin, X. Zhao, S. Du, and C. Zou, "Hypergraph learning: Methods and practices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2548–2566, 2020.

[61] C. Berge, *Hypergraphs: combinatorics of finite sets*. Elsevier, 1984, vol. 45.

[62] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," *Advances in neural information processing systems*, vol. 19, 2006.

[63] A. Banerjee, A. Char, and B. Mondal, "Spectra of general hypergraphs," *Linear Algebra and its Applications*, vol. 518, pp. 14–30, 2017.

[64] S. Bai, F. Zhang, and P. H. Torr, "Hypergraph convolution and hypergraph attention," *Pattern Recognition*, vol. 110, p. 107637, 2021.

[65] Q. Tan, N. Liu, X. Huang, S.-H. Choi, L. Li, R. Chen, and X. Hu, "S2gae: Self-supervised graph autoencoders are generalizable learners

with graph masking," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023, pp. 787–795.
[66] Q. Tan, N. Liu, X. Huang, R. Chen, S.-H. Choi, and X. Hu, "Mgae: Masked autoencoders for self-supervised learning on graphs," *arXiv preprint arXiv:2201.02534*, 2022.
[67] L. Getoor and C. P. Diehl, "Link mining: a survey," *Acm Sigkdd Explorations Newsletter*, vol. 7, no. 2, pp. 3–12, 2005.
[68] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," *Advances in neural information processing systems*, vol. 31, 2018.
[69] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *arXiv preprint arXiv:1611.07308*, 2016.
[70] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.
[71] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," *arXiv preprint arXiv:1809.10341*, 2018.
[72] C. Mavromatis and G. Karypis, "Graph infoclust: Leveraging cluster-level node information for unsupervised graph representation learning," *arXiv preprint arXiv:2009.06946*, 2020.
[73] Z. Zhou, L. Zhuo, X. Fu, J. Lv, Q. Zou, and R. Qi, "Joint masking and self-supervised strategies for inferring small molecule-mirna associations," *Molecular Therapy-Nucleic Acids*, vol. 35, no. 1, 2024.
[74] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann *et al.*, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic acids research*, vol. 44, no. W1, pp. W90–W97, 2016.
[75] M. B. Sadeghi, A. Nakhaee, R. Saravani, and S. Sargazi, "Significant association of lxrβ (nr1h2) polymorphisms (rs28514894, rs2303044) with type 2 diabetes mellitus and laboratory characteristics," *Journal of Diabetes & Metabolic Disorders*, vol. 20, pp. 261–270, 2021.
[76] H. Katsuki, "Nuclear receptors of nr1 and nr4 subfamilies in the regulation of microglial functions and pathology," *Pharmacology Research & Perspectives*, vol. 9, no. 6, p. e00766, 2021.
[77] M.-E. Tremblay, M. R. Cookson, and L. Civiero, "Glial phagocytic clearance in parkinson's disease," *Molecular neurodegeneration*, vol. 14, pp. 1–14, 2019.
[78] X. Liu, Z. Wang, Y. Yang, Q. Li, R. Zeng, J. Kang, and J. Wu, "Rab1a mediates proinsulin to insulin conversion in β-cells by maintaining golgi stability through interactions with golgin-84," *Protein & cell*, vol. 7, no. 9, pp. 692–696, 2016.
[79] Y. Gao, G. R. Wilson, S. E. Stephenson, K. Bozaoglu, M. J. Farrer, and P. J. Lockhart, "The emerging role of rab gtpases in the pathogenesis of parkinson's disease," *Movement Disorders*, vol. 33, no. 2, pp. 196–207, 2018.
[80] U. Galicia-Garcia, A. Benito-Vicente, S. Jebari, A. Larrea-Sebal, H. Siddiqi, K. B. Uribe, H. Ostolaza, and C. Martín, "Pathophysiology of type 2 diabetes mellitus," *International journal of molecular sciences*, vol. 21, no. 17, p. 6275, 2020.
[81] H. Li, F. Zeng, C. Huang, Q. Pu, E. R. Thomas, Y. Chen, and X. Li, "The potential role of glucose metabolism, lipid metabolism, and amino acid metabolism in the treatment of parkinson's disease," *CNS Neuroscience & Therapeutics*, 2023.
[82] C. Zhuo, F. Zhao, H. Tian, J. Chen, Q. Li, L. Yang, J. Ping, R. Li, L. Wang, Y. Xu *et al.*, "Acid sphingomyelinase/ceramide system in schizophrenia: implications for therapeutic intervention as a potential novel target," *Translational Psychiatry*, vol. 12, no. 1, p. 260, 2022.
[83] Y. Lin, L. Ran, X. Du, H. Yang, and Y. Wu, "Oxysterol-binding protein: new insights into lipid transport functions and human diseases," *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, p. 159365, 2023.
[84] Y. Zhang, X. Gao, X. Bai, S. Yao, Y.-Z. Chang, and G. Gao, "The emerging role of furin in neurodegenerative and neuropsychiatric diseases," *Translational Neurodegeneration*, vol. 11, no. 1, pp. 1–28, 2022.
[85] M. Watanabe and S. Hatakeyama, "Trim proteins and diseases," *The journal of biochemistry*, vol. 161, no. 2, pp. 135–144, 2017.
[86] M. Logotheti, O. Papadodima, N. Venizelos, A. Chatziioannou, F. Kolisis *et al.*, "A comparative genomic study in schizophrenic and in bipolar disorder patients, based on microarray expression profiling meta-analysis," *The Scientific world journal*, vol. 2013, 2013.
[87] S. Gladyck, S. Aras, M. Hüttemann, and L. I. Grossman, "Regulation of cox assembly and function by twin cx9c proteins—implications for human disease," *Cells*, vol. 10, no. 2, p. 197, 2021.
[88] P. M. Abruzzo, C. Panisi, and M. Marini, "The alteration of chloride homeostasis/gabaergic signaling in brain disorders: could oxidative stress play a role?" *Antioxidants*, vol. 10, no. 8, p. 1316, 2021.
[89] Y.-T. Hsu, Y.-G. Chang, and Y. Chern, "Insights into gabaaergic system alteration in huntington's disease," *Royal Society Open Biology*, vol. 8, no. 12, p. 180165, 2018.
[90] N. Hoffmann and J. Peters, "Functions of the (pro) renin receptor (atp6ap2) at molecular and system levels: pathological implications in hypertension, renal and brain development, inflammation, and fibrosis," *Pharmacological Research*, vol. 173, p. 105922, 2021.
[91] J. R. Liddell, "Targeting mitochondrial metal dyshomeostasis for the treatment of neurodegeneration," *Neurodegenerative Disease Management*, vol. 5, no. 4, pp. 345–364, 2015.

Saikat Biswas received his Master degree in Computer Science and Engineering from Jadavpur University, Kolkata. He is a doctoral student at Advanced Technology Development Centre, IIT Kharagpur. His research focuses on data mining and computational biology.

Vibhanshu Ranjan received his integrated dual degree (Btech + Mtech) in Electrical Engineering, IIT Kharagpur. He has done many projects on deep learning and computational biology.

Pabitra Mitra received his Ph.D. degree in Computer Science from Indian Statistical Institute, Kolkata. He is a professor at the Department of Computer Science and Engineering, IIT Kharagpur. His research focuses on machine learning, pattern recognition, data mining, information retrieval, and computational biology. He is a senior member of IEEE.

Krothapalli Sreenivasa Rao received his Ph.D. degree in Computer Science and Engineering from IIT Madras. He is a professor at the Department of Computer Science and Engineering, IIT Kharagpur. His research focuses on signal processing, speech processing, machine learning, pattern recognition, and biomedical engineering. He is a senior member of IEEE.