BMC Bioinformatics

RESEARCH

Open Access

A deep learning framework for predicting disease-gene associations with functional modules and graph augmentation

Xianghu Jia[1†], Weiwen Luo[1†], Jiaqi Li[1†], Jieqi Xing[1], Hongjie Sun[1], Shunyao Wu[1*] and Xiaoquan Su[1*]

[†]Xianghu Jia, Weiwen Luo and Jiaqi Li have contributed equally, co-first authors.

*Correspondence: wushunyao@qdu.edu.cn; suxq@qdu.edu.cn

[1] College of Computer Science and Technology, Qingdao University, Qingdao 266071, Shandong, China

Abstract

Background: The exploration of gene-disease associations is crucial for understanding the mechanisms underlying disease onset and progression, with significant implications for prevention and treatment strategies. Advances in high-throughput biotechnology have generated a wealth of data linking diseases to specific genes. While graph representation learning has recently introduced groundbreaking approaches for predicting novel associations, existing studies always overlooked the cumulative impact of functional modules such as protein complexes and the incompletion of some important data such as protein interactions, which limits the detection performance.

Results: Addressing these limitations, here we introduce a deep learning framework called ModulePred for predicting disease-gene associations. ModulePred performs graph augmentation on the protein interaction network using L3 link prediction algorithms. It builds a heterogeneous module network by integrating disease-gene associations, protein complexes and augmented protein interactions, and develops a novel graph embedding for the heterogeneous module network. Subsequently, a graph neural network is constructed to learn node representations by collectively aggregating information from topological structure, and gene prioritization is carried out by the disease and gene embeddings obtained from the graph neural network. Experimental results underscore the superiority of ModulePred, showcasing the effectiveness of incorporating functional modules and graph augmentation in predicting disease-gene associations. This research introduces innovative ideas and directions, enhancing the understanding and prediction of gene-disease relationships.

Keywords: Gene-disease associations, Deep learning, Graph augmentation, Protein complexes, Graph neural networks

Introduction

Gene mutations or genetic abnormalities play a pivotal role in the pathogenesis of various diseases. Consequently, uncovering the associations between genes and diseases is imperative to elucidate the underlying molecular mechanisms and enhance healthcare. While linkage analysis and genome-wide association studies are capable of detecting biomarkers, such as single nucleotide polymorphisms (SNPs), by examining genetic

variations within human populations, these approaches are time and resource-intensive due to the necessity of analyzing numerous false positives [1]. Moreover, these methods primarily focus on direct connections between genotypes and phenotypes, thereby overlooking the complex interactions between molecules [2].

Recent years, computational methods rooted in molecular networks have emerged as a prominent approach to complement and enhance linkage analysis and genome-wide

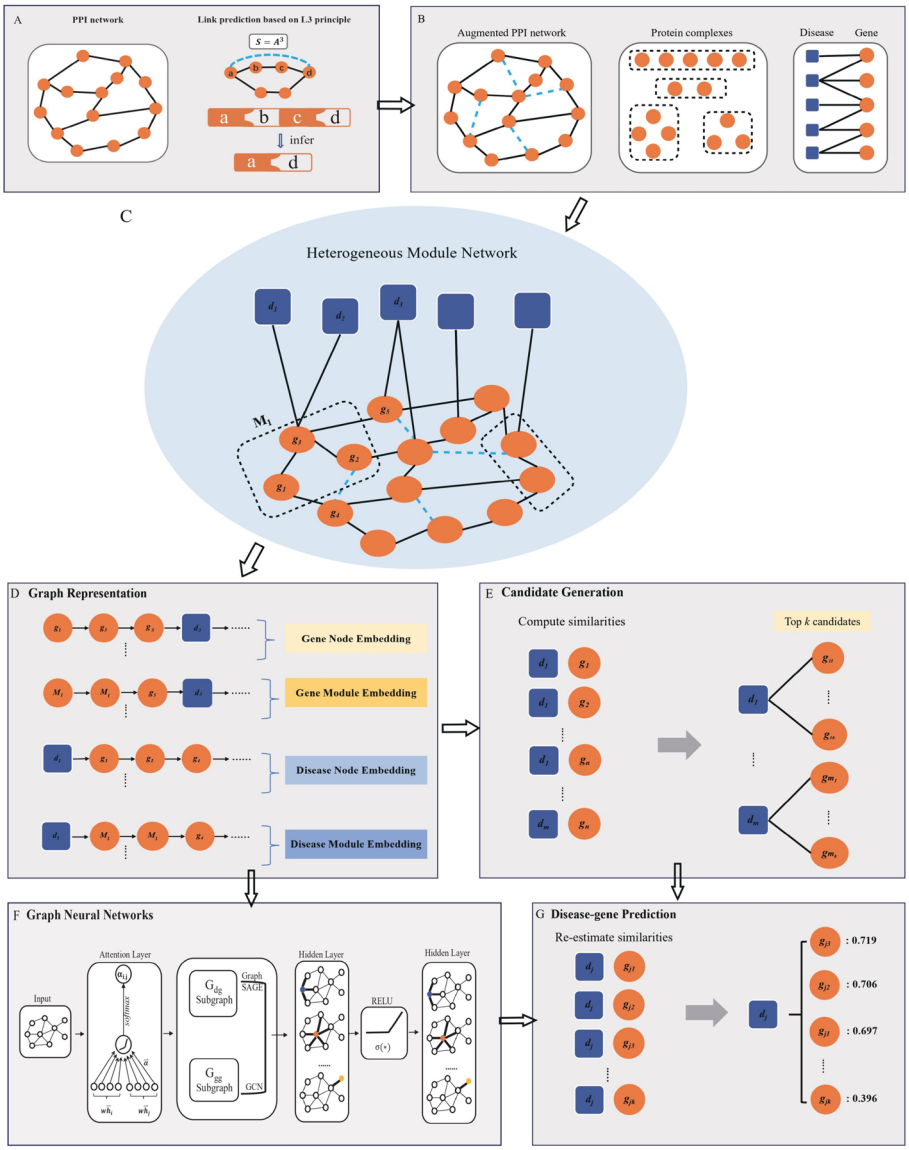

Fig. 1 An overview of our proposed approach. Firstly, Data augmentation was performed on the protein–protein interaction (PPI) network with L3 principle (**A**). Then, by integrating augmented PPI network, protein complexes and disease-gene associations (**B**), a heterogeneous module network was built (**C**). Subsequently, initial low-dimensional embeddings were obtained by graph representation (**D**) for the heterogeneous module network and candidate genes were generated for each disease (**E**). Furthermore, a graph neural network was constructed to learn better representations by collectively aggregating information from topological structure (**F**). Finally, for each disease, the candidate genes were scored and re-ranked based on the embeddings generated by the graph neural network (**G**)

association studies, providing valuable insights into disease gene prediction [3–5]. The primary objective is to extract topological features that precisely capture the intricate connections between genes and diseases, including measures of topological similarity between genes and diseases [6–8], as well as other artificially extracted features [9–11]. Notably, graph embedding methods such as node2vec and graph neural networks like graph convolutional network (GCN) have witnessed extensive application in gene-disease association mining, showcasing commendable performance by automatically discovering potent latent features [12, 13]. Despite significant strides in existing research, certain issues impede detection performance, including the oversight in investigating cooperative relationships among molecules. For instance, in cellular activities, proteins often depend on collaborative interactions within protein complexes to execute specific functions [14]. Additionally, the effectiveness of disease gene prediction faces substantial hindrance due to the incompleteness of existing molecular networks, notably the protein interaction network, which lacks experimental validation for numerous interactions.

This paper introduces a novel paradigm centered on modules to encapsulate cooperative relationships among molecules, particularly focusing on protein complexes. We present ModulePred, an advanced deep learning framework designed for the purpose of mining gene-disease associations. To tackle the issue of data incompleteness, we initiate the process by conducting data augmentation on the protein interaction network through L3-based link prediction algorithms (Fig. 1A). L3-based link prediction algorithms integrate biological motivations into the prediction of protein–protein interactions, surpassing the performance of general-purpose algorithms [15]. Subsequently, the establishment of a heterogeneous module network (Fig. 1C) unfolds, seamlessly integrating disease-gene associations, augmented protein interactions, and protein complexes (Fig. 1B). Within this framework, a sophisticated graph embedding method is devised to harness the cooperative relationships intrinsic to the heterogeneous module network (Fig. 1D), subsequently deploying this method to generate candidate genes for each disease (Fig. 1E). Furthermore, a graph neural network is engineered to glean enhanced representations by collectively aggregating information from the topological structure (Fig. 1F). Ultimately, low-dimensional disease and gene embeddings are harnessed for gene prioritization (Fig. 1G).

Materials and methods

Graph data augmentation based on L3 principle

Even with significant advancements in high-throughput mapping techniques, a considerable number of human protein–protein interactions (PPIs) remain unknown compared to those that have been experimentally documented [16]. Network-based link prediction algorithms are gaining momentum as valuable computational tools for predicting undetected interactions. Such state-of-the-art algorithms rely on the triadic closure principle, which assumes that the number of paths of length two between two nodes is correlated with the likelihood of them also being directly connected. However, the triadic closure principle inadequately characterize PPIs, thereby failing to guarantee the correctness and reliability of predictions. Figure 1A illustrates that protein a and protein c share a path of length 2, indicating a potential interaction based on the triadic closure principle. PPIs often require complementary interfaces [17, 18]. As a result, protein a and protein

c exhibit similar interfaces, as illustrated by their identical shapes in Fig. 1A. It is notable that such an interface does not typically guarantee that protein a and protein c interact with each other [15].

To address the aforementioned issue, Kovács et al.[15] proposed a novel link prediction predictor based on the L3 principle, positing that proteins linked by multiple paths of length three are more likely to have a direct link. As shown in Fig. 1A, an additional interaction partner of protein c (protein d) and protein a have a complementary interface, suggesting a possible direct interaction. Such an interaction can be predicted by using paths of length three (L3). In this paper, we adopted the L3 principle to perform data augmentation on the protein interaction network. Three L3 scores are assigned to each node pair, x and y (Eqs. 1–3)

$$L_3^{CN}(x,y) = \sum_{u,v} a_{xu}a_{uv}a_{vy} \tag{1}$$

$$L_3^{RA}(x,y) = \sum_{u,v} a_{xu}a_{uv}a_{vy}\left(\frac{1}{k_u} + \frac{1}{k_v}\right) \tag{2}$$

$$L_3^{AA}(x,y) = \sum_{u,v} a_{xu}a_{uv}a_{vy}\left(\frac{1}{logk_u} + \frac{1}{logk_v}\right) \tag{3}$$

where k_u represents the degree of node u while a_{xu} is a binary variable. $a_{xu} = 1$ if node x interacts with node u interacts, otherwise $a_{xu} = 0$. L_3^{RA} And L_3^{AA} are degree-normalized versions of L_3^{CN}, derived from the insights obtained from RA (Resource Allocation) and AA (Adamic-Adar) [19]. When performing data augmentation, taking protein x as an example, first calculate similarity scores with all remaining nodes (excluding those already connected to x). Then, select the top l nodes with the highest similarity to x for L_3^{CN}, L_3^{RA}, and L_3^{AA} respectively. The selected node sets are denoted as S_{CN}, S_{RA}, and S_{AA}. Lastly, create edges between x and each node in the set $S = S_{CN} \cup S_{RA} \cup S_{AA}$.

Graph representation for the heterogeneous module network and candidates generation

As illustrated in Fig. 1C, a heterogeneous module network, denoted as $G = (V, E)$, was constructed by integrating disease-gene associations, augmented protein–protein interactions, and protein complexes (Fig. 1B). In this network, the node set V, consists of disease and gene nodes, with $V = V_d \cup V_g$. And the edge set E, includes disease-gene associations and protein–protein interactions, $E = E_{dg} \cup E_{gg}$. For simplicity, protein nodes are referred to as gene nodes, and protein interactions are represented as gene interactions. Certain nodes, such as x and y, exhibit cooperative relationships and belong to a module, denoted as M_1. This can be expressed as $x \in M_1$, $y \in M_1$, or $M_1 = \{x, y\}$. M_1 is a member of the module set M that comprises of protein complexes.

In this study, Node2vec [20], a prevalent network embedding algorithm, was introduced to extract low-dimensional node representations from the heterogeneous module network. Firstly, we utilized random walks to generate multiple neighbor sequences for each node. It should be noted that two types of sequences were generated for each node: the conventional node sequences Q^n and enhanced sequences

Q^m that incorporate both nodes and modules. As depicted in Fig. 1D, the sequence $q_1^n = g_1 \rightarrow g_3 \rightarrow g_5 \rightarrow d_3 \ldots$ is a walk sequence starting from g_1 that only contains node. By replacing gene nodes with their corresponding module numbers (both g_1 and g_3 belong toM_1, so they are both replaced withM_1), the sequence q_1^n can be transformed into $q_1^m = M_1 \rightarrow M_1 \rightarrow g_5 \rightarrow d_3 \ldots$. Here, $q_1^n \in Q^n$ and$q_1^m \in Q^m$. Then, all the sequences of Q^n were treated as texts, where nodes were considered as words, and the skip-gram model, a typical natural language processing model, was applied to learn the node embeddings. Similarly, all the sequences of Q^m were provided to the skip-gram model to learn the module embeddings. If a node does not belong to any module, its node embeddings were used as its module embeddings.

For each disease, we computed cosine similarities between its node embedding and the embeddings of all gene nodes. Then, we selected the top-k genes with the highest similarity as candidates for each disease (Fig. 1E). In the disease gene prediction stage, we focused only on calculating similarities between each disease and its candidate genes, significantly reducing the computational complexity.

Graph neural networks for the heterogeneous module network

A graph neural network was constructed based on the graph representation, aimed at improving the learning of low-dimensional node representations by aggregating information from the topological structure. The embedding vectors obtained from the graph representation served as initial node features for the graph neural network. In the graph neural network architecture (Fig. 1F), a graph attention network was initially employed to assign different weights to neighbors for updating node information. Subsequently, two graph convolutional layers were applied to protein interactions, while two GraphSage layers were used for disease-gene associations.

The heterogeneous module network employed the Graph attention network (GAT) to compute the hidden states of each node through a self-attention strategy. This can be defined by Eqs. 4 and 5:

$$H_i^1 = \sum_{j \in N_i} \alpha_{i,j} W^0 H_j^0 \tag{4}$$

$$\alpha_{i,j} = softmax_j(e_{ij}) = softmax_j(LeakyReLU(\overrightarrow{a}^T [W^0 H_i^0 || W^0 H_j^0])) \tag{5}$$

where N_i represents the neighborhood set of node i, W^0 is a trainable weight matrix, H_j^0 is the initial features of node j obtained from graph representation, and H_i^1 denotes the embedding vector of node i obtained by GAT. A shared attentional mechanism $a : {}^F \times {}^F \rightarrow$ (F represents the number of the node features output by the layer) is performed on the nodes to compute attention coefficients $e_{ij} = a(W^0 H_i^0, W^0 H_j^0)$ that represent the importance of node j 's features to node i. $\alpha_{i,j}$ is calculated by normalizing e_{ij} with the softmax function. $\overrightarrow{a}$ is a weight vector to parameterize the single-layer feedforward neural network that forms the attention mechanism a. $[W^0 H_i^0 || W^0 H_j^0]$ signifies the concatenation of $W^0 H_i^0$ and $W^0 H_j^0$, and $LeakyReLU$ is the activation

function. Specifically, $H_j^0 = [H_j^{node}||H_j^{module}]$, where H_j^{node} and H_j^{module} represent the node embedding and module embedding of node j obtained from the graph representation, respectively.

For the subgraph G_{gg}, the convolution operation was conducted by the graph convolutional layer. Graph convolutional layer can be defined as *Eq.* 6:

$$H_i^{k+1} = \sigma\left(\sum_{j\in N_i} \frac{1}{c_{ji}} H_j^k W^k + b^k\right) \tag{6}$$

where $c_{ji} = \sqrt{|N_j|} \times \sqrt{|N_i|}$, b^k is a trainable bias matrix, and W^k is a trainable weight matrix. The activation function σ, set as RELU in this paper, is applied to the layer. H_j^{k+1} ($k \geq 1$) represents the embedding vector of node j in the $k+1_{th}$ layer, and H_j^1 captures the information of node j obtained by GAT.

GraphSage layer was adopted for the subgraph G_{dg}. In contrast to the graph convolutional layer that utilizes the full neighborhood set, GraphSage layer samples a specific proportion of neighbors to aggregate information. The embedding process of GraphSAGE is defined by Eqs. 7 and 8:

$$H_{N_i'}^{k+1} = AGG_{k+1}\left(\left\{H_j^k, \forall j \in N_i'\right\}\right) \tag{7}$$

$$H_i^{k+1} = \sigma(W^{k+1} \cdot [H_i^k||H_{N_i'}^{k+1}]) \tag{8}$$

where $N_i\prime$ represents a subset from the neighborhood set N_i. The aggregation function, denoted as AGG_{k+1}, was chosen as the mean aggregator in this study, and hence GraphSage takes the mean over neighbors of node i according to Eq. 7. Different with the graph convolutional layer, GraphSAGE concatenates the node representation with the mean aggregation of neighbor nodes as shown in Eq. 8, which avoids node information loss.

The outputs of the various convolutional layers were aggregated to incorporate information from all types of edges for each node. In this study, two layers were constructed for GCN and Graphs sage, which has demonstrated strong performance in prior research [21, 22]. Our ablation experiments also demonstrated that setting the number of layers to 2 for both GraphSage and GCN can achieve good results. Please refer to Sect. "Ablation study" and Supplementary Figs. S1 and S2.

Training and prediction

Denote the embedding of node i obtained from the graph neural network as H_i. To evaluate the strength of the association for a disease-gene pair (d_i, g_j), we employed cosine similarity (Eq. 9) as a measure:

$$score_{ij} = \frac{\widetilde{H}_i \cdot \widetilde{H}_j}{\left|\widetilde{H}_i\right|\left|\widetilde{H}_j\right|} \tag{9}$$

where $\widetilde{H}_i = [H_i||H_i^{node}]$, H_i^{node} represents the node embedding obtained from node2vec and $|\widetilde{H}_i|$ is the norm of$\widetilde{H}_i$.

During the training phase, negative samples were randomly selected from all unconnected pairs between diseases and genes. Due to the fact that the connected gene-disease

pairs are significantly less than the unconnected gene-disease pairs, we set the number of negative samples to be p times the number of positive samples. To learn the parameters, the margin loss function was adopted, defined by Eq. 10:

$$Loss(y_{ij}, \widehat{y}_{ij}) = Max(0, 1 - \widehat{y}_{ij} \cdot y_{ij}) \tag{10}$$

where $\widehat{y}_{ij} = score_{ij}$, and y_{ij} represents the true relationship between gene node i and disease node j. Specially, $y_{ij} = 1$ if there exists an association between i and j, otherwise $y_{ij} = 0$.

During the prediction phase, scores were solely computed for the associations between each disease and its candidate genes. Afterwards, the candidate genes were ranked for each disease based on their respective scores.

Results

Datasets

The heterogeneous module network consists of two types of nodes that represent genes and disease, two types of links corresponding to disease-gene associations and protein–protein interactions, and one type of modules (protein complexes). The disease-gene associations and 213,888 protein–protein interactions were downloaded from the literature [23], which sourced the data from the DisGeNet [24] database. A total of 2822 protein complexes were collected from Human Protein Reference Database [25].

In accordance with the experimental methodology of the prior research [23], the disease-gene associations were classified into two distinct groups. The first group, denoted as the internal dataset, contained 130,820 disease-gene associations involving 13,074 diseases and 8947 genes, which was used for cross validation. The second group comprised 10,066 disease-gene associations involving 1186 diseases and 2552 genes. Termed as the external dataset, this group was collected from DisGeNet that integrated animal model data, which was used to assessment the capacity to discover new candidate associations.

Experimental setting

We adopted the experimental settings proposed by Yang et al. [23]. To validate the effectiveness of our method, we conducted a tenfold cross validation on the 130,820 curated associations. Additionally, we used 10,066 associations from animal model as an external dataset for each fold. The parameter l in graph data augmentation is set to 10, resulting in a total of 243,379 newly added interactions. The hyperparameters were tuned with the help of cross validation. Specially, for the node2vec, we set the window size, the walk length, the number of walks, the in–out parameter, the embedding size and the iteration number to 5, 64, 10, 0.3, 128 and 10, respectively. For GAT, we set the size of hidden units for *GAT* to (256, 128), and the number of heads in multi-head attention to 2. The learning rate, epoch number and size of hidden units for *GCN* and *GraphSage* were set to 0.0009, 10 and (128, 64, 8), respectively. Moreover, the number of negative samples was set to be 50 times ($p = 50$) greater than the number of positive samples.

In the experiments, Precision, Recall, F1-score (F1) and Association Precision (AP) were employed to evaluate the performance of gene prioritization. Denote the true pathogenic genes of the disease d in the test set as *T(d)*, and record the top i genes with the

highest predicted probabilities for the disease d as $P_i(d)$. Precision, Recall, F1-score in Top@i can be defined as follows:

$$Prec = \frac{1}{|D|}\sum_{d\in D}\frac{|T(d)\bigcap P_i(d)|}{|P_i(d)|} \tag{11}$$

$$Recall = \frac{1}{|D|}\sum_{d\in D}\frac{|T(d)\bigcap P_i(d)|}{|T(d)|} \tag{12}$$

$$F1 = \frac{1}{|D|}\sum_{d\in D}\frac{2|T(d)\bigcap P_i(d)|}{|P_i(d)|+|T(d)|} \tag{13}$$

To assess the overall performance, the association precision (AP) is defined as follows:

$$AP = \frac{\sum_{d\in D}|T(d)\bigcap P_k(d)|}{\sum_{d\in D}\min(|P_k(d)|,10)} \tag{14}$$

Here, D is the disease set and k is set as the number of true pathogenic genes in the test for each disease. If the number of pathogenic genes for a certain disease is greater than 10, then set k as 10. The Eq. 14 imposes restrictions the list length of candidate genes, focusing solely on the top 10 candidate genes for each disease. This is because the exploration of gene-disease associations is essentially a ranking problem, and during cell experiments, animal model studies, and clinical trials, candidates are typically selected from the top-ranked genes. Additionally, AUC was utilized to evaluate the performance.

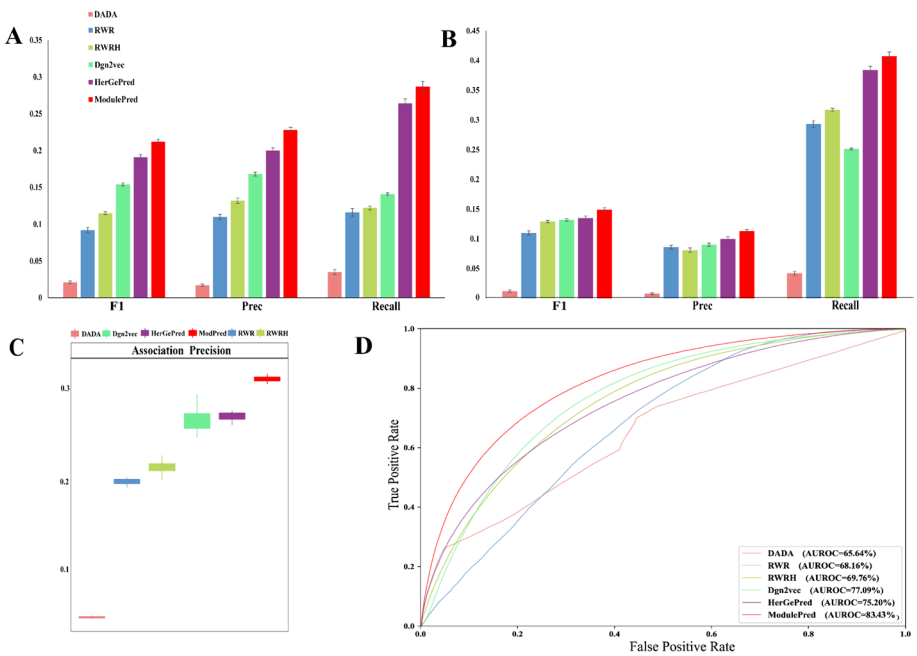

Fig. 2 Cross validation performance comparison with state-of-the-art methods on the internal dataset. **A** The average F1, Precision and Recall of Top-3 predicted genes. **B** The average F1, Precision and Recall of Top-10 predicted genes. **C** AP performance. **D** ROC curves for disease gene prediction. Error bars represent the distribution of tenfold cross validations

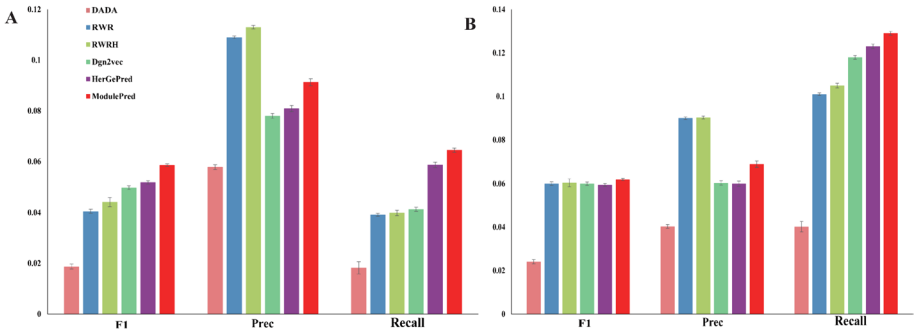

Fig. 3 Performance comparison with state-of-the-art methods on the external dataset. **A** The average F1, Precision and Recall of Top-3 predicted genes. **B** The average F1, Precision and Recall of Top-10 predicted genes. Error bars represent the distribution of tenfold cross validations

Performance comparisons with state-of-the-art methods

To validate the superiority of our approach, we compared ModulePred with three state-of-the-art methods including DADA [26], RWR [27], RWRH [28], Dgn2vec [29] and HerGePred [23]. As depicted in Fig. 2, our approach demonstrated superior performance compared to these competitive methods. In terms of Top@3, ModulePred exhibited the highest Precision, Recall and F1 (Fig. 2A). Similarly, for Top@10, ModulePred significantly outperformed the other methods across the three metrics (Fig. 2B). When evaluating the overall performance using the Association Precision (AP), HerGePred outperformed the other baseline methods. However, our approach, ModulePred, showed remarkable improvement over HerGePred, with an increase of approximately 4 percentage points in AP (from 0.259 to 0.306; Fig. 2C) and 7 percentage points in AUC (from 0.752 to 0.834; Fig. 2D).

To evaluate the capability for discovering new disease genes, we further assess the performance on the external dataset, as depicted in Fig. 3. In the Top@3 scenario, ModulePred outperformed other methods in terms of F1 and Recall, despite its Precision being lower than that of RWR and RWRH (Fig. 3A). Moreover, the performance of the methods in the Top@10 scenario was found to be similar to that in the Top@3 scenario (Fig. 3B). It is important to note that the performance on the external dataset in Fig. 3 was notably lower than that on the internal dataset in Fig. 2. This discrepancy arises from the fact that both the external and internal datasets were evaluated using the same prediction results. For example, assume that disease d is associated with genes g_1, g_2, g_3, g_4 and g_5 in the internal dataset, and with genes g_6 and g_7 in the external dataset. In a fold of cross-validation, the training set includes two gene-disease associations (d, g_1) and (d, g_2), while the test set includes (d, g_3), (d, g_4) and (d, g_5). An algorithm predicts the top 3 candidate genes most likely associated with disease d as g_3, g_4 and g_5. In the Top@3 scenario, the algorithm achieves 100% precision, recall and F1 score in predicting disease d. Since the top 3 candidate genes have no intersection with the external dataset, the algorithm completely fails to discover new genes in the external dataset, leading to a bias in its performance on the external dataset.

Ablation study

We compared the proposed ModulePred method with three ablations, namely GNN-M, GNN* and GNN. Theses variants were compared as follows:

(1) GNN*-M is the complete ModulePred method which uitlizes the augmented protein interaction network and applies graph representation with module information.
(2) GNN-M is an ablation of ModulePred that applies graph embedding solely on the original protein interaction network.
(3) GNN* is an ablation of ModulePred that uses the augmented protein interaction network without modules and performs graph embedding using the traditional node2vec approach.
(4) GNN is an ablation of GNN* that uses the original protein interaction network without protein complexes.

As depicted in Fig. 4, the incorporation of protein complexes allowed GNN-M to surpass GNN in all the evaluation metrics. Similarly, GNN* utilzied the augmented protein–protein interaction network to investigate the connections between diseases and genes, resulting in significant notable enhancements across all evaluation metrics compared to GNN. Notably, the impact of data augmentation had a greater impact on the AP index compared to module information (Fig. 4C). ModulePred, which integrated both

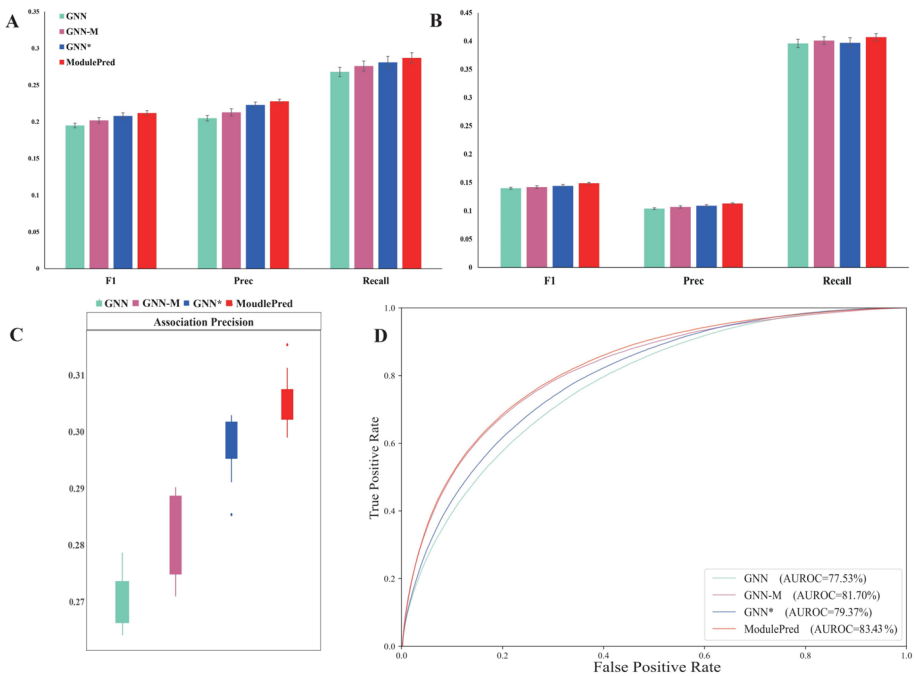

Fig. 4 Cross validation performance comparison with three ablations on the internal dataset. **A** The average F1, Precision and Recall of Top-3 predicted genes. **B** The average F1, Precision and Recall of Top-10 predicted genes. **C** AP performance. **D** ROC curves for disease gene prediction. Error bars represent the distribution of tenfold cross validations

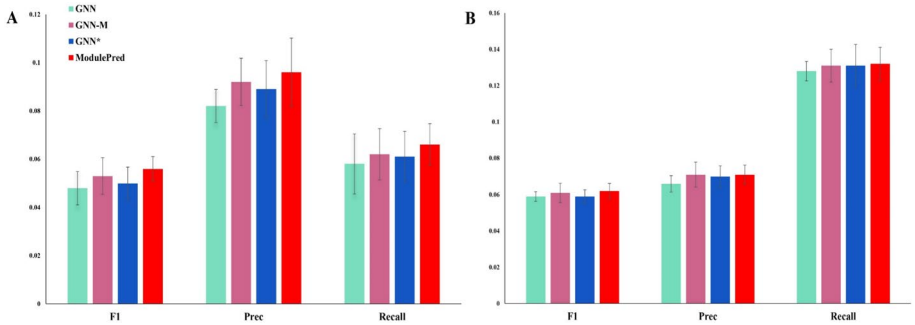

Fig. 5 Performance comparison with three ablations on the external dataset. **A** The average F1, Precision and Recall of Top-3 predicted genes. **B** The average F1, Precision and Recall of Top-10 predicted genes. Error bars represent the distribution of tenfold cross validations

module information and augmented protein interactions, made substantial progress when compared to GNN* and GNN-M (Fig. 4C).

Furthermore, we performed an analysis on the external dataset (Fig. 5), once again confirming the superiority of ModulePred over three ablations. This reinforced the potential of our approach in uncovering novel disease-gene associations. Both GNN-M and GNN* consistently exhibited better performance than GNN. However, GNN-M outperformed better than GNN* on the external dataset, showcasing a deviation from their performance on the internal dataset.

We further conducted three additional ablation experiments. Figure S1 indicates that the network structure of ModulePred (utilizing GAT for processing heterogeneous networks, GraphSage for processing gene-disease associations, and GCN for processing protein–protein interactions) can achieve good performance. Figure S2, suggests that setting the number of GAT layers to 1, GraphSage to 2 and GCN to 2 in ModulePred is an optimal parameter configuration. Moreover, Figure S3 demonstrates that setting the number of l in graph data augmentation to 10 can achieve optimal performance.

Table 1 Top 10 predicted genes for IPAH

Rank	Gene	Reference
1	MIR204	PMID: 30,854,934
2	CBLN2	PMID: 27,770,446
3	OTSC1	NA
4	EIF2AK4	PMID: 31,711,431
5	ENG	PMID: 30,312,106
6	PYCR1	NA
7	RTEL1	PMID: 30,523,160
8	LBR	NA
9	B3GAT3	NA
10	TGFBR3	PMID: 11,282,888

Table 2 Top 10 predicted genes for Hypothyroidism

Rank	Gene	Reference
1	LHX3	PMID: 12,244,277
2	OTX2	PMID: 26,416,826
3	GALE	NA
4	MAGEL2	PMID: 33,570,896
5	BRAF	PMID: 21,512,141
6	GLI2	PMID: 25,484,916
7	FANCB	PMID: 28,588,452
8	CDKN1C	NA
9	NDST1	NA
10	PAH	NA

Case study

To further elucidate the biological insights of our approach, we conducted two case studies in order to identify disease genes related to hypothyroidism and Idiopathic Pulmonary Arterial Hypertension (IPAH). The predicted genes were ranked based on their scores (refer to Eq. 9 for details). Furthermore, we manually searched published biomedical literature to obtain final confirmations.

IPAH is a progressive and potentially life-threatening condition characterized by elevated blood pressure in the pulmonary arteries without any discernible underlying cause, requiring thorough investigation and management from a medical perspective [17]. Among the top 10 genes predicted by ModulePred (Table 1), an impressive 6 associations were substantiated by previous publications, supported by their corresponding PubMed Unique Identifier (PMID). For instance, the top-ranked gene MIR204 has been reported to exhibit abnormal expression in relation to the onset and progression of IPAH [30].

Hypothyroidism is a multifaceted endocrine disorder characterized by diminished production or action of thyroid hormones, resulting in a variety of physiological disruptions that necessitate investigation and management from an endocrinological perspective. Recent studies have identified several genes associated with hypothyroidism [31–33]. As presented in Table 2, our ModulePred achieved high prediction accuracy rates of 100%, 80%, 86% for the top 2, top 5 and top 7 genes, respectively. For instance, OTX2 Mutations have been linked to developmental abnormalities in both the central nervous system and the thyroid, resulting in hypothyroidism [34]. Similarly, defects in GLI2 can disrupt normal thyroid development and function, potentially leading to reduce thyroid hormone levels [35].

Conclusion

In this article, a deep learning framework called ModulePred is presented for predicting disease-gene associations. ModulePred achieves competitive predictive performance by employing graph augmentation on the protein interaction network and graph embedding for the heterogeneous module network. Experimental results on the DisGeNet dataset substantiate the efficacy of ModulePred in discovering disease-gene associations. Furthermore, the ablation study highlights the greater impact of graph augmentation

on the performance of ModulePred compared to the graph embedding for the module network.

Abbreviations

SNPs	Single nucleotide polymorphisms
GCN	Graph convolutional network
PPI	Protein–Protein interaction
GAT	Graph attention network
F1	F1-score
AP	Association precision
IPAH	Idiopathic pulmonary arterial hypertension
PMID	PubMed unique identifier

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05841-3.

Supplementary Material 1.

Acknowledgements
We thank Mr. Yi Zhao from Qingdao University for the support of computing resources.

Author contributions
S.W. and X.J. conceived the idea. X.J., W.L. and J.L. implemented the algorithm and codes. X.J., W.L., H.S. and J.X. performed the analysis. W.L. and J.X. prepared figures. S.W. and X.S. wrote the manuscript. All contributed the proofread.

Funding
XS acknowledges support of Grant No. 2021YFF0704500 from National Key Research and Development Program of China, Grant No. 32070086 from National Natural Science Foundation of China, Shandong Province Youth Entrepreneurial Talent Introduction and Training Program, and Shandong Province Taishan Scholars Youth Experts Program. SW acknowledges support of Grant No. ZR2019PF012 from Shandong Provincial Natural Science Foundation of China.

Availability of data and materials
All datasets and code in this work are available at https://github.com/qdu-bioinfo/ModulePred. All other relevant data is available upon request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 14 March 2024 Accepted: 12 June 2024
Published online: 14 June 2024

References

1. Yoon S, Nguyen HCT, Yoo YJ, et al. Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. Nucleic Acids Res. 2018;46(10):160.
2. Ata SK, Wu M, Fang Y, et al. Recent advances in network-based methods for disease gene prediction. Brief Bioinform. 2020. https://doi.org/10.1093/bib/bbaa303.
3. Ghiassian SD, Menche J, Barabasi AL. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Comput Biol. 2015;11(4):e1004120.
4. Himmelstein DS, Lizee A, Hessler C, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife. 2017. https://doi.org/10.7554/eLife.26726.
5. Valdeolivas A, Tichit L, Navarro C, et al. Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics. 2019;35(3):497–505.
6. Lin CH, Konecki DM, Liu M, et al. Multimodal network diffusion predicts future disease-gene-chemical associations. Bioinformatics. 2019;35(9):1536–43.
7. Jiang B, Kloster K, Gleich DF, et al. AptRank: an adaptive PageRank model for protein function prediction on bi-relational graph. Bioinformatics. 2017;33(12):1829–36.

8. Lotfi Shahreza M, Ghadiri N, Mousavi SR, et al. A review of network-based approaches to drug repositioning. Brief Bioinform. 2018;19(5):878–92.
9. Jowkar GH, Mansoori EG. Perceptron ensemble of graph-based positive-unlabeled learning for disease gene identification. Comput Biol Chem. 2016;64:263–70.
10. Chen X, Yan CC, Zhang X, et al. Long non-coding RNAs and complex diseases: from experimental results to computational models. Brief Bioinform. 2017;18(4):558–76.
11. Li Y, Patra JC. Integration of multiple data sources to prioritize candidate genes using discounted rating system. BMC Bioinform. 2010;11(Suppl 1):S20.
12. Yang K, Wang R, Liu G, et al. HerGePred: heterogeneous network embedding representation for disease gene prediction. IEEE J Biomed Health Inform. 2019;23(4):1805–15.
13. Cinaglia P, Cannataro M. Identifying candidate gene-disease associations via graph neural networks. Entropy (Basel). 2023;25(6):909.
14. Zhang J, Zhong C, Huang Y, et al. A method for identifying protein complexes with the features of joint co-localization and joint co-expression in static PPI networks. Comput Biol Med. 2019;111:103333.
15. Kovacs IA, Luck K, Spirohn K, et al. Network-based prediction of protein interactions. Nat Commun. 2019;10(1):1240.
16. Luck K, Sheynkman GM, Zhang I, et al. Proteome-scale human interactomics. Trends Biochem Sci. 2017;42(5):342–54.
17. Keskin O, Tuncbag N, Gursoy A. Predicting protein-protein interactions from the molecular to the proteome level. Chem Rev. 2016;116(8):4884–909.
18. Szilagyi A, Grimm V, Arakaki AK, et al. Prediction of physical protein-protein interactions. Phys Biol. 2005;2(2):S1-16.
19. Lu L, Zhou T. Link prediction in complex networks: a survey. Physica A: Stat Mech Appl. 2010;390(6):1150–70.
20. Grover A, Leskovec J. node2vec: scalable feature learning for networks. KDD. 2016;2016:855–64.
21. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv:1609.02907. 2016.
22. Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. Adv Neural Info Proc Syst 2017;30.
23. Yang K, Wang R, Liu G, et al. HerGePred: heterogeneous network embedding representation for disease gene prediction. IEEE J Biomed Health Inform. 2019;23(4):1805–15.
24. Pinero J, Bravo A, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45(11):D833–9.
25. Chan W. The UniProt Knowledgebase (UniProtKB): a freely accessible, comprehensive and expertly curated protein sequence database. Genet Res. 2010;92(1):78–79.
26. Erten S, Bebek G, Ewing RM, et al. DA DA: degree-aware algorithms for network-based disease gene prioritization. BioData Min. 2011;4(1):1–20.
27. Fernández P. Google's pagerank and beyond: the science of search engine rankings. Math Intell. 2008;30(1):68–9.
28. Cao Z, Wei F, Dong L, et al. Ranking with recursive neural networks and its application to multi-document summarization. Proceed AAAI Conf Artif Intell. 2015. https://doi.org/10.1609/aaai.v29i1.9490.
29. Liu Y, Guo Y, Liu X, et al. Pathogenic gene prediction based on network embedding. Brief Bioinform. 2021;22(4):bbaa353.
30. Estephan LE, Genuardi MV, Kosanovich CM, et al. Distinct plasma gradients of microRNA-204 in the pulmonary circulation of patients suffering from WHO Groups I and II pulmonary hypertension. Pulm Circ. 2019;9(2):2045894019840646.
31. Hwangbo Y, Park YJ. Genome-wide association studies of autoimmune thyroid diseases, thyroid function, and thyroid cancer. Endocrinol Metab (Seoul). 2018;33(2):175–84.
32. Teumer A, Chaker L, Groeneweg S, et al. Genome-wide analyses identify a role for SLC17A4 and AADAT in thyroid hormone regulation. Nat Commun. 2018;9(1):4455.
33. Stoupa A, Adam F, Kariyawasam D, et al. TUBB1 mutations cause thyroid dysgenesis associated with abnormal platelet physiology. EMBO Mol Med. 2018. https://doi.org/10.15252/emmm.201809569.
34. Schoenmakers N, Alatzoglou KS, Chatterjee VK, et al. Recent advances in central congenital hypothyroidism. J Endocrinol. 2015;227(3):R51–71.
35. Ma D, Marion R, Punjabi NP, et al. A de novo 10.79 Mb interstitial deletion at 2q13q14.2 involving PAX8 causing hypothyroidism and mullerian agenesis: a novel case report and literature review. Mol Cytogenet. 2014;7(1):85.

Publisher's Note