NeuroImage 292 (2024) 120594

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

BPI-GNN: Interpretable brain network-based psychiatric diagnosis and subtyping

Kaizhong Zheng [a], Shujian Yu [b,c,*], Liangjun Chen [a], Lujuan Dang [a], Badong Chen [a,*]

[a] National Key Laboratory of Human–Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China
[b] Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
[c] Machine Learning Group, UiT - Arctic University of Norway, Tromsø, Norway

ARTICLE INFO

Dataset link: http://fcon_1000.projects.nitrc.org/indi/abide, http://rfmri.org/REST-meta-MDD, https://bicr-resource.atr.jp/srpbsfc

Keywords:
Psychiatric diagnosis
Psychiatric subtypes
Graph neural network (GNN)
Prototype learning
Brain network
Functional magnetic resonance imaging (fMRI)

ABSTRACT

Converging evidence increasingly suggests that psychiatric disorders, such as major depressive disorder (MDD) and autism spectrum disorder (ASD), are not unitary diseases, but rather heterogeneous syndromes that involve diverse, co-occurring symptoms and divergent responses to treatment. This clinical heterogeneity has hindered the progress of precision diagnosis and treatment effectiveness in psychiatric disorders. In this study, we propose BPI-GNN, a new interpretable graph neural network (GNN) framework for analyzing functional magnetic resonance images (fMRI), by leveraging the famed prototype learning. In addition, we introduce a novel generation process of prototype subgraph to discover essential edges of distinct prototypes and employ total correlation (TC) to ensure the independence of distinct prototype subgraph patterns. BPI-GNN can effectively discriminate psychiatric patients and healthy controls (HC), and identify biological meaningful subtypes of psychiatric disorders. We evaluate the performance of BPI-GNN against 11 popular brain network classification methods on three psychiatric datasets and observe that our BPI-GNN always achieves the highest diagnosis accuracy. More importantly, we examine differences in clinical symptom profiles and gene expression profiles among identified subtypes and observe that our identified brain-based subtypes have the clinical relevance. It also discovers the subtype biomarkers that align with current neuro-scientific knowledge.

1. Introduction

Psychiatric disorders are one of the leading causes of extensive social and economic burden for healthcare systems worldwide (Wittchen et al., 2011) and severely compromise the well-being of those affected (Hyman, 2008). Despite decades of research, unified or definitive biomarkers still remain uncertain in psychiatry (Goodkind et al., 2015). One possible cause is that current psychiatric diagnosis is mainly based on clinical symptoms and signs rather than the underlying biological mechanisms. For example, patients are diagnosed by major depressive disorder (MDD) when they exhibit at least five of nine clinical symptoms (such as depressed mood, anhedonia and cognitive impairments, etc.) (Drysdale et al., 2017), which leads to the high clinical heterogeneity among patients with the same diagnosis (Jacobi et al., 2004). Due to such clinical heterogeneity, researchers fail to obtain reliable biomarkers through traditional case-control studies (all patients with a same diagnosis compared to healthy controls) (Hawco et al., 2019). More importantly, it has hindered the progress of treatment effectiveness and outcome in psychiatric disorders (Wu et al., 2020).

To address this problem, the Research Domain Criteria (RDoC) initiative has been released (Insel et al., 2010) and "precision medicine for psychiatry" project has launched. The core idea of them is to identify subtypes of psychiatric disorders based on the underlying biological and cognitive measurements without relying solely on traditional symptom-based diagnosis (Insel and Cuthbert, 2015). So far, several studies have begun leveraging resting-state functional magnetic resonance imaging (fMRI) (a particularly useful modality) to investigate the biologically meaningful subtypes of psychiatric disorders (Clementz et al., 2016; Drysdale et al., 2017). fMRI is a noninvasive neuroimaging technique (Matthews and Jezzard, 2004), which easily quantifies functional connectivity (FC) computed by the pairwise correlations of fMRI time series as features to investigate the neurobiology and psychiatric subtypes in diverse patient populations. Most of the existing neuroimaging studies investigating psychiatric subtypes use ensemble hybrid frameworks that include feature selection (e.g., canonical correlation analysis (CCA) Hardoon et al., 2004 and AutoEncoder Hinton

[*] Corresponding authors.
E-mail addresses: kzzheng@stu.xjtu.edu.cn (K. Zheng), yusj9011@gmail.com (S. Yu), liangjunchen@xjtu.edu.cn (L. Chen), danglj@xjtu.edu.cn (L. Dang), chenbd@mail.xjtu.edu.cn (B. Chen).

https://doi.org/10.1016/j.neuroimage.2024.120594
Received 3 November 2023; Received in revised form 24 March 2024; Accepted 27 March 2024
Available online 1 April 2024

and Salakhutdinov, 2006, etc.) and unsupervised approaches (e.g., hierarchical clustering Nielsen and Nielsen, 2016 and k-means clustering Hartigan et al., 1979). Specifically, researchers firstly use feature selection approaches to obtain low-dimensional representations or a relatively small number of FCs and then adopt unsupervised learning methods to these low-dimensional biological features to identify subtypes of psychiatric disorders (Chang et al., 2021). However, current existing two-stage subtype approaches for psychiatric disorder easily cause suboptimal solutions because it is difficult to guarantee that the feature selection and unsupervised learning methods used are optimal and most suitable (Chang et al., 2021). Furthermore, due to the lack of ground truth of downstream task, these frameworks could obtain inconsistent results and unreliable or even inaccurate predictions such as inconsistent numbers of subtypes (Feczko et al., 2019).

Recently, graph neural networks (GNNs) (Hamilton et al., 2017) have gained increased attention in the domain of psychiatric diagnosis, due to their powerful ability of graph representation. In these studies, researchers regard brain as a graph, with nodes defined as brain regions of interest (ROIs) and edges defined as FC between these ROIs. Despite a tremendous improvement in performance (Li et al., 2021b; Cui et al., 2022), most of existing diagnostic models are trained on a dataset from a homogeneous or single site with a small sample sizes (less than 100), which could lead to over-fitting and spurious performance. Moreover, recent advances in exploring neurological biomarkers at both the group-level (Cui et al., 2022) and individual-level (Li et al., 2021b) have shown promise for psychiatric disorders. However, integrating these biomarkers into clinical practice remains challenging due to the unpredictable onset and high clinical heterogeneity in psychiatric disorders (Jacobi et al., 2004). Translating these biomarkers into practical clinical tools requires innovative GNN architectures capable of providing subtype-level explanations, which can offer insights into the biological and clinical heterogeneity inherent in psychiatric disorders.

Prototype learning is a type of case-based reasoning (Kolodner, 1992; Schmidt et al., 2001) and facilitates predictions for new instances by comparing them to a set of learned exemplar cases. So far, the concept of prototype learning has been integrated into image recognition (Fig. 1) to enhance interpretability, enabling the provision of subtype-level explanations. For example, ProtoPNet (Chen et al., 2019; Rymarczyk et al., 2020) utilizes a fusion of prototype learning and convolutional neural networks (CNNs) to acquire prototypical parts within a specific class and produce intuitive image explanations. Nevertheless, there are currently no compelling precedents for the application of prototype learning to graph classification task or the brain network analysis.

To address above technical issues, we develop a novel GNN architecture (as shown in Fig. 2) for psychiatric diagnosis and subtyping which is able to discriminate between psychiatric patients and healthy controls, and obtain biologically meaningful subtypes of psychiatric disorders. More importantly, to further validate the clinical relevance of our identified brain-based subtypes, we investigate differences in clinical symptom, brain pattern and gene expression profiles among identified subtypes and associations between clinical profiles, gene expression profiles, and dominant brain connections in each identified subtype. We term our architecture the **B**rain **P**rototype **I**nterpretable **G**raph **N**eural **N**etwork (BPI-GNN[1]) and evaluate it on three real-world, large-scale datasets of brain disease.

To summarize, our main contributions are fourfold:

- We first propose a new GNN architecture for both psychiatric diagnosis and subtyping. In other words, BPI-GNN not only demonstrates the capability to distinguish between psychiatric patients and healthy controls, but it also identifies biologically meaningful subtypes of psychiatric disorders.
- In terms of methodology, our BPI-GNN addresses above technical issues:
    1. Our framework design facilitates model interpretability by incorporating prototype learning to graph classification tasks and brain network analysis, which provides subtype-level explanations.
    2. We introduce a novel prototype subgraph generation process tailored for brain network analysis, enabling the identification of informative edges for distinct prototypes. Unlike many explainable brain network model, which often prioritize node selection, such as BrainGNN (Li et al., 2021b), our approach recognizes the paramount importance of edges (i.e., functional connectivities) in psychiatric diagnosis (Wang et al., 2021a).
    3. We employ the total correlation (TC), a measure used to assess redundancy or dependency among a set of multiple random variables, to ensure the independence of distinct prototype subgraph patterns.
- In terms of psychiatric diagnosis, we use BPI-GNN against 11 state-of-the-art (SOTA) baselines on three multi-site, large-scale psychiatric datasets, i.e., SRPBS dataset, ABIDE and REST-meta-MDD. Our model achieves the overwhelming classification performance in all datasets.
- In terms of psychiatric subtyping, we obtain biologically meaningful subtypes in patients with autism spectrum disorder (ASD), major depressive disorder (MDD) and schizophrenia (SZ) which are in part consistent with previous clinical and neuroimaging findings. More importantly, we examine clinical symptom profiles and gene expression profile differences among identified subtypes and observe that our identified brain-based subtypes have the clinical relevance.

2. Background knowledge

2.1. GNNs (graph neural networks) for psychiatric diagnosis

2.1.1. Graph neural networks

Graph neural networks (GNNs) leverage the message-passing mechanism to efficiently propagate and aggregate information along the edges of an input graph, enabling the acquisition of expressive node representations (Hamilton et al., 2017; Welling and Kipf, 2016; Xu et al., 2018). The GNN architecture consists of L layers, each comprising three fundamental steps. (1) First, at lth GNN layer, a message $m_{ij}^{l} = \text{Message}\left(h_i^{l-1}, h_j^{l-1}\right)$ is computed for each edge (i, j), where h_i^{l-1} and h_j^{l-1} correspond to the representations of nodes i and j in the previous layer, respectively. (2) Second, for each node i, GNN aggregates the received messages from its neighborhood $\mathcal{N}(i)$, using an aggregation function $m_i^l = \text{Aggregation}\left(\left\{m_{ij}^l | j \in \mathcal{N}(i)\right\}\right)$. (3) Finally, GNN updates the vector representation of each node i by applying the function $h_i^l = \text{Update}\left(m_i^l, h_i^{l-1}\right)$, which takes the aggregated message and the current node representation as inputs. The final node embedding, denoted as $z_i = h_i^L$, is derived from the hidden representation obtained from the last layer of GNN. After obtaining node embedding, GNN adopts READOUT function to learn the representation of the entire graph $h_G = \text{READOUT}\left(\left\{z_i | i \in G\right\}\right)$, where h_G is the representation of graph G.

In this study, we use sum-pooling (Xu et al., 2018) as READOUT function to learn graph embedding $h_G = \text{SUM}\left(\left\{z_i | i \in G\right\}\right)$.

2.1.2. GNN interpretability

Although GNNs have shown remarkable effectiveness, they are black-box models that lack interpretability, making it difficult to understand the underlying mechanisms behind their predictions. So far, there has been a surge of interest in explaining the predictions of

[1] https://github.com/ZKZ-Brain/BPI-GNN.

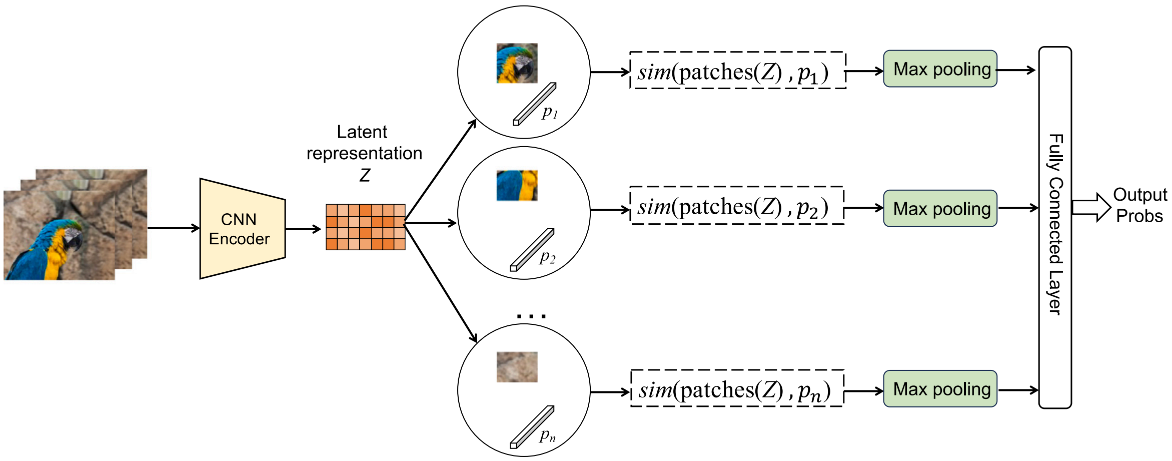


Fig. 1. Traditional prototype learning for image recognition. given an input image I, framework extracts image representation $Z = f(I)$ using convolutional neural networks (CNNs) f and learns n prototype vectors $P = \{p_i\}_{i=1}^{n}$. Subsequently, framework calculates the distance between the jth prototype p_j and all patches of Z, which are then inverted to obtain similarity scores $sim(Z, p_j) = \max_{z \in \text{patches}(Z)} \log\left(\frac{\|z-p_j\|_2^2+1}{\|z-p_j\|_2^2+\epsilon}\right)$, where ϵ is set to a small value e.g., 1e-4. Then these similarity scores are followed by global max pooling to result in a single similarity score. Finally, n similarity scores are sent to the fully connected layer to produce the output probabilities.

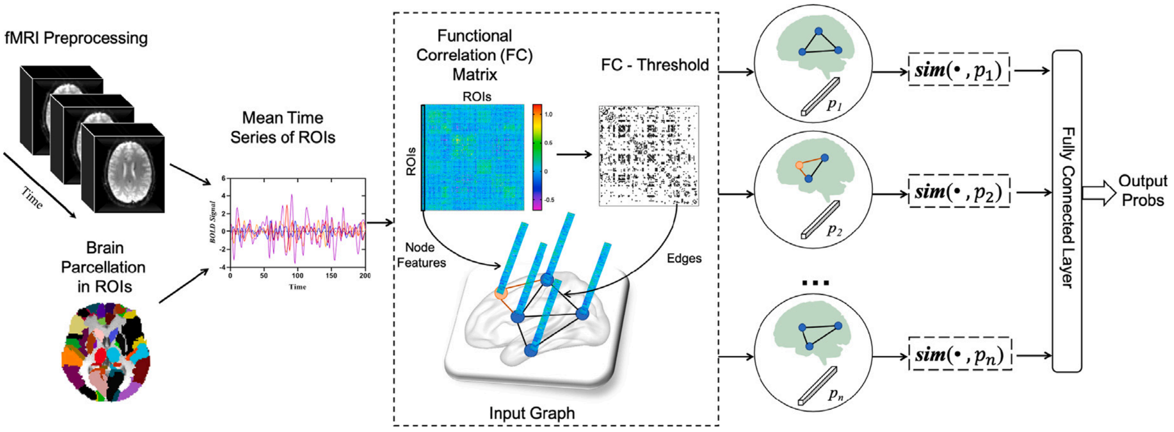


Fig. 2. Prototype learning for brain network analysis (e.g., BPI-GNN). The resting-state fMRI data undergo preprocessing and are subsequently partitioned into regions of interest (ROIs) using an atlas. Functional connectivity (FC) matrices are then generated through Pearson correlation between ROIs. These FC matrices are utilized to construct brain functional graphs (Gallo et al., 2023). Then the BPI-GNN generates a set of prototype subgraphs and learns n prototype vectors $P = \{p_i\}_{i=1}^{n}$. Subsequently, BPI-GNN calculates the similarity between the jth prototype p_j and the jth prototype subgraphs. Finally, n similarity scores are sent to the fully connected layer to produce the output probabilities.

GNNs (Yuan et al., 2020). The perturbation-based method (Ying et al., 2019) is currently the most widely adopted approach, which employs distinct mask generators to identify crucial subgraph structures and features. Subsequently, these mask generators are evaluated and optimized based on the performance of the subgraphs on a well-trained GNN.

However, most existing approaches are *post-hoc*, requiring the creation of a separate interpretive model to explain the well-trained GNN (Zhang et al., 2022). In addition, such explanations are generally unreliable, inaccurate, and can potentially mislead the entire model decision process (Rudin, 2018). To address these issues, researchers have proposed *built-in* interpretable models that generate explanations directly from the models themselves without the post-training of an auxiliary network. For example, ProtGNN (Zhang et al., 2022) combines with prototype learning and GNNs to provide inherent interpretability. In this study, BrainProtGNN is also a *built-in* interpretable GNN which could provide edge explanation across psychiatric subtypes.

2.1.3. GNN interpretability for psychiatric diagnosis

Recently, GNNs have been applied in the field of psychiatric diagnosis. A recent study (Li et al., 2021b) introduces BrainGNN, incorporating ROI-aware graph convolutional layers to pinpoint pivotal regions of interest (ROIs) in autism diagnosis. Additionally, researchers propose IBGNN (Cui et al., 2022) to discriminate between individuals with bipolar disorders (BD) and healthy controls, utilizing a global explanation mask to accentuate disorder-specific biomarkers. In our recent research, we introduce BrainIB (Zheng et al., 2022) where we harness the renowned Information Bottleneck (IB) principle to pinpoint the most informative edges in the context of psychiatric diagnosis.

However, current existing models only provide group-level and individual-level biomarkers which cannot take into account clinical heterogeneity, making it challenging for them to have a meaningful impact in real-world clinical applications. In this study, BPI-GNN could provide subtype-level explanations which offer insights into the biological and clinical heterogeneity inherent in psychiatric disorders.

2.2. Prototype learning

Prototype learning, a type of case-based reasoning (Kolodner, 1992; Schmidt et al., 2001), facilitates predictions for new instances by comparing them to a set of learned exemplar cases, known as prototypes. Prior researches (Chen et al., 2019; Rymarczyk et al., 2020), exemplified by ProtoPNet, utilizes a fusion of prototype learning and convolutional neural networks (CNNs) to acquire prototypical parts within a specific class and produce intuitive image explanations. ProtoPNet consists of a regular CNN f, prototype layer g_P and the fully connected layer. Specifically, given an input image I, ProtoPNet extracts image representation $Z = f(I)$ using f and learn k prototype vectors $P = \{p_i\}_{i=1}^{k}$ for each class. Subsequently, the jth prototype unit g_{p_j} in the prototype layer g_P calculates the distances between the jth prototype p_j and all patches of Z, which are then inverted to obtain similarity scores using the following equation:

$$g_{p_j}(z) = \max_{z \in \text{patches}(Z)} \log\left(\frac{\left\|z - p_j\right\|_2^2 + 1}{\left\|z - p_j\right\|_2^2 + \epsilon}\right), \tag{1}$$

where ϵ is set to a small value e.g., 1e-4.

Thus, an activation map is generated by each prototype unit g_{p_j}, containing similarity scores that reflect the strength of a prototypical part within the image. The activation map produced by each prototype unit g_{p_j} is subsequently subjected to global max pooling, resulting in a single similarity score. Finally, k similarity scores produced by the prototype layer g_P are sent to the fully connected layer with softmax function to produce the output probabilities for each class.

However, so far prototype learning is not yet explored for explaining GNNs and brain network analysis. In this study, we leverage prototype learning to identify prototypes (i.e., subtypes) of psychiatric disorders, which facilitates understanding clinical heterogeneity within psychiatric populations.

Note that, the general idea of Prototype Learning has recently been extended to GNNs (Zhang et al., 2022). However, the ProtGNN exhibits certain limitations, and our method distinguishes itself from it. First, our BPI-GNN enforces constraints on the independence of distinct prototype subgraph patterns, ensuring strict adherence to theoretical principles. This is an aspect absent in the ProtGNN, which could potentially result in similar subgraph patterns between prototypes. Second, our method utilizes a novel generation process of prototype subgraphs, eliminating the need for an auxiliary neural network and thereby reducing uncertainty induced by such process. Finally, The prototype subgraph generation process of ProtGNN suffers from an excessive number of training parameters and spatial allocation issues, making it applicable only for small-scale graph datasets with dozens of nodes. In contrast, our approach can be applied to brain network datasets comprising hundreds of nodes.

2.3. *Total correlation*

Securing the trustworthiness and validity of the gleaned subtype-level interpretations presents a formidable challenge within the confines of this model framework. Aligned with the overarching model structure of this study, this complexity is streamlined to assure the independence among prototype subgraph patterns. For information theory, minimizing total correlation (TC) (Watanabe, 1960) among a set of multiple random variables is a common approach to ensure their independence, which can be easily computed without any auxiliary neural network. Thus, we are inspired to consider using correlation (TC) to ensure independence among prototype subgraph patterns. Specifically, given the L-dimensional components of the random variable $Z = \{Z^1; Z^2; \ldots; Z^L\}$, TC can be defined as the Kullback–Leibler divergence from the joint distribution $\Pr(Z^1, Z^2, \ldots, Z^L)$ to the independent distribution of $\prod_{i=1}^{L} \Pr(Z^i)$:

$$\begin{aligned} TC(Z) &= D_{KL}\left(\Pr\left(Z^1, Z^2, \ldots, Z^L\right) \parallel \prod_{i=1}^{L} \Pr\left(Z^i\right)\right), \\ &= \left[\sum_{i=1}^{L} H\left(Z^i\right)\right] - H\left(Z^1, Z^2, \ldots, Z^L\right), \end{aligned} \tag{2}$$

where $H(Z^i)$ is the information entropy of variable Z^i and $H(Z^1, Z^2, \ldots, Z^L)$ denotes the joint entropy of the variable set $\{Z^1; Z^2; \ldots; Z^L\}$.

Table 1
Notations used in the paper.

Notations	Description
N	number of participants
n	number of nodes
k	number of prototypes
$\mathcal{V}$	node set
$\mathcal{E}$	edge set
$\mathcal{G}$	graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
A	graph adjacency matrix representing the graph structure ($A \in \{0,1\}^{n\times n}$)
X	node feature matrix ($X \in \mathbb{R}^{n\times n}$)
ρ	Pearson's correlation coefficient
ϕ	GNN Encoder of GraphVAE
θ_1	Decoder of GraphVAE
θ_2	Prototype subgraph sampling module
f	Prototype layer
φ	Basic classifier
Z	Node embedding through ϕ
$z^{\{i\}}$	Disentangled factor
p	Learned prototype vectors
h	Prototype graph embedding
TC	Total correlation
H	Entropy and joint entropy
K	Gram matrix
$\tilde{K}$	Normalized gram matrix
e_{ij}	Edge selection probability
z_i	Node embedding of node i

3. Methods

3.1. *Framework of BPI-GNN*

3.1.1. *Notations*

Fig. 2 depicts the pipeline for constructing the brain functional graph from rs-fMRI raw data. Initially, the resting-state fMRI data are subjected to preprocessing procedures, followed by parcellation of the brain into n regions of interest (ROIs) based on the automated anatomical labeling (AAL) atlas. Subsequently, the mean time series of each ROI is computed from the preprocessed fMRI data, and functional connectivity (FC) matrices are obtained by calculating the Pearson correlation between the mean time series of the ROIs. Based on FC, we define an undirected graph $\mathcal{G} = (A, X)$, where A denotes the graph adjacency matrix representing the graph structure ($A \in \{0,1\}^{n\times n}$) and X denotes the node feature matrix. Specifically, A is a binarized FC matrix, where only the top 20-percentile absolute values of the correlations are transformed into ones, while the rest are set to zeros. For node feature X, X_r for node r is defined as $X_r = \left[\rho_{r1}, \ldots, \rho_{rn}\right]^{\mathrm{T}}$, where ρ_{rl} is the Pearson's correlation coefficient for node r and node l. It is noteworthy that in this study, only functional connectivity values are considered as node features, which is a common practice in brain network analysis (Gallo et al., 2023). All the notations are listed in the Table 1.

3.1.2. *Overall workflow of BPI-GNN*

The Fig. 3 depicts the workflow of BPI-GNN, which comprises four critical components: a multi-head graph variational autoencoder (GraphVAE), a prototype subgraph generator, a prototype layer f and a basic classifier φ. We employ a two-step training strategy to jointly optimize generative performance and diagnostic accuracy.

In the stage I training, BPI-GNN is able to learn prototype subgraph embeddings. Specifically, given an input graph $\mathcal{G}$, the modified GraphVAE is responsible for learning latent factors $Z = \left[z^{\{1\}}, z^{\{2\}}, \ldots, z^{\{k\}}\right]$, where k represents a pre-determined number of prototypes. Using Z, the decoder is able to reconstruct graph feature matrix X and graph adjacency matrix A with separate heads. Afterward, another linear decoder generates the prototype subgraph $\mathcal{G}_{\text{sub}}^{k}$, which is fed into a graph encoder ϕ to obtain the prototype subgraph embedding h_k.

In the stage II training, BPI-GNN is able to learn prototype vectors which can be understood as the latent representation of different

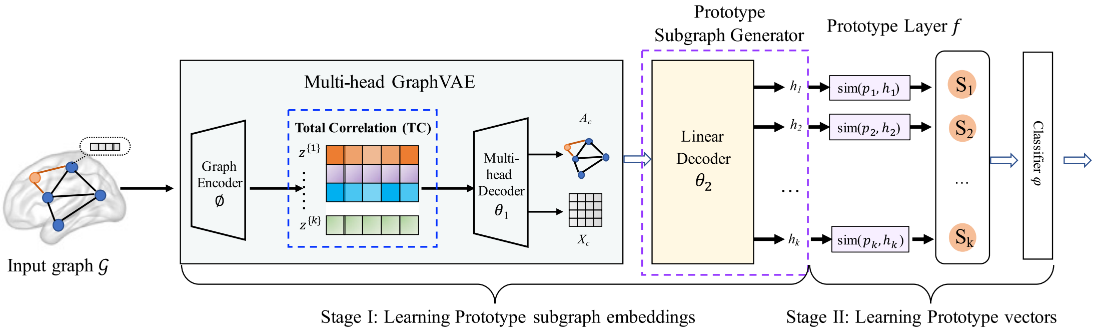


Fig. 3. The overall architecture of our proposed BPI-GNN. The model consists of four modules: multi-head GraphVAE, prototype subgraph generator, prototype layer and a basic classifier φ. The training procedures of BPI-GNN include two stages. In the stage I training, given an input graph $\mathcal{G} = \{A, X\}$, multi-head GraphVAE learns (disentangled) latent factors $Z = \left[z^{\{1\}}, z^{\{2\}}, \ldots, z^{\{k\}}\right]$. Another linear decoder generates the prototype subgraph $\mathcal{G}^k_{\text{sub}}$, which is fed to the graph encoder ϕ to obtain the prototype subgraph embedding h_k. In the stage II training, the prototype layer calculates the similarity scores between the prototype embeddings and according prototype vectors. These similarity scores are then used by the basic classifier φ to compute the output probabilities, enabling graph classification.

subtypes within psychiatric populations. In the prototype layer, the network learns k prototype vectors $P = \{p_i\}_{i=1}^{k}$. For each prototype vector, its shape is equal to the dimensions of prototype subgraph embedding h_k. Subsequently, the prototype layer computes the similarity scores between the prototype subgraph embeddings and according prototype vectors. For kth prototype subgraph embedding h_k and prototype vector p_k, the similarity score is defined as:

$$sim\left(p_k, h_k\right) = \log\left(\frac{\left\|p_k - h_k\right\|_2^2 + 1}{\left\|p_k - h_k\right\|_2^2 + \epsilon}\right), \tag{3}$$

where ϵ is a small value (1e-4) added to prevent division by zero. Finally, the basic classifier φ computes output probabilities using k similarity scores.

3.1.3. Stage I: Learning prototype subgraph embedding

Given an input graph $\mathcal{G} = (A, X)$ with n nodes, where $A \in \mathbb{R}^{n\times n}$ is the adjacency matrix and $X \in \mathbb{R}^{n\times n}$ is the node feature matrix, we employ GraphVAE (Simonovsky and Komodakis, 2018) like architecture to learn k disentangled factors $Z = \left[z^{\{1\}}, z^{\{2\}}, \ldots, z^{\{k\}}\right]$, where $z^{\{k\}} \in \mathbb{R}^{\tilde{d}\times n}$, $\tilde{d} = d(l)/k$ and $d(l)$ is the dimension of lth hidden layer. The graph encoder of our modified GraphVAE is a basic GCN, where the output Z of the lth layer can be computed as:

$$Z^l = \sigma\left(\tilde{A}Z^{l-1}W^{l-1}\right), \tag{4}$$

where $\tilde{A}$ is the normalized adjacency matrix, $\tilde{A} = A + I$ with I being the identity matrix, D is a diagonal matrix of the degree of nodes, and σ is the sigmoid activation function.

In the decoder of our modified GraphVAE, we use separate heads: a multi-layer perceptron (MLP) to reconstruct X, and a linear inner product decoder to recover A. Specifically, we formulate the reconstruction procedure as:

$$A_c = \sigma\left(ZZ^T\right), X_c = \text{MLP}(Z), \tag{5}$$

where A_c is reconstructed adjacency matrix, X_c is reconstructed node features and Z is the output of the last layer of the graph encoder.

The objective of our multi-head GraphVAE is to minimize the reconstruction error and maximize the compression of the latent variable Z. The objective is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{GraphVAE}} = \mathbb{E}\left[\left\|X - X_c\right\|_F\right] + \mathbb{E}\left[\left\|A - A_c\right\|_F\right] \\ - \mathbb{E}\left[D_{KL}\left[q(Z|A, X) \parallel p(Z)\right]\right],\end{aligned} \tag{6}$$

where $\|\|_F$ is the Frobenius norm, $q(Z|A, X)$ is the graph encoder model, and $p(Z)$ is an isotropic Gaussian prior distribution for Z.

In addition, to ensure the independence among latent factors $\{z^{\{1\}}, z^{\{2\}}, \ldots, z^{\{k\}}\}$, we resort to a total correlation (TC) term:

$$\begin{aligned}TC(Z) &= D_{KL}\left(\Pr\left(z^{\{1\}}, z^{\{2\}}, \ldots, z^{\{k\}}\right) \parallel \prod_{i=1}^{k}\Pr\left(z^{\{i\}}\right)\right), \\ &= \left[\sum_{i=1}^{k} H\left(z^{\{i\}}\right)\right] - H\left(z^{\{1\}}, z^{\{2\}}, \ldots, z^{\{k\}}\right),\end{aligned} \tag{7}$$

where H denotes entropy and joint entropy. If all latent vectors are independent, TC will be zero.

Here, we employ a matrix-based R'enyi's α-order entropy functional (Giraldo et al., 2014; Yu et al., 2019) to estimate the various entropy terms in Eq. (7). This newly proposed estimator can be computed easily without requiring density estimation or any auxiliary neural network, and it is differentiable, making it well-suited for deep learning applications. More details refer to Appendix.

Next, we further leverage another linear inner product decoder θ_2 (Li et al., 2021a) and graph encoder ϕ to generate prototype subgraph embedding h_k. Fig. 4 demonstrates the procedure of prototype subgraph generator. Specifically, given $z^{\{k\}} \in \mathbb{R}^{n\times\tilde{d}}$, we employ another linear inner product decoder θ_2 (Li et al., 2021a) to generate prototype subgraph $\mathcal{G}^k_{sub}$:

$$\mathcal{G}^k_{sub} = \sigma\left(z^{\{k\}}{z^{\{k\}}}^T\right), \tag{8}$$

where σ is the sigmoid function.

Then the trained prototype subgraph $\mathcal{G}^k_{sub}$ is fed into the graph encoder ϕ to output the prototype subgraph embedding h_k.

Furthermore, we further leverage one regularization term $\mathcal{L}_{\text{mask}}$ to encourage the compactness of the explanation and the discreteness of $\mathcal{G}_{sub}$:

$$\mathcal{L}_{\text{mask}} = \sum M - (M\log(M) + (1 - M)\log(1 - M)). \tag{9}$$

Thus, the overall loss of stage I is defined as:

$$\mathcal{L}_1 = \mathcal{L}_{\text{GraphVAE}} + \lambda_1 TC + \lambda_2\mathcal{L}_{\text{mask}}, \tag{10}$$

where λ_1 and λ_2 are hyper-parameters.

3.1.4. Stage II: Learning prototype vectors and completing the classification task

In the stage II training, BPI-GNN could learn prototype vectors which can be understood as the latent representation of different subtypes within psychiatric populations and provide a more comprehensive explanation of the reasoning process. Specifically, we compute similarity scores between the corresponding prototype vector and prototype subgraph embedding in the prototype layer. To determine

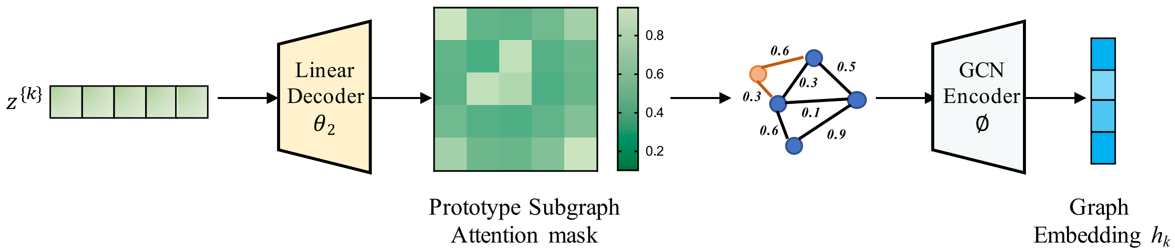


Fig. 4. The procedure of prototype subgraph generator. Given $z^{\{k\}} \in \mathbb{R}^{n\times \hat{d}}$, we employ another linear decoder θ_2 to generate Prototype Subgraph Attention mask. Then the trained Prototype Subgraph Attention mask is fed into the graph encoder ϕ to output the prototype subgraph embedding h_k.

which prototype subgraph is most similar with each prototype, the optimization objective is defined as:

$$\mathcal{L}_{\text{sim}} = \max_{\mathcal{G}_{\text{sub}}} \sum_{i=1}^{k} sim\left(p_k, \phi\left(\mathcal{G}_{\text{sub}}\right)\right), \tag{11}$$

where p_k is the kth learned prototype vector with the same dimension as the prototype subgraph embedding h_k, $\mathcal{G}_{\text{sub}}$ is the selected prototype subgraph and $sim(\cdot)$ represents similarity scores.

Finally, k similarity scores produced by the prototype layer f are sent to the basic classifier φ to produce the output probabilities for each class, where φ is the fully connected layer with softmax function. In summary, we define the optimization objective for stage II as follows:

$$\mathcal{L}_2 = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{CE}(\varphi \circ f \circ \theta_2 \circ \phi(\mathcal{G}_i), Y_i) + \lambda_3 \mathcal{L}_{\text{sim}}, \tag{12}$$

where $\mathcal{L}_{CE}$ represents cross entropy loss and λ_3 is hyper-parameter.

3.1.5. *Training procedures*

In summary, the training procedures of BPI-GNN is shown in Algorithm 1. We first perform GraphVAE and prototype subgraph generator to learn prototype subgraph embeddings by optimizing Eq. (10). After convergence of the stage I training, indicated by a reconstruction error below a predetermined threshold or exceeding a set number of training epochs, we learn prototype vectors and obtain prediction for classification. This is followed by the optimization of Eq. (12).

Algorithm 1: Training pipeline of BPI-GNN

```
Input: Training graphs $\mathcal{G}_{train} = \{\mathcal{G}_i, y_i\}$, Initialize $\{\phi, \theta_1, f, \varphi\}$,
       Initialize prototype vectors $P = \{p_i\}_{i=1}^{k}$
Output: The trained $\{\phi, \theta_1, f, \varphi\}$, The trained prototype vectors
        $P$, Prediction $\tilde{Y}$
while not stop criteria or converge do
    Perform $\phi, \theta_1$ to generate $\{z^{\{1\}}, z^{\{2\}}, \dots, z^{\{k\}}\}$;
    Compute total correlation (TC) in Eq. (7);
    Perform linear decoder $\theta_2$ to generate $\mathcal{G}_{\text{sub}}$;
    Learn prototype subgraph embedding set $H = \{h_i\}_{i=1}^{k}$;
    Update $\phi, \theta_1$ by minimizing Eq. (10).
end
while not stop criteria or converge do
    Compute similarity scores between $P$ and $H$ via $f$;
    Perform $\varphi$ to generate prediction $\tilde{Y}$ using similarity scores;
    Update $f$, $\varphi$ and $P$ by minimizing Eq. (12).
end
return $\tilde{Y}$, $\mathcal{G}_{\text{sub}}$, trained $P$, trained $\{\phi, \theta_1, f, \varphi\}$
```

3.2. *Subtype analysis from BPI-GNN*

In addition to performing psychiatric classification tasks, BPI-GNN also identifies different subtypes i.e., prototypes. Initially, we determine the number of prototypes by selecting the best performance setting, with k selected from the set $\{2, 3, 4\}$. Subsequently, the subtype of each individual is determined based on the similarity score. Specifically, if the similarity score between the prototype vector p_i and the prototype subgraph embedding h_i of an individual is higher than all other subtypes, we assign that individual to the ith subtype. To delineate the characteristics of different subtypes, we utilize two-sample t tests to investigate the differences in clinical profiles across subtypes. We consider statistical significance at $p < 0.05$ with false discovery rate (FDR) correction for multiple comparisons.

3.3. *Interpretation from BPI-GNN*

To assess the interpretability of BPI-GNN, we conducted additional analyses to investigate the ability of prototype subgraphs to interpret the neural mechanisms underlying different subtypes. BPI-GNN could capture the important prototype subgraph structure in each subject. To compare the differences between the subgraphs of different subtypes, we calculate the average prototype subgraph and select the top 50 edges to generate the dominant prototype subgraph.

3.4. *Association analysis between clinical profiles and brain connections*

We further investigate the relationship between subtype-differentiated clinical profiles and subtype-differentiated brain connectivity using the non-parametric Spearman correlation method. Our findings are reported with statistical significance at a threshold of $p < 0.05$ with false discovery rate (FDR) correction for multiple comparisons.

4. Experiments

4.1. *Datasets and data preprocessing*

In this study, three psychiatric datasets are used: the Autism Brain Imaging Data Exchange I (ABIDE) (Di Martino et al., 2014), Rest-meta-MDD (Yan et al., 2019) and Japanese strategic research program for the promotion of brain science (SRPBS) (Tanaka et al., 2021). ABIDE is a retrospective multicenter neuroimaging consortium for autism spectrum disorders (ASD) and openly shares more than 1000 resting-state fMRI data collected from 17 different international centers.[2] In this study, a total of 528 patients with ASD and 536 typically developed (TD) individuals are used. Rest-meta-MDD is the largest resting-state fMRI database for major depressive disorder (MDD) to date collected from 25 cohorts in China.[3] According to the exclusion criteria, 1604 subjects including 828 patients with MDD and 776 healthy controls are used. SRPBS is a multi-disorder MRI dataset including 1410 participants collected at 11 sites.[4] In the current study, we use 184 participants including 92 patients with schizophrenia and 92 healthy controls. We demonstrate the demographic and clinical characteristics of three psychiatric datasets in Table 2.

[2] http://fcon_1000.projects.nitrc.org/indi/abide/
[3] http://rfmri.org/REST-meta-MDD/
[4] https://bicr-resource.atr.jp/srpbsfc/

Table 2
Demographic and clinical characteristics.

Characteristic	ABIDE		Rest-meta-MDD		SRPBS	
	ASD	TD	MDD	HC	Schizophrenia	HC
Sample Size	528	536	828	776	92	92
Age	17.0 ± 8.4	17.2 ± 7.6	34.3 ± 11.5	34.4 ± 13.0	39.6 ± 10.4	38.0 ± 12.4
Sex (M/F)	464/64	471/65	301/527	318/458	47/45	60/32

ABIDE, Rest-meta-MDD and SRPBS are followed by an uniform standard preprocessing pipeline using the Statistical Parametric Mapping (SPM),[5] Data Processing Assistant for Resting-State fMRI (DPARSF)[6] and Graph Theoretical Network Analysis (GRETNA),[7] respectively. The standard preprocessing pipeline comprises multiple steps. The initial 10 volumes are discarded, and slice timing and head motion are corrected. The deformation parameters obtained from registering the fMRI images to the Montreal Neurological Institute (MNI) template are utilized to standardize the resting-state fMRI data into a common space. Furthermore, a Gaussian filter with a half maximum width of 6 mm is employed to smooth the functional images. Subsequently, a temporal band-pass filter with a range of 0.01–0.08 Hz is applied to the resulting fMRI images. Finally, the effects of head motion, white matter, cerebrospinal fluid signals, and linear trends are removed. Here, we adopt the Friston 24-parameter model to regress out head motion effects.

After preprocessing, we extract mean time courses of cortical and subcortical regions from all datasets using the automated anatomical labeling (AAL) atlas. The atlas defines 116 regions in total, including 90 cerebrum regions and 26 cerebellum regions. Functional connectivity (Fisher's r-to-z transformed Pearson's correlation) between all brain regions are estimated and the resulting 116×116 symmetric functional connectivity matrix are used to generate brain functional graph.

4.2. Control of site differences and covariates

After generating the functional connectivity matrix, we utilize the ComBat harmonization method (Johnson et al., 2007; Fortin et al., 2018; Yu et al., 2018) to account for site differences and covariates in functional connectivity. This approach allows us to retain biological variability while eliminating the variation introduced by site. The fMRI data dare assumed to come from m completely different multi-sites, with a total of n participants. The Combat model for each functional connectivity can be expressed as follows:

$$\mathrm{FC}_{ij} = \mathrm{const} + X_{ij}^{T}\beta + \gamma_i + \delta_i \epsilon_{ij}, \tag{13}$$

where FC_{ij} is defined as the FC value for the participant j at site i, const is the average FC value across all subjects from all sites, X is a design matrix for the covariates of interest ($p \times n$, p is the number of covariates), β is the vector of coefficients associated with X, γ_i and δ_i are the additive and multiplicative site effects of site i and we further assume that the residual terms ϵ_{ij} follows a normal distribution with mean zero. Here, the covariates include sex, age and head motion in the ABIDE and SRPBS datasets, while the covariates include sex, age, education and head motion in the REST-meta-MDD dataset. For the sex indicator, we set 0 to represent male and 1 to represent female.

Subsequently, the site effect parameters γ_i and δ_i are estimated using the Empirical Bayes. Thus, the final ComBat-harmonized functional connectivity is defined as:

$$\mathrm{FC}_{ij}^{ComBat} = \frac{\mathrm{FC}_{ij} - \widehat{\mathrm{const}} - X_{ij}\hat{\beta} - \gamma_i^{*}}{\delta_i^{*}} + \widehat{\mathrm{const}} + X_{ij}\hat{\beta}, \tag{14}$$

in which $\widehat{\mathrm{const}}$ is the estimated average FC values, γ_i^{*} and δ_i^{*} represent the estimated site effect parameters. More details are described in previous studies (Johnson et al., 2007; Fortin et al., 2018; Yu et al., 2018).

Table 3
Range of hyper-parameters and final specification for BPI-GNN.

Hyper-parameter	Range examined	Final specification
#GNN Layers	[2,3,4,5]	2
#Hidden Dimensions	[64,128,256,512]	128
Learning Rate	[1e−2,1e−3,1e-4]	1e−3
Batch Size	[32,64]	32
Weight Decay	[1e−3,1e−4]	1e−4
λ_1	[1e−4,5e−4,1e−3,5e−3,1e-2]	1e−3
λ_2	[1e−4,5e−4,1e−3,5e−3,1e-2]	1e−4
λ_3	[1e−4,5e−4,1e−3,5e−3,1e-2]	1e-3

4.3. Baselines

To demonstrate the effectiveness and superiority of BPI-GNN, we evaluate its performance against 11 popular traditional machine learning (ML) and deep learning (DL) models on three psychiatric datasets: ABIDE, REST-metaMDD, and SRPBS. The selected competitors include four traditional psychiatric classifiers (i.e., SVM Jakkula, 2006 with linear and RBF kernel, random forest (RF) Rigatti, 2017, and LASSO Roth, 2004), three representative graph neural networks (GNNs; i.e., GCN Welling and Kipf, 2016, GAT Veličković et al., 2017, and GIN Xu et al., 2018), and two state-of-the-art (SOTA) *built-in* interpretable GNNs (i.e., SIB Yu et al., 2021 and ProtGNN Zhang et al., 2022). We also include two SOTA GNNs specifically designed for brain networks: BrainGNN (Li et al., 2021b) and BrainIB (Zheng et al., 2022). Here, we use random data splitting strategy (train/validation/test sets is 80%, 10% and 10% data) to assess performance of BPI-GNN and baselines.

4.4. Experimental setup

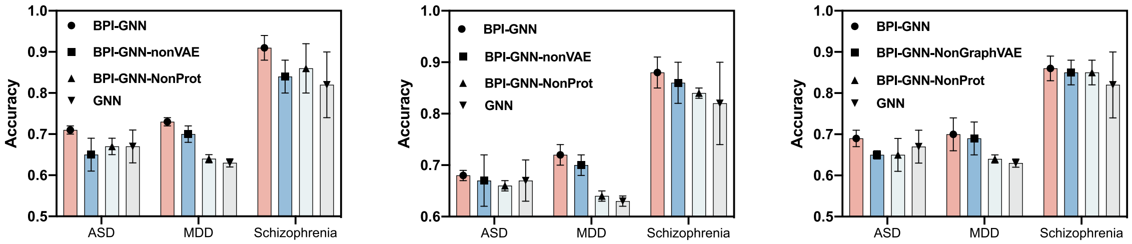
We trained and tested the BPI-GNN with PyTorch 1.12.1 (Paszke et al., 2019) and PyTorch Geometric 2.1.0 (Fey and Lenssen, 2019). During the training, the number of epoch is set to 350 and dropout ratio is set to 0.5. Table 3 shows the range of hyper-parameters that are examined and the final specification of all hyper-parameters are used to obtain the final results. The hyper-parameters are set through a grid search or based on the recommended settings of related work.

For the baselines, We train each model with 350 epochs. For traditional psychiatric classifiers including SVM with linear and RBF kernel, RF, and LASSO, we first perform two-sample t tests between patient group and HC group with FC network to obtain abnormal FC connections. Then these connections are concatenated as a long feature vector, and are sent to classifiers. For GIN, GAT and GCN, we use the recommended hyperparameters of related work to train the models. For SIB and BrainIB, the weight β of the mutual information term $I\left(\mathcal{G}, \mathcal{G}_{sub}\right)$ is selected from $\{0.0001, 0.1\}$. For ProtGNN, the hyperparameters λ_1, λ_2 and λ_3 are set to 0.10, 0.05, and 0.01, respectively, according to recommended setting.

4.5. Hyperparameter discussion and ablation study

To investigate the impact of the number of prototypes k on performance, we conduct a hyperparameter tuning with $k \in \{2, 3, 4\}$ using the train and validation sets. Additionally, we perform an ablation study to assess the potential contributions of various components within BPI-GNN. Specifically, we compare the classification accuracy of four variants of BPI-GNN, namely, the original model, BPI-GNN-NonVAE

[5] https://www.fil.ion.ucl.ac.uk/spm/software/spm12/
[6] http://rfmri.org/DPARSF/
[7] https://www.nitrc.org/projects/gretna



(a) $k = 2$ (b) $k = 3$ (c) $k = 4$

Fig. 5. Ablation study and the influence of prototypes number k on performance on three psychiatric datasets including ABIDE, REST-meta-MDD and SRPBS.

(BPI-GNN without modified GraphVAE), BPI-GNN-NonProt (BPI-GNN without prototype subgraph sampling module and prototype layer), and GNN (BPI-GNN without modified GraphVAE, prototype subgraph sampling module and prototype layer). It is worth noting that BPI-GNN-NonVAE omits the decoder of modified GraphVAE and the optimization of Eq. (6). The results of the ablation study are illustrated in Fig. 5.

In accordance with Fig. 5, it is evident that the value of k has an impact on the performance of BPI-GNN. Based on the optimal performance, we set the value of k to 2 for ABIDE dataset, 2 for REST-meta-MDD dataset, and 2 for SRPBS dataset. For ablation study, we observe that BPI-GNN outperforms BPI-GNN-NonProt on all datasets, indicating that the ability to identify distinct prototypes contributes to the improved performance. Moreover, the superior performance of BPI-GNN over BPI-GNN-NonVAE suggests that the GraphVAE is effective and crucial in the model.

5. Results

5.1. Evaluation on classification performance

Table 4 demonstrates the classification performances in terms of Accuracy, F1-score and Matthew's Correlation Coefficient (MCC) on three psychiatric datasets (i.e., ABIDE, REST-meta-MDD and SRPBS). Each model is independently run five times, and the mean and standard deviation of the metrics are reported.

Extensive experiments demonstrate that BPI-GNN outperforms all baseline models in terms of all evaluation metrics on all datasets, indicating that BPI-GNN has substantial advantages for brain network analysis. Furthermore, the performance of BPI-GNN demonstrates a remarkable superiority over the alternative methods (two sample t tests, $p < 0.05$) on the REST-meta-MDD dataset. The improvement in performance of BPI-GNN can be attributed to three factors. Firstly, BPI-GNN is a *built-in* interpretable deep learning model that eliminates the need for feature selection. Secondly, BPI-GNN leverages the prototype mechanism to better understand the underlying characteristics of psychiatric subtypes. Finally, as a graph neural network, BPI-GNN can effectively handle the topological and non-linear information within complex brain network structures, which gives it an edge over traditional psychiatric classifiers.

5.2. Interpretable analysis

5.2.1. Interpretation in ABIDE

Table 5 demonstrate demographic and clinical data for each subtype on ABIDE. Significantly attenuated ADOS_RRB in subtype1 compared with subtype2 is observed in patients with ASD.

Fig. 6 illustrates the dominant prototype subgraph G_{dsub} comparison between healthy controls and patient groups in ABIDE. In this visualization, the color of each node represents a distinct brain network, while the size of each edge reflects its weight in the dominant subgraph. The ROI nodes defined in each dataset are mapped onto nine commonly used brain networks, including the visual network (VN), somatomotor network (SMN), dorsal attention network (DAN), ventral attention network (VAN), limbic network (LIN), frontoparietal network (FPN), default mode network (DMN), cerebellum (CBL), and subcortical network (SBN).

Common ("shared") brain connections within G_{dsub} are observed in both ASD subtypes, including connectivity within SMN, SBN, LIN, CBL, and VN, as well as connectivity between DMN and FPN. Furthermore, subtype-specific patterns of G_{dsub} are also observed. Specifically, subtype1 of ASD exhibits tight interactions within FPN, involving the right orbital frontal lobe and bilateral inferior parietal gyrus, whereas these connections are absent in subtype2. In addition, connections within and between DAN (bilateral supramarginal) in subtype1 are less than that of subtype2.

5.2.2. Interpretation in REST-meta-MDD

Table 6 demonstrate demographic and clinical data for each subtype on REST-meta-MDD. Total scores for HAMD do not exhibit significant differences between subtypes in patients with MDD, suggesting that the differentiation is not based on the severity of illness.

Next, we conduct further analysis using two-sample t tests to investigate the differences in scores for each item of the HAMD between subtypes (see Fig. 7). Our results indicate significant differences in three symptom measures, namely suicide, retardation, and general somatic.

Fig. 8 illustrates the dominant prototype subgraph G_{dsub} comparison between healthy controls and MDD groups in REST-meta-MDD dataset. There are common ('shared') brain connections within G_{dsub} between both subtypes, which involves connections within VN, FPN SMN, CBL, FPN and etc. Patterns within SBN (connections between bilateral pallidum) of subtype1 are significantly more than that of subtype2, while subtype2 shows more connections with DAN (connections between bilateral superior parietal gyrus).

5.2.3. Interpretation in SRPBS

Table 7 demonstrate demographic and clinical data for each subtype on SRPBS dataset. Total scores PANSS do not exhibit significant differences between subtypes in patients with schizophrenia.

Furthermore, we use two-sample t tests to investigate the differences in scores for each item of the PANSS between subtypes (see Fig. 9). We observe two significantly different gene profiles including somatic concern and guilt feeling.

Fig. 10 illustrates the dominant prototype subgraph G_{dsub} comparison between healthy controls and schizophrenia groups in SRPBS dataset. We observe significant differences in the brain connections within G_{dsub} for both subtypes, which could be attributed to the limited sample size of this dataset. Specifically, subtype1 of schizophrenia shows tight interactions within the limbic network, particularly in the right medial superior orbital frontal lobe. In contrast, subtype2 of schizophrenia exhibits tight interactions within the left insula.

Table 4
The classification performance and standard deviations of BPI-GNN and the baselines on three psychiatric datasets. The best and second best performances are in bold and underlined, respectively.

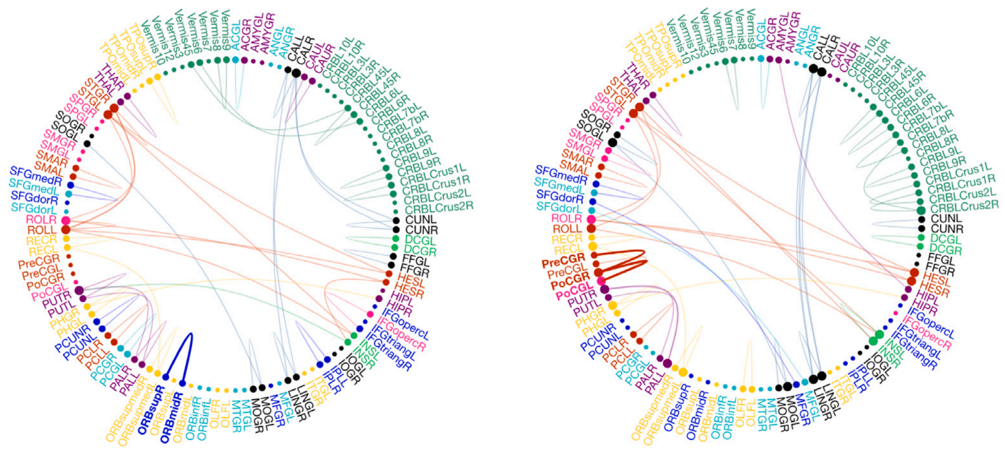
Method	ABIDE			Rest-meta-MDD			SRPBS		
	Accuracy	F1	MCC	Accuracy	F1	MCC	Accuracy	F1	MCC
RBF-SVM	0.69 ± 0.01	0.69 ± 0.01	0.39 ± 0.02	0.66 ± 0.03	0.64 ± 0.03	0.32 ± 0.05	0.87 ± 0.02	0.89 ± 0.01	0.76 ± 0.04
Linear-SVM	0.67 ± 0.05	0.67 ± 0.05	0.34 ± 0.11	0.63 ± 0.02	0.61 ± 0.03	0.34 ± 0.14	0.87 ± 0.02	0.88 ± 0.02	0.75 ± 0.03
RF	0.64 ± 0.01	0.64 ± 0.01	0.29 ± 0.05	0.60 ± 0.02	0.56 ± 0.02	0.20 ± 0.03	0.84 ± 0.10	0.84 ± 0.09	0.67 ± 0.20
LASSO	0.65 ± 0.03	0.64 ± 0.01	0.29 ± 0.05	0.61 ± 0.02	0.59 ± 0.02	0.22 ± 0.04	0.79 ± 0.06	0.79 ± 0.07	0.58 ± 0.13
GAT	0.68 ± 0.03	0.69 ± 0.04	0.37 ± 0.07	0.63 ± 0.04	0.61 ± 0.07	0.24 ± 0.06	0.84 ± 0.11	0.84 ± 0.10	0.70 ± 0.19
GIN	0.67 ± 0.04	0.67 ± 0.04	0.37 ± 0.07	0.66 ± 0.03	0.67 ± 0.01	0.31 ± 0.06	0.82 ± 0.08	0.80 ± 0.09	0.65 ± 0.17
GCN	0.66 ± 0.06	0.65 ± 0.08	0.30 ± 0.01	0.63 ± 0.01	0.60 ± 0.05	0.26 ± 0.02	0.81 ± 0.08	0.81 ± 0.10	0.62 ± 0.17
SIB	0.65 ± 0.01	0.62 ± 0.01	0.29 ± 0.02	0.64 ± 0.04	0.65 ± 0.03	0.28 ± 0.09	0.70 ± 0.03	0.69 ± 0.04	0.43 ± 0.07
ProtGNN	0.65 ± 0.03	0.68 ± 0.02	0.29 ± 0.06	0.61 ± 0.02	0.59 ± 0.03	0.23 ± 0.04	0.74 ± 0.15	0.79 ± 0.09	0.55 ± 0.19
BrainGNN	0.67 ± 0.05	0.66 ± 0.05	0.33 ± 0.10	0.61 ± 0.03	0.59 ± 0.07	0.22 ± 0.06	0.88 ± 0.02	0.83 ± 0.07	0.73 ± 0.04
BrainIB	0.70 ± 0.01	0.71 ± 0.02	0.40 ± 0.03	0.67 ± 0.01	0.65 ± 0.02	0.35 ± 0.01	0.86 ± 0.06	0.84 ± 0.08	0.73 ± 0.13
BPI-GNN	**0.71** ± **0.01**	**0.72** ± **0.01**	**0.41** ± **0.02**	**0.73** ± **0.01**[a]	**0.72** ± **0.01**[a]	**0.51** ± **0.02**[a]	**0.91** ± **0.03**	**0.92** ± **0.01**	**0.83** ± **0.06**

[a] Denotes significantly outperforming (two sample t tests, $p < 0.05$) all the alternative methods.

Table 5
Demographic features of the two subtypes of patients with autism spectrum disorder.

	Subtype1 (N = 284)	Subtype2 (N = 244)	Statistics	P-value
Age (years)	16.4 (8.0)	17.7 (8.8)	t = −1.810	0.071
Sex, male/female	249/35	215/29	χ^2 = 0.024	0.878
FIQ	105.28 (16.58) (N = 265)	105.91 (17.24) (N = 230)	t = −0.410	0.682
VIQ	104.75 (14.82) (N = 235)	103.78 (19.24) (N = 208)	t = 0.565	0.573
PIQ	104.89 (17.36) (N = 235)	105.86 (16.81) (N = 208)	t = −0.592	0.554
ADI-R_Social	19.82 (5.04) (N = 120)	19.98 (5.75) (N = 96)	t = −0.220	0.826
ADI-R_Communication	15.88 (4.44) (N = 120)	16.10 (4.76) (N = 96)	t = −0.350	0.727
ADI-R_RRB	6.53 (4.45) (N = 120)	6.03 (2.60) (N = 96)	t = 1.448	0.149
ADOS_Social	7.80 (2.63) (N = 120)	8.03 (2.88) (N = 96)	t = −0.613	0.541
ADOS_Communication	3.72 (1.54) (N = 120)	3.56 (1.52) (N = 96)	t = 0.732	0.465
ADOS_RRB	1.98 (1.38) (N = 120)	2.50 (1.72) (N = 96)	t = −2.481	0.014

All data are shown as mean (s.d.) or ratios. FIQ, Full-scale Intelligence Quotient; VIQ, Verbal Intelligence Quotient; PIQ, Performance Intelligence Quotient; ADI-R, Autism Diagnostic Interview-Revised; ADOS, Autism Diagnostic Observation Schedule; RRB, Restricted and Repetitive Behaviors.



(a) Subtype1 in ABIDE (b) Subtype2 in ABIDE

Fig. 6. Subtype-specific brain network connections on ABIDE dataset. The colors of brain neural systems are described as: visual network (VN), somatomotor network(SMN), dorsal attention network (DAN), ventral attention network (VAN), limbic network (LIN), frontoparietal network (FPN), default mode network (DMN), cerebellum (CBL) and subcortical network (SBN), respectively.

5.2.4. Association between clinical profiles and brain network connections

We further investigate association between clinical profiles and dominant prototype subgraph on three psychiatric datasets. However, we only observe significant association in MDD subtype1 on REST-meta-MDD, while no significant association are observed in ABIDE and SRPBS datasets. Fig. 11 demonstrates significant association between FC of dominant prototype subgraph and subtype-differentiated HAMD-17 scores in MDD subtype1. As can be seen, retardation is positively correlated with FC between bilateral putamen (r = 0.316, p < 0.0001, FDR correction), FC between bilateral pallidum (r = 0.407, p < 0.0001, FDR correction) and bilateral thalamus (r = 0.316, p < 0.0001, FDR correction).

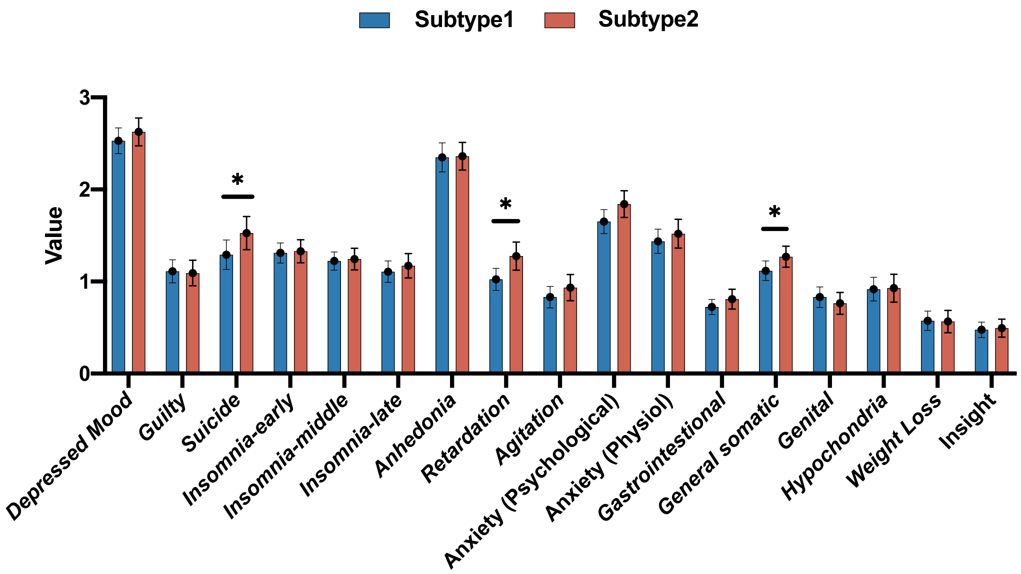
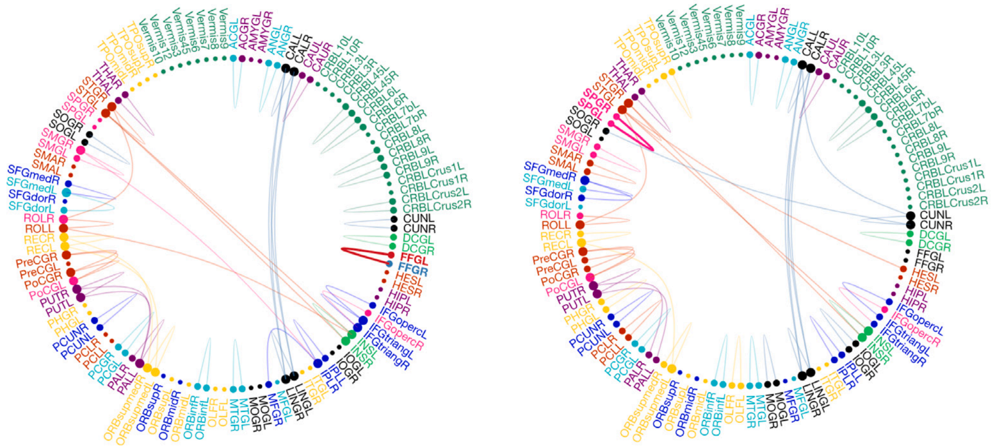


Fig. 7. Subtype-specific clinical profiles for depression symptoms (HAMD-17) that exhibit the significant variations across clusters ($P < 0.05$, two-sample t tests, false discovery rate corrected). The asterisk denotes a significant difference from the mean symptom severity rating between distinct subtypes ($P < 0.05$), and the error bars represent the standard error of the mean. $^{*}P < 0.05$.



(a) Subtype1 in REST-meta-MDD

(b) Subtype1 in REST-meta-MDD

Fig. 8. Subtype-specific brain network connections on REST-meta-MDD dataset. The colors of brain neural systems are described as: visual network (VN), somatomotor network(SMN), dorsal attention network (DAN), ventral attention network (VAN), limbic network (LIN), frontoparietal network (FPN), default mode network (DMN), cerebellum (CBL) and subcortical network (SBN), respectively.

Table 6
Demographic features of the two subtypes of patients with major depressive disorder.

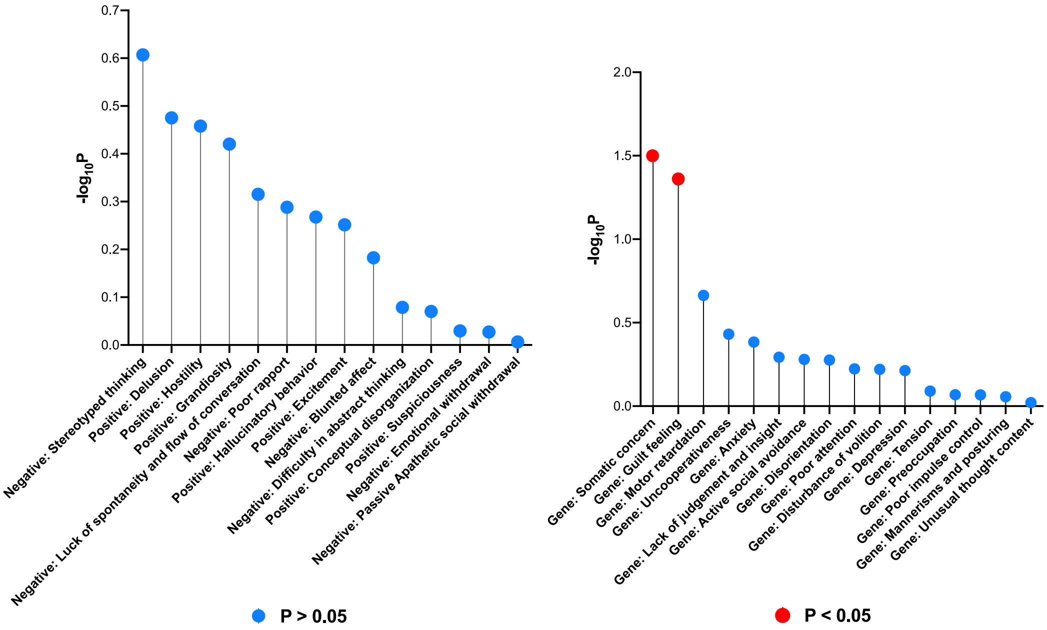
	Subtype1 (N = 456)	Subtype2 (N = 373)	Statistics	P-value
Age (years)	34.50 (11.63)	34.13 (11.28)	t = 0.464	0.643
Sex, male/female	177/279	124/249	χ^2 = 2.754	0.097
Education (years)	12.03 (3.41)	11.99 (3.37)	t = 0.161	0.872
Illness Duration (Months)	41.64 (64.03) (N = 335)	38.57 (61.48) (N = 271)	t = 0.597	0.551
HAMD	20.59 (7.23) (N = 335)	21.54 (6.12) (N = 271)	t = −1.718	0.086

All data are shown as mean (s.d.) or ratios. HAMD, Hamilton Depression Scale.

Table 7
Demographic features of the two subtypes of patients with schizophrenia.

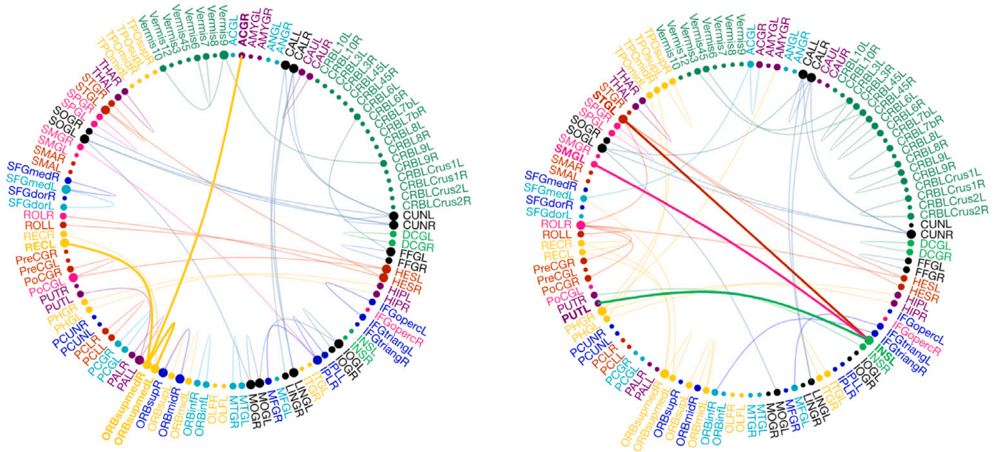
	Subtype1 (N = 59)	Subtype2 (N = 33)	Statistics	P-value
Age (years)	40.32 (10.19)	38.33 (10.48)	t = 0.879	0.382
Sex, male/female	29/30	18/15	χ^2 = 0.246	0.620
Illness Duration (years)	13.69 (9.08)	15.23 (9.84) (N = 31)	t = −0.730	0.468
PANSS_Total	58.02 (17.69) (N = 57)	59.87 (18.28) (N = 30)	t = −0.453	0.652

All data are shown as mean (s.d.) or ratios. PANSS, Positive and Negative Syndrome Scale.



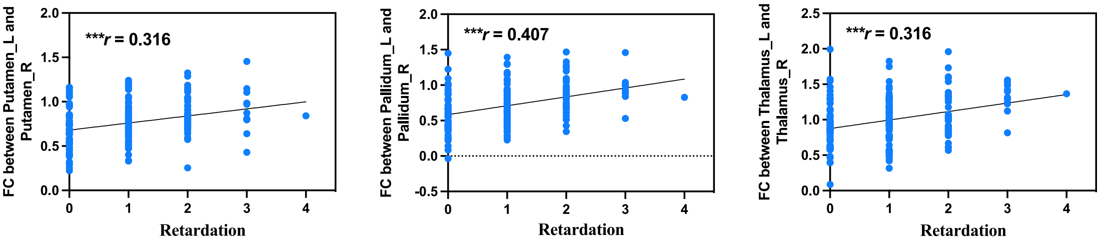
(a) PANSS symptom profiles (b) PANSS gene profiles

Fig. 9. Subtype-specific PANSS positive/negative symptom profiles and PANSS gene profiles that exhibit the significant variations across clusters (P < 0.05, two-sample t tests, false discovery rate corrected). The red circle denotes a significant difference (P < 0.05). PANSS, Positive and Negative Syndrome Scale.



(a) Subtype1 in SRPBS (b) Subtype2 in SRPBS

Fig. 10. Subtype-specific brain network connections on SRPBS dataset. The colors of brain neural systems are described as: visual network (VN), somatomotor network(SMN), dorsal attention network (DAN), ventral attention network (VAN), limbic network (LIN), frontoparietal network (FPN), default mode network (DMN), cerebellum (CBL) and subcortical network (SBN), respectively.



(a) MDD Subtype1 (b) MDD Subtype1 (c) MDD Subtype1

Fig. 11. Significant association between FC of dominant prototype subgraph and subtype-differentiated HAMD-17 scores. $^{***}P$ < 0.0001.

6. Discussion

6.1. The model

In this study, we develop a novel graph neural network architecture (BPI-GNN) for psychiatric diagnosis and subtyping. BPI-GNN includes (i) GraphVAE and Prototype layer that automatically identify prototype representations to correspond to subtypes; (ii) novel prototype subgraph generator that obtains the most informative edges in the brain of distinct subtypes; (iii) novel regularization terms (TC loss) that ensures the independence among distinct prototypes. BPI-GNN outperforms alternative machine learning methods such as SVM, LASSO, and GNN in terms of classification performance on three psychiatric datasets (i.e., ABIDE, REST-meta-MDD and SRPBS), suggesting the robustness of BPI-GNN.

BPI-GNN has the potential to address the issues of poor reproducibility and lack of interpretability in using GNN for the diagnosis of mental disorders. The issue of poor reproducibility is mainly due to the small sample sizes used and not accounting for clinical heterogeneity. Specifically, most previous diagnostic classifiers only used small sample sizes from a single center to train their models, resulting in over-fitting and poor generalization capacity during deployment. For example, despite the significant improvement in performance shown in Pitsik et al. (2023), Andreev et al. (2023), the sample sizes were limited, with only 49 healthy controls and 35 patients with MDD used in Pitsik et al. (2023), and 35 MDD patients and 50 healthy controls used in Andreev et al. (2023). Despite the use of large, multi-site sample sizes in a recent study (Gallo et al., 2023), the presence of clinical heterogeneity still results in an accuracy of only 62%. In this study, BPI-GNN uses three large, multi-site psychiatric datasets and offers insights into the biological and clinical heterogeneity inherent in psychiatric disorders.

For interpretability, although GNNs have demonstrated impressive efficacy, they inherently lack interpretability as black-box models, thus impeding their utility in disorder analysis. To address this concern, significant endeavors have been directed toward enhancing the interpretability of GNNs and their application in psychiatric diagnosis. However, most existing approaches are post-hoc (Zhang et al., 2022), requiring the creation of a separate interpretive model to explain the well-trained GNN. In addition, most of classifiers only could identify the important nodes for diagnostic diseases (Li et al., 2021b; Cui et al., 2022). Our proposed model, BPI-GNN, is a *built-in* interpretable GNN and could provide edge explanation. It is worth noting that edges (i.e. functional connectivities) play a more significant role in psychiatric diagnosis (Wang et al., 2021a).

Furthermore, this paper has the advantage of BPI-GNN being able to automatically identify psychiatric subtypes based on its performance. The importance of identifying psychiatric subtypes has been recognized for a long time, but few attempts have been made to do so. Typically, researchers use feature selection approaches to obtain low-dimensional representations or a relatively small number of features, and then adopt unsupervised learning methods on these features to identify subtypes of psychiatric disorders. However, this method faces two challenges: (1) how to ensure optimal feature selection, and (2) how to identify the number of subtypes. Our approach using prototype learning may provide a new pathway for addressing these problems.

6.2. Interpretation of our findings

In ABIDE dataset, BPI-GNN successfully identifies 2 subtypes. Subtype1 of ASD (53.8% of ASD sample) have significantly lower RRB scores of ADOS compared with that of subtype2 (46.2% of ASD sample). RRB is usually used to predict the prognosis of ASD (Troyb et al., 2016) and refers to a range of behaviors and activities that are characterized by repetition, inflexibility, invariance, inappropriateness, and lack of specific purpose (Langen et al., 2011). Our study advances the current clinical conceptualizations in autism research by establishing a linkage between functional brain features and the long-recognized heterogeneity of RRB features. Biological differences between subtypes including FPN involved in cognitive control (D'Souza et al., 2021) and DAN involved in attention to salient events (Ptak and Schnider, 2010) are associated with RRB. These results are consistent with a previous study (Guo et al., 2022), where the dynamic functional connectivity of ASD subtype2 could predict the ADOS stereotypic behavior score and ASD subtype2 exhibited higher weights for DAN. Furthermore, BPI-GNN could successfully identify two subtypes in the REST-meta-MDD dataset. Subtype1 of MDD (55% of MDD sample) have significantly different suicide, retardation, general somatic scores of HAMD compared with that of subtype2 (45% of MDD sample), which are related to depressive degree, cognitive deficits, and physiological symptoms. In addition, MDD subtype1 is characterized by patterns of SBN, while MDD subtype1 is characterized by patterns of DAN. This results is inline with a recent study (Wang et al., 2021b), in which MDD subtype1 was characterized by hyperconnectivity within the attention network, while MDD subtype2 was characterized by hypoconnectivity within the SBN. For SRPBS dataset, subtype1 shows tight interactions of orbital frontal lobe involve in affective processing (Kazama and Bachevalier, 2009), while subtype2 exhibits tight interactions of left insula. Note that, altered insula-related functions have been observed in schizophrenia, including the processing of visual and auditory emotional information, pain, and neuronal representations of the self (Wylie and Tregellas, 2010).

6.3. Limitations

Our study has some limitations that should be taken into account. First, while we discuss some variations of hyperparameters in Section 4.4, there are still many other hyperparameters that should be explored, such as the values of λ_1, the number of GNN layers, and different readout operations. Further research on these variations could enhance the effectiveness and robustness of our method. Second, we only consider FC to construct functional graph, even though it involves dynamic alternations in neural activity over time. Finally, incomplete clinical information can potentially affect the results. For instance, in ASD patients, there are 284 patients in subtype1, but only 120 of them have available ADI-R data.

7. Conclusions

In this study, we present BPI-GNN, a novel GNN framework based on prototype learning for psychiatric diagnosis and subtyping. To our knowledge, this is the first work to utilize prototype learning for psychiatric diagnosis and subtyping. BPI-GNN outperforms other state-of-the-art methods on three challenging psychiatric datasets and effectively identifies biologically meaningful subtypes and subtype-specific brain network connections. Furthermore, we also validate the rationality of our discovered subtypes with clinical and genetic profiles analysis. Results highlight the potential for this approach to contribute to the development of biologically informed diagnostic classifications and treatment guidelines for specific psychiatric cohorts.

CRediT authorship contribution statement

Kaizhong Zheng: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Shujian Yu:** Conceptualization, Methodology, Supervision, Validation, Writing – review & editing, Formal analysis. **Liangjun Chen:** Validation, Writing – review & editing. **Lujuan Dang:** Validation, Writing – review & editing. **Badong Chen:** Writing – review & editing, Conceptualization, Formal analysis, Funding acquisition, Supervision, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and code availability

Deidentified and anonymized data were contributed from studies approved by local Institutional Review Boards. All study participants provided written informed consent at their local institution. Data of the ABIDE are available at: http://fcon_1000.projects.nitrc.org/indi/abide. Data of the REST-meta-MDD project are available at: http://rfmri.org/REST-meta-MDD. Data of the SRPBS are available at: https://bicr-resource.atr.jp/srpbsfc. Codes used during the current study are available at GitHub repositor (https://github.com/ZKZ-Brain/BPI-GNN).

Acknowledgments

This work was supported by the National Natural Science Foundation of China with grant numbers (62088102, U21A20485, 62311540022) and the Arctic University of Norway under grant 309439.

Appendix. Additional details of total correlation

According to the matrix-based Rényi's α-order entropy functional, we have:

Definition 1. Let $\kappa : \chi \times \chi \mapsto \mathbb{R}$ be a real valued positive definite kernel that is also infinitely divisible (Bhatia, 2006). Given $\{\mathbf{x}_i\}_{i=1}^n \in \chi$, each $\mathbf{x}_i$ can be a real-valued scalar or vector, and the Gram matrix $K \in \mathbb{R}^{n\times n}$ computed as $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$, a matrix-based analogue to Rényi's α-entropy can be given by the following functional:

$$\begin{aligned} H_\alpha(\tilde{K}) &= \frac{1}{1-\alpha}\log_2\left(\mathrm{tr}(\tilde{K}^\alpha)\right) \\ &= \frac{1}{1-\alpha}\log_2\left(\sum_{i=1}^n \lambda_i(\tilde{K})^\alpha\right), \end{aligned} \tag{A.1}$$

where $\alpha \in (0,1)\cup(1,\infty)$. $\tilde{K}$ is the normalized K, i.e., $\tilde{K} = K/\mathrm{tr}(K)$. $\lambda_i(\tilde{K})$ denotes the ith eigenvalue of $\tilde{K}$.

Definition 2. Given a collection of n samples $\{s_i = \left(x_1^i, x_2^i, \dots, x_m^i\right)\}_{i=1}^n$, each sample contains $m\,(m \geq 2)$ measurements $x_1 \in \chi_1$, $x_2 \in \chi_2$, …, $x_m \in \chi_m$ obtained from the same realization. Given positive definite kernels $\kappa_1 : \chi_1 \times \chi_1 \mapsto \mathbb{R}$, $\kappa_2 : \chi_2 \times \chi_2 \mapsto \mathbb{R}$, …, $\kappa_m : \chi_m \times \chi_m \mapsto \mathbb{R}$, a matrix-based analogue to Rényi's α-order joint-entropy among m variables can be defined as:

$$H_\alpha(K_1, K_2, \dots, K_m) = H_\alpha\left(\frac{K_1 \circ K_2 \circ \dots \circ K_m}{\mathrm{tr}(K_1 \circ K_2 \circ \dots \circ K_m)}\right), \tag{A.2}$$

where $(K_1)_{ij} = \kappa_1(x_1^i, x_1^j)$, $(K_2)_{ij} = \kappa_2(x_2^i, x_2^j)$, …, $(K_m)_{ij} = \kappa_m(x_m^i, x_m^j)$, and $\circ$ denotes the Hadamard product.

Now, in the training of BPI-GNN, suppose we obtain $\left\{z_i^{\{1\}}, z_i^{\{2\}}, \dots, z_i^{\{k\}}\right\}_{i=1}^B$ in a mini-batch of B samples, we first need to evaluate three Gram matrices $K_{z\{1\}} = \kappa(z_i^{\{1\}}, z_j^{\{1\}}) \in \mathbb{R}^{B\times B}$, $K_{z\{2\}} = \kappa(z_i^{\{2\}}, z_j^{\{2\}}) \in \mathbb{R}^{B\times B}$, …, $K_{z\{k\}} = \kappa(z_i^{\{k\}}, z_j^{\{k\}}) \in \mathbb{R}^{B\times B}$ associated with $z^{\{1\}}$, $z^{\{2\}}$, …, $z^{\{k\}}$ respectively. Here, κ is a Gaussian kernel with kernel width σ, e.g., $\kappa(\alpha_i, \alpha_j) = \exp(-\frac{\|\alpha_i - \alpha_j\|^2}{2\sigma^2})$. For value of σ, we evaluate the 10 nearest distances of each sample and take the mean. We choose σ as the average of mean values for all samples.

Then, we normalize $K_{z\{1\}}$, $K_{z\{2\}}$, …, $K_{z\{k\}}$ by their trace to obtain $\tilde{K}_{z\{1\}}$, $\tilde{K}_{z\{2\}}$, …, $\tilde{K}_{z\{k\}}$, i.e.,

$$\tilde{K}_{z\{k\}} = K_{z\{k\}}/\mathrm{tr}(K_{z\{k\}}). \tag{A.3}$$

According to Definition 1, the entropy of variables p_k can be evaluated as (α is a hyperparameter which is set to 1.01):

$$H_\alpha(z^{\{k\}}) = \frac{1}{1-\alpha}\log_2\left(\mathrm{tr}(\tilde{K}_{z\{k\}}^\alpha)\right). \tag{A.4}$$

Meanwhile, according to Definition 2, the joint entropy term $H\left(z^{\{1\}}, z^{\{2\}}, \dots, z^{\{k\}}\right)$ can be evaluated as:

$$H_\alpha(z^{\{1\}}, z^{\{2\}}, \dots, z^{\{k\}}) = H_\alpha\left(\frac{K_{z\{1\}} \circ K_{z\{2\}} \circ \dots \circ K_{z\{k\}}}{\mathrm{tr}(K_{z\{1\}} \circ K_{z\{2\}} \circ \dots \circ K_{z\{k\}})}\right). \tag{A.5}$$

Given latent vectors $Z = \left[z^{\{1\}}, z^{\{2\}}, \dots, z^{\{k\}}\right]$, we resort to a total correlation (TC) term:

$$\begin{aligned} TC\,(Z) &= D_{KL}\left(\Pr\left(z^{\{1\}}, z^{\{2\}}, \dots, z^{\{k\}}\right) \,\|\, \prod_{i=1}^k \Pr\left(z^{\{i\}}\right)\right), \\ &= \left[\sum_{i=1}^k H\left(z^{\{i\}}\right)\right] - H\left(z^{\{1\}}, z^{\{2\}}, \dots, z^{\{k\}}\right). \end{aligned} \tag{A.6}$$

By plugging Eqs. (A.1)–(A.5) into Eq. (A.6), we obtain:

$$TC_\alpha\left(z^{\{1\}}, z^{\{2\}}, \dots, z^{\{k\}}\right) = \left[\sum_{i=1}^k H_\alpha\left(z^{\{i\}}\right)\right] - H_\alpha\left(z^{\{1\}}, z^{\{2\}}, \dots, z^{\{k\}}\right). \tag{A.7}$$

References

Andreev, A.V., Kurkin, S.A., Stoyanov, D., Badarin, A.A., Paunova, R., Hramov, A.E., 2023. Toward interpretability of machine learning methods for the classification of patients with major depressive disorder based on functional network measures. Chaos 33 (6).

Bhatia, R., 2006. Infinitely divisible matrices. Amer. Math. Monthly 113 (3), 221–235.

Chang, M., Womer, F.Y., Gong, X., Chen, X., Tang, L., Feng, R., Dong, S., Duan, J., Chen, Y., Zhang, R., et al., 2021. Identifying and validating subtypes within major psychiatric disorders based on frontal–posterior functional imbalance via deep learning. Mol. Psychiatry 26 (7), 2991–3002.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K., 2019. This looks like that: deep learning for interpretable image recognition. In: Advances in Neural Information Processing Systems, vol. 32.

Clementz, B.A., Sweeney, J.A., Hamm, J.P., Ivleva, E.I., Ethridge, L.E., Pearlson, G.D., Keshavan, M.S., Tamminga, C.A., 2016. Identification of distinct psychosis biotypes using brain-based biomarkers. Am. J. Psychiatry 173 (4), 373–384.

Cui, H., Dai, W., Zhu, Y., Li, X., He, L., Yang, C., 2022. Interpretable graph neural networks for connectome-based brain disorder analysis. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII. Springer, pp. 375–385.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry 19 (6), 659–667.

Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R.N., Zebley, B., Oathes, D.J., Etkin, A., et al., 2017. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat. Med. 23 (1), 28–38.

D'Souza, J.F., Price, N.S., Hagan, M.A., 2021. Marmosets: A promising model for probing the neural mechanisms underlying complex visual networks such as the frontal–parietal network. Brain Struct. Funct. 1–16.

Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A.M., Nigg, J.T., Fair, D.A., 2019. The heterogeneity problem: Approaches to identify psychiatric subtypes. Trends Cogn. Sci. 23 (7), 584–601.

Fey, M., Lenssen, J.E., 2019. Fast graph representation learning with PyTorch geometric. In: International Conference on Learning Representations.

Fortin, J.-P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., et al., 2018. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167, 104–120.

Gallo, S., El-Gazzar, A., Zhutovsky, P., Thomas, R.M., Javaheripour, N., Li, M., Bartova, L., Bathula, D., Dannlowski, U., Davey, C., et al., 2023. Functional connectivity signatures of major depressive disorder: machine learning analysis of two multicenter neuroimaging studies. Mol. Psychiatry 1–10.

Giraldo, L.G.S., Rao, M., Principe, J.C., 2014. Measures of entropy from data using infinitely divisible kernels. IEEE Trans. Inform. Theory 61 (1), 535–548.

Goodkind, M., Eickhoff, S.B., Oathes, D.J., Jiang, Y., Chang, A., Jones-Hagata, L.B., Ortega, B.N., Zaiko, Y.V., Roach, E.L., Korgaonkar, M.S., et al., 2015. Identification of a common neurobiological substrate for mental illness. JAMA Psychiatry 72 (4), 305–315.

Guo, X., Zhai, G., Liu, J., Cao, Y., Zhang, X., Cui, D., Gao, L., 2022. Inter-individual heterogeneity of functional brain networks in children with autism spectrum disorder. Mol. Autism 13 (1), 1–13.

Hamilton, W., Ying, Z., Leskovec, J., 2017. Inductive representation learning on large graphs. In: Advances in Neural Information Processing Systems, vol. 30.

Hardoon, D.R., Szedmak, S., Shawe-Taylor, J., 2004. Canonical correlation analysis: An overview with application to learning methods. Neural Comput. 16 (12), 2639–2664.

Hartigan, J.A., Wong, M.A., et al., 1979. A k-means clustering algorithm. Appl. Stat. 28 (1), 100–108.

Hawco, C., Buchanan, R.W., Calarco, N., Mulsant, B.H., Viviano, J.D., Dickie, E.W., Argyelan, M., Gold, J.M., Iacoboni, M., DeRosse, P., et al., 2019. Separable and replicable neural strategies during social brain function in people with and without severe mental illness. Am. J. Psychiatry 176 (7), 521–530.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. science 313 (5786), 504–507.

Hyman, S.E., 2008. A glimmer of light for neuropsychiatric disorders. Nature 455 (7215), 890.

Insel, T.R., Cuthbert, B.N., 2015. Brain disorders? precisely. Science 348 (6234), 499–500.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C., Wang, P., 2010. Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. Am. J. Psychiatry 167 (7), 748–751.

Jacobi, F., Wittchen, H.-U., Hölting, C., Höfler, M., Pfister, H., Müller, N., Lieb, R., 2004. Prevalence, co-morbidity and correlates of mental disorders in the general population: Results from the German Health Interview and Examination Survey (GHS). Psychol. Med. 34 (4), 597–611.

Jakkula, V., 2006. Tutorial on Support Vector Machine (Svm), vol. 37, (no. 2.5), School of EECS, Washington State University, p. 3.

Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8 (1), 118–127.

Kazama, A., Bachevalier, J., 2009. Selective aspiration or neurotoxic lesions of orbital frontal areas 11 and 13 spared monkeys' performance on the object discrimination reversal task. J. Neurosci. 29 (9), 2794–2804.

Kolodner, J.L., 1992. An introduction to case-based reasoning. Artif. Intell. Rev. 6 (1), 3–34.

Langen, M., Durston, S., Kas, M.J., Van Engeland, H., Staal, W.G., 2011. The neurobiology of repetitive behavior:…and men. Neurosci. Biobehav. Rev. 35 (3), 356–365.

Li, J., Shao, H., Sun, D., Wang, R., Yan, Y., Li, J., Liu, S., Tong, H., Abdelzaher, T., 2021a. Unsupervised belief representation learning in polarized networks with information-theoretic variational graph auto-encoders. In: Proceedings of ACM Conference.

Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L.H., Ventola, P., Duncan, J.S., 2021b. Braingnn: Interpretable brain graph neural network for fmri analysis. Med. Image Anal. 74, 102233.

Matthews, P.M., Jezzard, P., 2004. Functional magnetic resonance imaging. J. Neurol. Neurosurg. Psychiatry 75 (1), 6–12.

Nielsen, F., Nielsen, F., 2016. Hierarchical clustering. In: Introduction to HPC with MPI for Data Science. Springer, pp. 195–211.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32.

Pitsik, E.N., Maximenko, V.A., Kurkin, S.A., Sergeev, A.P., Stoyanov, D., Paunova, R., Kandilarova, S., Simeonova, D., Hramov, A.E., 2023. The topology of fMRI-based networks defines the performance of a graph neural network for the classification of patients with major depressive disorder. Chaos Solitons Fractals 167, 113041.

Ptak, R., Schnider, A., 2010. The dorsal attention network mediates orienting toward behaviorally relevant stimuli in spatial neglect. J. Neurosci. 30 (38), 12557–12565.

Rigatti, S.J., 2017. Random forest. J. Insurance Med. 47 (1), 31–39.

Roth, V., 2004. The generalized LASSO. IEEE Trans. Neural Netw. 15 (1), 16–28.

Rudin, C., 2018. Please stop explaining black box models for high stakes decisions. Stat 1050, 26.

Rymarczyk, D., Struski, Ł., Tabor, J., Zieliński, B., 2020. Protopshare: Prototype sharing for interpretable image classification and similarity discovery. arXiv preprint arXiv:2011.14340.

Schmidt, R., Montani, S., Bellazzi, R., Portinale, L., Gierl, L., 2001. Cased-based reasoning for medical knowledge-based systems. Int. J. Med. Inform. 64 (2–3), 355–367.

Simonovsky, M., Komodakis, N., 2018. Graphvae: Towards generation of small graphs using variational autoencoders. In: Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27. Springer, pp. 412–422.

Tanaka, S.C., Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., et al., 2021. A multi-site, multi-disorder resting-state magnetic resonance image database. Sci. Data 8 (1), 227.

Troyb, E., Knoch, K., Herlihy, L., Stevens, M.C., Chen, C.-M., Barton, M., Treadwell, K., Fein, D., 2016. Restricted and repetitive behaviors as predictors of outcome in autism spectrum disorders. J. Autism Develop. Disorders 46, 1282–1296.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y., 2017. Graph attention networks. In: International Conference on Learning Representations.

Wang, L., Lin, F.V., Cole, M., Zhang, Z., 2021a. Learning clique subgraphs in structural brain network classification with application to crystallized cognition. Neuroimage 225, 117493.

Wang, Y., Tang, S., Zhang, L., Bu, X., Lu, L., Li, H., Gao, Y., Hu, X., Kuang, W., Jia, Z., et al., 2021b. Data-driven clustering differentiates subtypes of major depressive disorder with distinct brain connectivity and symptom features. Brit. J. Psychiatry 219 (5), 606–613.

Watanabe, S., 1960. Information theoretical analysis of multivariate correlation. IBM J. Res. Develop. 4 (1), 66–82.

Welling, M., Kipf, T.N., 2016. Semi-supervised classification with graph convolutional networks. In: J. International Conference on Learning Representations. ICLR 2017.

Wittchen, H.-U., Jacobi, F., Rehm, J., Gustavsson, A., Svensson, M., Jönsson, B., Olesen, J., Allgulander, C., Alonso, J., Faravelli, C., et al., 2011. The size and burden of mental disorders and other disorders of the brain in Europe 2010. Eur. Neuropsychopharmacol. 21 (9), 655–679.

Wu, W., Zhang, Y., Jiang, J., Lucas, M.V., Fonzo, G.A., Rolle, C.E., Cooper, C., Chin-Fatt, C., Krepel, N., Cornelssen, C.A., et al., 2020. An electroencephalographic signature predicts antidepressant response in major depression. Nature Biotechnol. 38 (4), 439–447.

Wylie, K.P., Tregellas, J.R., 2010. The role of the insula in schizophrenia. Schizophrenia Res. 123 (2–3), 93–104.

Xu, K., Hu, W., Leskovec, J., Jegelka, S., 2018. How powerful are graph neural networks? In: International Conference on Learning Representations.

Yan, C.-G., Chen, X., Li, L., Castellanos, F.X., Bai, T.-J., Bo, Q.-J., Cao, J., Chen, G.-M., Chen, N.-X., Chen, W., et al., 2019. Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. Proc. Natl. Acad. Sci. 116 (18), 9078–9083.

Ying, Z., Bourgeois, D., You, J., Zitnik, M., Leskovec, J., 2019. Gnnexplainer: Generating explanations for graph neural networks. In: Advances in Neural Information Processing Systems, vol. 32.

Yu, S., Giraldo, L.G.S., Jenssen, R., Principe, J.C., 2019. Multivariate extension of matrix-based Rényi's α-order entropy functional. IEEE Trans. Pattern Anal. Mach. Intell. 42 (11), 2960–2966.

Yu, M., Linn, K.A., Cook, P.A., Phillips, M.L., McInnis, M., Fava, M., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., Sheline, Y.I., 2018. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. Hum. Brain Map. 39 (11), 4213–4227.

Yu, J., Xu, T., Rong, Y., Bian, Y., Huang, J., He, R., 2021. Recognizing predictive substructures with subgraph information bottleneck. IEEE Trans. Pattern Anal. Mach. Intell..

Yuan, H., Yu, H., Gui, S., Ji, S., 2020. Explainability in graph neural networks: A taxonomic survey. arXiv preprint arXiv:2012.15445.

Zhang, Z., Liu, Q., Wang, H., Lu, C., Lee, C., 2022. Protgnn: Towards self-explaining graph neural networks. Associ. Adv. Artif. Intell. 36 (8), 9127–9135.

Zheng, K., Yu, S., Li, B., Jenssen, R., Chen, B., 2022. Brainib: Interpretable brain network-based psychiatric diagnosis with graph information bottleneck. arXiv preprint arXiv:2205.03612.