Incorporating multi-scale module kernel for disease-gene identification in biological networks

Ju Xiang, Kaixin Zeng, Shengkai Chen, Xiangmao Meng, Ruiqing Zheng, Ying Zheng, Yahui Long, and Min Li*

Abstract: Biomedical data mining plays a crucial role in studying diseases, with disease-gene identification being one of the most prominent areas of research in this field. Many biomolecule networks are known to have multi-scale module structures, which may be helpful for studying complex diseases, but the mining and utilization of multi-scale module structure is an open issue. Therefore, we present a kind of novel hybrid network-based method (HyMSMK) for disease-gene identification through incorporating multi-scale module kernel in biomolecule networks. We first apply exponential sampling to construct multi-scale module profile containing local to global structural information, where modules at different scales are extracted from comprehensive interactome by multi-scale modularity optimization. Then, the multi-scale module profile is preprocessed by the relative information content, and is used to generate multi-scale module kernel, which is further preprocessed by kernel sparsification. We design multiple schemes for incorporating multi-scale module kernel to discover potential disease-related genes. We investigate the performance of these schemes by experimental evaluations, show the positive effect of kernel sparsification on reducing the requirement for space and time, and confirm the superior performance of our method compared to other state-of-art network-based baselines. The study demonstrates the utility of multi-scale module kernel in discovering disease genes, which could provide insights for the research of relevant issues.

Key words: Network biology, Biomolecule networks, multi-scale modules, Disease-gene identification

1 Introduction

- Ju Xiang, Kaixin Zeng, Shengkai Chen and Ying Zheng are with the School of Computer Science and Engineering, Changsha University of Science and Technology, Changsha, 410076, China. E-mail: xiang.ju@foxmail.com; KAIXINZ1999@163.com; skchen102@foxmail.com; zhengying@csust.edu.cn.
- Ju Xiang, Ruiqing Zheng, Yahui Long and Min Li are with the School of Computer Science and Engineering, Central South University, Changsha, 410083, China. E-mail: xiang.ju@foxmail.com; rqzheng@csu.edu.cn; Long_Yahui@bii.a-star.edu.sg; limin@mail.csu.edu.cn.
- Yahui Long is with the Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore. E-mail: Long_Yahui@bii.a-star.edu.sg.
- Xiangmao Meng is with the School of Computer Science, Xiangtan University, Xiangtan 411105, China. E-mail: mxmanhui@xtu.edu.cn.

* To whom correspondence should be addressed.

Manuscript received: year-month-day; accepted: year-month-day

Biomedical data mining plays a vital role in advancing biomedical research, particularly with the rapid growth of biomedical data such as protein-protein interactions (PPI) [1], disease-related genes [2], and disease-related phenotypes/symptoms [3, 4], which has driven the evolution of network medicine [5-7]. To

uncover hidden knowledge within these interconnected datasets, numerous network-based mining strategies have been developed. Among these, identifying disease-related genes stands out as one of the most significant areas of focus in the field [8-11].

The occurrence and development of a lot of diseases are caused by the dysfunctions of genes. Genes as well as their products typically achieve biological functions by synergistic interactions, forming complex biological molecular networks. The dysfunctions of some genes can lead the local perturbation in biomolecular networks, resulting in the phenotypes or symptoms of certain disease [12]. Therefore, a lot of network-based methods have been presented to infer disease-related genes [10, 13, 14]. Generally, there are more interactions or correlations among genes related to the same disease, because they may have more similar functions [15]. Both direct interaction and network distance or closeness are natural strategies for inferring potential disease-related genes [16, 17], while network propagation is applied more widely to disease-gene identification and related research [13, 14, 16, 18-21], since it can utilize the information of neighborhood around seeds in biomolecule networks more effectively. However, because of data noise in biomolecule networks, these methods still have room for improvement, which could be achieved through approaches like network enhancement [9, 22].

The modular structure is one of common properties of biological networks, which can offer valuable insights for the research of complex diseases [23, 24]. Some studies have analyzed network modules for diseases [23, 25, 26], but the task of mining network modules remains an open challenge [27-30]. For example, we found that the intrinsic resolution limitations may exist in many algorithms, leading to some unexpected results [31-33]. Certain algorithms may produce large-scale network modules that include numerous genes unrelated to the disease. Conversely, some algorithms might generate numerous fragmented network modules, potentially causing the missing of candidate genes that could be relevant to this disease. Moreover, many complex diseases often are related to several modules, as the associated genes may have diverse functions and contribute to the disease through various mechanisms [34]. These modules, while not entirely independent, may be related at a higher functional level or scale. Modules at different scales can offer valuable insights at various levels of analysis. Thus, a multi-scale approach to module analysis could provide more insights for studying these diseases.

Biomolecular networks commonly have multi-scale module structures, consisting of sub-networks that are to recursively split into smaller ones, forming a kind of hierarchical structure [35]. For example, some complexes are composed of several sub-complexes, while function annotations of genes and proteins also have hierarchical organization, suggesting the presence of functional modules across various scales [36-38]. Some studies have shown the significance of these multi-scale modules in understanding biological processes. For instance, Lewis et al. explored the relationship between multiscale modules and functions, highlighting how different scales of modules are linked to specific proteins and processes [39]. Similarly, Wang et al. demonstrated that the hierarchical structure of identified modules roughly corresponds to the hierarchy of Gene Ontology (GO) [40]. In summary, multi-scale module structures offer rich information spanning from local to global structures, which can be valuable for disease research. However, effectively utilizing multi-scale module structures for disease-gene identification is still a challenging task.

To utilize the multi-scale module structures to enhance the capability of identifying disease-related genes, a kind of novel hybrid method called HyMSMK is proposed by incorporating multi-scale module kernel (MSMK) in biomolecule networks. The rest of the work is arranged as follows. Section 2 introduces datasets, including the pre-computation methods. Then, our hybrid methods for incorporating MSMK will be introduced, including the mining of multi-scale module profile, the construction of MSMK, and multiple schemes for incorporating MSMK. Section 3 will introduce experimental settings. Subsequently, we investigate the performance of these incorporating schemes, analyze the effect of kernel sparsification on reducing the requirement for space and time, and show

the performance of our method compared to other state-of-art network-based algorithms. Additionally, we perform case studies on several diseases, which identify candidate gene sets for these diseases and validate their relevance to diseases. Finally, we present the conclusions of this study.

2 Materials and Methods

Here, the datasets used in this study will be introduced, including datasets of disease genes, disease-related symptoms/phenotypes and PPIs, respectively. Then, we introduce the details of our methods: 1) methods estimating disease-disease associations, 2) methods mining multi-scale module profiles, 3) methods constructing MSMK, and 4) methods incorporating MSMK, leading to several variants (**Table 1**).

2.1 Datasets

In our experimental study, we utilize two datasets of disease-gene associations (DGA). Dataset 1: This integrated dataset combines disease-related genes from OMIM and GWAS [41, 42]. The MeSH system is used to standardize disease nomenclature from multiple sources, and diseases with 20 or more genes are included [26, 43]. Dataset 2: This dataset is derived exclusively from the OMIM database, a comprehensive resource on human diseases and their causal genes, including extensive information on human genes and genetic disorders [42].

To estimate the associations between diseases, we will extract the data of disease-related symptoms and phenotype abnormalities from human phenotype ontology (HPO) and literature [4, 44]. Complex interactions among proteins form complex protein-protein interaction (PPI) network [45, 46], which can uncover intrinsic gene-gene associations (GGA) and has become an important basis for the study of molecular mechanisms of diseases. Due to the wide existence of incompleteness of data, we will utilize a comprehensive high-quality interactome constructed by integrating multiple sources of PPIs (e.g., BioGRID, MINT, IntAct, and HPRD) [43].

2.2 Estimation of associations between diseases

2.2.1 Estimation based on co-occurrence between diseases and symptoms

Symptoms are the observable high-level manifestations of a disease and play a critical role in clinical diagnosis and treatment. Research has shown that symptom-based disease similarity is closely linked to molecular interactions [4]. In this study, we are to utilize disease-disease associations inferred by disease-related symptoms from text mining.

Specifically, the associations between diseases and symptoms are characterized by the co-occurrence of MeSH terms for diseases and symptoms, measured by the number of PubMed articles where these terms appear together [42]. A disease is characterized by a vector representation of its associated symptoms, $c_{\bullet,x} = (c_{1,x}, c_{2,x}, \ldots, c_{j,x}, \ldots, c_{n,x})^T$, where $c_{j,x}$ denotes the co-occurrence between symptom j and disease x. A co-occurrence matrix of MeSH terms for diseases and symptoms can be constructed by these vectors,

$$C = (c_{\bullet,1}, c_{\bullet,2}, \ldots, c_{\bullet,i}, \ldots, c_{\bullet,N})^T, \quad (1)$$

Symptom overlap between diseases can inherently indicate associations between them. However, various symptoms, such as pain and stupor, often vary in their prevalence across diseases, and publication biases may favor certain diseases. To address this heterogeneity, the strength of the association between a disease x and a symptom j is typically calculated by the term frequency-inverse document frequency (*TF-IDF*) method [47]. Specifically, a normalized co-occurrence matrix can be obtained by,

$$\hat{C} = C \; D_{TFIDF}, \quad (2)$$

where the diagonal matrix D_{TFIDF} has diagonal elements: $(D_{TFIDF})_{\mathrm{jj}} = \log(N/N_j)$, N is the number of diseases and $N_j \geq 1$ is the number of diseases having symptom j. The similarity scores between diseases can be calculated using the normalized co-occurrence matrix,

$$S_D = D_D \hat{C} \hat{C}^T D_D, \quad (3)$$

where D_D is a diagonal matrix with diagonal elements: $(D_D)_{ii} = (\hat{C}\hat{C}^T)_{ii}^{-1/2}$. The similarity scores

between diseases are between 0 and 1. It is important to note that the above calculation does not consider the inherent correlations between symptoms.

2.2.2 Estimation based on human phenotype ontology

Diseases are generally linked to a group of abnormal phenotypes. The Human Phenotype Ontology (HPO) database offers annotations of human phenotypes for diseases, represented by a collection of HPO terms, which are organized into a directed acyclic graph (DAG) [44]. Unlike the co-occurrence data of diseases/symptoms discussed above, HPO offers more detailed information about the phenotypic abnormalities associated with diseases, offering an alternative approach to uncovering disease-disease associations.

For diseases in OMIM [42], we assess the disease-disease associations by using dataset from HPO, since it provides phenotype annotations for OMIM disease entries. The similarity scores between diseases can be determined by the number of shared HPO terms. However, not all HPO terms are equally informative. Diseases sharing rare HPO terms are generally more similar than those sharing common ones. To address this, the relevance of HPO terms is quantified using the relative information content (IC) of the terms [48],

$$IC(t) = -\log f_t, \tag{4}$$

where the frequency of HPO term t is denoted as f_t. The similarity between two terms t_1 and t_2 is derived by,

$$HPS(t_1, t_2) = \max_{v \in anc(t_1) \cap anc(t_2)} IC(v), \tag{5}$$

where $anc(t_1)$ is the ancestor terms of t_1 on the HPO structure. The similarity score between diseases can be calculated by,

$$S_D(d_x, d_y) = \frac{1}{2}\left(\frac{1}{|T_x|}\sum_{t_1 \in T_x} \max_{t_2 \in T_y} HPS(t_1, t_2) + \frac{1}{|T_y|}\sum_{t_2 \in T_y} \max_{t_1 \in T_x} HPS(t_1, t_2)\right). \tag{6}$$

where diseases d_x and d_y are characterized by two sets (T_x and T_y) of HPO terms. The similarity scores between diseases are determined by the overall rarity of their shared HPO terms. This measurement of disease-disease similarity incorporates not only the human phenotype annotations but also their inherent relationships within the DAG structure of HPO.

2.3 Multi-scale modularity optimization with exponential sampling for mining multi-scale module profile

The mining of multi-scale module structure is an open issue [27, 49], while it is widely present in biomolecule networks. Modularity optimization (MO) has been widely applied in many fields, including bioinformatics, though there are many network module mining algorithms. Therefore, we will extract modules at different scales by MO, along with exponential sampling (ES) (see **Fig. 1**) [50, 51]. Generally, MO will split a network into module partition by maximizing modularity function Q, that is, finding optimal solution of Q [50-52].

For a given module partition of a network, the modularity Q, which is widely used as target function of network module mining, can be written as,

$$Q = \sum_s \left(\frac{k_s^{in}}{2M} - \gamma\left(\frac{k_s}{2M}\right)^2\right), \tag{7}$$

where the total degree of module s is calculated by,

$$k_s = \sum_{i,j} A_{ij}\delta(c_i, s), \tag{8}$$

and the total inner degree of module s is calculated by,

$$k_s^{in} = \sum_{ij} A_{ij}\delta(c_i, s)\delta(c_j, s), \tag{9}$$

Here, m is the number of edges in the network; A denotes the adjacent matrix of the network; c_i is the module index of node i; γ denotes the resolution parameter. More concisely, it can be written as $Q = \sum_s(e_{ss} - \gamma a_s^2)$, where e_{st} indicates the fraction of edges between modules s and t, and e_{ss} indicates the fraction of ends of edges on module s. Given different γ-values, one can obtain modules at different scales by feeding a network into effective algorithms for MO [49, 53]. Generally, the smaller γ-value will discover larger modules, while the larger γ-value will discover smaller modules. A set of single-node modules will be discovered if the γ-value is very large, e.g., $\gamma > 2m/k_{min}^2$, where k_{min} is the minimum of node degrees [54]. In this study, the Louvain algorithm for MO was used, since it is a fast and efficient algorithm widely used in literature.

To extract a set of meaningful module partitions from local to global scales, according to our previous

work, we first define a meaningful range of the values for resolution parameter $\gamma \in [\gamma_{min}, \gamma_{max}]$, which should be able to include all possible resolution samplings. Given a network, γ_{min} and γ_{max} can be denoted as $\gamma_{min} = \max_{\#[H(\gamma)]=1} \gamma$ and $\gamma_{max} = \min_{\#[H(\gamma)]=N} \gamma$, where $\#[H(\gamma)]$ is the number of modules in the partition $H(\gamma)$ and N is the number of nodes in the network. Generally, it is difficult to get the accurate boundaries. Therefore, we use semi-empirical boundaries according to previous research [55]. The value of the resolution parameter is in continuous and real resolution space ($\gamma \in [\gamma_{min}, \gamma_{max}]$). Continuously scanning all resolution parameter values is meaningless because there must be a lot of information redundancy, and the computer cannot directly process it. To extract meaningful module partitions at different scales, it is necessary to select a series of discrete resolution parameter values for appropriate scales and then extract network module partitions at different scales. Here, we apply the ES strategy to extract a group of discrete γ-values in the continuous space, because it can sample meaningful scales reasonably, based on our previous studies [51, 53]. The ES strategy will obtain a group of γ-values equally spaced on a logarithmic scale, i.e., $log\,\gamma_{h+1} - log\,\gamma_h = C$. According to the set of sampled γ-values, we apply the MO algorithms to extract a set of corresponding module partitions. Consequently, we have formed a framework of multi-scale MO with ES (i.e., MO-ES) for extracting multi-scale modules.

By the MO-ES framework, we can reasonably generate a series of meaningful network partitions, which can cover the information from the local scales to global scales. For convenience of the quantitative description of module partitions extracted at different scales, we provide the following definitions.

Definition 1. We define the following module partition matrix for scale h by,

$$H^{(h)} = \begin{bmatrix} H_{1,1}^{(h)}, & \dots & H_{1,G_h}^{(h)} \\ \vdots & \ddots & \vdots \\ H_{N,1}^{(h)} & \dots & H_{N,G_h}^{(h)} \end{bmatrix}, \qquad (10)$$

where $H_{i,t}^{(h)}$=1 if module t in partition h contains node i, otherwise $H_{i,t}^{(h)}$=0.

Definition 2. We define the multi-scale module profile by,

$$H = (H^{(1)}, H^{(2)}, \dots, H^{(h)}, \dots), \qquad (11)$$

which incorporates multiple module partition matrices at different scales.

Definition 1 can quantitatively characterize module partition at each scale. Definition 2 is used to effectively organize a series of module partitions extracted by MO-ES. It forms the multi-scale module profile, including information from local scales to global scales that may be useful for disease-gene identification. The multi-scale module profile can be viewed as a matrix for multi-scale features, each of which corresponds a column recording the information of the feature.

2.4 Construction of multi-scale module kernel

2.4.1 Multi-scale module kernel

The multi-scale module profile can be viewed as a kind of feature matrix containing multi-scale information, which may be useful for uncovering associations between genes. However, like the DSA associations as well as HPA annotations mentioned above, some of these features may not be more informative than others. Genes should be regarded as more similar if they share more infrequent features than others. This is because infrequent features tend to be related to local-scale modules at higher resolutions, usually exhibiting higher edge densities and greater functional homogeneity. Consequently, genes belonging to a module of this kind are likely to be more like each other. To account for this, we use the relative information content (IC) to process these features. A similarity matrix for genes will be constructed by using the processed features. For simplicity in computation, we introduce several definitions.

Definition 3. We define a diagonal matrix (M_P) of multi-scale module profile by assigning the diagonal element,

$$(M_P)_{i,i} = -log\,(\textstyle\sum_k g(H_{k,i})/N), \qquad (12)$$

where $g(\cdot)$ is defined as a counting function; $g(t) = 1$ when $t > 0$, while $g(t) = 0$ when $t \leq 0$.

Definition 4. We define a multi-scale module

kernel by,

$$K_M = M_K H M_P^2 H^T M_K = M_K H_P H_P^T M_K, \quad (13)$$

where $H_P = HM_P$ is a pre-processed multi-scale module profile by the relative information content; M_K denotes a diagonal matrix with the diagonal element,

$$(M_K)_{i,i} = (H_P H_P^T)_{i,i}^{1/2}. \quad (14)$$

The multi-scale module kernel has the values of elements ranging from 0 to 1, representing a kind of similarity scores between genes. This kernel reveals novel potential associations between genes from the standpoint of multi-scale module structure, offering valuable insights that may aid in identifying disease-associated genes.

2.4.2 Kernel sparsification

The multi-scale module kernel captures potential relationships between genes that are derived from multi-scale module profile. However, it is costly as a full matrix in terms of storage and computation. To address this, we convert it into a sparse matrix by reserving only associations between a gene and its k-nearest neighbors (knn). This kernel sparsification can be defined as follows.

Definition 5. We define a ranking vector for the neighbor genes of gene i, based on the descending order of the similarity scores between this gene and its neighbor genes, by,

$$\vec{r}_i = (r_{i1}, r_{i2}, \dots, r_{ij}, \dots)^T, \quad (15)$$

where r_{ij} denotes the ranking index of neighbor j of gene i.

Definition 6. We define a ranking matrix of genes as the union of the ranking vectors of genes by,

$$R = (\vec{r}_1, \vec{r}_2, \dots, \vec{r}_i, \dots). \quad (16)$$

Finally, we generate a new gene-gene association (nGGA) network $G = (V, E, W)$ based on the full matrix of kernel and its ranking matrix, where V, E and W respectively indicate the sets of gene nodes, edges and weights,

$$E = \{(i,j) | R_{ij} \le k \text{ or } R_{ji} \le k\} \text{ and } W = \{S_{ij} | (i,j) \in E\}. \quad (17)$$

where k denotes the number of the nearest neighbors and $S_{ij} = (K_M)_{ij}$ indicates the similarity between nodes i and j. The nGGA network can be denoted by an adjacent matrix A_M with elements indicating the weights of edges.

The kernel sparsification can be characterized by the number of k-nearest neighbors, the degree of which is inversely related to the number of k-nearest neighbors retained during the sparsification process. This approach helps filter out noise from the data while preserving the essential gene-gene associations identified through multi-scale module profile.

Table 1 Schemes incorporating multi-scale module kernel.

Scheme	Description
CPF	Two heterogeneous networks are constructed, one of which includes multi-scale module kernel, and two ranking lists from them are incorporated by rank aggregation.
FS	A heterogeneous network with the fusion of GGA and nGGA is applied.
MG	A heterogeneous network with a multi-graph of GGA and nGGA is applied.

2.5 Approaches incorporating multi-scale module kernel

Gene-gene associations provide crucial information for network-based approaches to identifying disease genes. The kernel of multi-scale modules outlined above offers a novel perspective on relationships between genes, derived by using the multi-scale modular structure of the network. This may provide complementary information that aids in the identification of disease genes. In this context, we construct three kinds of hybrid approaches that incorporate the multi-scale module kernel, including rank aggregation model, fusion network model, and multi-graph model (see **Table 1**).

2.5.1 Hybrid approaches based on rank aggregation

The approaches will generate two dual-layer heterogeneous networks of diseases and genes (see **Fig. 1** and **Algorithm 1**). In HN_A, the original gene-gene associations (GGA) are linked to disease-disease associations (DDA) via gene-disease associations

(DGA), while in HN_B, nGGAs of multi-scale module kernel are connected to DDAs through DGAs. Next, two lists of prediction genes associated with a disease under study will be generated by conducting the random walk on two heterogeneous networks, separately. Finally, a probabilistic model for rank aggregation is designed to combine two lists of prediction genes. To conduct the random walk on the dual-layer heterogeneous network, we generate a transition matrix based on two intra-layer transition matrices and two inter-layer transition matrices.

Algorithm 1 HyMSMK(A_G,A_D,A_{GD},$p_G^{(0)}$,$p_D^{(0)}$,k,RA)

Input:

A_G, A_D, A_{GD}: gene-gene, disease-disease and gene-disease matrices;

$p_G^{(0)}$, $p_D^{(0)}$: initial probabilities of genes and diseases;

k: # of k-nearest neighbors;

RA: probabilistic model for rank aggregation.

Output: r_G: list of gene rankings

1: $H \leftarrow$ multi-scale module profile from *interactome* by MO-ES
2: $K_M \leftarrow$ multi-scale module kernel from H
3: $A_M \leftarrow$ sparse matrix from K_M by kernel sparsification (k)
4: $T_M \leftarrow$ intra-layer transition matrix of A_M
5: $T_G \leftarrow$ intra-layer transition matrix of A_G
6: $T_D \leftarrow$ intra-layer transition matrix of A_D
7: $T_{GD} \leftarrow$ inter-layer transition matrix of A_{GD}
8: $T_{DG} \leftarrow$ inter-layer transition matrix of A_{GD}^T
9: $T_A \leftarrow$ normalized matrix of $\begin{bmatrix} (1-\beta)T_G & \beta T_{GD} \\ \beta T_{DG} & (1-\beta)T_D \end{bmatrix}$
10: $p_A \leftarrow \alpha[I-(1-\alpha)T_A]^{-1}\begin{bmatrix} p_G^{(0)} \\ p_D^{(0)} \end{bmatrix}$
11: $r_A \leftarrow$ gene rankings from the descending order of p_A
12: $T_B \leftarrow$ normalized matrix of $\begin{bmatrix} (1-\beta)T_M & \beta T_{GD} \\ \beta T_{DG} & (1-\beta)T_D \end{bmatrix}$
13: $p_B \leftarrow \alpha[I-(1-\alpha)T_B]^{-1}\begin{bmatrix} p_G^{(0)} \\ p_D^{(0)} \end{bmatrix}$
14: $r_B \leftarrow$ gene rankings from the descending order of p_B
15: $r_G \leftarrow$ fusion rankings of r_A and r_B by RA
16: **return** r_G

Definition 7. We define the intra-layer transition matrix for the layer of gene-gene network as,

$$T_G = A_G D_G^{-1}, \tag{18}$$

where A_G denotes the adjacent matrix of GGA or nGGA; D_G denotes a diagonal matrix with diagonal elements $A_G^T\mathbf{1}$ (note that $\mathbf{1}$ is an all-one vector).

Definition 8. We define the intra-layer transition matrix for the layer of disease-disease network as $T_D = A_D D_D^{-1}$, where A_D denotes the adjacent matrix of DDA, and D_D denotes a diagonal matrix with diagonal elements $A_D^T\mathbf{1}$.

Definition 9. We define the inter-layer transition matrix from diseases to genes as,

$$T_{GD} = A_{GD} D_{GD}^{-1}, \tag{19}$$

where A_{GD} denotes the matrix of gene-disease associations: $A_{GD}(j,i) = 1$ when gene j is known to be related to disease i, and $M_{GD}(j,i) = 0$ otherwise; D_{GD} denotes a diagonal matrix with diagonal elements $A_{GD}^T\mathbf{1}$.

Definition 10. We define the inter-layer transition matrix from genes to diseases as $T_{DG} = A_{DG} D_{DG}^{-1}$, where $A_{DG} = A_{GD}^T$ and D_{DG} denotes a diagonal matrix with diagonal elements $A_{DG}^T\mathbf{1}$.

Definition 11. We define the transition matrix of the heterogeneous network as,

$$T_H = A_H D_H^{-1}, \tag{20}$$

where D_H denotes a diagonal matrix with diagonal elements $A_H^T\mathbf{1}$, and A_H is the raw transition matrix of the heterogeneous network,

$$A_H = \begin{pmatrix} (1-\beta)T_G & \beta T_{GD} \\ \beta T_{DG} & (1-\beta)T_D \end{pmatrix}, \tag{21}$$

where β denotes the probability of jumping between layers.

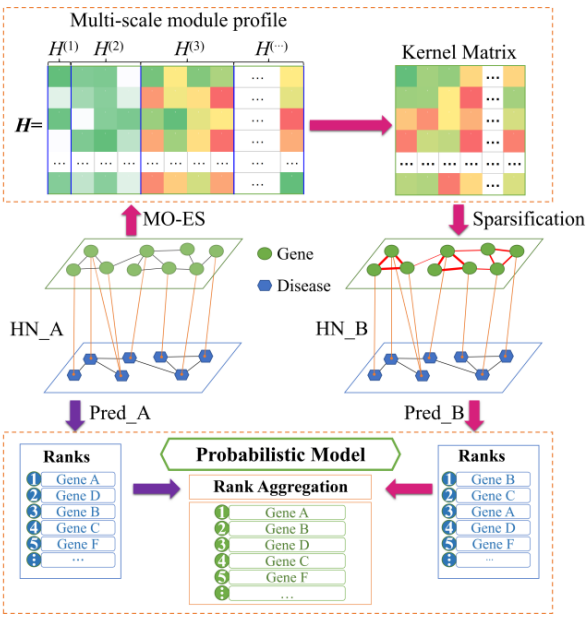

Fig. 1. Framework of our method for incorporating multi-scale module kernel by rank aggregation.

By the transition matrix T_H , one can easily obtain stationary probability vectors of walker arriving at different nodes (genes and diseases). One can calculate its closed solution by,

$$\begin{pmatrix} p_G \\ p_D \end{pmatrix} = \alpha\left[I - (1-\alpha)T_H\right]^{-1}\begin{pmatrix} p_G^{(0)} \\ p_D^{(0)} \end{pmatrix}, \quad (22)$$

where $p_G^{(0)}$ and $p_D^{(0)}$ are the initial probability vectors of gene layer and disease layer, respectively; $p_D^{(0)}(i) = 1$ for disease i under study, while 0 for other diseases; α denotes the restart probability. In order to avoid matrix inversion, one can calculate the stationary probability vectors based on an iterative process,

$$\begin{pmatrix} p_G^{(t)} \\ p_D^{(t)} \end{pmatrix} = (1-\alpha)T_H\begin{pmatrix} p_G^{(t-1)} \\ p_D^{(t-1)} \end{pmatrix} + \alpha\begin{pmatrix} p_G^{(0)} \\ p_D^{(0)} \end{pmatrix}, \quad (23)$$

where $\mathrm{p}_\mathrm{G}^{(\mathrm{t})}$ and $\mathrm{p}_\mathrm{D}^{(\mathrm{t})}$ denote the probability vectors of walker arriving at genes and diseases at the t-th time. Based on the descending order of stationary probability, one can obtain a ranking list of genes to identify genes related to the disease under study.

Model for rank aggregation. After obtaining two ranking lists of genes (**Fig. 1**), we generate the final results through a kind of rank aggregation, which is an interesting topic in its own right [56]. Assuming that two lists of genes mentioned above are each obtained based on different features, we can construct a rank-aggregation (RA) model based on Bayes theorem,

$$P(g_i|r_A, r_B) = P(g_i)P(r_A, r_B|g_i)/P(r_A, r_B) \propto P(g_i|r_A)P(g_i|r_A) \quad (24)$$

where $P(g_i)$ denotes the prior probability of gene i being associated with a disease; $P(g_i|r_A)$ denotes the conditional probability of gene i under known ranking r_A. To calculate the values of the conditional probability function (CPF), the formal definition of CPF is needed. For each list of prediction genes, a higher ranking of gene often implies a higher probability of being related to the disease. Here, three kinds of CPF's definitions are designed based on gene rankings:

$$P(g_i|r\ ) \propto 1 - r_i\ /N,\ r_i^{-1}, \text{ and } e^{-r_i}, \quad (25)$$

which are denoted respectively by CPF1, CPF2 and CPF3. Here, r_i denotes the ranking of g_i. The CPF's definitions are applied to the rankings r_A and r_B from different heterogeneous networks, respectively. The hybrid approaches with different CPF's definitions correspond to different variants of our method.

2.5.2 Other approaches incorporating multi-scale module kernel

In addition to the above approaches based on rank aggregation, we have also considered other approaches incorporating multi-scale module kernel, e.g., based on multigraph model and fusion network model.

For approaches based on multi-graph model (MG), a gene-gene multi-graph (MG) need to be generated first [21]. Here, we use GGA and nGGA to construct MG, and generate a multi-graph heterogeneous network by linking MG and DDA with DGA. Then, the list of predictive genes is obtained by the random walk in the multi-graph heterogeneous network.

In order to conduct the random walk on heterogeneous network of multi-graph, the transition matrix $\mathrm{T_G}$ of GGA is calculated by Eq.(18), and similarly the transition matrix $\mathrm{T_M}$ of nGGA is calculated by,

$$T_M = A_M D_M^{-1}, \quad (26)$$

where A_M is the adjacent matrix of nGGA, and D_M is a diagonal matrix with diagonal elements $A_M^T\mathbf{1}$. A transition matrix of the multi-graph is calculated by,

$$\hat{T}_G = \left(T_G + T_M\right)\hat{D}_G^{-1}, \quad (27)$$

where $\hat{D}_G$ denotes a diagonal matrix and its diagonal element $(\hat{D}_G)_{ii} = 2$ for node i located at both the two networks; or $(\hat{D}_G)_{ii} = 1$ for node i isolated in either of networks. By substituting T_G by $\hat{T}_G$ in Eq.(21) and Eq.(20), the transition matrix of the multi-graph heterogeneous network can be constructed. After the iterative process, stationary probability vector of genes can be obtained to identify genes related to the studied disease.

Moreover, we also consider the hybrid approaches based on fusion network model as comparison. We start by constructing a fusion network (FS) that combines edges from both GGA and nGGA. Next, we link the FS network with DDA via DGA to form a dual-layer heterogeneous network. Finally, we apply the

previously mentioned random walk to this network to obtain a list of predictive genes associated with the studied disease.

3 Experimental Results

This section provides a systematic evaluation of performance for our HyMSMK method, including its three variants using different schemes. First, the experimental settings are introduced. Then, through traditional cross-validation experiments, we analyze the effects of different integration schemes and kernel sparsification, compare our method against state-of-the-art network-based algorithms across multiple disease-gene datasets. Finally, we present multiple case studies that demonstrate the disease relevance of predicted candidate genes.

3.1 Experimental settings

In this study, we employ two evaluation schemes, two types of control sets, and three kinds of standard metrics to evaluate the performance of methods.

(1) Two evaluation schemes. We employ classical 5-fold cross-validation (5FCV) to evaluate our method. For 5FCV, applied to Dataset1 and Dataset2, the known disease-related genes are randomly divided into five subsets. Each subset is used as the test set once, and the rests are to form the training set. Additionally, we demonstrate the effectiveness of our method through case studies.

(2) Two types of control sets. Both a test set and a control set for candidate genes are needed to construct a candidate gene set for evaluation. Two types of control sets will be used, including the artificial linkage-interval control set denoted as ALICS and the whole-genome control set denoted as WGCS. ALICS has 99 nearest genes for each test gene on the same chromosome [16]. All unknown genes are used as the control set in WGCS, excluding those in the training and test sets.

(3) Three kinds of standard metrics. After obtaining the predictive list of candidate genes, this study will use three standard metrics (Recall, Precision and AUPRC) to evaluate the performance of predictive algorithms. a) Recall measures the proportion of known disease genes identified within the ranking list of top-k candidate genes, with regard to the test set. b) Precision indicates the likelihood of identified known disease genes within the ranking list of top-k candidate genes. Both Recall and Precision change with the length of the top-k ranking list, allowing for a straightforward comparison of the algorithms' local performance. c) To assess overall algorithm performance, we calculate the area under the Precision-Recall curve (AUPRC), where the PRC curve plots Precision on the *y*-axis and Recall on the *x*-axis.

3.2 Comparison between different schemes incorporating multi-scale module kernel

We have designed different schemes to incorporate multi-scale module kernel, leading to several variants of our method. The predictive performance of the schemes will be compared under different scenarios (**Fig. 2**). For both control sets (ALICS and WGCS), the CPF2 scheme outperforms all others across the three kinds of evaluation metrics in Dataset1, and it can yield better or comparative results than other schemes in Dataset2 (see Supplemental materials). Therefore, CPF2 is selected as the default scheme for the subsequent studies.

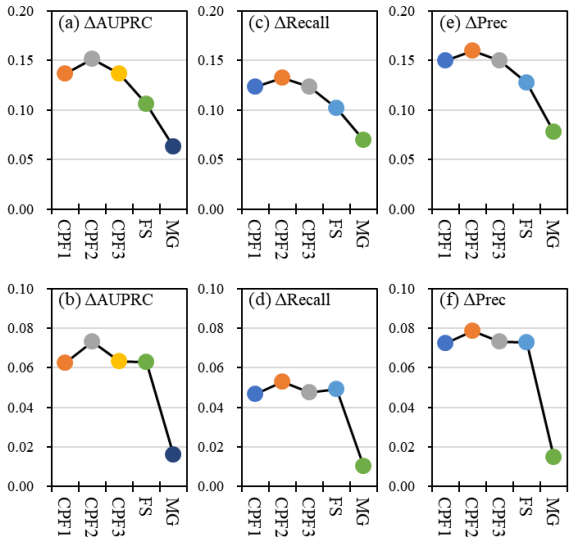

Fig. 2. Performance increase of HyMSMK with distinct schemes incorporating multi-scale module kernel, relative to baseline, in Dataset1. (a)-(b) ΔAUPRC, (c)-(d) ΔRecall, (e)-(f) ΔPrec. (a), (c) and (e) correspond to the ALICS control set, while (b), (d) and (f) correspond to the WGCS control set.

Additionally, the schemes based on k aggregation generally perform better than MG and FS. This may be because FS and MG can introduce significant interference when fusing the original GGA and nGGA networks. The MG scheme will switch the random walkers between networks with equal probability, and the predictive scores from each network are to be integrated immediately after each step. The FS scheme directly combines the GGA and nGGA networks into one and results in the loss of unique information for each. By contrast, the CPF schemes treat each network as an independent individual, minimizing interference between them.

3.3 Effect of kernel sparsification

The kernel sparsification process is helpful for improving the storage and computing of multiscale module kernel and filtering out data noises. In this section, we analyze the effect of the kernel sparsification process.

The results show that there are clear and consistent upward trends under both control sets (**Fig. 3**). The decrease of sparseness, that is, the increase of k-nearest neighbors (knn), will lead to the increasing performance of our method. It is more susceptible to small knn-values. A possible reason is that knn increasing from a low value rapidly introduces more strongly correlated information, leading to a faster improvement in performance. However, for very large knn-values, it can only add some neighbors with low similarity scores to each gene, causing the performance improvement to level off. Based on the performance trends, we can see that a reasonable choice of knn is 1000 in this study, striking a good balance between performance and speed, as it represents a turning point in the performance curve. Moreover, the performance of our method also improves rapidly as knn increases and stabilizes at a high level in Dataset2 (see Supplemental materials). Setting knn=1000 proves to also be an effective choice for our method, despite a slight decrease in Recall at this point. Notably, the AUPRC continues to improve as knn increases. It is important to mention that our method with knn=0 is equivalent to the reference method, indicating that our method with knn=1000 significantly outperforms the reference method.

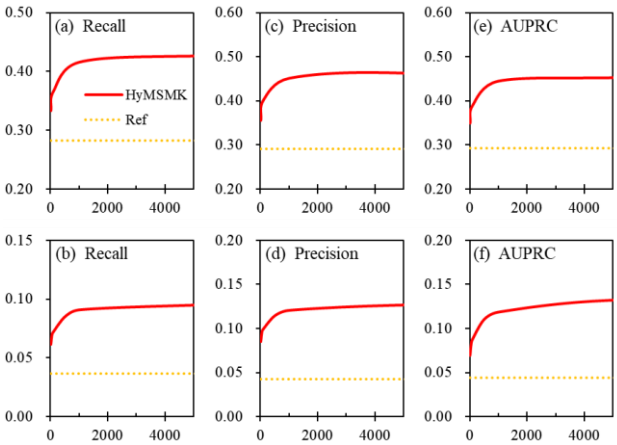

Fig. 3. Effect of kernel sparsification on predictive performance: (a)-(b) Recall, (c)-(d) Prec, (e)-(f) AUPRC (Dataset1). (a), (c) and (e) correspond to the ALICS control set, while (b), (d) and (f) correspond to the WGCS control set.

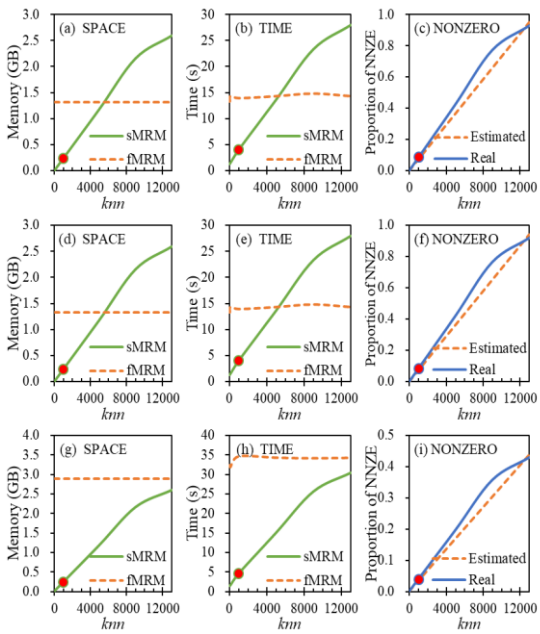

Fig. 4. Space and time requirements of HyMSMK under the sparse matrix represention mode (sMRM) and full matrix represention mode (fMRM), as well as the real and estimated proportion of non-zero elements (pNNZE), as a function of knn. (a)-(c) Test in homogeneous network nGGA; (d)-(f) Test in Dataset1; (g)-(i) Test in Dataset2.

Here, the requirements of space and time for our method under different levels of kernel sparsity will be evaluated, using both sparse matrix representation mode (sMRM) and full matrix representation mode (fMRM). We can see that usage of memory in sMRM increases approximately linearly with knn, and time consumption follows a similar trend, growing alongside space requirements (**Fig. 4**). Notably, when knn is small (e.g., knn≤1000), the memory usage of sMRM is significantly lower than that of fMRM. For example, for the default value of knn=1000 derived in the above analysis of performance, the memory usage of MSMK after the sparsification process is about 0.24GB under sMRM, while it is about 1.34 GB under fMRM. As a result, the memory usage of HyMSMK (estimated) and its time consumption are significantly reduced under sMRM, much less than that in fMRM.

We note that the memory usage of HyMSMK using sparsified kernel under sMRM is similar to that under fMRM when knn is close to 5000, and it will become larger with the increase of knn (see **Fig. 4**d-e). This is because the sparsity of the kernel matrix will decrease with the increase of knn, gradually becoming non-sparse, and the forced sparse representation of non-sparse matrix will lead to the increase of memory usage and time consumption, due to the inherent mechanism of sMRM. This is also the case in Dataset2 (**Fig. 4**g-h), but the critical point of knn (where the memory/time usage of sMRM is close to that of fMRM) is delayed. As we see, the memory/time usage of sMRM is still less than that of fMRM when knn is close to the number of nodes in gene network. This is because the size of DDA in Dataset2 is much larger than that in Dataset1, and its DDA and DGA are sparse.

In addition, we can see that the growth rate of space consumption slows down when knn is very large, and this is also the case for the growth rate of time consumption. This is because the space and time consumptions are closely related to the number of non-zero elements (NNZE) in matrix. For example, **Fig. 4**c shows that the growth rate of NNZE already slows down when knn is very large, while the growth rate for small value of knn is quicker than the estimates using knn as mean degree of nodes.

In summary, the kernel sparsification process effectively reduces memory usage and time consumption, enhancing its storage efficiency and computational speed. Moreover, HyMSMK utilizing sparsified kernel (e.g., knn=1000) maintains stable and high performance in disease-gene identification.

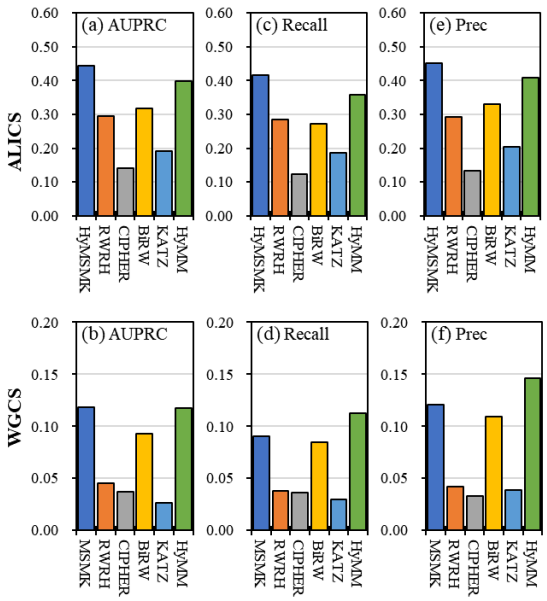

Fig. 5. For Dataset1, performance comparison to other algorithms. (a)-(b) AUPRC, (c)-(d) Recall, and (e)-(f) Precision. The first and second rows correspond to the ALICS and WGCS control sets, respectively.

3.4 Comparison to other algorithms

We compare the performance of HyMSMK to network-based baselines including RWRH [57], CIPHER [58], BiRW [59], KATZ [60] and HyMM[55], by the 5FCV experiments. **Fig. 5** demonstrates that our method outperforms the basic baseline algorithms (RWRH, CIPHER, BiRW and KATZ) under ALICS and WGCS control sets. In details, HyMSMK improves the AUPRC-, Recall-, and Prec-values by 41%, 47%, and 37%, respectively, under ALICS, and by 27%, 7%, and 10%, respectively, under WGCS, when compared to the best of the basic baselines. Further, we compare HyMSMK to another multiscale module-based algorithm called HyMM. The results show that HyMSMK outperforms HyMM under ALICS control

sets, and they have similar values of AUPRC under WGCS control sets (see **Fig. 5**). The Recall and Precision curves further confirm the strong local performance of our method (**Fig. 6**).

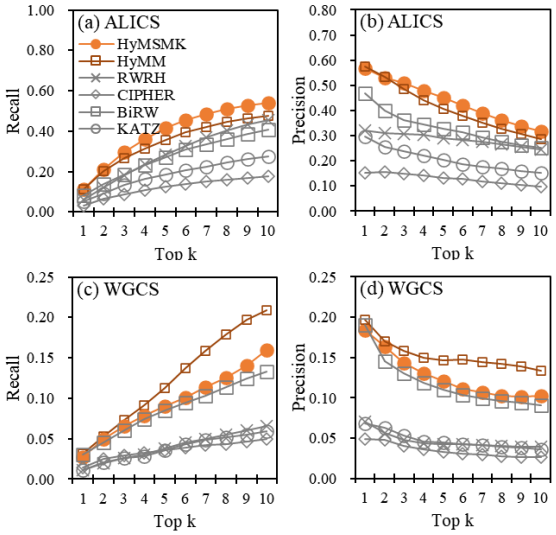

Fig. 6. For Dataset1, comparison of local performance (Recall and Precision) of HyMSMK to other algorithms. (a)-(b) ALICS, (c)-(d) WGCS.

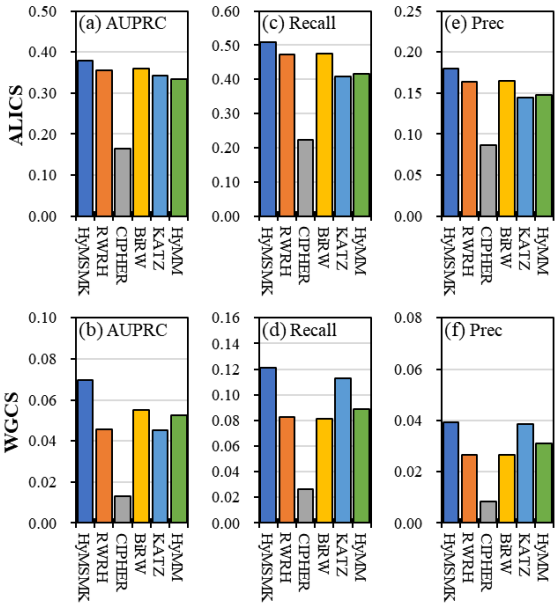

Fig. 7. For Dataset2, performance comparison to other algorithms. (a)-(b) AUPRC, (c)-(d) Recall, (e)-(f) Prec. The first and second rows correspond to the ALICS and WGCS control sets, respectively.

The results in **Fig. 7** show that HyMSMK also surpasses all the comparison algorithms under both control sets, in Dataset2. HyMSMK improves AUPRC, Recall, and Precision by 5%, 7%, and 9% under ALICS, and by 27%, 47%, and 48% under WGCS, respectively, when compared to the best of the comparison algorithms. In this dataset, our HyMSMK in terms of AUPRC, Recall, and Precision has consistently better performance than the multiscale module-based algorithm HyMM under both control sets. **Fig. 8** further corroborates HyMSMK's excellent local performance. Overall, HyMSMK proves to be a consistently strong choice under both control sets.

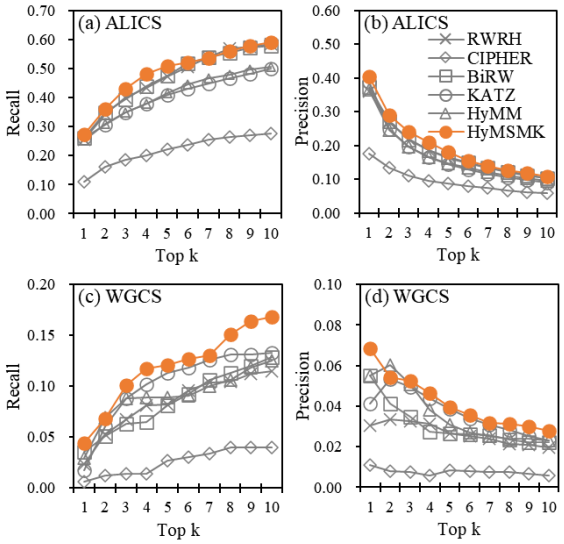

Fig. 8. For Dataset2, comparison of local performance (Recall and Precision). (a)-(b) ALICS, (c)-(d) WGCS.

3.5 Case studies

We perform case studies for several diseases to demonstrate the effectiveness of our method further. For each disease, a list of candidate genes is generated using the OMIM database as the training set. The disease relevance of these predicted genes is then evaluated through literature validation and differential expression analysis.

Table 2 presents the predicted top-k candidate genes for Alzheimer's disease (AD, #104300). Among these, 13 genes have supporting literature records related to AD, including 5 genes with AD-related entries in the DisGeNet database (APOE, TF, PSEN1, MTHFR, and MAOB). Additionally, 7 genes are supported by AD-related literature but are not listed in DisGeNet (EIF4G1, SNCA, SNCB, MAOA,

MTHFD1, CP, and AOC3). For instance, the APOE ε2 allele is one of the genetic protective factors in AD [61, 62]; Sutovsky et al showed that the APOE4 homozygotes and the APOE4 carriers with TT genotype of MTHFR rs1801133 have the highest risk of AD [63]. Jouini et al observed that the plasma concentrations of transferrin (TF), ceruloplasmin (CP), and iron in AD group were significantly lower than that of adults with normal cognition, and there was a strong correlation between most of them[64]. Zhang et al [65] found that a mutation (S169del) of PSEN1 can affect the APP processing and Aβ generation, and the neuritic plaque formation and the deficits of learning and memory can be promoted in AD model mice. This suggests that this mutation (S169del) may be an important site of AD treatment as a potential target. Florentinus-Mefailoski demonstrated that an increase in the mean precursor strength of peptides from various plasma proteins, including EIF4G1, is associated with AD [66]. α-Synuclein and tau aggregations are the neuropathological molecular feature of Parkinson's disease and Alzheimer's disease, respectively, but they also may co-occur in patients. Bassil showed that the accumulation and spread of tau in mice were weakened due to the lack of endogenous α-Synuclein, but the seeding or spreading ability of α-syn was not affected by the lack of tau. This indicates the important role of α-Synuclein in modulating tau pathological burden and spreads in the brains of patients [67]. Bergström reported the increased levels of synuclein beta (SNCB) in the cerebrospinal fluid (CSF) of AD patients compared to controls [68]. Similarly, Quartey et al. discovered that MAO-A activity increased in the cortical regions of AD patients but not in hippocampal samples, whereas MAO-B activity showed an increase in both regions [69]. Jiang et al showed that the polymorphisms of Methylenetetrahydrofolate reductase (MTHFR) gene: C677T and A1298C were related to AD in the Chinese population [70]; and Zuin et al showed that MTHFR C677T polymorphism had the association trend with late-onset Alzheimer's disease in Italian population [71]. Bi et al showed that the G1958A A allele of MTHFD1 may be a weak risk factor of early onset AD [72].

Stroke and AD have evident interrelationships mainly because the vascular system dysfunction (VSD) and their pathologies have important effects on cerebral vessels, together with increased SSAO levels. The VSD can directly lead to the occurrence and progression of AD, and there are signs of dysfunction of blood-brain barrier (BBB) in the early stages of AD. Solé et al showed that the expression of SSAO/VAP-1 was related to endothelial activation through changing the release of pro-inflammatory and pro-angiogenic neuroproteins, and revealed the important role of vascular SSAO/VAP-1 in the BBB dysfunction related to AD progression, which provides new insights to the study of AD's therapeutic targets [73].

Table 2 For Alzheimer's disease, candidate genes generated by our method.

Rank	Symbol	DisGeNet	Druggable	DGIdb	Literature Validation
1	TFR2	-	-	0	-
2	APOE	Y	Tier 3A	20	Yes
3	TF	Y	Tier 3A	12	Yes
4	PSEN1	Y	Tier 1	10	Yes
5	EIF4G1	-	-	1	Yes
6	SNCA	-	Tier 2	33	Yes
7	GBA	-	Tier 1	32	-
8	SNCB	-	-	0	Yes
9	MAOA	-	Tier 1	47	Yes
10	MTHFR	Y	-	26	Yes
11	MTHFD1	-	-	3	Yes
12	SPACA3	-	Tier 3B	0	-
13	MTR	-	-	9	-
14	AOC1	-	Tier 2	3	-
15	CP	-	Tier 3A	20	Yes
16	MTRR	-	-	3	-
17	MAOB	Y	Tier 1	68	Yes
18	AOC3	-	Tier 1	6	Yes
19	IFIT3	-	-	0	-
20	AOC2	-	Tier 2	0	-

Table 1 For Breast cancer, candidate genes generated by our method.

Rank	Symbol	Differential Expression			DisGeNet	Druggable (Science 2017)	DGIdb
		Sig	log2FC	Trend			#Drug
1	BRCA1	Y	0.49	Up	Y	Tier 2	97
2	RAD51C	Y	0.25	Up	Y	-	0
3	NTHL1	Y	0.57	Up	-	-	0
4	SEC23B	Y	0.33	Up	-	-	0
5	MLH1	-	-	-	Y	-	0
6	SMAD7	-	-	-	-	-	2
7	MSH2	Y	0.23	Up	Y	-	1
8	RBBP8	-	-	-	-	-	0
9	XRCC2	Y	0.8	Up	Y	-	0
10	PMS2	Y	0.15	Up	Y	-	2
11	RAD51D	-	-	-	Y	-	0
12	CTNNB1	-	-	-	-	Tier 2	20
13	TOPBP1	-	0.19	Up	-	-	0
14	PATZ1	Y	0.29	Up	-	-	0
15	PRKN	-	-	-	-	-	0
16	TPX2	Y	1.64	Up	-	-	0
17	OPCML	-	0.41	Up	-	Tier 3A	1
18	MSH6	Y	0.27	Up	Y	-	2
19	NRAS	Y	0.21	Up	-	-	48
20	BLM	Y	0.98	Up	-	Tier 1	100

Furthermore, we analyze the top-k candidate genes predicted for other diseases, evaluating their associations with specific conditions through gene differential expression analysis [74] and the DisGeNet database [2]. shows the predicted top-20 candidate genes for breast cancer (#114480), of which 12 are significantly differentially expressed in breast cancer, and 8 have breast cancer-related entries in the DisGeNet database, suggesting a strong likelihood of their associations with the disease. **Table 2** presents the predicted top-20 candidate genes for lung cancer (#211980), with 10 genes significantly differentially expressed and 7 recorded in the DisGeNet database as associated with lung cancer, indicating their potential relevance to the disease. **Table 5** lists the predicted top-20 candidate genes for prostate cancer (#176807), among which 7 are significantly differentially expressed in prostate cancer, and 2 genes have prostate cancer-related entries in the DisGeNet database, implying a probable link to the disease.

Table 2 For Lung cancer, candidate genes generated by our method.

Rank	Symbol	Differential Expression			DisGeNet	Druggable (Science 2017)	DGIdb
		Sig	log2FC	Trend			#Drug
1	TP53	-	-	-	Y	Tier 1	383
2	DICER1	-	-	-	-	-	0
3	RB1	-	-	-	Y	Tier 2	23
4	BAP1	-	-	-	Y	-	6
5	TSC1	-	-0.24	Down	-	-	3
6	TSC2	-	-	-	-	-	1
7	MLH1	Y	-0.3	Down	Y	-	0
8	POU2F1	Y	0.45	Up	-	-	0
9	MSH2	Y	1.13	Up	-	-	1
10	MUC5B	-	-	-	-	Tier 3B	0
11	SFTPA2	Y	-3.18	Down	-	Tier 3B	0
12	HRAS	Y	0.97	Up	Y	Tier 1	38
13	FAS	-	-	-	-	Tier 3A	19
14	NRAS	Y	0.38	Up	Y	-	48
15	CHEK2	Y	1.87	Up	-	Tier 1	29
16	SHOC2	Y	-0.35	Down	-	-	0
17	RASGRP2	Y	-2.5	Down	-	-	0
18	RAP1GDS1	-	-	-	-	-	0
19	ERCC5	Y	-0.33	Down	-	Tier 2	1
20	AKT1	-	-	-	Y	Tier 1	78

Table 3 For prostate cancer, candidate genes generated by our method.

Rank	Symbol	Differential Expression			DisGeNet	Druggable (Science 2017)	DGIdb
		Sig	log2FC	Trend			#Drug
1	TP53	Y	0.35	Up	-	Tier 1	383
2	DHH	-	-	-	-	Tier 3A	0
3	EPHB2	-	-	-	Y	Tier 1	4
4	RNASEL	Y	-0.42	Down	Y	-	2
5	KLF4	-	-0.39	Down	-	-	1
6	APOM	Y	-0.41	Down	-	Tier 3A	0
7	AFP	-	-	-	-	Tier 1	12
8	MSH3	Y	0.25	Up	-	-	0
9	MSH6	-	-	-	-	-	2
10	MLH3	-	-	-	-	-	0
11	ANPEP	-	-	-	-	Tier 2	13
12	HIP1R	-	-	-	-	-	0
13	CLTB	Y	-0.22	Down	-	-	0
14	CLTA	Y	0.2	Up	-	-	0
15	PNO1	-	-	-	-	-	0
16	CDO1	-	-	-	-	-	2
17	RPL23AP32	-	-	-	-	-	0
18	CLTC	Y	0.26	Up	-	-	0
19	GOT1	-	-	-	-	-	7
20	LGALSL	-	-	-	-	-	0

4 Conclusion and discussion

Disease-gene discovery is a critical area of biomedical data mining, driven by the high cost and time required for biomedical experiments. The modular structure of biological networks offers valuable information for disease research, which is embedded in multi-scale module structures. However, effectively leveraging the information from multi-scale module structures to identify disease-associated genes remains a challenge. To address this, we have proposed the HyMSMK method of incorporating the multi-scale module kernel. It extracts the multi-scale modules at different scales from a comprehensive interactome through multi-scale modularity optimization and exponential sampling, and then constructs the multi-scale module profile, which can capture multi-scale information from local structure to global ones. The multi-scale module kernel is constructed using the multi-scale module profile, combined with the relative information content (IC), which is further preprocessed by kernel sparsification. Multiple schemes for incorporating multi-scale module kernel are designed to explore more effective approaches for discovering potential disease-related genes.

Through a series of experiments, we have evaluated our HyMSMK method under both ALICS and WGCS control sets systematically and examined the effects of various incorporating schemes and kernel sparsification. The results across various scenarios demonstrated the strong performance of our method, affirming the utility of multi-scale module structure in disease-gene identification. Among the schemes, CPF2 is recommended as it consistently delivers reliable results across various scenarios. Additionally, case studies for specific diseases validate the relevance of candidate genes identified by our method, through literature support and differential expression analysis, further emphasizing our method's effectiveness.

Multiscale module mining is still an open issue in complex networks, including biological networks. It is based on the given network that the mining of multi-scale modules are conducted. So, the extraction and optimization of multi-scale modules would inevitably be affected by the scale of the data or the sparsity of the network. Our previous theoretical research for module mining algorithms found that existing algorithms based on global target function optimization have two types of resolution limitations. The first type of limitation is that when the network size is very large, some small network modules may be hidden and cannot be directly discovered by the algorithms. The second type of limitation is that although it is possible to identify

modules at different scales through algorithms with adjustable resolution, when identifying small-scale modules by increasing resolution, some large modules may experience cracking issues. That is to say, under certain conditions, it is impossible to identify both large-sized and small-sized modules simultaneously. There is still a lot of research to this day in this field of module mining. The analysis and mining of biological networks from the perspective of modules can alleviate the problem that the study of a single node or edge is more easily affected by network noise. Furthermore, compared to traditional hard partitioning methods of network modules, multi-scale module mining can extract multi-scale information from local to global levels. The impact of network sparsity will further be reduced due to its more detailed network mining.

The method in this work has been proved to be able to achieve good performance, while there is still room for further improvement. To utilize the information in multi-scale module structure, we provide an effective strategy in this article, but it does not mean that the information has been fully and effectively utilized. We use the extracted multi-scale information to construct a multi-scale module kernel for disease-gene prediction, but it may also serve as a feature input for machine learning algorithms such as support vector machines, random forests, and neural networks to train machine learning models.

Module structures widely exist in various biological networks, such as PPI networks, metabolic networks, and disease-disease networks. In this work, we only consider the multi-scale modules in GGA networks, while disease-gene prediction may benefit from the mining of multi-scale modules in DDA networks. In disease networks, similar diseases may aggregate into network modules, and there are also varying degrees of similarity between different diseases. From the perspective of different degrees of similarity, disease modules at different levels can be extracted. These pieces of information have the potential to improve disease networks and provide more valuable information for disease-gene prediction.

The HyMSMK method uses MO-ES to extract multiscale modules due to its stable performance in our previous research, while it is possible to enhance disease-gene prediction based on multiscale modules by developing more effective module identification algorithms. For instance, Hu et al proposed a method for clustering analysis by exploiting a variety of network motifs which are a kind of small patterns recurring in a network and showed that taking account of higher-order structures could obtain new insight into the analysis of biological networks [75].

In this work, we only use data for disease-related genes/phenotypes/symptoms and gene-gene networks. In past decades, a large amount of omics data from genomics to transcriptomics and proteomics has been generated continuously [76]. The occurrence of diseases can be reflected at different levels of the organism, from genes to protein complexes and pathways. In theory, the utilization of more relevant omics data, e.g., gene mutations, methylations and transcription factors, may improve the mining of biological knowledge. Disease-related research will benefit from the accumulation in the amount and types of omics data, enhancing the ability of disease-gene prediction[77].

Personalized study of diagnosis and treatment is an inevitable trend in the future development of precision medicine. The availability of patient-level datasets containing both genotypic and phenotypic data offers new opportunities to explore individual pathogenic genes, though most research currently focuses on disease classes or their subcategories. With advancements in single-cell sequencing technologies, vast amounts of cell-level multi-omics data are being generated at an unprecedented rate [76, 78]. Integrating cell-level multi-omics data across various molecular levels remains challenging due to its inherent heterogeneity, while it has the potential to uncover the complexity of biological systems.

In conclusion, this study presents an efficient framework HyMSMK by incorporating the multi-scale module kernel into disease-gene identification, which may offer valuable insights for leveraging multi-scale module structures in the exploration of other bio-entity associations[79-82]. It can theoretically be generalized to explore associations between other biological entities, such as disease-metabolite associations, disease-

ncRNA associations and disease-drug associations, since the networks consisting of these bio-entities also have module structures.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under (Grant No.62472051), the Hunan Provincial Natural Science Foundation of China (Grant No. 2024JJ5006), the Research project of Changsha University of Science and Technology (097_000303919), the National Key Research and Development Program of China (Grant No. 2019YFA0706202), the National Natural Science Foundation of China under (Grant No.62202502) and the High Performance Computing Center of Central South University. .

References

[1] R. Oughtred, et al., The BioGRID interaction database: 2019 update, Nucleic Acids Research, 2019, 47(D1): D529-D541.
[2] J. Piñero, et al., DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, Nucleic Acids Research, 2017, 45(D1): D833-D839.
[3] S. Köhler, et al., The Human Phenotype Ontology in 2021, Nucleic Acids Research, 2020, 49(D1): D1207-D1217.
[4] X. Zhou, et al., Human symptoms–disease network, Nature Communications, 2014, 5: 4212.
[5] L.Y.-H. Lee, J. Loscalzo, Network Medicine in Pathobiology, The American Journal of Pathology, 2019, 189(7): 1311-1326.
[6] A.-L. Barabasi, et al., Network medicine: a network-based approach to human disease, Nat Rev Genet, 2011, 12(1): 56-68.
[7] I.F. do Valle, et al., Network medicine framework shows that proximity of polyphenol targets and disease proteins predicts therapeutic effects of polyphenols, Nature Food, 2021, 2(3): 143-155.
[8] Y. Moreau, L.-C. Tranchevent, Computational tools for prioritizing candidate genes: boosting disease gene discovery, Nat Rev Genet, 2012, 13(8): 523-536.
[9] P. Ruan, S. Wang, DiSNEP: a Disease-Specific gene Network Enhancement to improve Prioritizing candidate disease genes, Briefings in Bioinformatics, 2021, 22(4): bbaa241.
[10] Y. Liu, et al., Pathogenic gene prediction based on network embedding, Briefings in Bioinformatics, 2021, 22(4): bbaa353.
[11] S.K. Ata, et al., Recent advances in network-based methods for disease gene prediction, Briefings in Bioinformatics, 2021, 22(4): bbaa303.
[12] A. del Sol, et al., Diseases as network perturbations, Current Opinion in Biotechnology, 2010, 21(4): 566-571.
[13] A. Valdeolivas, et al., Random walk with restart on multiplex and heterogeneous biological networks, Bioinformatics, 2018, 35(3): 497–505.
[14] J. Xiang, et al., NIDM: network impulsive dynamics on multiplex biological network for disease-gene prediction, Briefings in Bioinformatics, 2021, 22(5): bbab080.
[15] L.I. Furlong, Human diseases through the lens of network biology, Trends in Genetics, 2013, 29(3): 150-159.
[16] S. Köhler, et al., Walking the Interactome for Prioritization of Candidate Disease Genes, The American Journal of Human Genetics, 2008, 82(4): 949-958.
[17] K. Hu, et al., Predicting disease-related genes by path structure and community structure in protein–protein networks, J. Stat. Mech., 2018, 2018(10): 100001.
[18] L. Cowen, et al., Network propagation: a universal amplifier of genetic associations, Nature Reviews Genetics, 2017, 18: 551.
[19] R. Jiang, Walking on multiple disease-gene networks to prioritize candidate genes, Journal of Molecular Cell Biology, 2015, 7(3): 214-230.
[20] H. Zhang, et al., Benchmarking network-based gene prioritization methods for cerebral small vessel disease, Briefings in Bioinformatics, 2021, 22(5): bbab006.
[21] Y. Li, J. Li, Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data, BMC Genomics, 2012, 13(7): S27.
[22] J. Xiang, et al., PrGeFNE: Predicting disease-related genes by fast network embedding, Methods, 2021, 192: 3-12.
[23] S. Choobdar, et al., Assessment of network module identification across complex diseases, Nature Methods, 2019, 16(9): 843-852.
[24] S.S. Bhowmick, B.S. Seah, Clustering and Summarizing Protein-Protein Interaction Networks: A Survey, IEEE Transactions on Knowledge and Data Engineering, 2016, 28(3): 638-658.
[25] S.K. Dwivedi, et al., Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder, Nature Communications, 2020, 11(1): 856.
[26] S.D. Ghiassian, et al., A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome, PLOS Computational Biology, 2015, 11(4): e1004120.
[27] S. Fortunato, D. Hric, Community detection in networks: A user guide, Phys. Rep., 2016, 659: 1-44.
[28] D. Jin, et al., A Survey of Community Detection Approaches: From Statistical Modeling to Deep Learning, IEEE Transactions on Knowledge and Data Engineering, 2023, 35(2): 1149 - 1170.
[29] Z. He, et al., Detecting Statistically Significant Communities, IEEE Transactions on Knowledge and Data Engineering, 2022, 34(6): 2711 - 2725.
[30] X. Liu, et al., Evaluation of Community Detection Methods, IEEE Transactions on Knowledge and Data Engineering, 2020, 32(9): 1736-1746.
[31] S. Fortunato, M. Barthélemy, Resolution limit in community detection, Proc. Natl. Acad. Sci. USA, 2007, 104(1): 36-41.
[32] J. Xiang, et al., Phase transition of Surprise optimization in community detection, Physica A, 2018, 491(C): 693-707.
[33] J. Xiang, et al., Community detection based on significance optimization in complex networks, J. Stat. Mech., 2017,

2017(5): 053213.
[34] A.-L. Barabasi, Z.N. Oltvai, Network biology: understanding the cell's functional organization, Nat Rev Genet, 2004, 5(2): 101-113.
[35] L.V. Schaffer, T. Ideker, Mapping the multiscale structure of biological systems, Cell Systems, 2021, 12(6): 622-635.
[36] T.G.O. Consortium, The Gene Ontology Resource: 20 years and still GOing strong, Nucleic Acids Research, 2018, 47(D1): D330-D338.
[37] M. Giurgiu, et al., CORUM: the comprehensive resource of mammalian protein complexes—2019, Nucleic Acids Research, 2018, 47(D1): D559–D563.
[38] X. Meng, et al., DPCMNE: detecting protein complexes from protein-protein interaction networks via multi-level network embedding, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 19(3): 1592-1602.
[39] A. Lewis, et al., The function of communities in protein interaction networks at multiple scales, BMC Syst. Biol., 2010, 4(1): 100.
[40] J. Wang, et al., A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 8(3): 607 - 620.
[41] E.M. Ramos, et al., Phenotype–Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources, European Journal of Human Genetics, 2014, 22(1): 144-147.
[42] J.S. Amberger, et al., OMIM.org: leveraging knowledge across phenotype–gene relationships, Nucleic Acids Research, 2018, 47(D1): D1038-D1043.
[43] J. Menche, et al., Uncovering disease-disease relationships through the incomplete interactome, Science, 2015, 347(6224): 841.
[44] S. Köhler, et al., Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources, Nucleic Acids Research, 2018, 47(D1): D1018-D1027.
[45] D. Guala, et al., Genome-wide functional association networks: background, data & state-of-the-art resources, Briefings in Bioinformatics, 2019, 21(4): 1224-1237.
[46] X. Peng, et al., Protein–protein interactions: detection, reliability assessment and applications, Briefings in Bioinformatics, 2016, 18(5): 798-819.
[47] G. Salton, et al., A vector space model for automatic indexing, Commun. ACM, 1975, 18(11): 613–620.
[48] P. Resnik, Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, Journal of artificial intelligence research 1999, 11 95-130.
[49] A. Singhal, et al., Multiscale community detection in Cytoscape, PLOS Computational Biology, 2020, 16(10): e1008239.
[50] M.E. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys Rev E 2004, 69(2): 026113.
[51] J. Xiang, et al., Multi-resolution community detection based on generalized self-loop rescaling strategy, Physica A, 2015, 432: 127-139.
[52] S. Rahiminejad, et al., Topological and functional comparison of community detection algorithms in biological networks, BMC Bioinformatics, 2019, 20(1): 212.
[53] J. Xiang, et al., Identifying multi-scale communities in networks by asymptotic surprise, J. Stat. Mech., 2019, 2019(3): 033403.
[54] J. Xiang, et al., Multi-resolution modularity methods and their limitations in community detection, Eur. Phys. J. B, 2012, 85(10): 1-10.
[55] J. Xiang, et al., HyMM: Hybrid method for disease-gene prediction by integrating multiscale module structure, Briefings in Bioinformatics, 2022, 23(3): bbac072.
[56] X. Li, et al., A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications, Briefings in Bioinformatics, 2017, 20(1): 178-189.
[57] Y. Li, J.C. Patra, Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network, Bioinformatics, 2010, 26(9): 1219-1224.
[58] X. Wu, et al., Network-based global inference of human disease genes, Molecular Systems Biology, 2008, 4: 189-189.
[59] M. Xie, et al., Network-based Phenome-Genome Association Prediction by Bi-Random Walk, Plos One, 2015, 10(5): e0125138.
[60] U.M. Singh-Blom, et al., Prediction and Validation of Gene-Disease Associations Using Methods Inspired by Social Network Analyses, Plos One, 2013, 8(5): e58977.
[61] A. Serrano-Pozo, et al., APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches, The Lancet Neurology, 2021, 20(1): 68-80.
[62] K.A. Zalocusky, et al., An Alzheimer's-disease-protective APOE mutation, Nature Medicine, 2019, 25(11): 1648-1649.
[63] S. Sutovsky, et al., Allelic Distribution of Genes for Apolipoprotein E and MTHFR in Patients with Alzheimer's Disease and Their Epistatic Interaction, J Alzheimers Dis, 2020, 77(3): 1095-1105.
[64] N. Jouini, et al., Impacts of Iron Metabolism Dysregulation on Alzheimer's Disease, Journal of Alzheimer's Disease, 2021, 80: 1439-1450.
[65] S. Zhang, et al., A presenilin-1 mutation causes Alzheimer disease without affecting Notch signaling, Molecular Psychiatry, 2020, 25(3): 603-613.
[66] A. Florentinus-Mefailoski, et al., The plasma peptides of Alzheimer's disease, Clinical Proteomics, 2021, 18(1): 17.
[67] F. Bassil, et al., α-Synuclein modulates tau spreading in mouse brains, Journal of Experimental Medicine, 2021, 218(1): e20192193.
[68] S. Bergström, et al., Multi-cohort profiling reveals elevated CSF levels of brain-enriched proteins in Alzheimer's disease, Annals of Clinical and Translational Neurology, 2021, 8(7): 1456-1470.
[69] M.O. Quartey, et al., Alzheimer Disease and Selected Risk Factors Disrupt a Co-regulation of Monoamine Oxidase-A/B in the Hippocampus, but Not in the Cortex, Front Neurosci, 2018, 12: 419.
[70] Y. Jiang, et al., Genetic effect of MTHFR C677T, A1298C, and A1793G polymorphisms on the age at onset, plasma

homocysteine, and white matter lesions in Alzheimer's disease in the Chinese population, Aging (Albany NY), 2021, 13(8): 11352-11362.
[71] M. Zuin, et al., Methylenetetrahydrofolate reductase C667T polymorphism and susceptibility to late-onset Alzheimer's disease in the Italian population, Minerva Med, 2021, 112(3): 365-371.
[72] X.H. Bi, et al., Association analysis of CbetaS 844ins68 and MTHFD1 G1958A polymorphisms with Alzheimer's disease in Chinese, J Neural Transm (Vienna), 2010, 117(4): 499-503.
[73] M. Solé, et al., Blood-brain barrier dysfunction underlying Alzheimer's disease is induced by an SSAO/VAP-1-dependent cerebrovascular activation with enhanced Aβ deposition, Biochim Biophys Acta Mol Basis Dis, 2019, 1865(9): 2189-2202.
[74] H.M. Dingerdissen, et al., BioMuta and BioXpress: mutation and expression knowledgebases for cancer biomarker discovery, Nucleic Acids Research, 2017, 46(D1): D1128-D1136.
[75] L. Hu, et al., HiSCF: leveraging higher-order structures for clustering analysis in biological networks, Bioinformatics, 2020, 37(4): 542-550.
[76] Y. Li, et al., Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine, Briefings in Bioinformatics, 2021, 22(5): bbab024.
[77] J. Yan, et al., Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data, Briefings in Bioinformatics, 2017, 19(6): 1370-1381.
[78] M. Forcato, et al., Computational methods for the integrative analysis of single-cell data, Briefings in Bioinformatics, 2021, 22(3): bbaa042.
[79] M. Chen, et al., Drug-Target Interactions Prediction Based on Signed Heterogeneous Graph Neural Networks, Chinese Journal of Electronics, 2024, 33(1): 231-244.
[80] X. Li, et al., Multiplex Networks and Pan-Cancer Multiomics-Based Driver Gene Identification Using Graph Neural Networks, Big Data Mining and Analytics, 2024.
[81] L. Peng, et al., Deciphering ligand–receptor-mediated intercellular communication based on ensemble deep learning and the joint scoring strategy from single-cell transcriptomic data, Computers in Biology and Medicine, 2023, 163: 107137.
[82] L. Peng, et al., Drug repositioning via Multi-view Representation Learning with Heterogeneous Graph Neural Network, IEEE Journal of Biomedical and Health Informatics, 2024: 1-12.