BMC Genomics

RESEARCH

Open Access

Identification of gene biomarkers for brain diseases via multi-network topological semantics extraction and graph convolutional network

Ping Zhang[1,5†], Weihan Zhang[4†], Weicheng Sun[5], Jinsheng Xu[5], Hua Hu[1*], Lei Wang[1,2*] and Leon Wong[3*]

Abstract

Background Brain diseases pose a significant threat to human health, and various network-based methods have been proposed for identifying gene biomarkers associated with these diseases. However, the brain is a complex system, and extracting topological semantics from different brain networks is necessary yet challenging to identify pathogenic genes for brain diseases.

Results In this study, we present a multi-network representation learning framework called M-GBBD for the identification of gene biomarker in brain diseases. Specifically, we collected multi-omics data to construct eleven networks from different perspectives. M-GBBD extracts the spatial distributions of features from these networks and iteratively optimizes them using Kullback–Leibler divergence to fuse the networks into a common semantic space that represents the gene network for the brain. Subsequently, a graph consisting of both gene and large-scale disease proximity networks learns representations through graph convolution techniques and predicts whether a gene is associated which brain diseases while providing associated scores. Experimental results demonstrate that M-GBBD outperforms several baseline methods. Furthermore, our analysis supported by bioinformatics revealed *CAMP* as a significantly associated gene with Alzheimer's disease identified by M-GBBD.

Conclusion Collectively, M-GBBD provides valuable insights into identifying gene biomarkers for brain diseases and serves as a promising framework for brain networks representation learning.

Keywords Gene biomarkers, Brain diseases, Gene-disease associations prediction, Multi-network topological semantics, Graph convolutional network

†Ping Zhang and Weihan Zhang contributed equally to this work.

*Correspondence:
Hua Hu
huhua@uzz.edu.cn
Lei Wang
wanglei@uzz.edu.cn
Leon Wong
huangliguang18@mails.ucas.ac.cn
[1] College of Information Science and Engineering, Zaozhuang University, Zaozhuang 277100, Shandong, China
[2] Guangxi Key Lab of Human-Machine Interaction and Intelligent Decision, Guangxi Academy of Sciences, Nanning 530007, China
[3] College of Big Data and Internet, Shenzhen Technology University, Shenzhen 518118, China
[4] CAS Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, The Innovative Academy of Seed Design, Chinese Academy of Sciences, Hubei Hongshan Laboratory, Wuhan 430074, China
[5] College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

Background

According to the Global Burden of Disease study, brain diseases have emerged as the leading cause of disability and the second leading cause of death since 2016 [1], imposing a substantial burden on individuals and society [2, 3]. As the intricate central nervous system organ, the brain orchestrates every bodily process. Sustaining a healthy brain is imperative for attaining longevity and overall well-being [4]. However, diagnosing and treating brain diseases pose complex challenges [5–8]. Numerous human brain diseases exhibit significant genetic components [9–11]. Identifying gene biomarkers associated with these conditions is crucial for elucidating their pathogenesis and facilitating drug development. Consequently, this can enable early clinical diagnosis and treatment.

Identification of gene biomarkers for diseases is typically achieved through linkage analysis [12, 13], large clinical cohorts [14, 15], and genome-wide association studies (GWAS) [16, 17]. However, these approaches are time-consuming and costly, particularly in the context of brain diseases. It should be noted that genes require complex regulation to perform biological functions and diseases rarely result from a single gene abnormality [18–20]. Several network-based strategies have been proposed for disease gene prediction and have successfully been applied to the study of brain diseases [21–27]. For instance, the MAGI method utilizes random walk techniques to integrate protein–protein interactions and co-expression networks during brain development to identify genes associated with autism and intellectual disability [22]. Another example is eMAGMA which incorporates genetic and expression networks into tissue-specific analyses to identify genes related to depression risk [28]. In addition to the molecular-based network studies mentioned above, several investigations have focused on brain functional connectivity (BFC) networks constructed using functional magnetic resonance imaging (fMRI). Nevertheless, it is important to note that these methods primarily focus on a single network without providing a comprehensive overview of information across multiple types of networks.

Integrating multiple types of networks allows for the combination of multi-dimensional information, compensating for the limitations of a single network [29, 30]. However, effectively leveraging diverse biological networks to identify disease-related genes remains challenging due to their spatial inconsistencies and high structural heterogeneity. Given the complexity of the brain and its requirement for precise gene biomarker prediction, a comprehensive fusion of multiple networks is necessary [31]. The BFC network reflects functional correlations between genes in the brain [32]. A framework called brainMI has been developed to enable consistent representation of BFC and molecular networks, facilitating predictions on gene-brain disease associations using machine learning approaches [7]. However, the gene network used by brainMI is solely an inference network derived from matrix multiplication. Consequently, this approach overlooks gene regulatory relationships and lacks comprehensiveness in terms of fusion. Therefore, it is crucial to fully consider transcription factor regulation when constructing a biologically meaningful gene network.

Regulatory interactions between transcription factors (TFs) and their targets constitute a gene regulatory network (GRN), which is pivotal for understanding the mechanisms underlying various biological processes [33–35]. With advancements in sequencing technologies, numerous large-scale projects have implemented bulk or single-cell RNA sequencing, resulting in an extensive collection of gene regulation data [34–36]. Hence, integrating TFs to enhance the accuracy of gene networks has become both feasible and increasingly urgent, particularly for complex brain diseases. Furthermore, from the perspective of constructing rugged networks, introducing intermediate/bridge nodes can effectively mitigate noise associated with network connections and minimize the presence of pseudo-edges within the network to some extent [37, 38]. Additionally, different diseases exhibit shared similarities that enable construction of a disease proximity network. Previous studies have demonstrated that genes associated with similar diseases are more likely to possess physical interactions among their protein products as well as display similar expression patterns [39, 40]. In conclusion, modeling the brain network as an association network comprising genes and diseases can effectively and directly reflect the correlation between brain diseases and genes implicated in causing these disorders. This approach can be regarded as a link prediction issue within complex networks. Identification of gene-level biomarkers for brain diseases will provide novel insights into causative genes identification, drug repositioning and disease taxonomy.

In recent years, deep learning methods, especially Graph Neural Networks (GNN) based methods, have been widely used in brain network studies [41–44]. It is advantageous to use GNNs due to their power to combine node features and graph structures through end-to-end feature combinations and model the adjacency relationship between nodes via message passing [45]. Among GNNs, Graph Convolutional Network (GCN) [46] stands out as a typical method that leverages structure information and performs convolution operations on graphs to aggregate neighboring node features. Given the diverse, informative, and complex nature of brain networks, it is reasonable and efficient to perform link prediction tasks by fusing multiple heterogeneous networks. Consequently, several methods have been proposed to employ GCN for learning latent patterns in brain networks for purposes such as brain disease classification or identification of related genes [47–50]. However, existing methods are limited by their usage of restricted diseases and networks within large and complex brain networks, thus hindering the potential for predicting related pathogenic genes.

In this study, we propose M-GBBD, a Multi-network representation learning framework for the identification of Gene Biomarkers in Brain Disease. We employ eleven brain networks and extract topological semantics using a joint optimizer with dual feature extraction channels to comprehensively capture brain features. By incorporating a disease proximity subgraph and gene-disease bipartite graph into a heterogeneous graph obtained by M-GBBD, we obtain a brain gene network with biological significance. The GCN is then utilized to learn representations of gene and neurodegenerative diseases from the heterogeneous graph, enabling the prediction of association scores between genes and brain diseases. Comprehensive experimental results demonstrate that M-GBBD achieves highly competitive performance compared to several baselines in terms of both dataset and model architecture. Importantly, the generalizability and accuracy of M-GBBD are confirmed by large-scale cohort GWAS studies, where we identify *CAMP* as a potential candidate gene associated with Alzheimer's disease.

Materials and methods

Overview of networks used in M-GBBD

This study employe four types of omics data, including genomics, transcriptomics, radiomics, and connectomics to construct distinct brain networks for training and testing our model. The genomic data include human genome sequence and gene annotation information as well as disease pathogenic variants, obtained from the Human Genome Resources at NCBI (version GRCh38) and DisGeNET database [51]. The transcriptomic data consist of two types of gene expression datasets downloaded from Allen Human Brain Atlas (AHBA) [52] and Genotype-Tissue Expression (GTEx) [53], along with gene regulatory data downloaded from Gene Regulatory Networks Database (GRNdb) [34]. Radiomic data comprise brain r-fMRI signals obtained from Human Connectome Project (HCP) [54]. Regarding connectomic data, we obtained the brain functional connectivity network framework developed by the Cole Neurocognition Lab [55] (Fig. 1).

A total of eleven brain networks are constructed in this study (Table 1 and Supplemental Notes): Gene regulatory network (G-T), TF-TF similarity network (T-T), TF and brain region matching network (T-R), Gene network based on regulatory relationships (G-G), Gene-region expression network (G-R), Brain region-region functional connectivity network (R-R), Brain parcel and region matching network (P-R), Brain parcel-parcel functional connectivity network (P-P), Gene-parcel expression network (G-P), Disease-disease similarity network (D-D) and Gene-disease association network (G-D).

Overview of M-GBBD

We model the identification of causative genes in brain diseases as a link prediction issue. M-GBBD is an end-to-end framework with three main components (Fig. 2): (i) constructing two types of brain heterogeneous graphs to comprehensively represent the brain functional connectivity and gene regulatory relationships, (ii) leveraging deep neural network (DNN) with the Kullback–Leibler (KL) divergence loss to learn topological semantics from the heterogeneous graphs, thereby generating an enhanced brain functional connectivity (eBFC)-based gene network with biological significance, and finally, (iii) integrating the eBFC-based gene network with the G-D and D-D networks to perform feature representation using graph convolution network(GCN).

To capture and integrate a richer set of structural information and features of the brain, we constructed two heterogeneous graphs. The first heterogeneous graph, denoted as $H_{GPR} \in \mathbb{R}^{(N_G+N_P+N_R)\times(N_G+N_P+N_R)}$, encompasses brain parcel-parcel functional connectivity, brain region-region functional connectivity and a gene network based on regulatory relationships. The second heterogeneous graph, referred to as $A_{GTR} \in \mathbb{R}^{(N_G+N_T+N_R)\times(N_G+N_T+N_R)}$, incorporates functional connectivity among brain regions, brain gene regulatory networks, and gene networks based on gene regulatory relationships. Mathematically, the two heterogeneous graphs can be represented by the following adjacency matrix:

$$A_{GPR} = \begin{bmatrix} M_{GG} & M_{GP} & M_{GR} \\ {M_{GP}}^{T} & M_{PP} & M_{PR} \\ {M_{GR}}^{T} & {M_{PR}}^{T} & M_{RR} \end{bmatrix} \tag{1}$$

$$A_{GTR} = \begin{bmatrix} M_{GG} & M_{GT} & M_{GR} \\ {M_{GT}}^{T} & M_{TT} & M_{TR} \\ {M_{GR}}^{T} & {M_{TR}}^{T} & M_{RR} \end{bmatrix} \tag{2}$$

where ${M_{GP}}^{T}$, ${M_{GR}}^{T}$, ${M_{PR}}^{T}$, ${M_{GT}}^{T}$ and ${M_{TR}}^{T}$ indicates the transpose of M_{GP}, M_{GR}, M_{PR}, M_{GT} and M_{TR}, respectively.

Graph topological semantics extraction

We employ a deep neural network (DNN) with the KL-divergence loss to extract topological semantics

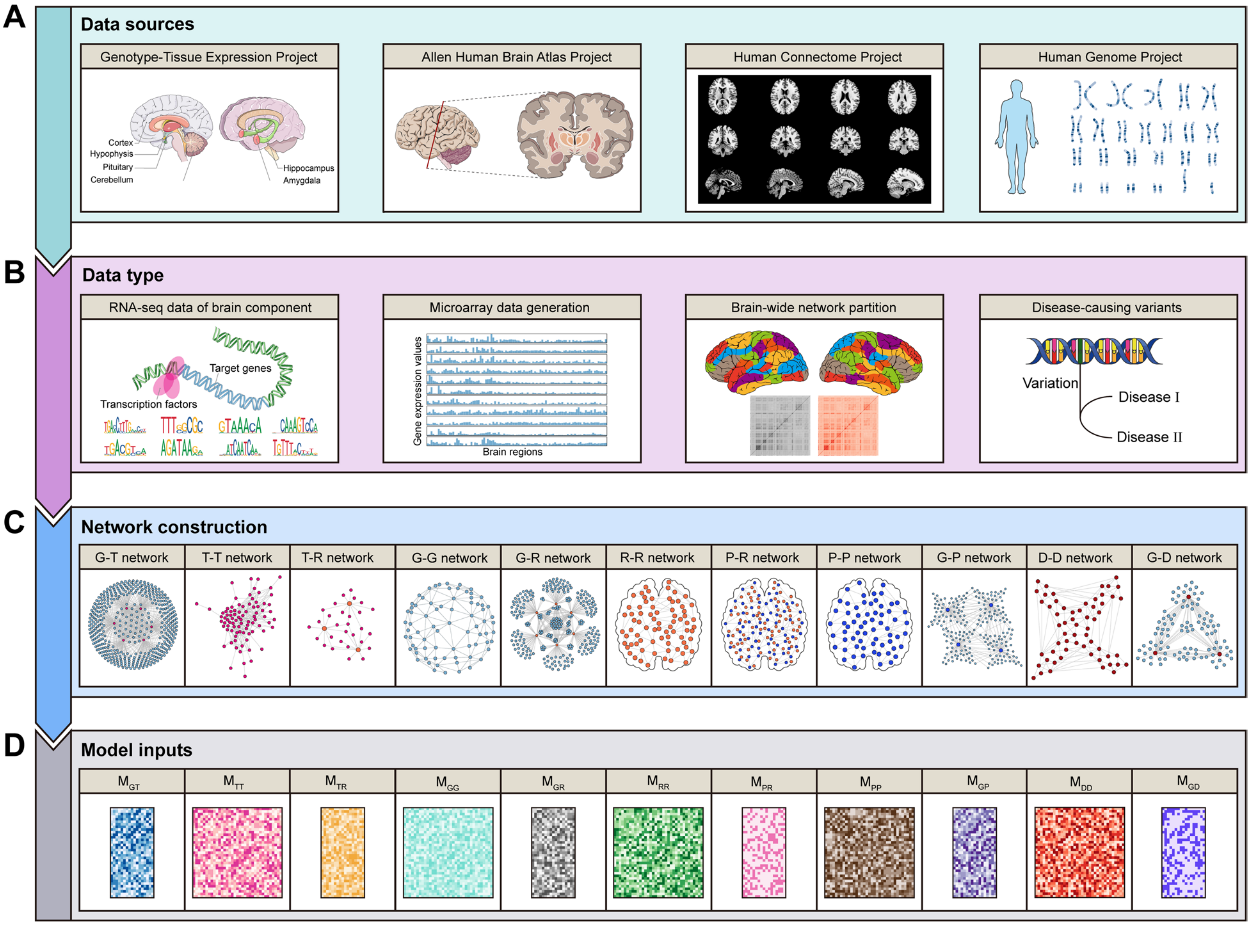


Fig. 1 Overview of the datasets used in M-GBBD. **A** The data sources for each project. **B** The types of raw data collected for each project. **C** Various brain networks constructed using the collected data. **D** Mathematical representations in the form of unique matrices are used to represent each brain network as inputs for M-GBBD

from two heterogeneous graphs. Specifically, we treat the feature maps of the two heterogeneous graphs A_{GPR} and A_{GTR} as two-dimensional representations and construct a joint optimizer with dual feature extraction channels. The input consists of these feature maps, which are then fed into a multi-layer DNN for dimensionality reduction and extraction of gene primary features along with their corresponding spatial distributions. Subsequently, we calculate the KL-divergence between the distributions of gene primary features to learn a common subspace that captures multiple heterogeneous information. During optimization, the DNN is iteratively trained using gradient backpropagation to enhance the representability of gene nodes, resulting in two final representation maps obtained through collaborative optimization of subspace and dual channels. These representation maps are utilized to derive an enhanced brain functional connectivity-based gene network (eBFC-based gene network), incorporating both brain functional

Table 1 Summary of brain networks used in this study

Brain networks	Mathematical representations	Nodes	Edges
G-T	$M_{G\text{-}T}$	14,923	123,045
T-T	$M_{T\text{-}T}$	728	15,714
T-R	$M_{T\text{-}R}$	4,430	2,687,652
G-G	$M_{G\text{-}G}$	14,195	105,824,555
G-R	$M_{G\text{-}R}$	17,897	52,549,890
R-R	$M_{R\text{-}R}$	3,702	13,704,804
P-R	$M_{P\text{-}R}$	4,420	2,933
P-P	$M_{P\text{-}P}$	718	515,524
G-P	$M_{G\text{-}P}$	14,913	5,706,390
D-D	$M_{D\text{-}D}$	10,392	992,230
G-D	$M_{G\text{-}D}$	24,587	588,178

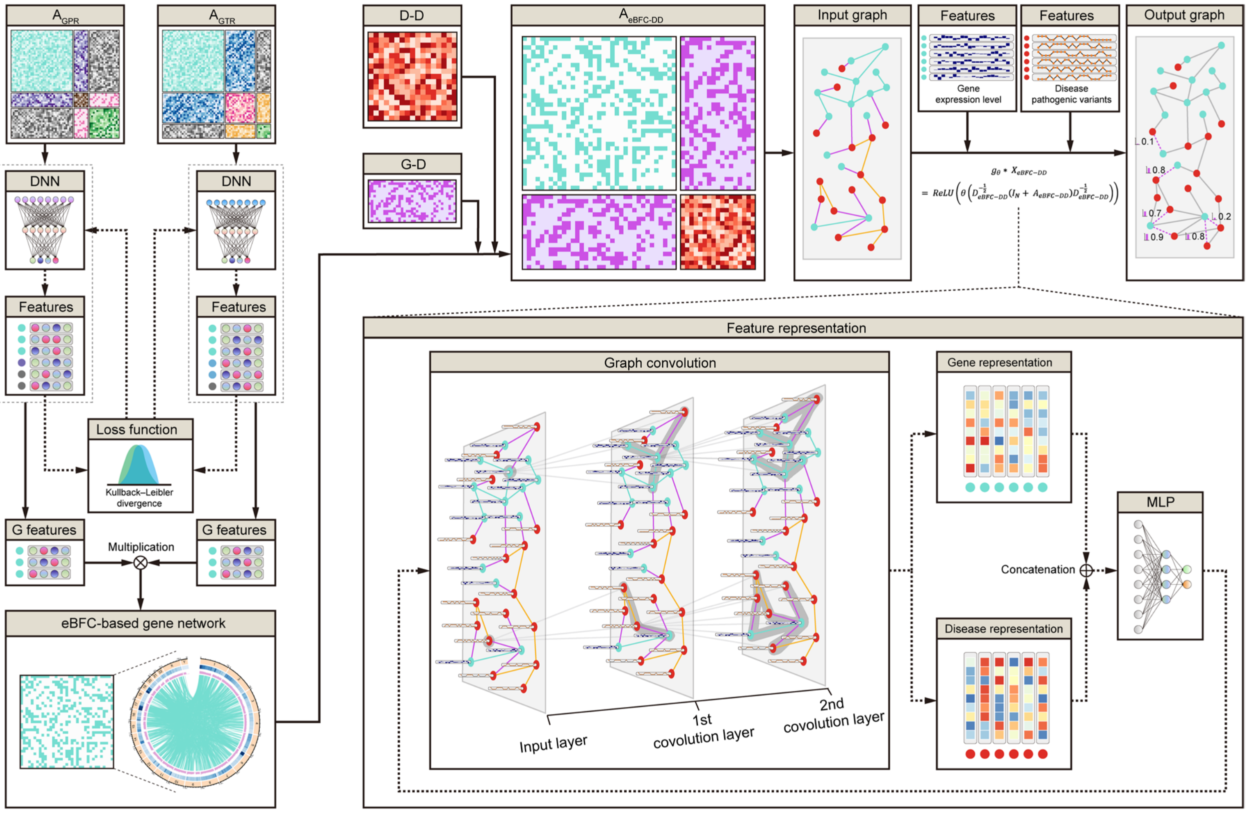


Fig. 2 Overview of the M-GBBD framework. The framework takes two brain heterogeneous graphs, namely A_{GPR} and A_{GTR} (top left) as input. To reduce dimensionality and extract gene primary features along with spatial distributions, a multi-layer DNN is employed. The Kullback–Leibler divergence is utilized to calculate and learn the distributions of common subspace. After iterative optimization, an eBFC-based gene network is obtained. By combining the eBFC-based gene network with the D-D network and G-D network, GCN is applied to learn representations of genes and diseases. Finally, these representations are fed into MLP for predicting gene-disease associations

connectivity and gene regulatory information. Following normalization based on previous studies [7], this eBFC-based gene network is further integrated into large-scale disease-disease networks to construct a bipartite graph named $G_{eBFC-DD}$. This step can be represented as follows:

$$z^{(i,j)} = w_3\alpha\left(w_2\alpha\left(w_1x^{(i,j)} + b_1\right) + b_2\right) + b_3 \tag{3}$$

where w_1, w_2 and w_3 represent the corresponding weight matrix, and b_1, b_2 and b_3 represent the bias vector for the three corresponding layers. $\alpha(\cdot)$ represents the activation function ReLU. The KL-divergence loss is defined as

$$D_{KL}(P_{GPR}||P_{GTR}) = \sum_{i=1}^{n} P_{GPR}(x_i)\log\left(\frac{P_{GPR}(x_i)}{P_{GTR}(x_i)}\right) \tag{4}$$

where P_{GPR} and P_{GTR} represent the distribution of different representations and

$$Z_{G-PR} = f\left(Z_G \cdot {Z_{PR}}^T\right) \tag{5}$$

$$Z_{G-TR} = f\left(Z_G \cdot {Z_{TR}}^T\right) \tag{6}$$

$$Z_G = Z_{G-PR} \cdot {Z_{G-TR}}^T \tag{7}$$

where Z_G denotes the representations of genes and Z_{PR} denotes the representations of TFs and brain regions from Z_{G-PR}. Z_{TR} denotes the representations of TFs and brain regions from Z_{G-TR}. $f(\cdot)$ denotes the dimension reduction operation.

Graph convolutional network

In general, graph-based deep learning approaches can be categorized into two types: spatial-based and spectral-based. Spatial-based methods learn node representation by iteratively aggregating information from neighboring nodes, which may result in over-smoothing of the node representation [56]. On the other hand, spectral-based methods rely on the spectrum of the graph Laplacian of the design matrix [46]. Compared with spatial-based methods [57–61], spectral-based methods generally exhibit better performance in graph learning [62, 63]. A representative example of a spectral-based method is modified Chebyshev polynomials, which simplifies parameters and avoids large computational burdens. Given the complexity and scale of our networks, employing a multilayer GCN that is spectral-based to learn gene and disease representations from brain networks is feasible.

Specifically, the input to a GCN is the graph $G_{eBFC-DD} = (\sqsubseteq, \mathcal{E})$, where $\sqsubseteq = (N_G, N_D)$ represents N_G gene nodes and N_D disease nodes, and $\mathcal{E}$ is a set of edges between nodes. The objective is to predict potential edges between gene-disease pairs that have not been previously identified in $G_{eBFC-DD}$. Denoting $G_{eBFC-DD}$ as an adjacency matrix $A_{eBFC-DD} \in \mathbb{R}^{(N_G+N_D)\times(N_G+N_D)}$, the features of both types of nodes are required. It should be noted that there are two types of nodes: gene nodes and disease nodes, which correspond to different types of features. For gene nodes, the features consist of gene expression levels at different brain sites based on RNA-seq results from 2,642 brain sites. Pathogenic variant genotypes are used as features for disease nodes, with a value of 1 indicating association with a variation and 0 otherwise. The raw data for both node types is encoded using stacked autoencoders (SAE) to ensure consistent feature dimensions. Denoting the dimensionality of SAE output as $C_{SAE} \in \mathbb{R}$, the final node feature matrix $X_{eBFC-DD} \in \mathbb{R}^{(N_G+N_D)\times C_{SAE}}$ can be obtained by concatenating SAE outputs for gene and disease nodes.

The graph convolution is defined on a graph as the product of the input signal and the filter g_θ in the Fourier domain. Here, denoting the symmetric normalized Laplacian matrix of $A_{eBFC-DD}$ as $L_{eBFC-DD} = U_{eBFC-DD}\Lambda_{eBFC-DD}{U_{eBFC-DD}}^t$, where $U_{eBFC-DD}$ represents the eigenvector matrix and $\Lambda_{eBFC-DD} = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_{N_G+N_D})$ denotes the diagonal matrix of eigenvalues. The Fourier transform of $X_{eBFC-DD}$ can be represented as ${U_{eBFC-DD}}^t X_{eBFC-DD}$. However, computing the eigenvector matrix and eigenvalue diagonal matrix becomes computationally expensive with an increasing scale of the graph. To reduce computational complexity, a modified GCN based on Chebyshev polynomials $T_K(x) = 2xT_{K-1}(x) - T_{K-2}(x)$ was used here for brain network feature representation. Consequently, we define and represent the filter g_θ as

$$g_\theta(\Lambda_{eBFC-DD}) = \sum_{K=0}^{K} \theta_K T_K\left(\widetilde{\Lambda}_{eBFC-DD}\right) \tag{8}$$

$$g_\theta * X_{eBFC-DD} = \sum_{K=0}^{K} \theta_K T_K\left(\widetilde{L}_{eBFC-DD}\right) X_{eBFC-DD} \tag{9}$$

where $\theta \in \mathbb{R}^K$ denotes the vector of Chebyshev coefficients, $\widetilde{\Lambda}_{eBFC-DD} = \frac{2\Lambda_{eBFC-DD}}{\lambda_{max}} - I_N$, $\widetilde{L}_{eBFC-DD} = \frac{2L_{eBFC-DD}}{\lambda_{max}} - I_N$, I_N denotes the identity matrix and K denotes the K^{th}-order neighborhood.

Given that Chebyshev polynomials are recursive [64], the formulation is simplified by restricting K = 1 [46] and introducing activation functions in each layer (l > 0) to enhance the power of the model. Finally, the graph convolution method used in this study can be represented as

$$g_\theta * X_{eBFC-DD} = \theta\left(D_{eBFC-DD}^{-\frac{1}{2}}(I_N + A_{eBFC-DD})D_{eBFC-DD}^{-\frac{1}{2}}\right) \tag{10}$$

$$g_\theta * X_{eBFC-DD} = ReLU\left(\theta\left(D_{eBFC-DD}^{-\frac{1}{2}}(I_N + A_{eBFC-DD})D_{eBFC-DD}^{-\frac{1}{2}}\right)\right) \tag{11}$$

$$\begin{bmatrix} H_G \\ H_D \end{bmatrix} = g_\theta * X_{eBFC-DD} \tag{12}$$

$$H_{GD} = H_G \oplus H_D \tag{13}$$

where $D_{eBFC-DD}$ denotes the diagonal matrix with diagonal entry $[D_{eBFC-DD}]_{i,j} = \sum_j [A_{eBFC-DD}]_{i,j}$, H_G denotes the embedding of genes and H_D represents the embedding of diseases. ⊕ denotes a concatenation operator and H_{GD} denotes the embedding of gene-disease pair.

The prediction of the gene-disease association scores is formulated as an end-to-end binary classifier in this study. After applying the GCN to obtain embedding vectors, they are concatenated and used as the input for a multi-layer perception (MLP). The association scores are computed using the sigmoid function applied to the output of the last hidden layer:

$$S = Sigmoid(W_{out} \cdot H_{GD} + b_{out}) \tag{14}$$

where $\mathcal{S}$ denotes the scores of gene-disease associations, W_{out} and b_{out} denote the weight matrix and the bias vector.

The cross-entropy loss $\mathcal{L}$ is adopted to optimize model parameters as

$$L = \sum_{i,j\in\mathcal{Y}\cup\mathcal{Y}^-}\left(y_{ij}\log\widehat{y}_{ij} + \left(1 - y_{ij}\right)\log\left(1 - \widehat{y}_{ij}\right)\right) \tag{15}$$

where y_{ij} represents the true label of the edges, which will be 1 or 0, $\mathcal{Y}$ and $\mathcal{Y}^-$ denote the sets of nodes contained in the positive edges set and negative edges set, respectively. Then, the whole model via back propagation algorithm in an end-to-end manner can be trained.

Experimental setting

The prediction model is tuned using five-fold cross-validation (5-CV). To evaluate the accuracy of M-GBBD, the receiver operating characteristic (ROC) curve is employed. The area under the ROC curve (AUC) served as the primary evaluation metric. Additionally, considering AUC's bias towards imbalanced datasets, we also utilize the precision-recall (PR) curve. The area under the PR curve (AUPR) is selected as another primary evaluation metric. Besides, other evaluation metrics such as accuracy (ACC), recall (REC), precision (PRE) and F1-score (F1) are also calculated.

Several hyperparameters are consisted in the model, including the learning rate of optimizer $L \in \{0.0002, 0.0004, 0.0006, 0.0008\}$, the hidden dimensionality of embeddings $H \in \{16, 32, 64, 128\}$, the dropout rate $D \in \{0.01, 0.05, 0.1, 0.3\}$, the Chebyshev filter size $K \in \{2, 3, 4, 5\}$, and the total training epochs $E \in \{200, 600, 1000, 2000\}$. The best obtained parameters are $L = 0.0004$, $H = 64$, $D = 0.05$, $K = 4$ and $E = 1000$.

After intersecting all datasets used in this study, a total of 14,195 genes were retained. As we have collected comprehensive human genome-wide gene information that includes consistent characterization and network structure information here, 20 known gene-disease associations related to two specific brain diseases (Alzheimer's disease and Parkinson's disease) have been pre-isolated by random selection for further demonstration. These pre-isolated associations are not involved in any training process to prevent data leakage, and thus ensuring objectivity. Finally, a total of 14,175 genes and 10,392 diseases formed a dataset consisting of 557,893 associations which participated in the subsequent training process.

Results

Overall performance

The eBFC-based gene network, which covers most genes in the human genome, has been derived through topological semantics extraction from A_{GPR} and A_{GTR}. It is essential to note that gene expression may be regulated through various mechanisms, resulting in one gene being associated with multiple diseases due to distinct regulatory pathways [65–67]. In other words, several common pathogenic genes can be identified across different diseases, with differential regulation of these genes being particularly prevalent among brain diseases [18, 20]. Therefore, it is more reasonable to use a link prediction paradigm for identifying pathogenic genes related to brain diseases. In our study, we constructed a disease-disease (D-D) network comprising 10,392 diseases in M-GBBD, enabling the prediction of associations between any given gene and disease within this network. Evaluation of M-GBBD performance shows that across all diseases considered, the mean values for AUC, AUPR, ACC, PRE, REC, and F1 of M-GBBD are found to be 0.891, 0.893, 0.729, 0.939, 0.489 and 0.643, respectively (Fig. 3A). Furthermore, the consistency observed in each cross-validation further supports the robustness of our finding (Fig. 3B and C). Among these 10,392 diseases, there are 2,102 kinds of diseases that are specifically associated with brain-related diseases. The AUC and AUPR values for each disease exhibit relatively similar trends (Fig. 3D). Notably, diseases linked to the brain demonstrate higher values for both AUC and AUPR compared to other non-brain related ailments (Fig. 3E), indicating that M-GBBD is sensitive to such diseases.

Furthermore, we have chosen four representative brain diseases (Alzheimer's disease, Parkinson's disease, Major depression, and Autism) for comprehensive investigation and discussion. These diseases are well-known for their high prevalence and significant impact on individuals, thus extensively studied by various models [5, 7, 8, 44, 68]. The performance evaluation metrics including AUC/AUPR/ACC/PRE/REC/F1 for the aforementioned diseases are as follows: 0.893/0.867/0.829/0.768/0.820/0.793 (Alzheimer's disease), 0.866/0.881/0.767/0.666/0.832/0.740 (Parkinson's disease), 0.883/0.864/0.797/0.699/0.813/0.752 (Major depressive disorder) and 0.887/0.844/0.746/0.632/0.846/0.723 (Autism), respectively (Fig. 3F). Notably, all these evaluation metrics surpass those of the previous method [7], demonstrating the excellent performance of M-GBBD.

Improved performance of multiscale disease network and eBFC-based gene network

To evaluate the performance across different combinations of multiscale disease network and eBFC-based gene networks, we conducted three comparative experiments. The first experiment aims to evaluate the predictive performance improvement of eBFC-based gene network compared to BFC-based gene network. To be specific, we use BFC-based and eBFC-based gene networks to train and predict associations between genes and four representative brain diseases using brainMI. The results demonstrate a significantly higher performance of brainMI when utilizing eBFC-based gene network compared to BFC-based gene network (Fig. 4A). On average, the AUC and AUPR values for disease prediction by brainMI using eBFC-based gene network increase by 0.038 and 0.041, respectively, in comparison with those obtained from BFC-based gene network (Fig. 4A). This indicates that eBFC-based gene network may encompass more comprehensive information than the BFC-based counterpart, thereby improving predictive performance.

The other two experiments are conducted to evaluate the performance improvement achieved by multiscale

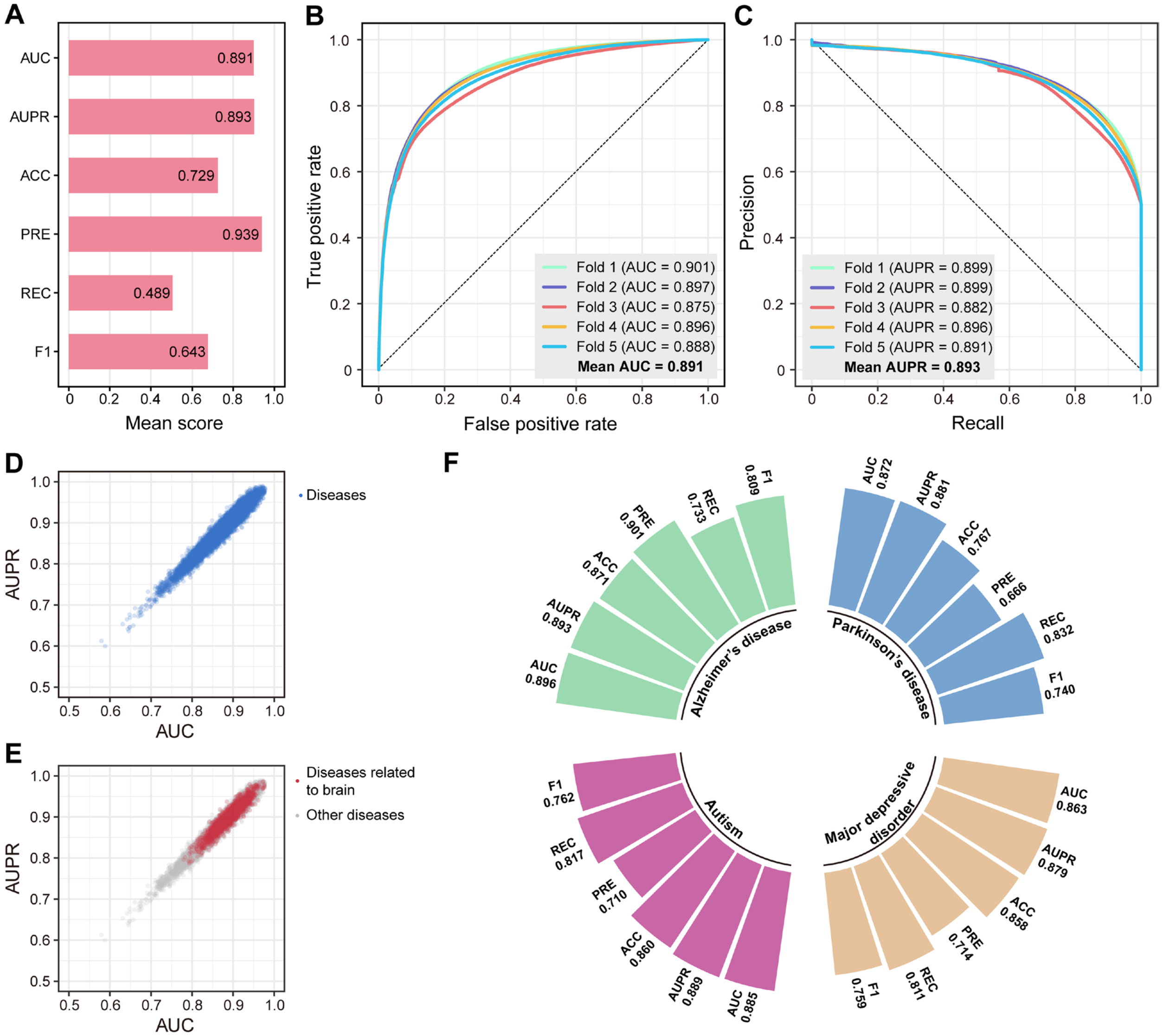


Fig. 3 Overall performance of M-GBBD. **A** Mean values of each evaluation metrics under 5-CV. **B** ROC curves of M-GBBD under 5-CV. **C** PR curves of M-GBBD under 5-CV. **D** Distribution of AUC and AUPR values for all diseases. **E** Distribution of AUC and AUPR values for disease related to brain. **F** Performance of M-GBBD on four representative diseases that related to brain

disease networks. Due to that brainMI employs a node classification strategy whereas M-GBBD utilizes a link prediction strategy, the D-D network cannot be directly utilized in brainMI experiments. Therefore, we constructed a small-scale disease proximity network (sDD) that includes only four diseases mentioned in brainMI using the same methodology as for the D-D network and performed experiments using M-GBBD. For clarity, we refer to the D-D network used in M-GBBD as the large-scale D-D network (lDD). By combining both sDD and lDD with two gene networks (BFC-based and eBFC-based), we aim to demonstrate whether lDD can indeed improve predictive performance significantly. Compared to sDD, when combined with BFC-based gene network, lDD exhibited an average increase of 0.034 in AUC and 0.032 in AUPR, respectively (Fig. 4B). When combined with the eBFC-based gene network, there is an average improvement of 0.048 in AUC and 0.049 in AUPR using lDD (Fig. 4C). These results consistently indicate that regardless of which gene network is employed, lDD consistently outperforms sDD. In summary, utilizing the biologically significant eBFC-based gene network along with a large-scale proximity network can achieve superior performance for predicting gene-disease associations within the brain compared to traditional single BFC-based gene network.

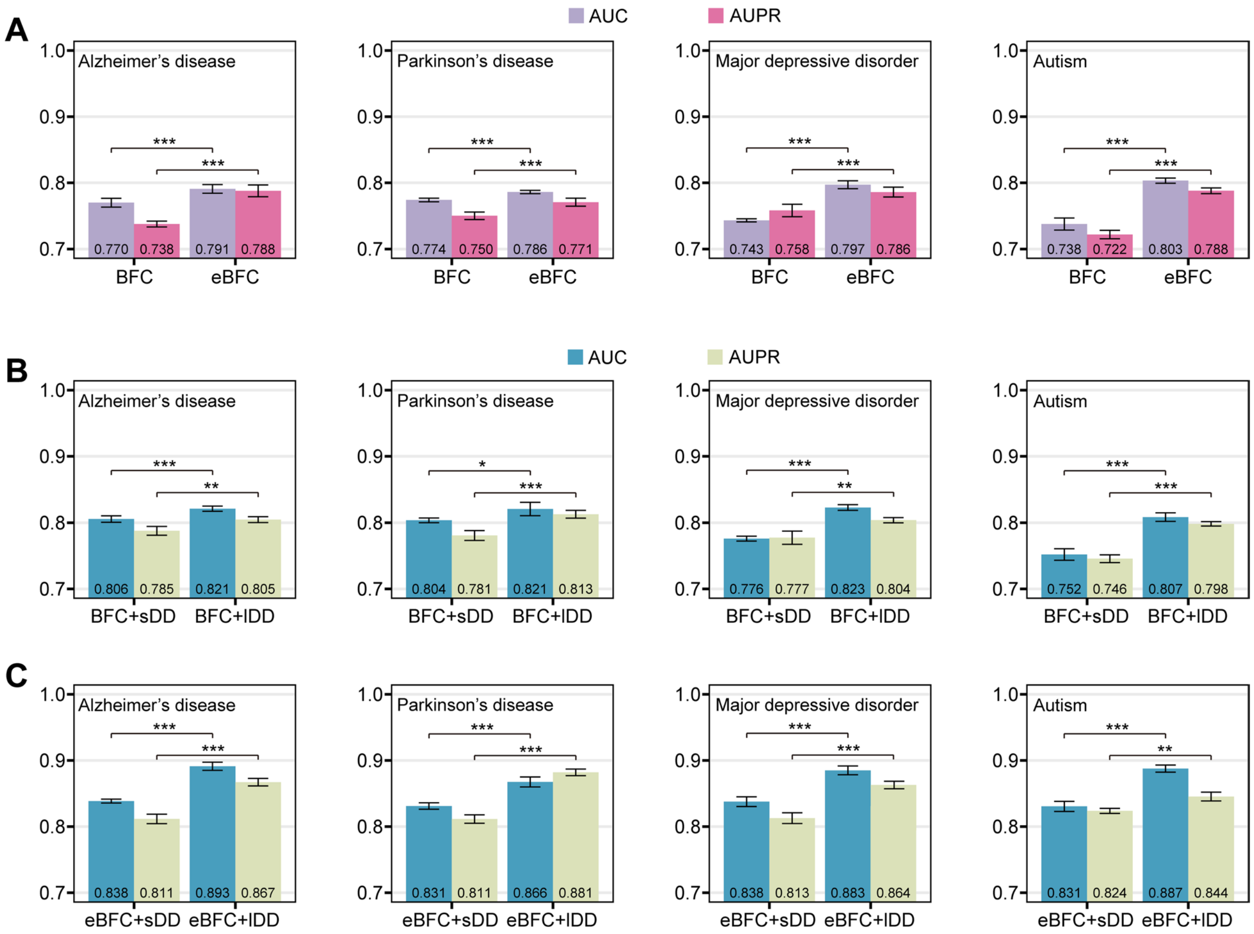


Fig. 4 Performance of M-GBBD and brainMI with different datasets on four representative brain related diseases. **A** Mean AUC and AUPR values of brainMI with only BFC- and eBFC-based gene networks under 5-CV. **B** Mean AUC and AUPR values of M-GBBD with BFC-based gene network and sDD/IDD under 5-CV. **C** Mean AUC and AUPR values of M-GBBD with eBFC-based gene network and sDD/IDD under 5-CV. Statistical significant was estimated using two-tailed Student's t-test. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$

Comparison with the state-of-the-art frameworks

Given the tedious and multilayered nature of brain disease diagnosis, graph-based methods offer an efficient approach to learn representations for identifying associations from vast amounts of data [69, 70]. To evaluate the performance improvement of the M-GBBD algorithm, three gene-disease prediction frameworks, including BiRW [71], PMFMDA [72] and MeSHHeading2vec [73], are compared with M-GBBD. These frameworks are all designed to predict associations between genes and diseases. BiRW utilizes a bi-random walk algorithm, while PMFMDA is based on matrix factorization, and MeSHHeading2vec employs graph embedding algorithms for relationship prediction tasks. Each framework was executed using default parameters and 5-CV. The evaluation metrics including AUC, AUPR, ACC, REC, PRE, and F1 were calculated for each framework in order to facilitate comparison.

The results show that M-GBBD outperforms all other frameworks in terms of evaluation metrics, except for REC (Fig. 5A). Although PMFMDA achieves the highest REC value, its PRE values are the lowest. Compared to other methods, M-GBBD shows an average improvement of 0.194 and 0.341 in AUC and AUPR respectively (Fig. 5B). This superior performance can be attributed to the GCN's ability to more effectively aggregate network information. Overall, with the benefit of the GCN and its end-to-end computational structure, our M-GBBD is a more suitable method for predicting associations between genes and disease in the brain.

Ablation analysis demonstrates the importance of multiple semantics extraction

To further investigate the contribution of critical components and evaluate the robustness of M-GBBD, we compared it with two variant methods, namely

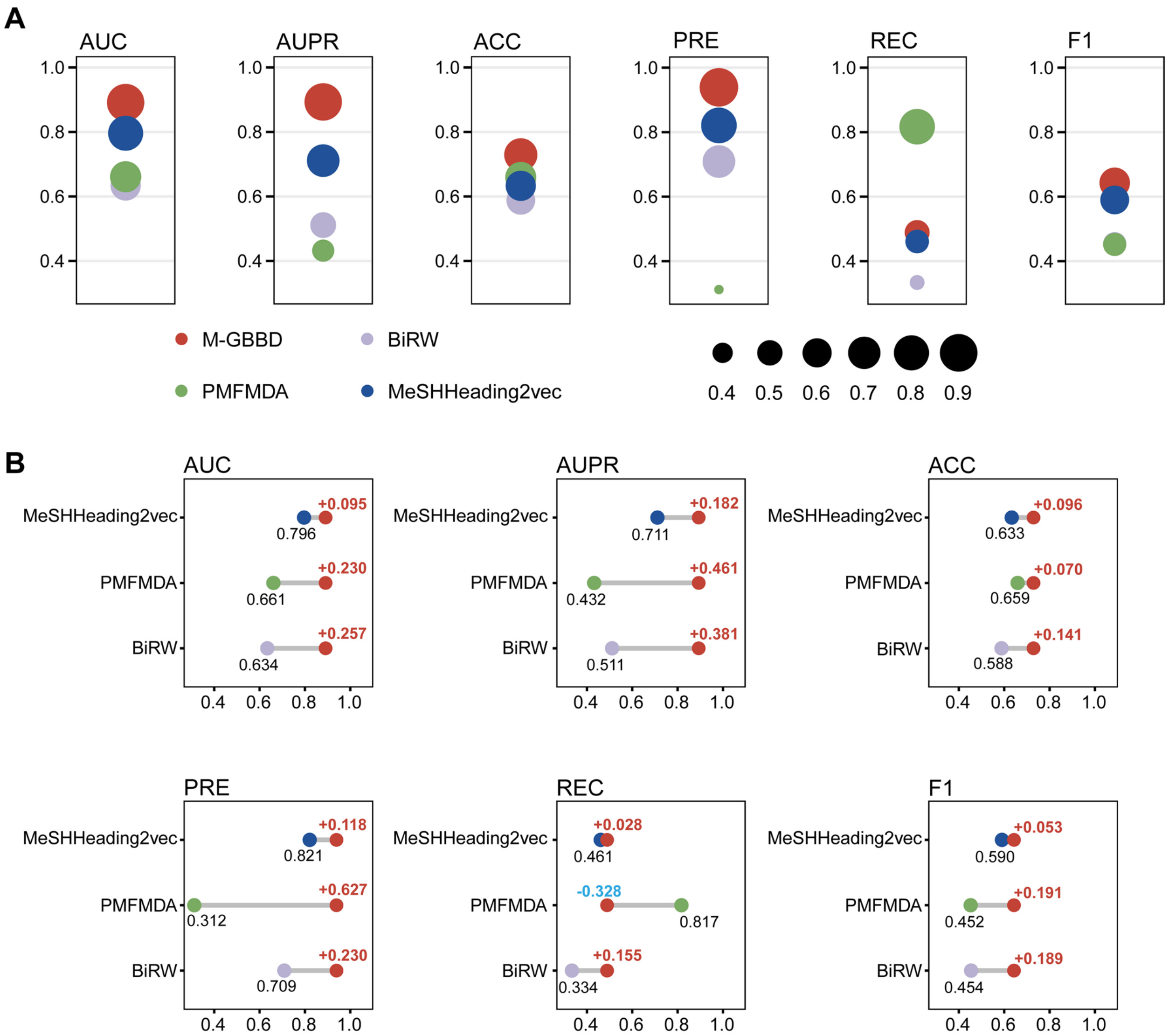
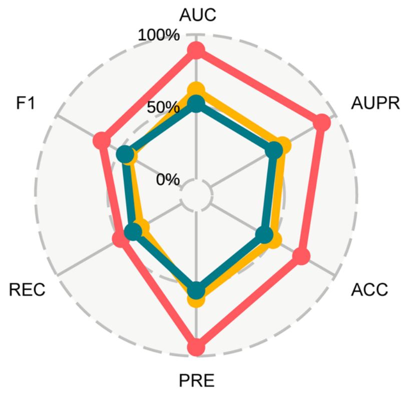


Fig. 5 Comparison on the performance of different gene-disease prediction frameworks. **A** Results of the six evaluation metrics for the four frameworks. **B** The difference in performance of M-GBBD relative to the other three frameworks. Colors of dots are same as in (**A**) and improvement/decline are indicated by red/blue bold numbers

M-GBBD-noGPR and M-GBBD-noGTR. The M-GBBD-noGPR method exclude the heterogeneous network comprising brain parcel-parcel functional connectivity, while the M-GBBD-noGTR method removed the heterogeneous network involving gene regulatory interactions. Following a 5-CV for each method, we obtain AUC values 0.891, 0.613 and 0.522 for M-GBBD, M-GBBD-noGPR and M-GBBD-noGTR respectively. Correspondingly, the AUPR values were found to be 0.893, 0.578 and 0.510 (Fig. 6). In addition, ACC, PRE, REC and F1 of M-GDAB are also superior to corresponding metrics of other methods (Fig. 6). Our ablation experiments results demonstrate that combining brain parcel-parcel functional connectivity with gene regulatory features forms a crucial foundation for performance improvement.

Case studies

To demonstrate the applicability of M-GBBD in predicting potential gene-disease associations in practical scenarios, we apply M-GBBD to predict genes associated with two brain diseases: Alzheimer's disease and Parkinson's disease. For each disease, five associated genes are randomly selected while their known twenty gene-disease associations for the two diseases are concealed to ensure these associations are



Methods	AUC	AUPR	ACC	PRE	REC	F1
🟥 M-GBBD	0.891	0.893	0.729	0.939	0.489	0.643
🟨 M-GBBD-noGPR	0.613↘	0.578↘	0.503↘	0.604↘	0.332↘	0.428↘
🟦 M-GBBD-noGTR	0.522↘	0.510↘	0.433↘	0.546↘	0.393↘	0.457↘

*↘ The arrow indicates this value is lower than standard M-GBBD

Fig. 6 Comparison on the performance of variant methods of M-GBBD. Five evaluation metrics for the three methods including the raw M-GBBD were calculated and compared. All metrics in the table are lower than M-GBBD, which is highlighted with blue arrows

pre-isolated. These associations are not considered during the semantics extracting and model training steps, which make the case study objective and reliable. Subsequently, M-GBBD was used to predict the gene-disease associations for these associated genes and report their association scores. The results are validated using the DisGeNET database, based on biological experiment reports, or further bioinformatics analysis of biological data.

In the DisGeNET database, *LRP6*, *F11*, *CXCL10*, *TCF4* and *IGF2* are identified as the top five genes associated with Alzheimer's disease, with association scores of 0.989, 0.981, 0.953, 0.938 and 0.914, respectively (Fig. 7A). Notably, all scores exceed the threshold of 0.9. Besides, *HAVCR2*, *CAMP*, *MRPS11*, *LPIN2* and *TMEM30B* are five genes without labeled associations in DisGeNET but exhibit association scores of 0.898, 0.809, 0.307, 0.233 and 0.102, respectively (Fig. 5A). Interestingly, *HAVCR2* and *CAMP* demonstrate higher scores compared to other genes, suggesting that M-GBBD has potential for predicting potential Alzheimer's disease-associated genes not yet annotated by DisGeNET. Further analysis is conducted to investigate the rationale behind the high scores of the two genes predicted by M-GBBD. According to a recent large-scale genome-wide association analysis for Alzheimer's disease based on more than one million individuals, significant associations between *HAVCR2* and Alzheimer's disease were found [68]. The variant site of locus 8 (rs6891966) in an intron of *HAVCR2* results in a significant differential expression level in brain tissue samples from patients compared to controls. This is consistent with the results obtained from M-GBBD, indicating an association between *HAVCR2* and Alzheimer's disease. The protein product of *CAMP* is a sequence with 170 amino acids and the high confidence structure model was predicted by AlphaFold (Fig. 7B and C) [74]. It exhibits antibacterial activity and binds to bacterial lipopolysaccharides (LPS) [75, 76]. Although direct experimental evidence supporting the association between *CAMP* and Alzheimer's disease is currently lacking, microarray analysis (GSE85426), which included 90 patients with Alzheimer's disease and 90 controls, revealed significant changes in *CAMP* expression levels (Fig. 7D and E). Furthermore, an epigenome-wide association study also found a CpG island located in a significant differentially methylated region of *CAMP* [77]. Therefore, it is reasonable for M-GBBD to identify *CAMP* as highly associated with Alzheimer's disease. Additionally, the microarray analysis also demonstrated significant differences in *HAVCR2* expression ($P<0.001$) (Fig. 7E), consistent with the original report [68]. Conversely, no significant differences were observed in the expression levels of *MRPS11*, *LPIN2* and *TMEM30B* and the three genes all received low scores (Fig. 7E). Both GWAS and microarray analysis results corroborate the accuracy and applicability of M-GBBD for predicting candidate gene biomarkers related to Alzheimer's disease.

In the case of Parkinson's disease, another severe neurodegenerative disorder, M-GBBD also demonstrated satisfactory performance. The DisGeNET database labels *NLRP1*, *MSC*, *PTK2B*, *TAC1* and *FOSL2* as genes associated with Parkinson's disease, with association scores of 0.964, 0.944, 0.907, 0.889 and 0.888, respectively (Fig. 8A). Except for *MUC19* which scored at 0.782,

all other unlabeled genes have association scores below 0.4 in M-GBBD. To further investigate the potential association between *MUC19* and Parkinson's disease, a GWAS summary based on data from 482,730 individuals and analyzing a total of 17,510,617 SNPs was collected [78]. The GWAS result revealed that there were significant associations between Parkinson's disease and eleven SNPs located within the gene body of *MUC19*

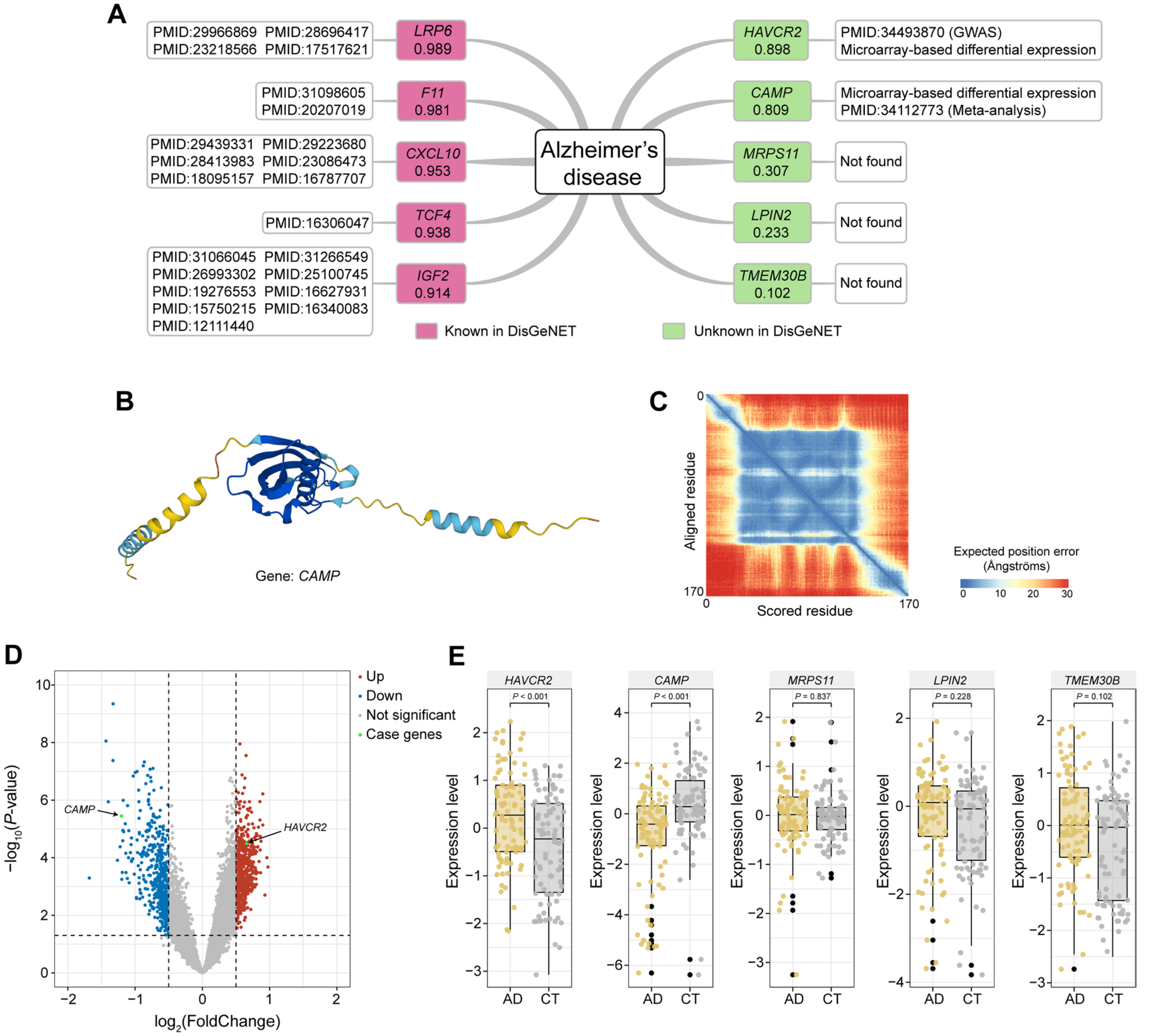


Fig. 7 Case study of Alzheimer's disease. **A** Gene-Alzheimer's disease associations predicted by M-GBBD, with corresponding scores. The pink box indicates that DisGeNET has recorded that this gene is associated to Alzheimer's disease, and the green box indicates that DisGeNET has no record that this gene is associated to Alzheimer's disease. The white boxes following the pink/green boxes is the evidence. **B** Three-dimensional structure of *CAMP* from AlphaFold. **C** Heatmap of the three-dimensional structure predicted aligned error. It means the AlphaFold's expected position error reside x, which the predicted and true structures are aligned on residue y. **D** The results of differential expression analysis from microarray (GSE85426). **E** The normalized expression values of each sample in microarray (GSE85426) for five genes that not recorded in DisGeNET. Statistically significant was estimated using two-tailed Student's t-test

(Fig. 8B), providing evidence for the relationship between *MUC19* and Parkinson's disease. According to detailed information of *MUC19* from the human genome, 5,125 potential variant sites are located in or neighbored by gene coding region. These variants were detected by genome sequencing (27.2%), exome sequencing (52.4%) or both (20.4%) in a previous study (Fig. 8C), and 40.9% of them will cause loss of function (nonsynonymous, splicing and frameshift) (Fig. 8D) that *MUC19* was assessed to have high association with Parkinson's disease by M-GBBD is sensible, as supported by GWAS results.

Discussion

The brain system is a complex network of regulatory molecules, in which their interactions contribute to the normal or disordered biological characteristics of the brain system. As attention towards brain diseases increases, various graph deep learning-based studies

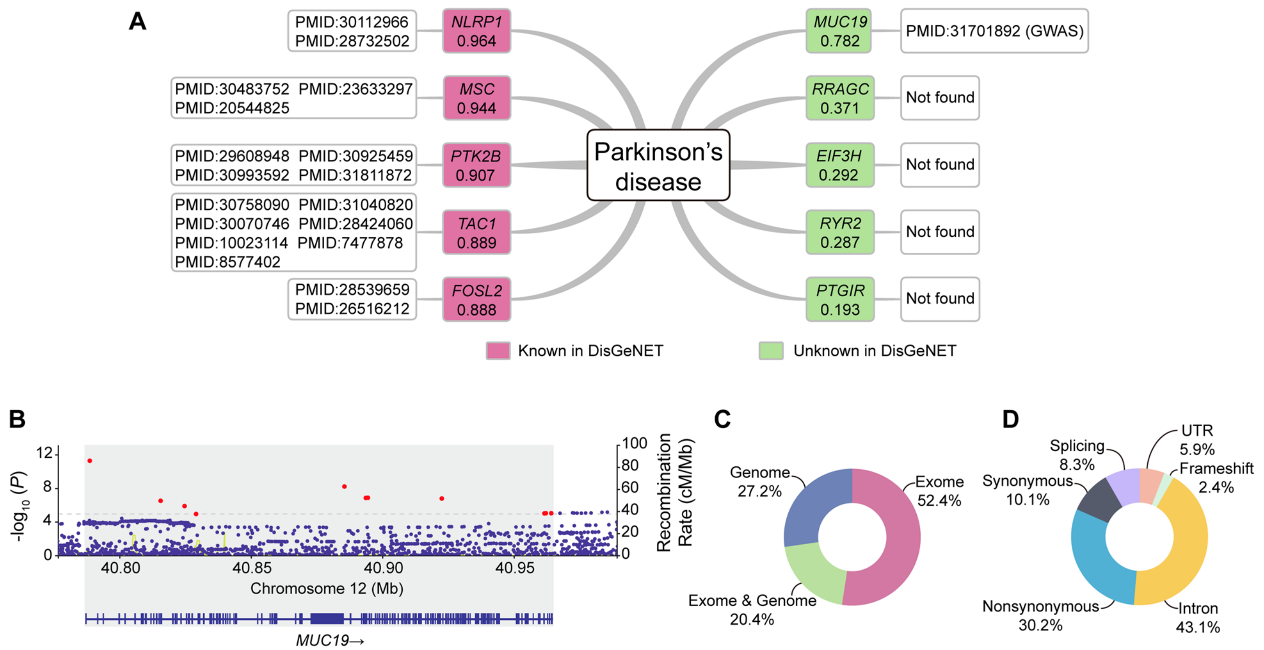


Fig. 8 Case study of Parkinson's disease. **A** Gene-Parkinson's disease associations predicted by M-GBBD, with corresponding scores. The pink box indicates that DisGeNET has recorded that this gene is associated to Parkinson's disease, and the green box indicates that DisGeNET has no record that this gene is associated to Parkinson's disease. The white boxes following the pink/green boxes is the evidence. **B** Manhattan plot of *MUC19*, the grey area indicates the range of the *MUC19*. Red dots are significant variants within *MUC19*. **C** The potential variant sites located or near coding region obtain from gnomAD browser. **D** The functional annotations of potential variants

have been proposed for brain gene biomarker identification. However, these studies have several shortcomings including limited diversity in biological network types, lack of an effective and biologically meaningful network fusion strategy, inadequate extraction of graph structure and node feature information, as well as unsatisfactory model performance and generalizability [7, 79–81]. Although we have partially addressed these limitations by developing a pioneering topological semantics extraction approach called M-GBBD to construct a biological meaningful brain gene network, this approach only extracts semantics from networks constructed using genomics, transcriptomics, radiomics, and connectomics data. Networks constructed using other omics data such as epigenomics, metabolomics and proteomics have not yet been used or discussed in this study. With advancements in molecular biology and biotechnology innovation, more comprehensive data will be easily obtained in the future. Admittedly, incorporating different types of brain networks into M-GBBD may further improve its predictive performance for associations between genes and brain diseases; however effective and accurate strategies for topological semantics extraction from brain networks that aim to obtain a gene network with rich semantics reflecting multiple biological meanings continue to pose challenges.

In addition, M-GBBD is a GCN model that follows the Transductive Learning paradigm [82], which takes a broad and global perspective on gene biomarker identification.

At the beginning of model training, the training set (nodes with edges and labels) and the node information of the test set (without edges) are available while the corresponding edge information remains unseen as these edges will be predicted in the subsequent model test phase. Although the true edges of the test set are unknown during training, additional information can be obtained from their node feature distribution, such as distribution aggregation, which resembles drug repositioning. While transductive learning can extract some additional information from all nodes and edge information in the training set to enhance model effectiveness, it also necessitates retraining and increased computation whenever new samples are received. In future work, we will further explore how to leverage inductive learning to improve identification accuracy of brain disease gene markers by considering brain network specificity.

Conclusions

In this study, we constructed and conducted topological semantics extraction of eleven brain networks to characterize the brain features from different perspectives. In contrast to existing methods that only focus on a single disease, we introduced a biologically meaningful disease network by incorporating common disease-causing variants. Our M-GBBD model captures both functional connectivity and gene regulation information through joint optimization and multi-channel feature extraction strategies, enabling us to obtain an informative brain gene network with superior performance compared to other methods. The extraction of different network topological semantics highlights the crucial role of utilizing multi-networks for studying brain diseases comprehensively. Extensive experiments demonstrated the accuracy of M-GBBD, while case studies showcased its excellent generalizability in accurately assessing the association between genes and brain diseases. The M-GBBD gave accurate and reasonable scores for all genes used in the case analysis. Notably, our analysis suggests a potential association between *CAMP* and Alzheimer's disease, which is further supported by in-depth bioinformatics analysis.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-09967-9.

Additional file 1.

Acknowledgements
All the authors agree to the Open Access Data Use Terms of HCP ConnectomeDB (https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms), the terms of Allen Insitute (https://alleninstitute.org/legal/terms-use/) and the accept the terms of GTEx (https://gtexportal.org/home/license). Parts of the figure were drawn by using pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/).

Authors' contributions
Ping Zhang and Weihan Zhang designed the methods and arranged the datasets. Ping Zhang, Weihan Zhang and Weicheng Sun implemented the methods and performed the analyses. Jinsheng Xu, Weicheng Sun and Weihan Zhang tested the methods. Ping Zhang and Weihan Zhang wrote the manuscripts. Hua Hu, Lei Wang and Leon Wong provided financial support and gave suggestions for improvement of the methods. All authors read and approved the final manuscript.

Funding
This work was supported in part by STI 2030-Major Projects, under Grant 2021ZD0200403, in part by the Guangxi Postdoctoral Special Funding Project, the Natural Science Foundation of Guangxi, under Grant 2022JJD170019, the National Natural Science Foundation of China, under Grants 62172355, the Guangxi Science and Technology Base and Talent Special Project under Grant 2021AC19394 and 2021AC19354.

Availability of data and materials
The human genome was obtained from National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/genome/guide/human/), the diseases information and causing variants were obtained from DisGeNET (https://www.disgenet.org/home/), the gene expression data of different brain regions and sites were obtained from Allen Human Brain Atlas (https://human.brain-map.org/static/download) and Genotype-Tissue Expression project (https://gtexportal.org/home/datasets), the gene regulatory information was obtained from Gene Regulatory Networks Database (http://www.grndb.com/download/), the r-fMRI data was obtained from the Human Connectome Project (https://db.humanconnectome.org/), the CAB-NP brain functional connectivity network was obtained from the ColeLab (https://github.com/ColeLab/ColeAnticevicNetPartition). The code and data of M-GBBD are available at https://github.com/Weihankk/M-GBBD.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 6 September 2023 Accepted: 3 January 2024
Published online: 14 February 2024

References

1. Feigin VL, Nichols E, Alam T, Bannick MS, Beghi E, et al. Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 2019;18:459–80.
2. Erskine HE, Moffitt TE, Copeland WE, Costello EJ, Ferrari AJ, et al. A heavy burden on young minds: the global burden of mental and substance use disorders in children and youth. Psychol Med. 2015;45:1551–63.
3. Wijeratne T, Fox S, World Brain Day. Join Us to "Move to End Parkinson's Disease": A World Federation of Neurology and International Parkinson and Movement Disorders Society Collaboration. Can J Neurol Sci. 2020;48(2021):56–8.
4. Chen CLH, Rundek T. Vascular brain health. Stroke. 2021;52:3700–5.
5. Cao J, Hou J, Ping J, Cai D. Advances in developing novel therapeutic strategies for Alzheimer's disease. Mol Neurodegener. 2018;13:64.
6. Erkkinen MG, Kim M-O, Geschwind MD. Clinical Neurology and Epidemiology of the Major Neurodegenerative Diseases. Cold Spring Harbor Perspect Biol. 2018;10(4):a033118.
7. Wang W, Han R, Zhang M, Wang Y, Wang T, et al. A network-based method for brain disease gene prediction by integrating brain connectome and molecular network. Brief Bioinform. 2022;23:bbab459.
8. Zhao T, Hu Y, Zang T, Wang Y. Integrate GWAS, eQTL, and mQTL data to identify Alzheimer's disease-related genes. Front Genet. 2019;10:1021.
9. Ciaranello RD, Ciaranello AL. Genetics of major psychiatric disorders. Annu Rev Med. 1991;42:151–8.
10. Liu B, Jiang T, Ma S, Zhao H, Li J, et al. Exploring candidate genes for human brain diseases from a brain-specific gene network. Biochem Biophys Res Commun. 2006;349:1308–14.
11. Masters CL, Beyreuther K. Alzheimer's disease: a clearer definition of the genetic components. Med J Aust. 1994;160:243–4.
12. Pavlidis P, Noble WS. Analysis of strain and regional variation in gene expression in mouse brain. Genome Biol. 2001;2:research0042.0041.
13. Quadri M, Mandemakers W, Grochowska MM, Masius R, Geut H, et al. LRP10 genetic variants in familial Parkinson's disease and dementia with Lewy bodies: a genome-wide linkage and sequencing study. Lancet Neurol. 2018;17:597–608.
14. Fratiglioni L, Launer LJ, Andersen K, Breteler MM, Copeland JR, et al. Incidence of dementia and major subtypes in Europe: a collaborative study of population-based cohorts. Neurologic diseases in the elderly research group. Neurology. 2000;54:S10–15.

15. Veturi Y, Lucas A, Bradford Y, Hui D, Dudek S, et al. A unified framework identifies new links between plasma lipids and diseases from electronic medical records across large-scale cohorts. Nat Genet. 2021;53:972–81.
16. Feng Y-CA, Cho K, Lindstrom S, Kraft P, Cormack J, et al. Investigating the genetic relationship between Alzheimer's disease and cancer using GWAS summary statistics. Hum Genet. 2017;136:1341–51.
17. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011;43:1066–73.
18. Anttila V, Bulik-Sullivan B, Finucane HK, Walters RK, Bras J, et al. Analysis of shared heritability in common disorders of the brain. Science. 2018;360:eaap8757.
19. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12:56–68.
20. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460:748–52.
21. Erten S, Bebek G, Koyutürk M. Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. J Comput Biol. 2011;18:1561–74.
22. Hormozdiari F, Penn O, Borenstein E, Eichler EE. The discovery of integrated gene networks for autism and related disorders. Genome Res. 2015;25:142–54.
23. Jiang R. Walking on multiple disease-gene networks to prioritize candidate genes. J Mol Cell Biol. 2015;7:214–30.
24. Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet. 2008;82:949–58.
25. Li Y, Patra JC. Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. Bioinformatics. 2010;26:1219–24.
26. Nitsch D, Gonçalves JP, Ojeda F, de Moor B, Moreau Y. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. BMC Bioinformatics. 2010;11:460.
27. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol. 2010;6:e1000641.
28. Gerring ZF, Gamazon ER, Derks EM. C. for the Major Depressive Disorder Working Group of the Psychiatric Genomics, A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression. PLoS Genet. 2019;15:e1008245.
29. Gao J, Li P, Chen Z, Zhang J. A survey on deep learning for multimodal data fusion. Neural Comput. 2020;32:829–64.
30. Zhang Y-D, Dong Z, Wang S-H, Yu X, Yao X, et al. Advances in multimodal data fusion in neuroimaging: overview, challenges, and novel orientation. Information Fusion. 2020;64:149–87.
31. Saha A, Kim Y, Gewirtz ADH, Jo B, Gao C, et al. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. Genome Res. 2017;27:1843–58.
32. Richiardi J, Altmann A, Milazzo A-C, Chang C, Chakravarty MM, et al. Correlated gene expression supports synchronous activity in brain networks. Science. 2015;348:1241–4.
33. Chen K, Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet. 2007;8:93–103.
34. Fang L, Li Y, Ma L, Xu Q, Tan F, et al. GRNdb: decoding the gene regulatory networks in diverse human and mouse conditions. Nucleic Acids Res. 2020;49:D97–103.
35. Kulkarni SR, Vaneechoutte D, Van de Velde J, Vandepoele K. TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information. Nucleic Acids Res. 2017;46:e31–e31.
36. Hu H, Miao Y-R, Jia L-H, Yu Q-Y, Zhang Q, et al. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. 2018;47:D33–8.
37. Guo Z-H, You Z-H, Huang D-S, Yi H-C, Chen Z-H, et al. A learning based framework for diverse biomolecule relationship prediction in molecular association network. Commun Biol. 2020;3:118.
38. Li G, Zhang P, Sun W, Ren C, Wang L. Bridging-BPs: a novel approach to predict potential drug–target interactions based on a bridging heterogeneous graph and BPs2vec. Brief Bioinform. 2022;23:bbab557.
39. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, et al. The human disease network. Proc Natl Acad Sci. 2007;104:8685–90.
40. Yang J, Wu S-J, Yang S-Y, Peng J-W, Wang S-N, et al. DNetDB: the human disease network database based on dysfunctional regulation mechanism. BMC Syst Biol. 2016;10:36.
41. Kawahara J, Brown CJ, Miller SP, Booth BG, Chau V, et al. BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. Neuroimage. 2017;146:1038–49.
42. Li X, Zhou Y, Dvornek N, Zhang M, Gao S, et al. BrainGNN: interpretable brain graph neural network for fMRI analysis. Med Image Anal. 2021;74:102233.
43. Tang H, Guo L, Fu X, Qu B, Ajilore O, et al. A hierarchical graph learning model for brain network regression analysis. Front Neurosci. 2022;16:963082.
44. Wein S, Malloni WM, Tomé AM, Frank SM, Henze GI, et al. A graph neural network framework for causal inference in brain networks. Sci Rep. 2021;11:8061.
45. Yue X, Wang Z, Huang J, Parthasarathy S, Moosavinasab S, et al. Graph embedding on biomedical networks: methods, applications and evaluations. Bioinformatics. 2019;36:1241–51.
46. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. 2016. arXiv preprint arXiv:1609.02907.
47. Bi X-A, Li L, Wang Z, Wang Y, Luo X, et al. IHGC-GAN: influence hypergraph convolutional generative adversarial network for risk prediction of late mild cognitive impairment based on imaging genetic data. Brief Bioinform. 2022;23:bbac093.
48. Bi X-A, Zhou W, Luo S, Mao Y, Hu X, et al. Feature aggregation graph convolutional network based on imaging genetic data for diagnosis and pathogeny identification of Alzheimer's disease. Brief Bioinform. 2022;23:bbac137.
49. Shan X, Cao J, Huo S, Chen L, Sarrigiannis PG, et al. Spatial–temporal graph convolutional network for Alzheimer classification based on brain functional connectivity imaging of electroencephalogram. Hum Brain Mapp. 2022;43:5194–209.
50. Wen G, Cao P, Bao H, Yang W, Zheng T, et al. MVS-GCN: A prior brain structure learning-guided multi-view graph convolution network for autism spectrum disorder diagnosis. Comput Biol Med. 2022;142:105239.
51. Piñero J, Saüch J, Sanz F, Furlong LI. The DisGeNET cytoscape app: exploring and visualizing disease genomics data, computational and structural. Biotechnol J. 2021;19:2960–7.
52. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012;489:391–9.
53. Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348:648–60.
54. Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TEJ, et al. The human connectome project: a data acquisition perspective. Neuroimage. 2012;62:2222–31.
55. Ji JL, Spronk M, Kulkarni K, Repovš G, Anticevic A, et al. Mapping the human brain's cortical-subcortical functional network organization. Neuroimage. 2019;185:35–57.
56. Bruna J, Zaremba W, Szlam A, LeCun Y. Spectral networks and locally connected networks on graphs. 2013. arXiv preprint arXiv:1312.6203.
57. Su X, Hu L, You Z, Hu P, Wang L, et al. A deep learning method for repurposing antiviral drugs against new viruses via multi-view nonnegative matrix factorization and its application to SARS-CoV-2. Brief Bioinform. 2022;23:bbab526.
58. Wang L, Wong L, Li Z, Huang Y, Su X, et al. A machine learning framework based on multi-source feature fusion for circRNA-disease association prediction. Brief Bioinform. 2022;23:bbac388.
59. Wong L, Wang L, You Z-H, Yuan C-A, Huang Y-A, et al. GKLOMLI: a link prediction model for inferring miRNA–lncRNA interactions by using Gaussian kernel-based method on network profile and linear optimization algorithm. BMC Bioinform. 2023;24:188.
60. Zhang H-Y, Wang L, You Z-H, Hu L, Zhao B-W, et al. iGRLCDA: identifying circRNA–disease association based on graph representation learning. Brief Bioinform. 2022;23:bbac083.
61. Zheng K, Zhang X-L, Wang L, You Z-H, Ji B-Y, et al. SPRDA: a link prediction approach based on the structural perturbation to infer disease-associated Piwi-interacting RNAs. Brief Bioinform. 2023;24:bbac498.

62. Ding Y, Tian L-P, Lei X, Liao B, Wu F-X. Variational graph auto-encoders for miRNA-disease association prediction. Methods. 2021;192:25–34.
63. Huang Y-A, Hu P, Chan KCC, You Z-H. Graph convolution for predicting associations between miRNA and drug resistance. Bioinformatics. 2019;36:851–8.
64. Hammond DK, Vandergheynst P, Gribonval R. Wavelets on graphs via spectral graph theory. Appl Comput Harmon Anal. 2011;30:129–50.
65. Lee PH, Anttila V, Won H, Feng Y-CA, Rosenthal J, et al. Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. Cell. 2019;179:1469–1482.e1411.
66. Liu J, Zhang C, Zhao Y, Yue X, Wu H, et al. Parkin targets HIF-1α for ubiquitination and degradation to inhibit breast tumor progression. Nat Commun. 2017;8:1823.
67. Pietzner M, Wheeler E, Carrasco-Zanini J, Cortes A, Koprulu M, et al. Mapping the proteo-genomic convergence of human diseases. Science. 2021;374:eabj1541.
68. Wightman DP, Jansen IE, Savage JE, Shadrin AA, Bahrami S, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. Nat Genet. 2021;53:1276–82.
69. Jabbar MA, Deekshatulu BL, Chandra P. Graph Based Approach for Heart Disease Prediction, in. New York, NY: Springer New York; 2013. p. 465–74.
70. Ata S K, Wu M, Fang Y, et al. Recent advances in network-based methods for disease gene prediction. Brief Bioinformatics. 2021;22(4):bbaa303.
71. Xie M, Xu Y, Zhang Y, Hwang T, Kuang R. Network-based phenome-genome association prediction by bi-random walk. PLoS ONE. 2015;10:e0125138.
72. Xu J, Cai L, Liao B, Zhu W, Wang P, et al. Identifying potential miRNAs–disease associations with probability matrix factorization. Front Genet. 2019;10:1234.
73. Guo Z-H, You Z-H, Huang D-S, Yi H-C, Zheng K, et al. MeSHHeading2vec: a new method for representing MeSH headings as vectors based on graph embedding algorithm. Brief Bioinform. 2020;22:2085–95.
74. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.
75. Li X, Li Y, Han H, Miller DW, Wang G. Solution structures of human LL-37 fragments and NMR-based identification of a minimal membrane-targeting antimicrobial and anticancer region. J Am Chem Soc. 2006;128:5776–85.
76. Wang G. Structures of human host defense cathelicidin LL-37 and its smallest antimicrobial peptide KR-12 in lipid micelles. J Biol Chem. 2008;283:32637–43.
77. Smith RG, Pishva E, Shireby G, Smith AR, Roubroeks JAY, et al. A meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights novel differentially methylated loci across cortex. Nat Commun. 2021;12:3517.
78. Nalls MA, Blauwendraat C, Vallerga CL, Heilbron K, Bandres-Ciga S, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 2019;18:1091–102.
79. Vilela J, Asif M, Marques AR, Santos JX, Rasga C, et al. Biomedical knowledge graph embeddings for personalized medicine: predicting disease-gene associations. Expert Syst. 2023;40:e13181.
80. Cinaglia P, Cannataro M. Identifying candidate gene-disease associations via graph neural networks. Entropy. 2023;25:909.
81. Suratanee A, Plaimas K. Gene association classification for autism spectrum disorder: leveraging gene embedding and differential gene expression profiles to identify disease-related genes. Appl Sci. 2023;13:8980.
82. Bousquet O. Transductive learning: Motivation, models, algorithms, in: University of New Mexico. 2002.

Publisher's Note