frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 22 September 2023
DOI 10.3389/fgene.2023.1222346

Check for updates

OPEN ACCESS

EDITED BY
Alessio Martino,
Guido Carli Free International University for Social Studies, Italy

REVIEWED BY
Zhiwei Ji,
Nanjing Agricultural University, China
Liang Qi,
Shandong University of Science and Technology, China

*CORRESPONDENCE
Zhe Liu,
✉ 20170049@cuz.edu.cn

RECEIVED 14 May 2023
ACCEPTED 11 September 2023
PUBLISHED 22 September 2023

CITATION
Shi W, Feng H, Li J, Liu T and Liu Z (2023), DapBCH: a disease association prediction model Based on Cross-species and Heterogeneous graph embedding. *Front. Genet.* 14:1222346. doi: 10.3389/fgene.2023.1222346

DapBCH: a disease association prediction model Based on Cross-species and Heterogeneous graph embedding

Wanqi Shi[1], Hailin Feng[1], Jian Li[1], Tongcun Liu[1] and Zhe Liu[2]*

[1]School of Mathematics and Computer Science, Zhejiang A & F University, Hangzhou, Zhejiang, China, [2]College of Media Engineering, Zhejiang University of Media and Communications, Hangzhou, Zhejiang, China

The study of comorbidity can provide new insights into the pathogenesis of the disease and has important economic significance in the clinical evaluation of treatment difficulty, medical expenses, length of stay, and prognosis of the disease. In this paper, we propose a disease association prediction model DapBCH, which constructs a cross-species biological network and applies heterogeneous graph embedding to predict disease association. First, we combine the human disease–gene network, mouse gene–phenotype network, human–mouse homologous gene network, and human protein–protein interaction network to reconstruct a heterogeneous biological network. Second, we apply heterogeneous graph embedding based on meta-path aggregation to generate the feature vector of disease nodes. Finally, we employ link prediction to obtain the similarity of disease pairs. The experimental results indicate that our model is highly competitive in predicting the disease association and is promising for finding potential disease associations.

KEYWORDS

heterogeneous network, biological information, comorbidity, cross-species, multiple networks

1 Introduction

It is important to understand the association between diseases for diagnosing, curing, and taking precautions against diseases (Jin et al., 2019). Since there may be direct or indirect causal relationships between diseases, patients experience multiple diseases at the same time. For example, end-stage renal disease (ESRD) is frequently present in HIV-1 patients, while chronic obstructive pulmonary disease (COPD) is often accompanied by lung cancer, osteoporosis, cachexia, and cardiovascular disease (Mlynarski et al., 2005; Decramer and Janssens, 2013). In addition, the tumor-related disease always indicates serious complications, which has a high comorbidity pattern with hypertension, respiratory diseases, and cerebrovascular diseases (Ko et al., 2016). Moreover, the main cause of death in patients with cirrhosis is hepatocellular carcinoma (HCC) (Ji et al., 2015). Currently, numerous treatment strategies have been investigated in relation to various diseases (Shao et al., 2013). However, the current medical research on comorbidity is not perfect, and there is still great uncertainty in the association between many diseases. Therefore, how to effectively analyze the relationship between diseases and discover potential comorbidity relationships has become a new research issue.

Previously, people employed traditional biological experimental methods to explore the association between diseases. It requires a lot of human resources and financial support, for the reason that these methods need to target a great quantity of genes to identify related diseases (Uffelmann et al., 2021). Fortunately, with the large-scale integration of various experimental data such as human gene annotations, disease phenotypes, and protein–protein interaction, data support has been provided to better elucidate the underlying biological mechanisms of complex diseases (Silverman et al., 2020; Graw et al., 2021). Although these biological data are complex, it becomes simple and clear to represent various biological data forms through the network, so network-based approaches have emerged as the main prediction method for the association between diseases (Wang et al., 2011; Ji et al., 2019; Yue et al., 2020).

The network-based models are based on the hypothesis of 'guilt-by-association,' where genes close to each other in physiology or functionally often participate in the same biological pathway (Aravind, 2000; Oliver, 2000). Since measuring the distance between candidate genes and known disease genes in the protein–protein interaction (PPI) network is a crucial component of the network-based model, numerous computerized methods have been devised (Sharan et al., 2007; Shlomi et al., 2008). Goh et al. (2007) built a network that links two different genes together if they are linked to the same disease. Tian et al. (2008) reconstructed a network by adding protein interactions, gene interactions, and gene expression correlations to calculate more accurate distances between genes. Ulitsky and Shamir (2007) further supplemented this by adding interactions from human cell cycle networks and yeast two-hybrid experiments. The aforementioned three models all utilize the direct protein interaction, but they all suffer from insufficient network data. Based on the assumption that diseases with an overlapping phenotype have potentially functionally similar genes, Wu et al. (2008), Li and Patra (2010), and Vanunu et al. (2010) have found that combining disease phenotype networks and PPI networks in priority tasks would result in better performance of the model. Wu et al. (2008) employed the linear regression method CIPHER to predict the disease gene associations by combining the PPI network and phenotype network, and analyze the relevance between the distribution of phenotypic similarity and gene compact distribution in the protein interaction network. Li and Patra (2010) proposed the RWRH model to apply the RWR algorithm to the heterogeneous network they constructed, connecting gene networks and phenotype similar networks using known gene phenotype associations. Luo and Liang (2015) further developed RWRHN on the basis of the RWRH algorithm, enabling it to carry out random walks on heterogeneous networks to predict potential candidate genes for genetic diseases. RWRHN is a random walk algorithm with resistance, and its main contribution is to predict the background of protein network reconstruction through linking, so as to obtain a more reliable PPI network. The heterogeneous network can describe the real world more precisely compared to homogeneous graphs, so more meta-pathway-based models (Jin et al., 2019; Xiong et al., 2019; Yang et al., 2019; Zhang et al., 2019; Deng et al., 2020; Zhang et al., 2020; Zhou et al., 2020; Ata et al., 2021; He et al., 2021) have been developed to better adapt to heterogeneous biological networks. Luo et al. (2016) proposed the RMLM and RMMSe methods for mining meta-path-based miRNA–target interactions by constructing networks. Jin et al. (2019) constructed a heterogeneous network by integrating the disease–gene association network, miRNA–gene association network, gene–disease association network, and protein–protein interactions networks, and then infer disease association by applying random walk and skip-gram based on the meta-path. Compared with the aforementioned methods, the Metapath2Vec algorithm they adopted can better preserve the structural and semantic interrelationships. However, the application of this method is limited by the existing large differences between vertexes or link attributes in heterogeneous networks. With the success of network embedding methods in analyzing various networks, researchers are increasingly using heterogeneous network representation learning methods to extract node (edge) embeddings in biological networks in order to more fully extract information from heterogeneous biological networks and thus obtain better disease association prediction performance. Yang et al. (2018) proposed an embedded representation model HerGePred based on the heterogeneous disease gene-related network. This model restarts random walk on the reconstructed heterogeneous disease gene network and obtains the low-dimensional vector representation of nodes in the network, which improves the prediction performance. Altabaa et al. (2022) proposed and evaluated the geneDRAGNN method using graph neural networks, which uses information from gene–gene interaction networks to predict disease associations.

Although the aforementioned model in the prediction of disease association study has made great progress, but there are still some limitations. Some of the models previously mentioned (Goh et al., 2007; Ulitsky and Shamir, 2007; Tian et al., 2008; Wu et al., 2008; Li and Patra, 2010; Vanunu et al., 2010) and their similar ones (Suratanee and Plaimas, 2015; Zhang et al., 2016; Iida et al., 2020) only employ protein–protein interactions and related genes to predict, so they will suffer from insufficient data. There are also some models (Jin et al., 2019; Zhang et al., 2019; Zhou et al., 2020; Ata et al., 2021) that ignore the content of nodes, or other models (Deng et al., 2020; He et al., 2021), only to consider both ends of the meta-paths and ignore the intermediate nodes of each meta-path, resulting in the loss of information on heterogeneous graphs. Moreover, some models (Xiong et al., 2019; Zhang et al., 2020) rely on a single meta-path to gain the target node's embedding in the heterogeneous graph, which may lose information about other meta-paths and lead to sub-optimal performance.

In this paper, we develop a disease association prediction model, which is based on a cross-species heterogeneous biological network and a heterogeneous graph embedding method by applying the meta-path-aggregated graph neural network (DapBCH). Our model has the following contributions.

1. We construct a cross-species heterogeneous network to alleviate the problem of insufficient data. The phenotype of model organisms can be applied to human phenotype studies (Liao and Zhang, 2008). Therefore, we apply mice as model organisms and effectively integrate their biological data into heterogeneous bioinformatics networks. Specifically, DapBCH combines the human disease–gene network, mouse gene–phenotype network, human–mouse homologous gene network, and human protein–protein interaction network to create a more complete bioinformatics network.

2. We apply a heterogeneous graph embedding method (MAGNN) (Fu et al., 2020) to extract the features of disease nodes, which can fully capture the node content and context structure of the heterogeneous biological network. Specifically, first, we project nodes of different dimensions into the same vector space to address the issue of various node types in heterogeneous biological networks. Second, we apply the attention mechanism aggregation in each meta-path to handle the problem that the aforementioned methods (Jin et al., 2019; Zhang et al., 2019; Deng et al., 2020; Zhou et al., 2020; Ata et al., 2021; He et al., 2021) only consider the neighbor nodes in the meta-path, while information about phenotype-related genes that are not connected to the target disease node is ignored. Third, we aggregate the potential vectors obtained from the four meta-paths to obtain the final node embedding, which tackles the problem of relying only on a single meta-path in the heterogeneous biological network and not making full use of various biological paths.
3. Furthermore, we scientifically verify the predicted comorbidities by reviewing the literature, demonstrating that DapBCH is effective in predicting disease associations.

Experiments have demonstrated that DapBCH can more accurately predict disease associations. Ablation experiments confirm the correctness of our method, that is, adding mouse phenotype association data and human–mouse homologous gene data, as well as selecting multiple meta-pathways, can improve the accuracy of disease association prediction. Moreover, scientific validation of our prediction of comorbidity demonstrates that our model can detect potential disease associations.

2 Materials and methods

2.1 Biological data

The main data used in this paper include the following: (1) the association between human diseases and genes; (2) mouse–gene phenotype association; (3) human–mouse homologous gene; (4) protein interaction group; and (5) a set of known positive disease–disease associations. Among them, the first three groups of data are obtained from the MGI database (Eppig, 2017), and the fourth group of data is obtained from the STRING database (Mering et al., 2003). The fifth dataset is obtained by integrating three manually checked datasets (Pakhomov et al., 2010; Suthram et al., 2010; Mathur and Dinakarpandian, 2012; Žitnik et al., 2013; Cheng et al., 2014) of disease pairs with high similarity. The specific dataset of this experiment also includes the mapping set of EntrezGene ID to GO ID for human gene, the score link dataset of human gene EntrezGene ID to mouse gene MGI ID, and the relationship between human gene GO to protein ID. The sources of all datasets are as follows:

Human disease and gene association: in this experiment, the association dataset from the mouse genome information (MGI) database is derived from MGI_DO.rpt.txt.

Mouse–gene phenotype association: in this experiment, as the mapping set from the mouse gene MGI ID to phenotype, it is derived from MGI_GenePheno.rpt of the MGI database.

Human–mouse homologous gene: in this experiment, the human gene EntrezGene ID is mapped to the mice gene MGI ID, and the data are obtained from HMD_HumanPhenotype.rpt of the MGI database.

A set of known positive disease–disease associations: in this experiment, one of the datasets that we integrated is from the linked disease pairs obtained by fusing molecular data by Žitnik et al. (2013). Another dataset is the confirmed similar disease pairs collected by Cheng et al. (2014). The last part of the dataset is the disease pairs extracted by Mathur and Dinakarpandian (2012) through literature validation. There are 73 diseases and 92 disease–disease associations in this collection of known disease associations.

Idmapping_selected.Tab: in this experiment, as the mapping set from EntrezGene ID to GO ID for human genes, it comes from the database of Georgetown University in the United States (Huang et al., 2003).

Protein interaction group: in this experiment, as the association dataset of human genes and proteins, it comes from the Final_GO_ProteinID_human.txt of the Gene Ontology Resource database (Consortium, 2004).

9606.protein.links.v11.5.txt: in this experiment, as the score link dataset of the protein interaction, it is derived from the STRING database.

2.2 Methods

Our model, DapBCH, constructs a heterogeneous biological network cross-species and applies the heterogeneous graph embedding method (MAGNN) to predict disease association. The key steps of this model are as follows: (1) network construction: apply the aforementioned biological data to generate node information and adjacency matrix, and then construct a heterogeneous graph neural network. The heterogeneous network includes human–disease gene, mouse–phenotype gene, human–mouse homologous gene, and protein–protein interactions; (2) heterogeneous graph embedding: first, we convert the contents of different types of nodes into the vector space of the same dimension, then we apply intra-meta-path aggregation, and finally apply inter-meta-path aggregation on the four meta-paths to generate the feature vector of disease nodes; and (3) network-based disease association prediction: apply link prediction using the acquired disease node feature vectors to obtain the similarity of disease pairs.

2.2.1 Network construction

We construct a heterogeneous biological network by combining four different networks: (a) human disease–gene association network, (b) mouse gene–phenotype association network, (c) human–mouse homologous gene network, and (d) the human protein–protein interactions network. We list the contents of each network of the heterogeneous biological network in Table 1.

Human disease–gene association network. We collect the experimentally validated human disease–gene associations from the MGI database. The genes are annotated using the Entrez IDs, and the diseases are represented using their OMIM identifier. The human disease–gene associations in database are picked out for

TABLE 1 Description of each network of the heterogeneous biological information network.

Network	Node		Association	Source
Human disease–gene network	Human diseases	2,958		
	Human genes	3,562	3,687	MGI
Mouse gene–phenotype network	Mouse genes	12,319		
	Mouse phenotype	8,801	77,644	MGI
Human–mouse homologous gene network			10,491	MGI
Human protein–protein interaction network	Proteins (genes)	13,281	229,524	STRING

TABLE 2 Experimental results (%) of using our method and four network embedding methods to identify the performance of the disease association.

Method	Metapath2vec	DeepWalk	HAN	HeteWalk	MAGNN
AUC	85.68	73.21	90.87	88.23	92.62
AP	85.86	72.15	90.54	87.81	93.13

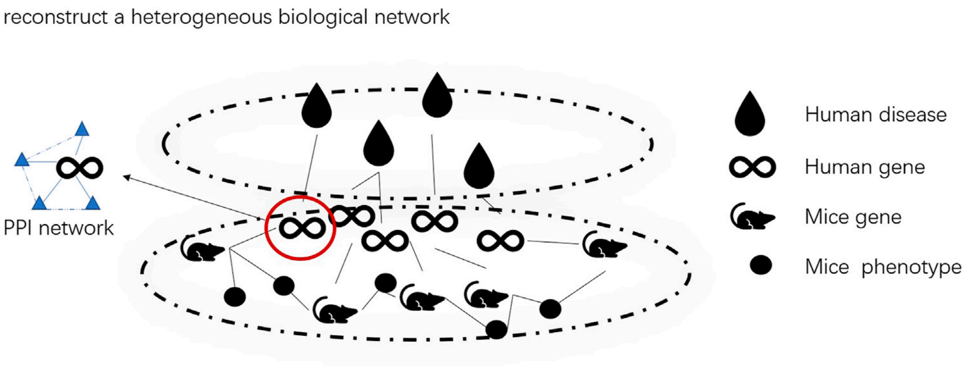

FIGURE 1
Heterogeneous biological network by assembling four types of networks.

providing information on humans. We artificially map Entrez to DO terms and annotate each DO term with its associations. Finally, we obtain a total of 3,687 gene–disease associations, linking 2,958 human diseases to 3,562 human genes.

Mouse gene–phenotype association network. We collect mouse gene–phenotype associations from the MGI database. Because our evaluation set uses human gene identifiers, we use human–mouse homologous genes to connect our mouse phenotypes to the network.

Human–mouse homologous gene network. We collect the human and mouse gene associations from HMD_HumanPhenotype.rpt in the MGI database, which include the mouse orthologs of human genes and human orthologs of mouse genes. We map each mouse gene to their human direct homologs and obtain 10,491 human genes, where the mouse direct homologs have phenotype associations.

The human protein–protein interaction network. Physical protein–protein interactions are extracted from the STRING database. We artificially map protein to GO terms based on Idmapping_selected.Tab, Final_GO_ProteinID_human.txt, and 9606.protein.links.v11.5.txt. We map the proteins to Idmapping_selected.Tab and screen out these entries that are not mapped to the database to obtain the desired association table of protein interactions relevant to this experiment. Furthermore, we extract the confidence score of the interaction group from the STRING database and delete the interactions with confidence less than 400. The protein–protein interaction network we obtained consists of 13,281 proteins and 229,534 interactions. For the extracted directly connected protein pairs, their corresponding coding genes are connected by unweighted edges in the PPI network, and we set the weight to 1.0.

The heterogeneous biological network across species we built is shown in Figure 1.

2.2.2 MAGNN

MAGNN is a graph neural network based on the meta-path for heterogeneous graph embedding. It mainly includes three steps: node content transformation, intra-meta-path aggregation, and inter-meta-path aggregation to generate node embedding. A key assumption of our model is that diseases (genes or phenotypes) that are physically or functionally close to each other in the network have higher similarity. If these two diseases are related, the genes related to these two diseases should be close to each other in the gene phenotype network or the protein network. This allows us to depend on the existing edges to dig unknown disease-related associations. Therefore, we select four meta-paths, namely, M1 (disease → human gene → disease), M2 (disease → human gene → human gene → disease), M3 (disease → human gene → mouse gene → human gene → disease), and M4 (disease → human gene → mouse gene → phenotype → mouse gene → human gene → disease), for extraction of disease node embeddings.

2.2.3 Heterogeneous graph embedding

Different network embedding techniques have been proposed for extracting embeddings in complex network structures (Li et al., 2015; Hamilton et al., 2017; Vaswani et al., 2017; Shi et al., 2018). Most of the existing network embedding methods can only be applied to homogeneous networks, where both node types and edge types are the same in their networks. However, in order to describe the disease association network more realistically and accurately, different genes and phenotypes are integrated into biological networks with different features in our work. Therefore, we need to first project node features of different types (e.g., disease and phenotype) in heterogeneous biological networks into the same latent vector space. Specifically, we designed a parametric weight matrix W. For a node $d \in \mathcal{D}_A$ of type $A \in \mathbf{A}$, we obtain

$$h_d' = W_A \cdot x_d^A, \tag{1}$$

where $h_d' \in \mathbb{R}^{d'}$ is the projected latent vector of node d, $x_d \in \mathbb{R}^{d_A}$ is the initial vector, and $W_A \in \mathbb{R}^{d' \times d_A}$ is the parameter weight matrix of A-type nodes.

In our work, in addition to the content of two disease nodes at the start and end of the meta-path, the intermediate nodes of the meta-path, such as the human gene node and the mice phenotype node, are also important for calculating the similarity of the two diseases. For example, in our disease network containing diseases, human genes, and proteins, M1 disease → human gene → disease (DGD) and M2 disease → human gene → human gene → disease (DGGD) are two meta-paths that describe different relationships. The DGD meta-path describes two diseases associated with the same gene, while the DGGD meta-path links diseases associated with a pair of genes coding for directly linked proteins. Therefore, we can consider a meta-path as a higher-order approximation between two nodes. Furthermore, the intra-meta-path aggregation and attention mechanisms of MAGNN are applied to each meta-path, which can fully capture the contextual structure and content of the nodes of our heterogeneous biological network.

Based on the meta-path P, we define the target disease node as d and the neighbor node as $g \in \mathcal{N}_d^P$. In addition, we define the corresponding meta-path instance as $P(d,g)$, and the intermediate nodes of the meta-path $P(d,g)$ as $\{m^{P(d,g)}\} = P(d,g) \backslash \{d,g\}$.

In the intra-meta-path aggregation, we need to convert the feature vectors of all nodes of the meta-path instance into a vector $h_{P(d,g)}$ through a special meta-path encoder. The meta-path instance is treated as a set in the normal mean and linear encoders, so the information embedded in the structure of the meta-path is ignored. Relational rotation (Sun et al., 2019) provides us a better way to extract meta-path information. The relational rotation encoder is defined as follows:

$$\begin{aligned} o_0 &= h_{t_0'} = h_g' \\ o_i &= h_{t_i}' + o_{i-1} \odot r_i \\ h_{P(d,g)} &= \frac{o_n}{n+1}, \end{aligned} \tag{2}$$

where R_i is the relation between node t_{i-1} and t_i. Given $h_{P(d,g)} = (t_0, t_1, \ldots, t_n)$, where $t_0 = d$ and $t_n = g$, let r_i represent the relation vector of R_i. h_{t_i}' and r_i are both complex vectors, and $\odot$ is the element-wise product.

Then, we need to define the parameterized attention (Velickovic et al., 2017) vector a_P of the meta-path P and apply the attention mechanism to perform weighted aggregation of all meta-path instances based on the meta-path P of the target node d.

$$\begin{aligned} e_{dg}^P &= LeakyReLU\left(a_P^T \bullet \left[h_d' \| h_{P(d,g)}\right]\right) \\ \alpha_{dg}^P &= \frac{exp\left(e_{dg}^P\right)}{\sum_{s \in N_d^P} exp\left(e_{ds}^P\right)} \\ h_d^P &= \sigma\left(\sum_{g \in N_d^P} \alpha_{dg}^P \bullet h_{P(d,g)}\right). \end{aligned} \tag{3}$$

Here, e_{dg}^P is the importance weight of the meta-path instance $P(d,g)$ for node d. α_{dg}^P is the result of softmax normalization of the importance weights related to all meta-path instances of meta-path P, that is, the normalized attention coefficient. Then, we apply α_{dg}^P and the vector representation $h_{P(d,g)}$ of the corresponding meta-path instance to perform weighted aggregation. Finally, the vector representation h_d^P based on the meta-path P of the target node d is the output through an activation function.

Through intra-meta-path aggregation, we finally obtain h_d^{Pi} that contains all the intermediate information on the P_i-meta-path about the target disease node d, where P_i includes four meta-paths, namely, M1 (DGD), M2 (DGGD), M3 (DGMMGD), and M4 (DGMPMGD).

One meta-path P can describe one composite relationship between two objects. However, the similarity between diseases is affected by many factors in bioinformatics, and diseases are not only related to genes but also related to phenotypes and proteins. Therefore, we consider multiple meta-paths and further inter-meta-path aggregation, using the attention mechanism to aggregate the potential vectors obtained from the four meta-paths to get the final node embedding so as to better capture the complex structural information on heterogeneous networks.

If the node of node type A has M meta-paths, the target node d_A will have a set of vectors: $\left\{h_d^{P_1}, h_d^{P_2}, \ldots, h_d^{P_M}\right\}$. Similarly, each node belonging to node type A will have such a set of vectors. Each meta-path vector of all nodes with node type A needs to be converted and averaged, respectively, as follows:

$$s_{P_i} = \frac{1}{|D_A|} \sum_{d \in D_A} tanh\left(M_A \bullet h_d^{P_i} + b_A\right), \tag{4}$$

where meta-path $P_i \in \mathcal{P}_A$, and M_A and b_A are learnable parameters.

In a heterogeneous biological network, different meta-paths are not equally important, so we need to assign appropriate weights to

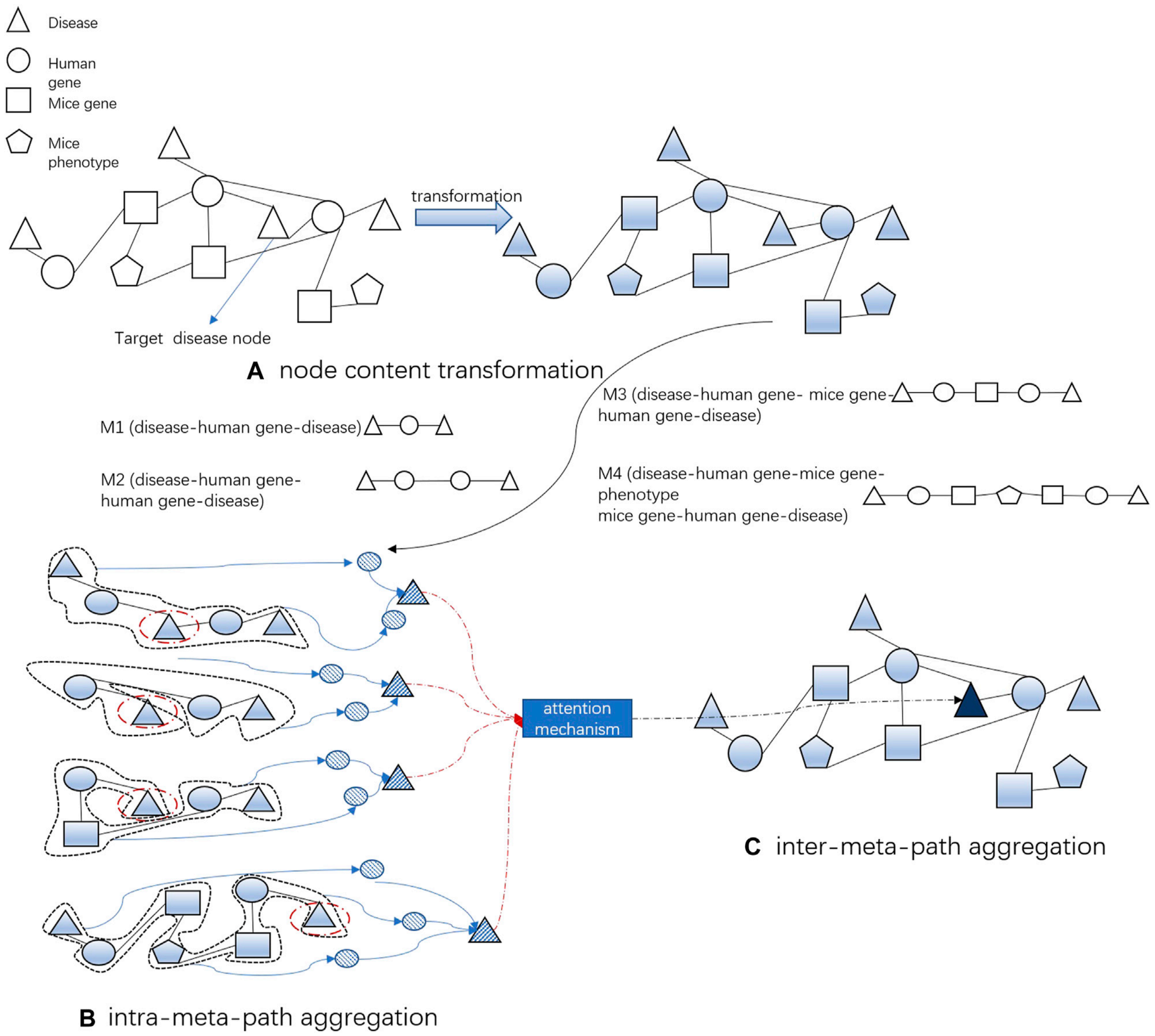
Disease
Human gene
Mice gene
Mice phenotype
transformation
Target disease node
A node content transformation
M1 (disease-human gene-disease)
M2 (disease-human gene-human gene-disease)
M3 (disease-human gene- mice gene-human gene-disease)
M4 (disease-human gene-mice gene-phenotype mice gene-human gene-disease)
attention mechanism
C inter-meta-path aggregation
B intra-meta-path aggregation

FIGURE 2
Process of DapBCH to obtain a disease node vector. **(A)** Project the different types of node features of biological networks into the same latent vector space. **(B)** Aggregate the context structure and content in the meta-path of biological networks into nodes. **(C)** Employ the attention mechanism to aggregate the potential vectors obtained from the four meta-paths to get the final node embedding.

different meta-paths by employing the attention mechanism to extract the information in the network more accurately.

We apply the attention mechanism to find the target node d_A new feature vector mixed with all meta-path information as follows:

$$\begin{aligned} e_{P_i} &= q_A^T \bullet s_{P_i}, \\ \beta_{P_i} &= \frac{exp\left(e_{P_i}\right)}{\sum_{P \epsilon P_A} exp\left(e_P\right)}, \\ h_d^{P_A} &= \sum_{P \epsilon P_A} \beta_P \bullet h_d^P, \end{aligned} \tag{5}$$

where q_A is the parameterized attention vector of type node A. β_{P_i} is the corresponding weight of meta-path P_i uniformly normalized to type node A. After calculating β_{P_i} corresponding to each meta-path, the corresponding target nodes d_A of h_d^P are weighted sum to obtain $h_d^{P_A}$, and finally $h_d^{P_A}$ containing the information of four meta-paths.

Finally, we output the disease node embedding in the required dimension through linear transformation and non-linear activation function:

$$h_d = \sigma\left(W_o \bullet h_d^{P_A}\right). \tag{6}$$

We select four meta-paths related to disease nodes in the heterogeneous biological network so that the node embedding of the target disease can aggregate the information on multiple meta-paths through inter-meta-path aggregation. After obtaining the vector representation of disease nodes by the aforementioned method, the vector representation of gene nodes and phenotype nodes is obtained in the same way for subsequent training of the model. Figure 2 simply demonstrates the embedding generation of a single disease node.

2.2.4 Disease association prediction

We apply link prediction to predict disease association. Link prediction is the analysis of existing network structures and known associations to uncover missing connections and predict possible connections. In generating the disease node embedding, we optimized our model by minimizing the loss function described in the following equation:

$$\mathcal{L} = -\sum_{(d_1,d_2)\in\Omega} \log\sigma\left(h_{T}\cdot h_{d_2}\right) - \sum_{\left(d_1',d_2'\right)\in\Omega^-} \log\sigma\left(-h_{d_1'}^{T}\cdot h_{d_2'}\right), \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function, Ω is the set of positive node pairs of known disease associations, and Ω^- is the complement of negative node pairs without disease associations. We optimize our model to reduce the score of unknown disease pairs and improve the score of known disease–disease associations. Since our model is an end-to-end training model, its parameters are continuously adjusted during the training process, allowing us to successfully complete the optimization task.

In a train-generated disease embedding h_{d_1} and another disease node embedding h_{d_2}, we calculate the probability of a disease–disease association as follows:

$$p_{d_1 d_2} = \sigma\left(h_{d_1}^{T}\cdot h_{d_2}\right). \tag{8}$$

Finally, the disease associations are ranked using the disease association probability $p_{d_1 d_2}$ predicted by the model. The higher the $p_{d_1 d_2}$ association probability, the higher the likelihood that the disease pair is associated.

2.3 Experimental indicators

In order to evaluate the experimental results and better compare and analyze the experimental results under different experimental settings, the area under the curve (AUC) and average precision (AP) are selected as experimental indicators.

AUC refers to the area covered by the ROC curve and is used in the visual performance of the evaluation model, the AUC value in the [0, 1] interval; the bigger the AUC value is showed, the better the performance of the model.

AP refers to the area enclosed by the P–R curve, reflecting the comprehensive performance between the model's accuracy in identifying positive examples and its coverage ability for positive examples. The AP value is in the range of [0,1]. The higher the AP value, the better the algorithm is in predicting the disease–disease association.

2.4 Comparison algorithm

The experimental comparison algorithms in this paper mainly include the following:

1) Metapath2vec (Dong et al., 2017): it is a classical heterogeneous network representation learning method that reconstructs the heterogeneous neighbors of a node through a single meta-path-guided random walk, which is then transmitted to the heterogeneous skip-gram model to obtain node embeddings for downstream tasks. Among other things, it obtains a single meta-path from each defined type of node by random walk, and this meta-path represents the structural information on this node. The model relies on manually pre-selecting a meta-path, so the four defined meta-paths are tested separately in the experiments, and the meta-path with the best results is demonstrated.
2) DeepWalk (Perozzi et al., 2014): it is also a pioneering algorithm. Perozzi et al. used deep learning for large-scale network analysis for the first time, combining the traditional random walk in the graph theory with the skip-gram model and using stochastic gradient descent to learn parameters, resulting in a simple and efficient network representation learning algorithm. We apply it to heterogeneous graphs by ignoring the heterogeneity of the graph structure and removing all node content features.
3) HAN (Wang et al., 2019): it uses an attention mechanism to integrate meta-path-specific node embedding, which learns a single vector representation of each node from several meta-path-based homogeneous graphs.
4) HeteWalk (Xiong et al., 2019): it is a representation learning method that generates node vectors through a heterogeneous skip-gram model, which is based on random walks guided by meta-paths and link weights. HeteWalk preserves the existing relationships by maximizing the conditional probability of each node pair appearing in the node sequence. In this case, the node sequence is created based on meta-paths.

3 Results

In this section, we conduct two experiments to validate our model. In the first experiment, we compare our method with four state-of-the-art network embedding methods. Through this experiment, it is confirmed that our method is more suitable than other methods for predicting disease association.

In the second experiment, we perform a series of ablation experiments. Our ablation experiments remove certain parts of the heterogeneous biological network to better understand the importance of this part of the network to the overall network. If worse result is obtained after the ablation experiment, it means that the part of the network worked. The first experiment not only evaluates the performance of each different meta-path but also evaluates the overall four meta-paths, and then compares and analyzes. The second setup is to remove mouse phenotype–gene association information in the heterogeneous network. The third experimental setup builds on the second experiment by continuing to remove the human–mice homologous gene network to generate experimental results for each model. The fourth experiment is set to delete the PPI network and compare the results with the whole network. The results of ablation experiments show the effectiveness of the model organism information and PPI information we added, and the integration of multiple different meta-paths can also help improve the prediction of disease association performance.

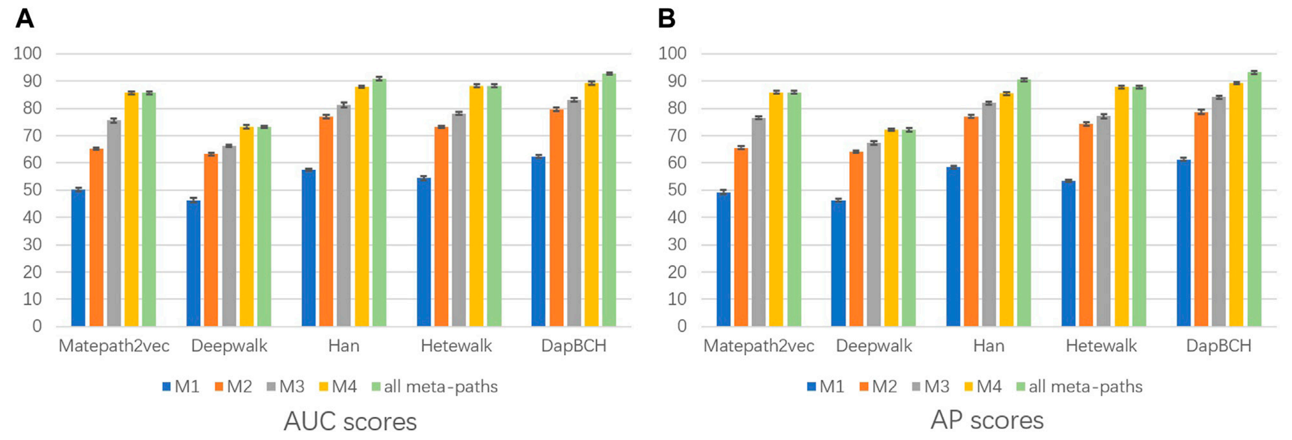

FIGURE 3
Performance of applying different meta-paths. The aforementioned figure illustrates the AUC and AP scores for disease associations predicted by four network embedding methods and our approach. The blue bars are generated by applying M1 meta-paths to each model, the orange bars indicate results from applying M2 meta-paths, the gray bars are generated by applying M3 meta-paths, the yellow bars are generated by applying M4 meta-paths, and the green bars indicate results generated by applying all meta-paths. **(A)** AUC scores. **(B)** AP scores.

3.1 Comparison experiment and ablation experiment

First, we compare our method with four network embedding methods, namely, Matepath2vec, DeepWalk, HAN, and HeteWalk. In each comparison experiment, we conduct a 10-fold cross-validation, that is, the unconnected disease pairs (unknown associations) are first partitioned into 10 equally sized folds, among which nine folds are selected as the training data and the remaining one fold is selected as the test data. After 10 folds are completed, 10 iterations of training and verification are performed. In this way, in each iteration, different folds of data are reserved for verification, and the remaining nine folds are applied for learning. The model learnt subsequently is used to predict the data in the verification fold (Singh and Panda, 2011). For Metapath2vec, DeepWalk, HAN, and HeteWalk, we follow the original settings in their previous experiments. The window size of Metapath2vec, DeepWalk, and HeteWalk is set to 5, the walk length to 100, the number of walks per node to 40, and the negative sample count to 5. Among them, the discard rate of HAN and our proposed model is set to 0.5, the number of attention heads is set to 8, and the dimension of the attention vector in the meta-path aggregation is set to 128. For fair comparison, the embedding dimension of all models compared to 64 is set. Since Metapath2vec requires us to select only one meta-path in each experiment, we apply the M4 meta-path for it in the experiment. Each embedding model is run independently 10 times, and the average value of each model is calculated as the final prediction result. We employ the AUC and AP to compare the performance of the models. We investigate whether the similarity of known disease pairs used for optimization could be prioritized in the model as a way to generate AUC values. We report the results of each embedding model run in Table 2.

Next, we perform ablation experiments and design four different setups to confirm the effectiveness of adding biological information and to demonstrate that multiple meta-paths can improve the accuracy of predictions. The four experimental settings are as follows:

1) The first experiment not only evaluates the performance of each different meta-path but also evaluates the overall four meta-paths, and then compares and analyzes. The M4 meta-path with better performance is applied to the Metapath2Vec model in that experiment. Figure 3 shows the performance results related to the prediction of disease association when selecting each single meta-path and all meta-paths.
2) The second experimental setup removes the mouse phenotype–gene association network in the heterogeneous network we conducted, that is, only M1, M2, and M3 meta-paths are selected in the multiple meta-path model. The M3 meta-path with better performance is applied to the Metapath2Vec model in this experiment. We show the performance results of removing the mouse gene–phenotype information network *vs.* the whole network in Figure 4.
3) The third experimental setup continues the second experiment by removing the human–mouse homologous gene network, that is, only M1 and M2 meta-paths are selected in the multiple meta-path model. The M2 meta-path is applied to the Metapath2Vec model in this experiment. Figure 5 shows the performance results of the network and the complete network without all mouse-related information.
4) The fourth experiment is set to remove the PPI network, that is, only M1, M3, and M4 meta-paths are selected in the multiple meta-path model. The M4 meta-path is applied to the Metapath2Vec model in this experiment. We show the performance results for the network with PPI network information removed *vs.* the full network in Figure 6.

As in the first comparative experiment, each embedded model runs independently for 10 times in each ablation experiment.

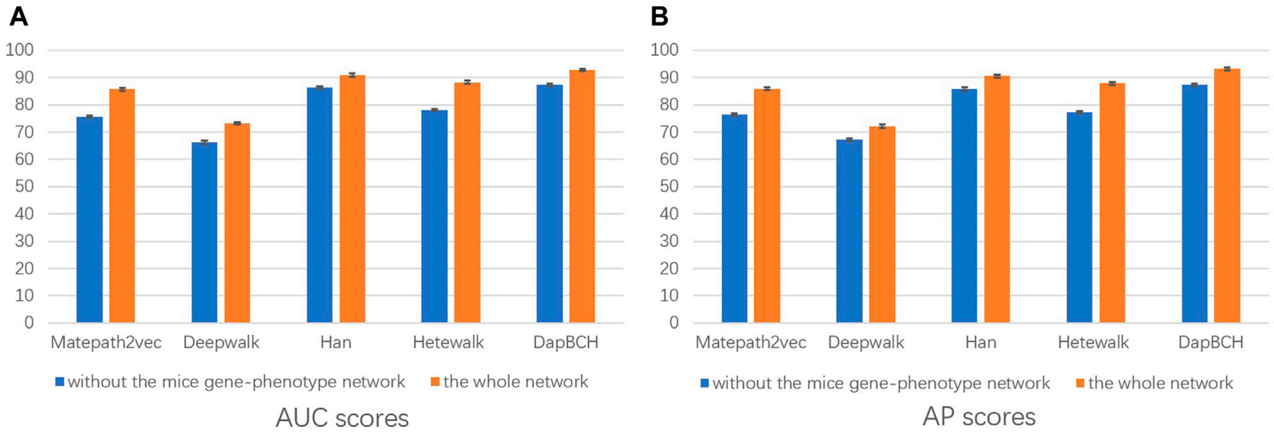

FIGURE 4
Performance of removing the mouse gene–phenotype information network prediction. The aforementioned figure illustrates the AUC and AP scores for disease associations predicted by the four network embedding methods and our methods. The blue bar represents the results of removing the mouse gene–phenotype information network, and the orange bar represents the results of the whole heterogeneous biological network. **(A)** AUC scores. **(B)** AP scores.

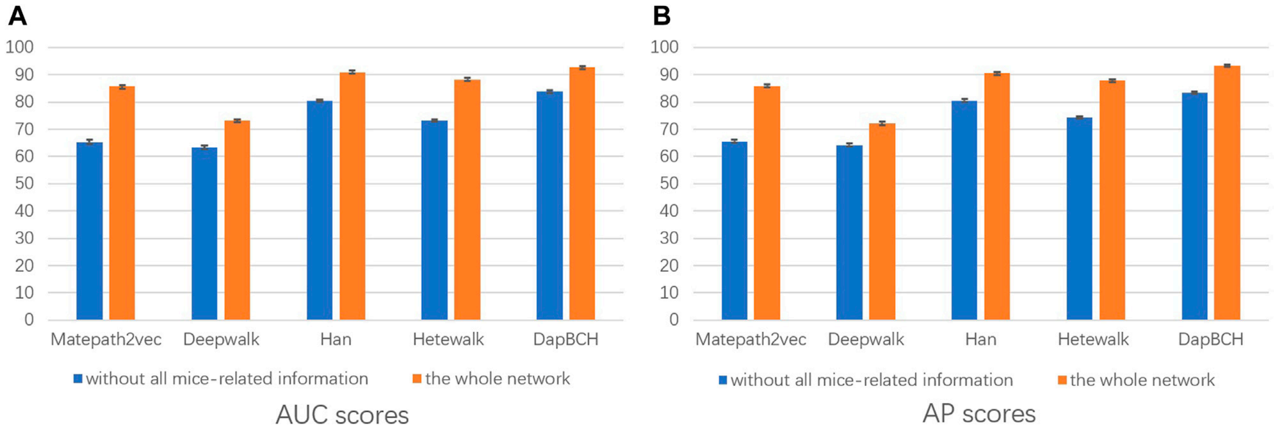

FIGURE 5
Performance of removing all mouse-related information network predictions. The aforementioned figure illustrates the AUC and AP scores for disease associations predicted by the four network embedding methods and our methods. The blue bars represent the results generated on the complete network without all mouse-related information, and the orange bars represent the results generated on the entire heterogeneous biological network. **(A)** AUC scores. **(B)** AP scores.

4 Discussion

The experimental results showed that compared with the previous four network embedding methods, our method has improved in predicting the disease association. Figure 3 shows that our method consistently outperforms the best baseline HAN obtaining the highest AUC score of 92.62% and AP score of 93.13% under different experimental settings. From the aforementioned results, we can conclude that our method can predict the disease associations more accurately. By applying node content transformation, intra-meta-path aggregation, and inter-meta-path aggregation to generate disease node embedding, more information in heterogeneous biological networks can indeed be obtained.

In the second experiment, we perform ablation experiments. Figure 3 shows that the performance of combining multiple meta-paths is significantly improved over the performance in using only one meta-path. Among them, M4, which combines mouse-related data, has the best prediction performance in each model. Moreover, the experiments show that MAGNN and HAN receive a better score than Metapath2vec. In other words, combining multiple meta-paths by inter-meta-path aggregation gives more accurate results than selecting only a single meta-path. Figure 4 shows that when the constructed heterogeneous biological network removes the mouse gene–phenotype network, the AUC of the DapBCH model decreases by 8.86% and AP decreases by 5.87%. Figure 5 shows that when removing the mouse gene–phenotype association network and the human–mouse homologous gene network, the AUC of DapBCH decreases by 9.41%

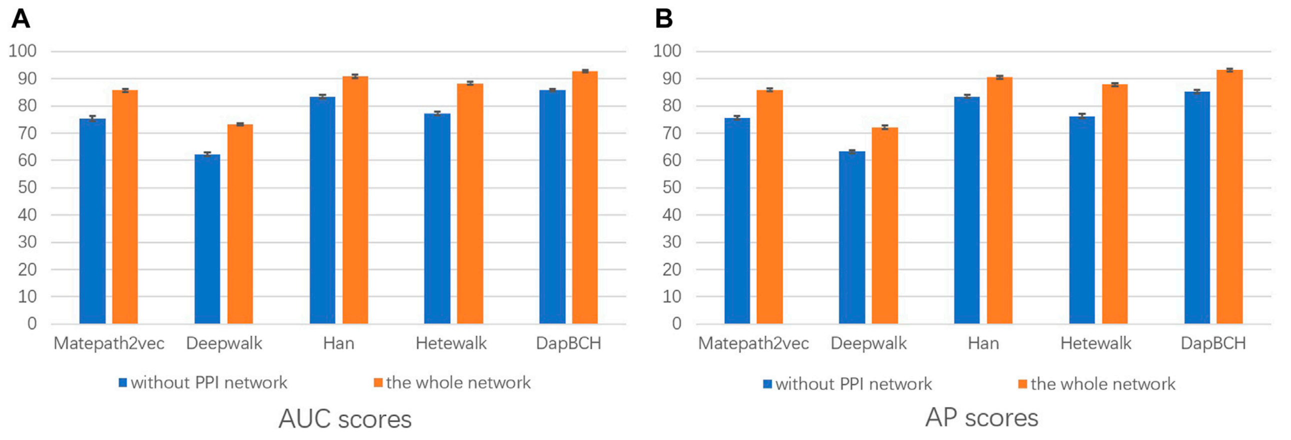

FIGURE 6
Performance of removing PPI network predictions. The aforementioned figure illustrates the AUC and AP scores for disease associations predicted by the four network embedding methods and our methods. The blue bars represent the results generated on the complete network without PPI network information, and the orange bars represent the results generated on the entire heterogeneous biological network. **(A)** AUC scores. **(B)** AP scores.

and AP decreases by 9.87%. Figure 4 and Figure 5 show that the increase in the mouse gene–phenotype association network and the human–mouse homologous gene network improves the performance results, which shows that the increased model organism information is helpful in improving the accuracy of predicting disease associations. Figure 6 shows that the results obtained by adding protein interaction information are more accurate in each model, indicating that applying the PPI network information network helps improve the model prediction performance. Thus, the ablation experiments confirm the correctness of our method that adding the mouse gene–phenotype association network, the human–mouse homologous gene network, and the PPI network, as well as selecting multiple meta-path combinations for predicting disease associations, can improve the accuracy.

5 Scientific verification of the predicted comorbidity

We apply the graph embedding method based on the meta-path to predict the disease association through four meta-paths in the heterogeneous biological network we constructed. We list the top 15 disease pairs with the highest similarity in one of the experimental results, excluding the known disease pairs for training, as shown in Table 3.

By searching the literature (Ottman et al., 2011; Erden and Bicakci, 2012; Al-Goblan et al., 2014; Iglay et al., 2016; Su et al., 2016; Guekht, 2017; Lopez et al., 2017; Chuang et al., 2018; Newcombe et al., 2018; Oganov et al., 2019; Baradaran et al., 2020; Choi, 2020; de Lucena et al., 2020; Benhammou et al., 2021; Maciejewska et al., 2021) related to the diseases listed in the table for analysis, we found the correlation between the pathogenesis of the disease pairs. We interpret and describe some of the disease pairs in Table 3. Alzheimer's disease–hypercholesterolemia: the pathogenesis of AD is associated with multiple complications and advanced age (Santiago and Potashkin, 2021). Schizophrenia, depression, epilepsy, sleep disorder, hypercholesterolemia, hypertension, and other pathological conditions may cause AD. Hypercholesterolemia and hypertension may impair functions such as verbal memory, verbal reasoning, and visual memory. In clinical treatment, it may help controlling these risk factors in patients diagnosed with AD (Goldstein et al., 2008). Obesity–sleep disorder: the decrease in sleep time and quality is related to the increase of weight and obesity. Sleep disorder and sleep deprivation will also worsen the development of obesity. Insomnia or other sleep disorders may

TABLE 3 Top 15 disease pairs with the highest similarity obtained from our model predicting the disease–disease association. The first column is the descending number of similarities between diseases. The second column represents disease 1 in the disease pair predicted by the model. The third column represents disease 2 of the disease pair.

Order	Disease 1	Disease 2
1	Alzheimer's disease	Hypercholesterolemia
2	Hepatitis	Liver disease
3	Obesity	Sleep disorder
4	Bipolar disorder	Epilepsy
5	Familial combined hyperlipidemia	Hypertension
6	Asthma	Epilepsy
7	Epilepsy	Anxiety disorder
8	Systemic scleroderma	Essential hypertension
9	Systemic scleroderma	Asthma
10	Polycystic kidney disease	Congestive heart failure
11	Allergic rhinitis	Asthma
12	Mood disorder	Anxiety disorders
13	Cryoglobulinemia	Hepatitis
14	COVID-19	Severe acute respiratory syndrome
15	Male infertility	Obesity

cause excessive consumption of human energy, resulting in weight gain (Hargens et al., 2013).

To sum up, diseases may be related through symptoms, onset period, and other ways. The association between diseases is not accidental, and the current disease may be a risk factor for another disease. Therefore, we can realize that the common mechanism of finding comorbidity is helpful for the early intervention, prevention, and control measures and late treatment of the disease. Finally, the 15 pairs of disease listed in Table 3 can be confirmed to have some comorbidity in the relevant literature, which shows that our model is an effective method for predicting the association of diseases.

6 Conclusion

Understanding the association between diseases is of great significance in disease prevention, diagnosis, and treatment. In this paper, we construct a crossing species heterogeneous biological network, which consists of human genetic disease association, mouse genetic–phenotype association, human–mouse homologous genes, and protein-interacting groups, and apply the meta-path-based graph embedding method to predict the disease association. The experimental results show that compared with other previous predicting models, the AUC score of 92.62% and the AP score of 93.13% achieve the best performance. Through ablation experiments, we prove that integrating the information on model organisms into the network can improve the effectiveness of inferring the disease association. The combination of mouse genes and phenotypes achieves the best prediction results on the dataset.

Therefore, our main contributions lie in the following aspects: (1) we employ mice as a model organism to efficiently integrate its biological data into a heterogeneous bioinformatics network to predict the disease association; (2) we propose a disease association prediction model, DapBCH, which applies the graph embedding method to the aforementioned bioinformatics network, and compares its performance with four other network representation models in predicting the disease association; (3) it turns out that our integration of cross-species information (mice genes and phenotypes) can improve the predictability of disease in the network; and (4) it turns out that the multiple meta-paths and aggregated information on our model are helpful in predicting disease associations.

Our research on predicting the disease–disease association can be extended to solve practical clinical problems. By providing analytical and computational support to assess the risk of disease development and predict disease progression, we can advance clinical decision-making on possible treatments. As for future work, our method is more reliable in the homogeneous aggregation category classification with fewer disease examples and overlapping features. Therefore, we plan to combine low-cost and highly disease-sensitive heterogeneous network data to predict more specific disease associations, such as using disease–miRNA associations to predict lung cancer associations with other diseases.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

WS: methodology, investigation, data preparation, and writing–original draft preparation; HF: formal analysis, validation, and supervision; JL: writing–review and editing, conceptualization, and validation; TL: formal analysis, validation, and supervision; and ZL: writing–review and editing, project administration, and supervision. All authors contributed to the article and approved the submitted version.

Funding

The paper was supported by commonweal projects in Zhejiang Province (Nos LGG21A010001 and LGG22F020010) and the Key R&D Projects in Zhejiang Province (2022C02009 and 2022C02044).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Al-Goblan, A. S., Al-Alfi, M. A., and Khan, M. Z. (2014). Mechanism linking diabetes mellitus and obesity. *Diabetes, metabolic syndrome Obes. targets Ther.* 7, 587–591. doi:10.2147/DMSO.S67400

Altabaa, A., Huang, D., Byles-Ho, C., Khatib, H., Sosa, F., and Hu, T. (2022). "genedragnn: gene disease prioritization using graph neural networks," in 2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (IEEE), Ottawa, Canada, 15-17 August.

Aravind, L. (2000). Guilt by association: contextual information in genome analysis. *Genome Res.* 10, 1074–1077. doi:10.1101/gr.10.8.1074

Ata, S. K., Wu, M., Fang, Y., Ou-Yang, L., Kwoh, C. K., and Li, X.-L. (2021). Recent advances in network-based methods for disease gene prediction. *Briefings Bioinforma.* 22, bbaa303. doi:10.1093/bib/bbaa303

Baradaran, A., Ebrahimzadeh, M. H., Baradaran, A., and Kachooei, A. R. (2020). Prevalence of comorbidities in covid-19 patients: a systematic review and meta-analysis. *Archives Bone Jt. Surg.* 8, 247–255. doi:10.22038/abjs.2020.47754.2346

Benhammou, J. N., Moon, A. M., Pisegna, J. R., Su, F., Vutien, P., Moylan, C. A., et al. (2021). Nonalcoholic fatty liver disease risk factors affect liver-related outcomes after direct-acting antiviral treatment for hepatitis c. *Dig. Dis. Sci.* 66, 2394–2406. doi:10.1007/s10620-020-06457-2

Cheng, L., Li, J., Ju, P., Peng, J., and Wang, Y. (2014). Semfunsim: a new method for measuring disease similarity by integrating semantic and gene functional association. *PloS one* 9, e99415. doi:10.1371/journal.pone.0099415

Choi, M. E. (2020). Autophagy in kidney disease. *Annu. Rev. physiology* 82, 297–322. doi:10.1146/annurev-physiol-021119-034658

Chuang, Y.-W., Yu, T.-M., Huang, S.-T., Sun, K.-T., Lo, Y.-C., Fu, P.-K., et al. (2018). Young-adult polycystic kidney disease is associated with major cardiovascular complications. *Int. J. Environ. Res. Public Health* 15, 903. doi:10.3390/ijerph15050903

Consortium, G. O., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., et al. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids Res.* 32, D258–D261. doi:10.1093/nar/gkh036

de Lucena, T. M. C., da Silva Santos, A. F., de Lima, B. R., de Albuquerque Borborema, M. E., and de Azevêdo Silva, J. (2020). Mechanism of inflammatory response in associated comorbidities in covid-19. *Diabetes & Metabolic Syndrome Clin. Res. Rev.* 14, 597–600. doi:10.1016/j.dsx.2020.05.025

Decramer, M., and Janssens, W. (2013). Chronic obstructive pulmonary disease and comorbidities. *Lancet Respir. Med.* 1, 73–83. doi:10.1016/S2213-2600(12)70060-7

Deng, L., Yang, J., and Liu, H. (2020). "Predicting circrna-disease associations using meta path-based representation learning on heterogenous network," in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE), Seoul, Korea, Dec. 16 2020 to Dec. 19 2020, 5.

Dong, Y., Chawla, N. V., and Swami, A. (2017). "metapath2vec: scalable representation learning for heterogeneous networks," in Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, Halifax NS Canada, August 13 - 17, 2017, 135–144.

Eppig, J. T. (2017). Mouse genome informatics (mgi) resource: genetic, genomic, and biological knowledgebase for the laboratory mouse. *ILAR J.* 58, 17–41. doi:10.1093/ilar/ilx013

Erden, S., and Bicakci, E. (2012). Hypertensive retinopathy: incidence, risk factors, and comorbidities. *Clin. Exp. Hypertens.* 34, 397–401. doi:10.3109/10641963.2012.663028

Fu, X., Zhang, J., Meng, Z., and King, I. (2020). "Magnn: metapath aggregated graph neural network for heterogeneous graph embedding," in Proceedings of The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020, 2331–2341.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proc. Natl. Acad. Sci.* 104, 8685–8690. doi:10.1073/pnas.0701361104

Goldstein, F. C., Ashley, A. V., Endeshaw, Y., Hanfelt, J., Lah, J. J., and Levey, A. I. (2008). Effects of hypertension and hypercholesterolemia on cognitive functioning in patients with alzheimer disease. *Alzheimer Dis. Assoc. Disord.* 22, 336–342. doi:10.1097/wad.0b013e318188e80d

Graw, S., Chappell, K., Washam, C. L., Gies, A., Bird, J., Robeson, M. S., et al. (2021). Multi-omics data integration considerations and study design for biological systems and disease. *Mol. Omics* 17, 170–185. doi:10.1039/d0mo00041h

Guekht, A. (2017). Epilepsy, comorbidities and treatments. *Curr. Pharm. Des.* 23, 5702–5726. doi:10.2174/1381612823666171009144400

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. *Adv. neural Inf. Process. Syst.* 30.

Hargens, T. A., Kaleth, A. S., Edwards, E. S., and Butner, K. L. (2013). Association between sleep disorders, obesity, and exercise: a review. *Nat. Sci. sleep* 5, 27–35. doi:10.2147/NSS.S34838

He, M., Huang, C., Liu, B., Wang, Y., and Li, J. (2021). Factor graph-aggregated heterogeneous network embedding for disease-gene association prediction. *BMC Bioinforma.* 22, 165–215. doi:10.1186/s12859-021-04099-3

Huang, H., Barker, W. C., Chen, Y., and Wu, C. H. (2003). i proclass: an integrated database of protein family, function and structure information. *Nucleic Acids Res.* 31, 390–392. doi:10.1093/nar/gkg044

Iglay, K., Hannachi, H., Joseph Howie, P., Xu, J., Li, X., Engel, S. S., et al. (2016). Prevalence and co-prevalence of comorbidities among patients with type 2 diabetes mellitus. *Curr. Med. Res. Opin.* 32, 1243–1252. doi:10.1185/03007995.2016.1168291

Iida, M., Iwata, M., and Yamanishi, Y. (2020). Network-based characterization of disease–disease relationships in terms of drugs and therapeutic targets. *Bioinformatics* 36, i516–i524. doi:10.1093/bioinformatics/btaa439

Ji, Z., Meng, G., Huang, D., Yue, X., and Wang, B. (2015). Nmfbfs: A nmf-based feature selection method in identifying pivotal clinical symptoms of hepatocellular carcinoma. *Comput. Math. methods Med.* 2015, 846942. doi:10.1155/2015/846942

Ji, Z., Zhao, W., Lin, H.-K., and Zhou, X. (2019). Systematically understanding the immunity leading to crpc progression. *PLoS Comput. Biol.* 15, e1007344. doi:10.1371/journal.pcbi.1007344

Jin, S., Zeng, X., Fang, J., Lin, J., Chan, S. Y., Erzurum, S. C., et al. (2019). A network-based approach to uncover microrna-mediated disease comorbidities and potential pathobiological implications. *NPJ Syst. Biol. Appl.* 5, 41–11. doi:10.1038/s41540-019-0115-2

Ko, Y., Cho, M., Lee, J.-S., and Kim, J. (2016). Identification of disease comorbidity through hidden molecular mechanisms. *Sci. Rep.* 6, 39433–39438. doi:10.1038/srep39433

Li, Y., and Patra, J. C. (2010). Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. *Bioinformatics* 26, 1219–1224. doi:10.1093/bioinformatics/btq108

Li, Y., Tarlow, D., Brockschmidt, M., and Zemel, R. (2015). *Gated graph sequence neural networks. arXiv preprint arXiv:1511.05493.*

Liao, B.-Y., and Zhang, J. (2008). Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc. Natl. Acad. Sci.* 105, 6987–6992. doi:10.1073/pnas.0800387105

Lopez, A. Y., Wang, X., Xu, M., Maheshwari, A., Curry, D., Lam, S., et al. (2017). Ankyrin-g isoform imbalance and interneuronopathy link epilepsy and bipolar disorder. *Mol. psychiatry* 22, 1464–1472. doi:10.1038/mp.2016.233

Luo, J., Huang, C., and Ding, P. (2016). A meta-path-based prediction method for human mirna-target association. *BioMed Res. Int.* 2016, 7460740. doi:10.1155/2016/7460740

Luo, J., and Liang, S. (2015). Prioritization of potential candidate disease genes by topological similarity of protein–protein interaction network and phenotype data. *J. Biomed. Inf.* 53, 229–236. doi:10.1016/j.jbi.2014.11.004

Maciejewska, K., Czarnecka, K., and Szymański, P. (2021). A review of the mechanisms underlying selected comorbidities in alzheimer's disease. *Pharmacol. Rep.* 73, 1565–1581. doi:10.1007/s43440-021-00293-5

Mathur, S., and Dinakarpandian, D. (2012). Finding disease similarity based on implicit semantic similarity. *J. Biomed. Inf.* 45, 363–371. doi:10.1016/j.jbi.2011.11.017

Mering, C. v., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). String: a database of predicted functional associations between proteins. *Nucleic acids Res.* 31, 258–261. doi:10.1093/nar/gkg034

Mlynarski, W. M., Placha, G. P., Wolkow, P. P., Bochenski, J. P., Warram, J. H., and Krolewski, A. S. (2005). Risk of diabetic nephropathy in type 1 diabetes is associated with functional polymorphisms in rantes receptor gene (ccr5) a sex-specific effect. *Diabetes* 54, 3331–3335. doi:10.2337/diabetes.54.11.3331

Newcombe, E. A., Camats-Perna, J., Silva, M. L., Valmas, N., Huat, T. J., and Medeiros, R. (2018). Inflammation: the link between comorbidities, genetics, and alzheimer's disease. *J. neuroinflammation* 15, 276–326. doi:10.1186/s12974-018-1313-3

Oganov, R., Simanenkov, V., Bakulin, I., Bakulina, N., Barbarash, O., Boytsov, S., et al. (2019). Comorbidities in clinical practice. algorithms for diagnostics and treatment. *Cardiovasc. Ther. Prev.* 18, 5–66. doi:10.15829/1728-8800-2019-1-5-66

Oliver, S. (2000). Guilt-by-association goes global. *Nature* 403, 601–603. doi:10.1038/35001165

Ottman, R., Lipton, R. B., Ettinger, A. B., Cramer, J. A., Reed, M. L., Morrison, A., et al. (2011). Comorbidities of epilepsy: results from the epilepsy comorbidities and health (epic) survey. *Epilepsia* 52, 308–315. doi:10.1111/j.1528-1167.2010.02927.x

Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., and Melton, G. B. (2010). "Semantic similarity and relatedness between clinical terms: an experimental study," in *AMIA annual symposium proceedings* (United States: American Medical Informatics Association), 572.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "Deepwalk: online learning of social representations," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, August 24-27, 2014, 701–710.

Santiago, J. A., and Potashkin, J. A. (2021). The impact of disease comorbidities in alzheimer's disease. *Front. aging Neurosci.* 13, 631770. doi:10.3389/fnagi.2021.631770

Shao, H., Peng, T., Ji, Z., Su, J., and Zhou, X. (2013). Systematically studying kinase inhibitor induced signaling network signatures by integrating both therapeutic and side effects. *PLoS One* 8, e80832. doi:10.1371/journal.pone.0080832

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Mol. Syst. Biol.* 3, 88. doi:10.1038/msb4100129

Shi, C., Hu, B., Zhao, W. X., and Philip, S. Y. (2018). Heterogeneous information network embedding for recommendation. *IEEE Trans. Knowl. Data Eng.* 31, 357–370. doi:10.1109/tkde.2018.2833443

Shlomi, T., Cabili, M. N., Herrgård, M. J., Palsson, B. Ø., and Ruppin, E. (2008). Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.* 26, 1003–1010. doi:10.1038/nbt.1487

Silverman, E. K., Schmidt, H. H., Anastasiadou, E., Altucci, L., Angelini, M., Badimon, L., et al. (2020). Molecular networks in network medicine: development and applications. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 12, e1489. doi:10.1002/wsbm.1489

Singh, G., and Panda, R. K. (2011). Daily sediment yield modeling with artificial neural network using 10-fold cross validation method: a small agricultural watershed, kapgari, india. *Int. J. Earth Sci. Eng.* 4, 443–450.

Su, X., Ren, Y., Li, M., Zhao, X., Kong, L., and Kang, J. (2016). Prevalence of comorbidities in asthma and nonasthma patients: a meta-analysis. *Medicine* 95, e3459. doi:10.1097/MD.0000000000003459

Sun, Z., Deng, Z.-H., Nie, J.-Y., and Tang, J. (2019). *Rotate: Knowledge graph embedding by relational rotation in complex space. arXiv preprint arXiv:1902.10197.*

Suratanee, A., and Plaimas, K. (2015). Dda: a novel network-based scoring method to identify disease-disease associations. *Bioinforma. Biol. insights* 9, 175–186. BBI–S35237. doi:10.4137/BBI.S35237

Suthram, S., Dudley, J. T., Chiang, A. P., Chen, R., Hastie, T. J., and Butte, A. J. (2010). Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput. Biol.* 6, e1000662. doi:10.1371/journal.pcbi.1000662

Tian, W., Zhang, L. V., Taşan, M., Gibbons, F. D., King, O. D., Park, J., et al. (2008). Combining guilt-by-association and guilt-by-profiling to predict saccharomyces cerevisiaegene function. *Genome Biol.* 9, S7–S21. doi:10.1186/gb-2008-9-s1-s7

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., et al. (2021). Genome-wide association studies. *Nat. Rev. Methods Prim.* 1, 59–21. doi:10.1038/s43586-021-00056-9

Ulitsky, I., and Shamir, R. (2007). Identification of functional modules using network topology and high-throughput data. *BMC Syst. Biol.* 1, 8–17. doi:10.1186/1752-0509-1-8

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* 6, e1000641. doi:10.1371/journal.pcbi.1000641

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. neural Inf. Process. Syst.* 30.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *stat* 1050, 20.

Wang, X., Gulbahce, N., and Yu, H. (2011). Network-based methods for human disease gene prediction. *Briefings Funct. genomics* 10, 280–293. doi:10.1093/bfgp/elr024

Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., et al. (2019). "Heterogeneous graph attention network," in The world wide web conference. 2022, France, 25 – 29 April 2022.

Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (2008). Network-based global inference of human disease genes. *Mol. Syst. Biol.* 4, 189. doi:10.1038/msb.2008.27

Xiong, Y., Guo, M., Ruan, L., Kong, X., Tang, C., Zhu, Y., et al. (2019). Heterogeneous network embedding enabling accurate disease association predictions. *BMC Med. genomics* 12, 186–217. doi:10.1186/s12920-019-0623-3

Yang, K., Wang, R., Liu, G., Shu, Z., Wang, N., Zhang, R., et al. (2018). Hergepred: heterogeneous network embedding representation for disease gene prediction. *IEEE J. Biomed. health Inf.* 23, 1805–1815. doi:10.1109/JBHI.2018.2870728

Yang, K., Zhao, X., Waxman, D., and Zhao, X.-M. (2019). Predicting drug-disease associations with heterogeneous network embedding. *Chaos Interdiscip. J. Nonlinear Sci.* 29, 123109. doi:10.1063/1.5121900

Yue, X., Wang, Z., Huang, J., Parthasarathy, S., Moosavinasab, S., Huang, Y., et al. (2020). Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics* 36, 1241–1251. doi:10.1093/bioinformatics/btz718

Zhang, H., Liang, Y., Peng, C., Han, S., Du, W., and Li, Y. (2019). Predicting lncrna-disease associations using network topological similarity based on deep mining heterogeneous networks. *Math. Biosci.* 315, 108229. doi:10.1016/j.mbs.2019.108229

Zhang, W., Coba, M. P., and Sun, F. (2016). Inference of domain-disease associations from domain-protein, protein-disease and disease-disease relationships. BMC Syst. Biol. *Biomed. Cent.* 10, 4–80. doi:10.1186/s12918-015-0247-y

Zhang, Y., Lei, X., Fang, Z., and Pan, Y. (2020). Circrna-disease associations prediction based on metapath2vec++ and matrix factorization. *Big Data Min. Anal.* 3, 280–291. doi:10.26599/bdma.2020.9020025

Zhou, R., Lu, Z., Luo, H., Xiang, J., Zeng, M., and Li, M. (2020). Nedd: a network embedding based method for predicting drug-disease associations. *BMC Bioinforma.* 21, 387–412. doi:10.1186/s12859-020-03682-4

Žitnik, M., Janjić, V., Larminie, C., Zupan, B., and Pržulj, N. (2013). Discovering disease-disease associations by fusing systems-level molecular data. *Sci. Rep.* 3, 3202–3209. doi:10.1038/srep03202