Graph Self-supervised Learning with Application to Brain Networks Analysis

Guangqi Wen, Peng Cao*, Lingwen Liu, Jinzhu Yang, Xizhe Zhang, Fei Wang*, Osmar R. Zaiane

***Abstract*— The less training data and insufficient supervision limit the performance of the deep supervised models for brain disease diagnosis. It is significant to construct a learning framework that can capture more information in limited data and insufficient supervision. To address these issues, we focus on self-supervised learning and aim to generalize the self-supervised learning to the brain networks, which are non-Euclidean graph data. More specifically, we propose an ensemble masked graph self-supervised framework named BrainGSLs, which incorporates 1) a local topological-aware encoder that takes the partially visible nodes as input and learns these latent representations, 2) a node-edge bi-decoder that reconstructs the masked edges by the representations of both the masked and visible nodes, 3) a signal representation learning module for capturing temporal representations from BOLD signals and 4) a classifier used for the classification. We evaluate our model on three real medical clinical applications: diagnosis of Autism Spectrum Disorder (ASD), diagnosis of Bipolar Disorder (BD) and diagnosis of Major Depressive Disorder (MDD). The results suggest that the proposed self-supervised training has led to remarkable improvement and outperforms state-of-the-art methods. Moreover, our method is able to identify the biomarkers associated with the diseases, which is consistent with the previous studies. We also explore the correlation of these three diseases and find the strong association between ASD and BD. To the best of our knowledge, our work is the first attempt of applying the idea of self-supervised learning with masked autoencoder on the brain network analysis. The code is available at `https://github.com/GuangqiWen/BrainGSL`.**

***Index Terms*— graph embedding learning, graph self-supervised learning, brain networks, autism spectrum disorder (ASD)**

I. Introduction

Recent studies have shown that rs-fMRI based analysis for brain functional connectivity (FC) is effective in helping understand the pathology of brain diseases [1]–[7]. The functional brain network can be modeled as a graph where nodes denote the brain regions and the edges represent the connection strength between those regions [8]. Hence, the brain disease identification can be seen as a graph classification problem. The fundamental task of the graph classification problem is the representation learning of graph-structured data. Compared to shallow models [9], deep neural networks can learn hierarchical representations of functional brain networks and capture the complex interactions in the brain regions. Though recent studies have shown deep neural network architecture is promising on the medical imaging application [10]–[14] , these methods, particularly graph convolutional networks (GCNs), do not reveal obvious improvement on the brain disease diagnosis in the previous works [13], [14], including our recent study [15]. We hypothesize that such performance gap roots in two main limitations of deep neural networks: 1) limited training data and 2) insufficient supervision [16]–[18].

Manuscript received XX XXXX XXXX; revised XX XXXX XXXX; accepted XX XXXX XXXX. Date of publication XX XXXX XXXX; date of current XXXX X XXXX XXXX. This research was supported by the National Natural Science Foundation of China (No.62076059) and the Science Project of Liaoning province (2021-MS-105). Corresponding authors: Peng Cao and Fei Wang.

Guangqi Wen, Peng Cao, Lingwen Liu and Jinzhu Yang are with the College of Computer Science and Engineering, Northeastern University, Shenyang, China (e-mail: 2110658@stu.neu.edu.cn; caopeng@cse.neu.edu.cn; 2201839@stu.neu.edu.cn; yangjinzhu@cse.neu.edu.cn).

Xizhe Zhang is a professor in School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China (e-mail: zhangxizhe@njmu.edu.cn).

Fei Wang is with Early Intervention Unit, Department of Psychiatry, Affiliated Nanjing Brain Hospital, Nanjing Medical University, Nanjing, China (e-mail: fei.wang@yale.edu).

Osmar R. Zaiane is with the department of Computing Science at the University of Alberta, Canada (e-mail: zaiane@cs.ualberta.ca).

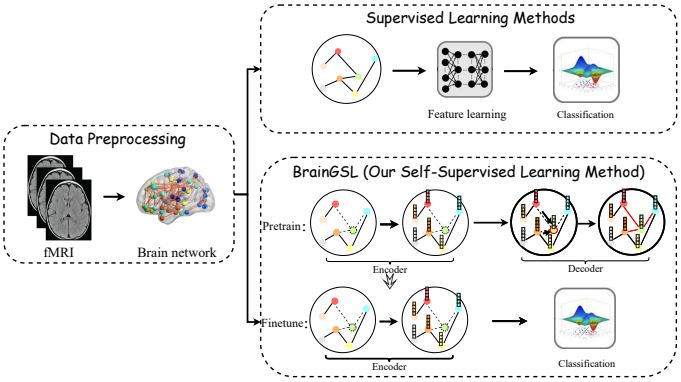


Fig. 1: Illustration of supervised learning methods and our self-supervised learning method for the graph embedding learning and classification.

TABLE I: Comparison of GMAEs [19] and our BrainGSLs for self supervised learning on the ABIDE dataset. The average graph classification performances of accuracy (ACC), AUC, sensitivity (SEN), specificity (SPEC) and parameter number of models are reported. The best results are bold.

Method	ACC(%)	AUC(%)	SEN(%)	SPEC(%)	Parameters
GMAEs [19]	55.1	56.8	30.1	**84.4**	8.05 M
BrainGSLs	**70.4**	**71.1**	**70.6**	69.6	**1.12 M**

Limited training data: The deep neural networks are data-driven and require a large amount of training data. For medical data where annotation cost is extremely high, the deep learning models are heavily restricted by the limited amount of brain disorder data. How to capture additional information from limited data is a challenge.

Insufficient supervision: Accurately diagnosing brain disorders remains a challenge since there exists heterogeneity in etiology, phenotypic for the brain disorders [20], [21]. The heterogeneous neurodevelopmental disorders are characterized by diverse deficits or impairments in behavioral features and communicative functioning. The heterogeneity of brain disorders indicates that subjects may belong to different subtypes of the same disorder, which produces diverse graph structures and data distribution. Therefore, the pure supervised learning scheme inevitably suffers from poor generalization due to the insufficient supervision.

Recently, self-supervised learning (SSL) provides a promising learning paradigm that obtains supervision from the data itself by

leveraging the underlying structure in the data, and has been widely studied in CV and NLP [22]–[28]. However, to the best of our knowledge, it has not yet been researched in the brain network analysis, where it is more urgent for the self-supervised learning to alleviate the issues of limited data and supervision. Motivated by the masked autoencoding [22], [23], it naturally comes into minds how to design a self-supervised learning with the masked autoencoding scheme on the brain network. In this study, we try to answer the question: "Is self-supervised learning able to exploit the structure information contained in a limited amount of brain networks to learn rich hidden representation?" It is non-trivial to transfer the pretext tasks designed for CV/NLP for brain network analysis. Most related to ours is the model of GMAE (Graph Masked Autoencoders with transformers) [19], which is a graph self-supervised learning method by simply applying the masked autoencoding idea on the graph data. A comparison of brain network classification results between our proposed graph self-supervised learning and GMAE on the ABIDE dataset is shown in Table I. It can be easily found that a large margin of about 15.3% and 14.3% improvements in terms of ACC and AUC, and more parameter efficiency (around 1/7 of the parameters) have been achieved. Therefore, it is necessary to take into the consideration of the topology structure in the brain network when developing a graph self-supervised learning. Two key aspects need to be solved in the self-supervised learning with masked autoencoder for the brain network data:

1. How to develop an encoder model for learning the latent representation by exploiting the inherent graph structure in the brain network?

2. How to design a reconstruction task for encouraging the encoder to capture richer representation during the graph self-supervised learning?

To address these issues, we focus on generalizing the self-supervised learning to the brain networks, which are non-Euclidean graph data. We consider the nodes (brain regions) as the tokens and propose a general graph self-supervised learning framework for brain network analysis, named BrainGSL, by leveraging the inherent graph structure for guiding both the encoding and decoding stages in the self-supervised learning, as illustrated in Fig. 1. Considering the sufficient topological information in graphs, the proposed BrainGSL consists of a topology-aware encoder for capturing the local connectivity of input graphs and modeling their latent representations, and a node-edge bi-decoder for predicting the embeddings of the masked nodes and the associated masked edges. To identify the biomarkers associated with the disorder and the connections between regions, our encoder is capable of exploiting the crucial nodes or edges via an attention mechanism. Due to the locality characteristic of the topology-aware encoder, we propose an ensemble framework, named BrainGSLs by integrating multiple random masked models to improve the generalization ability. In addition, the temporal information in rs-fMRI is also critical for understanding the underlying functional brain activities [29], [30]. Thus, we propose a signal representation learning module via the transformer block [31] to sufficiently capture the temporal information from the blood–oxygen-level-dependent (BOLD) signals in the downstream classification task. We evaluate our model on three brain diseases covering autism spectrum disorder (ASD), major depression disorder (MDD) and bipolar disorder (BD). Without bells and whistles, our model achieves state-of-the-art results on these diseases diagnosis. Empirical results suggest that our self-supervised learning presents a better way to sufficiently leverage intrinsically complex structure of brain networks and learn a general representation for brain network classification compared with previous traditional classification methods and deep learning methods. We summarize our main contributions as follows:

1. We propose a general graph self-supervised learning paradigm for brain network analysis, which is able to capture better feature representation in a limited data and provide richer potential representations for downstream tasks. To the best of our knowledge, our work is the first attempt of applying the idea of self-supervised learning with masked autoencoding on the brain network analysis by taking into the graph structure information. Our results shed new light on the importance of self-supervision learning for improving brain network classification performance.

2. We propose a topology-aware encoder to exploit the graph structure with a local-to-global strategy for better capturing the inherent graph representation. The encoder is a generic pre-trained model, which can be applied to other downstream related graph learning tasks.

3. The reconstruction task is essential for guiding the training of both the encoder and decoder in the self-supervised learning procedure. Different from the previous reconstruction tasks which focus on the token reconstruction in the self-supervised learning, we propose a more appropriate reconstruction task by leveraging the relationship among the nodes and edges.

4. We provide a new perspective to study the association of multiple psychiatric disorders including autism spectrum disorder (ASD), major depression disorder (MDD) and bipolar disorder (BD). These findings are in accordance with previous studies regarding the inherent association of multiple disorders.

5. Generalization and interpretability are crucial while developing any predictive models for clinical diagnosis. In our work, we evaluate the proposed model on the multiple datasets, including the public ABIDE dataset and the center NMU dataset with rs-fMRI data. The experimental results demonstrate the advantage of the proposed method in brain disorder diagnosis, involving the strong generalization performance and the ability of identifying the brain regions and connections highly related to clinical functions.

II. Related Work

A. Brain Disease Classification

Notable progress has been made by exploiting constructed correlation information between brain regions to capture high-level representations. Recently, deep learning methods achieve great success in identifying neuropsychology due that it is better at extracting nonlinear features from data than traditional methods.

Many works [32]–[34] establish the CNN-based model to extract hierarchical topological features of brain networks for brain disease identification. Although these works overcome the limitations of feature extraction of traditional machine learning methods, the structure information among brain regions was ignored and they can not capture the high-level topological representations. The structure information has been proven to be important for brain network learning [13].

In recent years, there is an increased interest in graph convolution networks (GCNs). GCNs combine the advantages of both graph theory and deep learning approaches and have the potential to learn spatial representations in non-Euclidean domains. Recently, several studies have introduced GCN into the field of fMRI analysis, and have demonstrated the effectiveness of GCN-based models for brain disease classification. For example, Yao [35] proposed a multi-scale triple graph convolutional network (MTGCN) for functional brain network analysis. They first construct multi-scale functional connections by employing multi-scale atlases from coarse to fine ROI analysis. Then, they proposed a triplet GCN model to capture the multi-scale graph representations, followed by a weighted fusion scheme for classification. Moreover, Li [36] proposed an ensemble

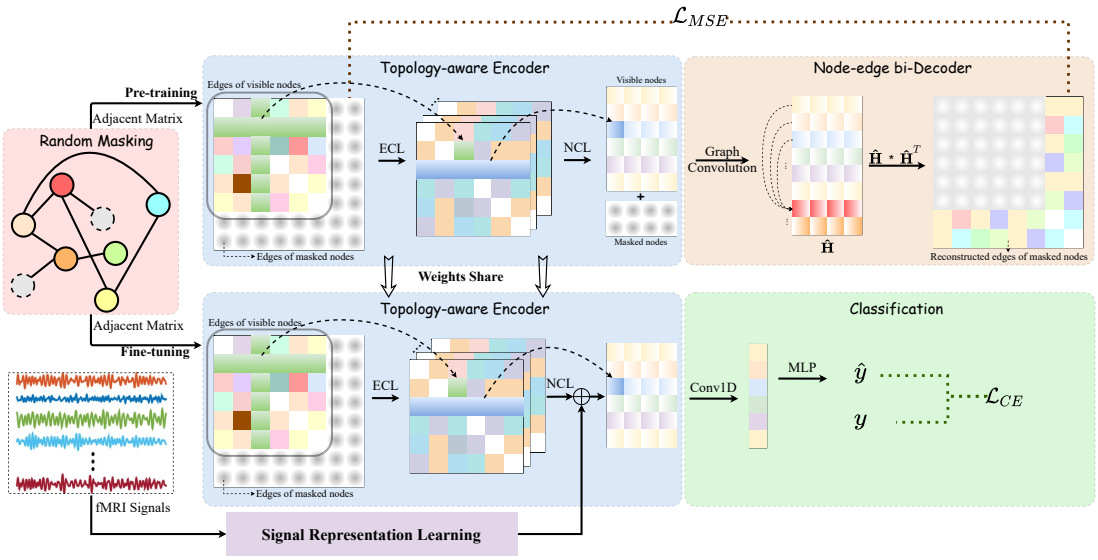


Fig. 2: Illustration of the proposed BrainGSL. Our model can be divided into a masked graph autoencoder (pre-training) and a graph classification model (fine-tuning). (1) masked graph autoencoder consists of a topology-aware encoder for latent node embedding learning and a node-edge bi-decoder for the reconstruction of the masked nodes and edges. The input of the encoder is a masked graph and the aim of the decoder is to reconstruct the masked edges by masked nodes and latent visible node. (2) Then, we combine the signal representation learning module and the pre-trained encoder to obtain the node embedding from two aspects of graph structure and BOLD signals for the graph classification under the supervision of the class labels.

framework with hierarchical graph convolution network, which can capture intrinsic correlations among subjects to improve graph embedding learning for disease diagnosis.

However, the most deep learning models do not reveal obvious improvement for the brain diseases classification due to the limited training data [16], [37]. To solve this issue, Yang et al. [37] propose GraphGAN++, which generates realistic brain networks by simultaneously preserving the global consistent distribution and local topology properties. Although the graph generation model alleviates the issue of limited data by generating extra training data, the distribution of the generated data is not exactly the same as the original data, which results in limited classification performance. Self-supervised learning can sufficiently utilize the structure in the data, alleviating the issue of limited data without causing extra issues.

B. Self-supervised Learning

1) Graph generation-based self-supervised learning: Graph generation-based self-supervised learning models [38], [39] focus on designing advanced pretext tasks (feature/structure reconstruction) for self-supervised training. GPT-GNN [39] is a graph generation-based self-supervised method, which modeled both the structure and attributes of the graph for capturing the structural and semantic properties of the graph. Zhang [19] propose Graph Masked Autoencoders (GMAEs), which is a self-supervised transformer-based model for learning graph representations. It takes partially masked graphs as input, and reconstructs the features of the masked nodes.

2) Masked Autoencoder: Masked Autoencoder (MAE) [22] is a pre-training method that has been proven to be effective on the image domain. MAE is an asymmetric encoder-decoder architecture consisting of multiple visual transformer layers which mask random patches of the input image with high masking ratio and reconstruct the missing pixels. Specifically, the input patches are divided into a visible subset and a masked subset. The encoder operates only on the visible patches and the decoder reconstructs the masked patches from the latent representation and position embedding of masked tokens. Due to the large amount of redundant information in images and the sparse semantics, MAE employs a large masking ratio to increase the difficulty of the pretext task and forces the network to learn high-level semantic features. However, simply applying the masked autoencoder idea on the brain networks is inappropriate because the brain networks contain intrinsically complex structure and rich associations.

III. Methods

A. Overview

The brain network usually leverages a graph structure to describe interconnections between brain regions by BOLD signals extracted from fMRI data. In our work, the brain network is constructed by Pearson's correlation coefficient (PCC). Formally, let $\mathbf{A} \in \mathbb{R}^{N \times N}$ represent the adjacent matrix of the brain network, where N is the node number for the graph. The major difference between our graph data of brain networks and the graph data in other graph domains is that the brain networks do not contain the initial node features, which hinders the GCNs for directly updating node embedding by aggregating from the neighborhood. Therefore, we propose a general local-to-global graph embedding learning strategy, which consists of three stages: edge embedding learning, node embedding learning and graph embedding learning. The embedding learning in our model starts from the learning of edge embedding, the initial value of which is obtained by PCC.

Edge Embedding Learning. Given a graph, the aim of edge embedding learning is to update the edge embedding by aggregating the associated edges. Specifically, two edges with common nodes are defined as associated edges.

Node Embedding Learning. The aim of node embedding learning is to capture the node representations by aggregating the associated edge embedding.

Graph Embedding Learning. With the learned embedding of nodes and edges, a learnable mapping function is used to transform the embeddings of all the nodes into a graph representation for the following graph classification task.

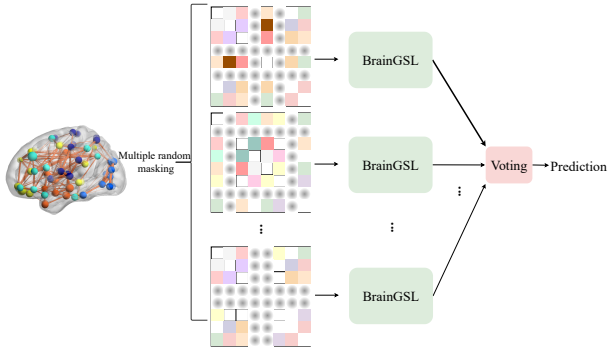


Fig. 3: The ensemble framework (BrainGSLs) by integrating multiple BrainGSL, named BrainGSLs.

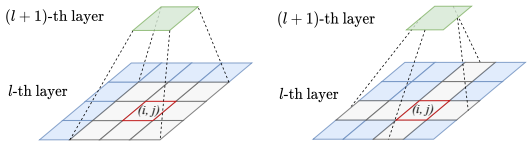


Fig. 4: Illustration of the comparison on spatial locality between images and graphs. The difference is that the locality in images denotes the pixels that are close together (for the (i,j)-th pixel, the locality is the 3×3 neighbourhood), while the locality in graph refers to the local connectivity structure associated with each edge (for the (i,j)-th edge, the locality is edge set of $e_{i\cdot}$ and $e_{\cdot j}$ associated the i-th node and j-th node).

Following the proposed strategy, we propose a graph self-supervised learning framework (BrainGSL) on brain network data for brain disorder classification. As illustrated in Fig. 2, the proposed BrainGSL consists of a topology-aware encoder, a node-edge bi-decoder, a signal representation learning module and a graph classifier. Similar to MAE, we randomly partition the nodes into two sets: visible nodes and masked nodes with a small random subset (e.g., 25%) being masked out. To sufficiently exploit the graph structure, we randomly sample nodes without replacement following a uniform distribution. The edges associated the masked nodes are also masked at the same time. The sparsity characteristic gives an opportunity for training a robust encoder. We design an asymmetric encoder-decoder architecture that consists of: 1) an encoder that takes only the visible nodes as input and learns the latent representations only for the visible tokens by modeling the correlation of the nodes, 2) a node-edge bi-decoder that takes the masked nodes and the latent visible node representations as inputs, and then produces predictions for the masked nodes and edges by graph convolution and link prediction. Moreover, we design a signal representation learning module to capture the rich temporal information in BOLD signals for better graph classification. Finally, the node embeddings yielded from the signal representation learning module and the pre-trained encoder are utilized for the graph classification. Our model allows the encoder to operate only on the partial, observed nodes by a local-to-global scheme and an asymmetric node-edge bi-decoder that reconstructs the full graph from the latent representation and mask tokens. Due to the random mask, the encoder is trained only on the visible nodes and associated edge in the pre-training, which causes randomization for the performance of the model. To eliminate the issues caused by random masking, we construct an ensemble self-supervised learning framework for brain network, named BrainGSLs. Each BrainGSL with different random masking is integrated to achieve stable and robust performance, which is illustrated in Fig. 3.

B. Mask Graph Autoencoder

1) Topology-aware Encoder: The graph convolution layers have been proven to be effective in node embedding learning in many works [14], [36]. The aim of the graph convolution layer is to update its own representation by aggregating the features from the neighbours of the nodes [40]. Although GCNs have shown great ability in modeling graphs, they are vulnerable to the noisy edges. The reason is that the message passing magnifies the negative effects of the noisy edges [37]. Moreover, only depending on the graph structure affects the node embedding learning by aggregating the neighborhood due to the densely noisy edges and lacking of node initial features [36]. To better learn the node embedding, the graph structure is desirable to be sufficiently leveraged. Hence, how to enable our encoder to preserve the topology locality is critical in the encoding stage. Motivated by BrainNetCNN [41], we propose a topology-aware encoder, consisting of an Edge Convolution Layer (ECL) for sufficiently exploiting the edge features, and a Node Convolution Layer (NCL) for modeling the node features. The detailed structures are described as follows.

Edge convolution learning, ECL: The graph topological information is crucial for graph embedding learning. To capture the essential topological representation, we propose an edge convolution layer to leverage the local connectivity in the graphs by aggregating the features of the connection associated with the nodes at the two ends of the e_{ij}. In contrast to the spatial locality in CNN, the locality in our method refers to the local connectivity structure associated with the edge, as illustrated in Fig. 4. Different from images, no location information exists in the adjacent matrix. Hence, the locality in the adjacent matrix denotes for a certain edge connecting two specific nodes. For i-th node and j-th node, the locality of e_{ij} is the set of $e_{i\cdot}$ and $e_{\cdot j}$. To capture the local information, our edge convolution layer involves multiple cross-shaped filters for the spatial locality in the graph domain. Let $\mathbf{M}$ indicate the feature map in encoder. We define that the input feature map $\mathbf{M}^{v(0)} = \mathbf{A}^{v}$, where $\mathbf{A}^{v}$ is the adjacent matrix of visible nodes. The cross-shaped filters in edge convolution layer involves a combination of $1\times N$ and $N\times 1$ basis filters, which are less computationally expensive filters with horizontal and vertical orientations. More specifically, the two basis filters are individually performed by an element-wise multiplication and the two outputs are added at each position e_{ij}. Our edge convolution layer is defined as:

$$\begin{aligned} M_{i,j}^{v(\ell)} &= \mathrm{ECONV}(M_{i,j}^{v(\ell-1)}, w_{e(H)}^{(\ell-1)}, w_{e(V)}^{(\ell-1)}) \\ &= \mathrm{ECONV}_H(M_{i,j}^{v(\ell-1)}, w_{e(H)}^{(\ell-1)}) \\ &+ \mathrm{ECONV}_V(M_{i,j}^{v(\ell-1)}, w_{e(V)}^{(\ell-1)}), \end{aligned} \tag{1}$$

where $M_{ij}^{v(\ell)}$ denotes the edge weight of the connection between i-th node and j-th node at (ℓ)-th layer. $w_{e(H)}^{(\ell-1)} \in \mathbb{R}^{1\times N}$ and $w_{e(V)}^{(\ell-1)} \in \mathbb{R}^{N\times 1}$ denote the learned vectors of the convolution kernel at $(\ell-1)$-th layer. Specifically, the filter in ECL computes the sum of edge weights over all the connections to the i-th node from the j-th node, $j = 1, ..., N$. For example, the correlation between i-th node and j-th node is updated from the structure in the neighborhood of i-th and j-th nodes. The association of the connected node pair contains the local structure information, which benefits for the node embedding learning. Therefore, our edge convolution is capable of enhancing the potentially important node pairs and generate the enhanced edge embedding. The ECL layer considers not only the explicit association of the edge that directly connecting the two nodes,

but also the implicit associations of the edges that individually contain the i-th node or j-th node.

Moreover, to alleviate the noisy edges, we introduce the spatial-attention mechanism. The graph edge scores are calculated by the graph spatial-attention module, which takes the output $\hat{\mathbf{M}}^v \in \mathbb{R}^{N\times N\times D_e}$ of the ECL as the input and adopts the spatial-wise attention to highlight the inter-spatial relationship of the graph related to the disease, where D_e is the number of channels. The enhanced edge embedding is defined as:

$$\hat{\mathbf{M}}_s^v = \hat{\mathbf{M}}^{\mathbf{v}} \otimes \sigma(\mathbf{S}), \tag{2}$$

where $\mathbf{S} \in \mathbb{R}^{N\times N\times 1}$ is the spatial-wise attention map, which is the output of the Conv2d operation for the input $\hat{\mathbf{M}}^v$. $\sigma(\cdot)$ is a sigmoid function to normalize the attention weights into [0, 1] and $\otimes$ denotes the element-wise multiplication. With the spatial attention, we can highlight the disease-specific connections and suppress the connections that are irrelevant to the disease, which is beneficial for the study of diseases.

Node convolution learning, NCL: With the edge embedding learned by edge convolution layers, we further learn the node embedding by aggregating the associated edges with the nodes with a learnable layer. More specifically, the node convolution learning (NCL) takes the enhanced edge embedding by ECL as the inputs, and maps them to generate a node embedding from a node-wise view by a 1D convolutional filter. The NCL is defined as,

$$\mathbf{H}_i^{v(\ell)} = \sum_{k=1}^{N} w_n^{(\ell-1)} M_{i,k}^{v(\ell-1)}, \tag{3}$$

where $\mathbf{H}_i^v \in \mathbb{R}^{N\times D_n}$ is i-th visible node embedding after NCL, $w_n^{(\ell-1)} \in \mathbb{R}^{1\times N}$ is the learned vector of the filter at $(\ell-1)$-th layer, and D_n is the number of channels in NCL.

Different from the graph convolution layer which updates a node's representation by aggregating its own features and its neighbors' features, our NCL achieves the node representation by aggregating the associated edge embeddings.

2) Decoder: The purpose of the decoder of masked autoencoder is to reconstruct the masked edges associated with the masked nodes. Hence, it is a link prediction problem, which aims to predict whether two nodes are likely to have a connection. The task of link prediction depends on the node embedding. Nonetheless, the masked nodes do not engage into the encoding stage, thereby it lacks node embedding for the masked nodes. To solve it, our bi-decoder consists of two reconstruction procedures: (i) masked nodes reconstruction and (ii) masked edges reconstruction. Noting that the decoder is only leveraged during pre-training to perform the graph reconstruction task (only the encoder is used to produce graph representations for the downstream classification). Therefore, the decoder can be flexibly designed in a manner that is independent of the encoder design.

Masked node reconstruction: Given the visible node embeddings obtained by the topology-aware encoder, the aim of masked node reconstruction is to obtain the masked node embedding learning from the visible nodes. To this end, we leverage a graph convolutional layer to aggregate the visible node features to update the masked node. Specifically, for a masked node v_i in the graph, the main purpose of the masked node reconstruction is to learn its representation by iteratively aggregating the representations of its neighbors. Considering the existence of a large amount of noisy connections in the graph structure, we binarize the adjacent matrix to alleviate it. The graph convolutional layer is defined as:

$$\mathbf{H}_i^{(\ell+1)} = \mathrm{AGGREGATE}\left(\left\{\mathbf{H}_j^{(\ell)} : v_j \in \mathcal{N}_i\right\}, \widetilde{\mathbf{A}}, \mathbf{W}^{(\ell)}\right), \tag{4}$$

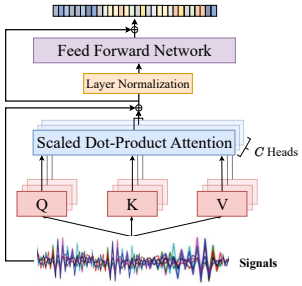


Fig. 5: Illustration of the signal representation learning module. Q, K and V are queries, keys and values of transformer block.

where $\mathbf{H}_i^{(\ell+1)}$ denotes i-th node embedding of $(\ell+1)$-th layer, $\mathcal{N}_i$ is the neighbor nodes of i-th node, D_g is the number of dimensions of node features, $\mathrm{AGGREGATE}(\cdot)$ is a GNN-based aggregator function, which can be the GCN or GAT, $\widetilde{\mathbf{A}}$ is a binarized adjacency matrix of the undirected graph without self-connections and $\mathbf{W}^l \in \mathbb{R}^{D_g\times D_g}$ is the trainable weight matrix. We employ the GCN as our aggregator function, which is defined as: $\mathbf{H}^{(\ell+1)} = \sigma(\widetilde{\mathbf{A}}\mathbf{H}^{(\ell)}\mathbf{W}^{(\ell)} + \mathbf{b}^{(\ell)})$, where $\sigma(\cdot)$ indicates the activation function and $\mathbf{b}$ is the bias. It is worth mentioning that the visible and masked node embeddings are all updated by the graph convolution layers. We binarize the adjacent matrix with a threshold because the densely connected graph contains many weak correlations (potential noise), which increases the difficulty of node embedding learning. In our work, the threshold is set to 0.15.

Masked edge reconstruction: With the node representations, we predict the edge $\hat{e}_{ij}$ between $\hat{v}_i \in V_m$ and $v_j \in V_s$ or $\hat{v}_i \in V_m$ and $\hat{v}_j \in V_m$, where V_m and V_s indicate the masked and visible node sets. The reconstruction process can be written as: $\hat{\mathbf{A}} = \mathbf{H}^{(\ell+1)}(\mathbf{H}^{(\ell+1)})^T$, where the $\hat{\mathbf{A}} \in \mathbb{R}^{N\times N}$ is the reconstruction graph.

3) Objective function: The reconstruction loss is given by the mean squared error (MSE) of masked edges between the reconstructed and original graph:

$$\mathcal{L}_{\mathrm{MGAE}} = \min_{\hat{\mathbf{A}}} \sum_{(i\in V_m \vee j\in V_m)\wedge i\neq j} \left\| \hat{A}_{i,j} - A_{i,j} \right\|_F^2, \tag{5}$$

where the $\hat{A}_{i,j}$ is the reconstructed edges connected with v_i and v_j.

In summary, compared with previous MAE for images, the proposed self-supervised learning with mask scheme possesses two distinctive properties:

(1) The nodes are considered as tokens, and the reconstruction task is the edge reconstruction instead of the node reconstruction.

(2) Rather than the transformer module used as the encoder and decoder in MAE, only local spatial convolution is unitized on the adjacent matrix in our masked graph autoencoder. Hence, our model can be substantially efficient.

C. Signal Representation Learning, SRL

Intuitively, exploring the temporal information in the BOLD signals is beneficial to enriching the graph representations and improving the classification performance. Hence, how to exploit the rich temporal information from BOLD signals is a key problem. Based on this issue, the signal representation learning module is proposed to capture the node embeddings in temporal domains to learn fine-grained representation for facilitating diagnosis performance, which is illustrated in Fig. 5. First, the BOLD signals are processed with a transformer encoder where time points in the fMRI time series

correspond to the sequence of transformer tokens. Formally, given the node signals $X \in \mathbb{R}^{T\times N}$, where T is the BOLD signal length, we calculate the query $\mathbf{Q}^c$, key $\mathbf{K}^c$ and value $\mathbf{V}^c$ for the c-th head through linear projection as follows:

$$\mathbf{Q}^c = XW_q^c, \mathbf{K}^c = XW_k^c, \mathbf{V}^c = XW_v^c, \tag{6}$$

where W_q^c, W_k^c, and $W_v^c \in \mathbb{R}^{N\times D_t^1}$ are linear projections, $c = 1, 2, ..., C$, and C is the number of heads in the multi-head attention mechanism. Then, the multi-head graph self-attention is computed as a scaled dot-product among W_q^c, W_k^c and W_v^c as follows:

$$\tilde{\mathbf{H}}_t^c = \text{Softmax}\left(\frac{\mathbf{Q}^c\left(\mathbf{K}^c\right)^T}{\sqrt{d}}\right)\mathbf{V}^c, \tag{7}$$

where d is the dimensionality of each attention head and $\tilde{\mathbf{H}}_t^c$ is the output of the c-th head. Then, we concatenate all single head outputs and obtain the final output $\tilde{\mathbf{H}}_t$. Following graph self-attention layers, the transformer block contains a fully connected feed-forward network, which consists of three linear transformations with a $\text{ReLU}(\cdot)$ activation in between, described as follows:

$$\begin{aligned}\bar{\mathbf{H}}_t &= \text{LayerNorm}(\tilde{\mathbf{H}}_t W_f^1 + X),\\ \bar{\mathbf{H}}_t &= \text{ReLU}(\bar{\mathbf{H}}_\mathbf{t} W_f^2)W_f^3,\\ \mathbf{H}_t &= (\bar{\mathbf{H}}_\mathbf{t})^T W_f^4,\end{aligned} \tag{8}$$

where $W_f^1 \in \mathbb{R}^{(C*D_t^1)\times N}$, $W_f^2 \in \mathbb{R}^{N\times D_t^2}$, $W_f^3 \in \mathbb{R}^{D_t^3\times N}$ and $W_f^4 \in \mathbb{R}^{T\times Dn}$ are linear transformations, and $\mathbf{H}_t$ is the learned node embedding from BOLD signals.

D. Graph Classification

With the signal representation learning module and the pre-trained topology-aware encoder, the node embeddings learned from two aspects of graph structure and BOLD signals can be gained in the downstream graph classification task. Then, we fuse the two node embeddings with a simple summation and employ a 1D convolutional layer to convert it into the final graph representation. Finally, the multi-layer perceptron (MLP) is used to predict the class of the input graphs. We fine-tune the graph classification model with cross-entropy loss, which is defined as:

$$\mathcal{L}_{CE} = \min \frac{1}{\mathcal{N}}\sum_{p}^{\mathcal{N}} -\left[y_p \cdot \log\left(\hat{y}_p\right) + \left(1-y_p\right)\cdot\log\left(1-\hat{y}_p\right)\right], \tag{9}$$

where y_p indicates the label of p-th graph, $\hat{y_p}$ is the prediction of p-th subject and $\mathcal{N}$ is the number of samples.

IV. Experiments

A. Datasets

To validate our approach, we conduct experiments on three different brain disease diagnosis tasks on the two datasets: ABIDE dataset [42] for ASD diagnosis, and the center NMU dataset for major depression disorder (MDD) and bipolar disorder (BD) diagnosis.

1) ABIDE dataset: The ABIDE is a public autism research database aggregating rs-fMRI and phenotypic data of 1112 subjects from 17 different acquisition sites [43]. In this work, the Connectome Computing System (CCS) [44] was leveraged to preprocess the images. The preprocessing included slice timing correction, correction for motion, and normalization of voxel intensity. We obtained 871 quality subjects including 403 individuals with ASD (54 females and 349 males, aged 17.07±7.96 years, range 7-58 years) and 468 normal controls (90 females and 378 males, aged 16.84±7.24 years, range 6-56 years).

TABLE II: The parameter settings of the training.

Parameter name	Value
Optimizer	SGD
Learning rate	0.005
Dropout rate	0.5
Weight decay	0.01
Batch size	32
Max training epochs for pre-training/classification	10/50
Masked ratio	25%
Edge/node convolution layers	2/1
Graph embedding learning layers	1
Threshold for adjacent matrix	0.15
Head number	2

2) Center NMU dataset: The center NMU (Nanjing Medical University) dataset is provided from Nanjing Medical University. In the process of data pre-processing, we deal with data by using dpabi [45] and divided the whole brain into 90 brain regions based on Automated Anatomical Labeling (AAL) for analysis. They included spatial normalization to Echo Planar Imaging (EPI) template of standard Montreal Neurological Institute (MNI) space (spatial resolution 3mm×3mm×3mm), spatial and temporal smoothing with a 6mm×6mm×6mm Gaussian kernel and filter processing with adopting 0.01-0.08Hz low-frequency fluctuations to remove interference signals. The dataset includes 246 health controls (152 females and 94 males, aged 26.89±6.14 years, range 14-51 years), 151 MDDs (108 females and 43 males, aged 16.97±5.01 years, range 12-51 years) and 126 BDs (83 females and 43 males, aged 17.24±4.03 years, range 12-39 years), who were scanned at a single site with identical inclusion and exclusion criteria.

B. Experimental Setup

In this work, we evaluated our BrainGSLs on ABIDE dataset and the center NMU dataset using a 10-fold stratified cross validation strategy. The parameters setting of our model is shown in Table II. With the above setting, we performed comprehensive experiments to evaluate the effectivity of our BrainGSLs for the brain network classification.

C. Results of Brain Disorder Identification

In this section, we aim to answer two questions:

Q1. Can BrainGSLs learn the potential node embedding via pre-training? And to what extent can the pre-training benefit the downstream classification tasks?

Q2. How does BrainGSLs perform compared to the state-of-the-art brain network classification methods, including the traditional connectivity-based traditional machine learning methods and deep learning methods?

1) Comparison with the state-of-the-art methods: To comprehensively evaluate the performance of brain network diagnosis, we compare our BrainGSL with state-of-the-art models. Specifically, the comparable methods can be grouped into two categories: traditional methods and deep learning methods.

Traditional methods:

FC+SVM/RF: Based on the correlation feature (the flattened Person correlation), a SVM-RFE with an RBF kernel or a random forest (RF) is trained for brain disorder diagnosis.

GRP: Gaussian random projection (GRP) [1] reduces the feature dimensionality by mapping Pearson correlation features to a random matrix, and then selects the best features as the inputs of linear SVM for classification.

NAG-FS [46]: The NAG-FS is a feature selection guided by brain network atlases. The aim of NAG-FS is to estimate representative and

centered brain network atlases for identifying discriminative brain connections between healthy and disordered populations.

MC-NFE [1]: The Multi-site Clustering and Nested Feature Extraction method is a general framework to model inter-site heterogeneity for functional connectivity by a nested SVD strategy.

Deep learning methods:

IN-Net [3]: IN-Net maps the FC matrix to a feature domain and then predicts it with a fully connected layer.

GroupINN [47]: GroupINN learns the node grouping to construct the common and clean graph structure for graph embedding learning. The GroupINN combines the graph structure learning and classification, resulting in better classification performance.

ST-GCN [30]: ST-GCN can capture the spatio-temporal representations of fMRI data to predict the age and gender of subjects. We expand and adjust the signal length of the fMRI data from ABIDE since the limitation of ST-GCN for processing equal-length signals.

BrainNetCNN [41]: BrainNetCNN is a convolutional neural network framework which consists of edge-to-edge, edge-to-node and node-to-graph convolutional filters that leverage the topological locality of brain networks. In our work, we take it as our baseline.

s-GCN [10]: s-GCN is a siamese GCN for identifying the patterns associated with the similarity between two graphs. The learned similarity metric can be properly captured through the graph structure.

BrainGNN [13]: BrainGNN is an end-to-end graph neural network-based framework that jointly learns ROI clustering and the whole-brain fMRI classification.

MVS-GCN [15]: MVS-GCN is a multi-view graph convolution network, which combines the graph structure learning and multi-task learning to improve the classification performance.

LSTM-ASD [48]: LSTM-ASD is a LSTM based model for the identification of individuals with ASD and health controls directly from the fMRI data.

ASD-DiagNet [49]: ASD-DiagNet is a joint learning method combining an autoencoder for reconstructing the input embedding with a single layer perceptron for classification.

Variant methods:

To better evaluate the effectiveness of our self-supervisd learning, we design four variants: TA-encoder, BrainGSL-AE, BrainGSL-GCN and BrainGSLs-JL.

TA-encoder: It is a graph classification model with the proposed topology-aware encoder.

BrainGSL-AE: We replace our proposed encoder-decoder framework with vanilla autoencoder for the pretext learning.

BrainGSL-GCN: We replace our proposed topology-aware encoder with vanilla GCN for the pretext learning.

BrainGSLs: We average the results of 20 different single random BrainGSL models.

BrainGSLs-JL: Following the end-to-end training strategy of ASD-DiagNet, BrainGSLs-JL is jointly trained for classification and reconstruction with 10 epochs and then fine-tuned for the classification with extra 50 epochs.

BrainGSLs-SRL: We incorporate the signal representation learning module into BrainGSLs.

Experimental results of ASD diagnosis on the ABIDE dataset, and MDD as well as BD diagnosis on the center NMU dataset, are reported in Table III and Table IV where the best results are boldfaced. In addition, we are also interested in the effectiveness of each component in the proposed model. Accordingly, we conduct an ablation study for BrainGSLs-SRL to investigate how each component affects the classification performance on ABIDE dataset in Table III. We highlight the following observations:

1. For a fair comparison, Our experimental results show that our BrainGSLs-SRL yields the best ACC and AUC results (ACC=71.3%, AUC=71.6%) on ABIDE dataset, compared to all the existing ASD diagnosis approaches, achieving a new state-of-the-art on the ABIDE dataset. Specifically, compared with the traditional methods such as SVM/MC-NFE, BrainGSLs-SRL achieves an improvement of 3.9%/2.9% and 4.7%/2.5% in terms of ACC and AUC, respectively. Similarly, compared with supervised deep learning methods such as TA-encoder/GroupINN, BrainGSLs-SRL also suggests an additional improvement of 5.7%/7.4% and 6.6%/7.6% in terms of ACC and AUC, respectively, which demonstrates the effectiveness of our model. Moreover, BrainGSLs-SRL achieves the best performance in other two diseases classification tasks (HC vs. MDD and HC vs. BD). These results demonstrate that our model is effective for the graph classification with the brain disorders diagnosis.

2. It can be seen that some traditional methods such as MC-NFE perform better than deep supervised learning methods such as ST-GCN from Table III. This is not surprising as most deep supervised learning methods are influenced by the limited data, resulting in poor classification performance.

3. Compared with the non-graph deep learning methods, graph-based neural network methods perform worse. For example, ASD-DiagNet/LSTM-ASD outperform GroupINN by 4.4%/4.6 and 4.6%/-

TABLE III: Comparison with the state-of-the-art methods on ABIDE dataset of CC200 atlas. The best results are bold. The average graph classification performances of accuracy (ACC), AUC, sensitivity (SEN) and specificity (SPEC) are reported.

Method	ACC(%)	AUC(%)	SEN(%)	SPEC(%)
FC + SVM	67.4	67.1	52.5	69.3
FC + RF	64.5	62.2	46.4	62.9
GRP [1]	59.4	61.1	60.9	57.1
NAG-FS [46]	61.8	59.9	58.8	60.7
MC-NFE [1]	68.4	69.3	70.1	63.6
IN-Net [3]	65.3	61.4	60.8	58.1
GroupINN [47]	63.9	63.2	61.5	57.4
ST-GCN [30]	57.3	51.7	54.8	48.9
sGCN [10]	67.5	64.3	64.7	60.1
BrianGNN [13]	61.8	60.8	61.7	60.8
LSTM-ASD [48]	68.5	-	-	-
MVS-GCN [15]	69.9	69.1	70.2	63.1
ASD-DiagNet [49]	68.3	67.8	60.3	67.8
GMAEs [19]	55.1	56.8	30.1	**84.4**
BrainNetCNN [41]	65.6	64.2	52.5	69.3
TA-encoder	66.5	66.1	59.7	68.3
BrainGSL-AE	67.4	67.0	65.1	67.3
BrainGSL-GCN	66.2	65.6	60.9	66.3
BrainGSL	68.6	68.1	65.9	67.6
BrainGSLs-JL	68.5	69.3	67.2	67.0
BrainGSLs	70.4	71.1	**70.6**	69.5
BrainGSLs-SRL	**71.3**	**71.6**	70.2	69.9

TABLE IV: Comparison with the state-of-the-art methods on MDD and BD identification. The best results are bold. We evaluate the models on two classification tasks: HC (health control) vs. MDD and HC vs. BD.

Method	HC vs. MDD		HC vs. BD	
	ACC(%)	AUC(%)	ACC(%)	AUC(%)
FC + SVM	68.1	67.5	73.2	70.8
FC + RF	63.4	59.8	67.7	60.2
TA-encoder	70.6	69.2	68.6	68.3
GroupINN	66.8	65.3	67.9	63.3
ASD-DiagNet	68.2	66.7	73.3	70.1
MVS-GCN	68.3	68.2	66.9	64.2
ST-GCN	58.1	52.3	67.1	57.5
BrainGSLs	75.5	74.4	75.3	71.8
BrainGSLs-SRL	**76.2**	**74.7**	**76.9**	**73.0**

in terms of ACC and AUC, respectively. The results indicate that GNNs are affected by the noisy edges.

4. By comparing the BrainNetCNN and the TA-encoder, we can find that TA-encoder has an improvement of 0.9% and 1.9% in terms of ACC and AUC, respectively. This result indicates the topology-aware encoder is more effective for node embedding learning due to the enhancement of important correlations by the spatial attention module.

5. We compared the self-supervised learning methods of GMAE [19], BrainGSL-AE and BrainGSL, all of which involve two branches: pre-training through self graph reconstruction and fine-tuning on downstream graph classification. They differ in how the self-supervised learning is performed. Specifically, BrainGSL-AE has no masking scheme and the aim of the decoder is only the edge construction, which is different from BrainGSL. GMAE is a masked node reconstruction autoencoder by simply borrowing the masked autoencoding idea through graph transformer layers on the graph data. It is worth mentioning that our method outperforms BrainGSL-AE and MGAE [19] by 1.2% and 13.5% on the ACC, respectively. The result confirms the necessity of developing more comprehensive self-supervised learning for brain networks. The results also demonstrate that our model is a successful application of masked autoencoding from CV and NLP domains to the brain network. Moreover, by comparing BrainGSL and BrainGSL-GCN, we can find that BrainGSL has an improvement of 2.4%/2.5% in terms of ACC/AUC, respectively. It proves the effectiveness of our topology-aware encoder on exploiting the node representations.

6. The comparison between BrainGSLs and BrainGSLs-JL indicates the training strategy is important for the self-supervised learning. Jointly training the two tasks of graph reconstruction and classification can not sufficiently capture the node embeddings.

7. BrainGSLs-SRL achieve better classification performance compared to BrainGSLs without the signal representation learning module on the ABIDE dataset (ACC: 70.4% $\rightarrow$ 71.3% and AUC: 71.1% $\rightarrow$ 71.6%). The result confirms the necessity of capturing the temporal representations from BOLD signals for graph classification in the graph embedding learning.

V. Results and Discussion

In this section, there are four questions need to be answered:

1. How much is our proposed method influenced by the key hyperparameters including the mask ratio in masked graph autoencoder, the pre-training epochs, the threshold value of binarizing graph structure and the number of heads in the signal representation learning?
2. Is our BrainGSLs repeatable with high classification performance on multiple brain atlases?
3. Can our model provide insights for the association among the common brain diseases?
4. Can our model provide discriminative brain regions or function connections?

A. Hyperparameter Sensitivity

In our model, there are three key hyperparameters including the mask ratio, the pre-training epochs and the threshold for binarizing graph structure. We design three experiments to systematically investigate the different factors on the classification performance.

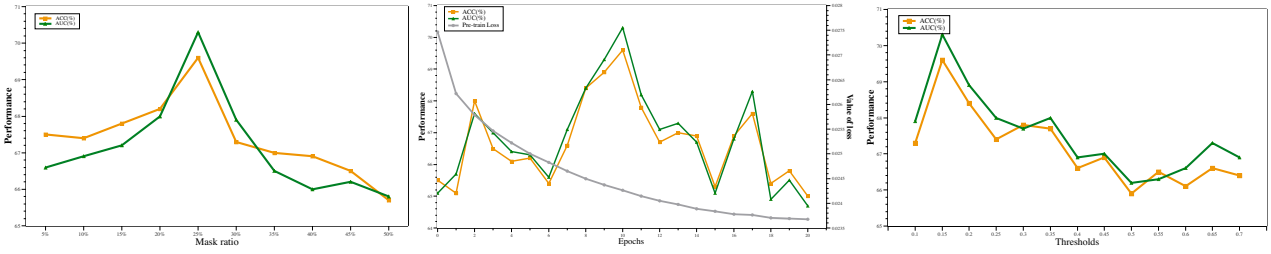
Fig. 6 (a) shows the influence of the variability of masking ratios on the classification. It can be found that the masking ratio in masked graph autoencoder has a significant impact on the classification performance, and the best results are achieved when masking ratio is 0.25. It is largely different from MAE [22], where the best masking ratio is 0.75. We also find that the classification performance of our model gradually decreases as the ratio increases. The reason is that the locality property of each visible node in our encoder is associated with all the available connections of the node, thereby the encoder depends on the number of the masked nodes. A higher masking ratio yields fewer nodes participating in the downstream graph classification task. We also evaluate the influence of the pre-training epochs on the classification performance and show the convergence curve of our model pre-training in Fig. 6 (b). It can be found that the performance improves with the pre-training epochs increasing until 10 epochs, which demonstrates that more pre-training epochs help capture accurate node embedding. Nevertheless, as the pre-training epochs further increases, the reconstruction task rapidly achieves convergence and tends to be overfitting, which decreases the downstream classification performance decreases. In our work, the termination condition of pre-training is that the value of loss decreases by no more than 1e-4. The pretext reconstruction task and downstream classification task are relevant, however, the embedding obtained by encoder is inappropriate for the downstream classification task when the reconstruction is overfitting. At last, to explore the effect of different thresholds on the masked node reconstruction during the graph structure binarization, we vary the threshold values in the range of $\{0.1, 0.15, ..., 0.7\}$ to observe the classification performance, which is shown in Fig. 6 (c). We find that the best result is achieved when $T = 0.15$ (nearly 85% edges are removed), which demonstrates that the adjacent matrix with higher sparse levels negatively influences the reconstruction of the masked nodes due to insufficient neighbourhood information.

In order to evaluate how the head number affects our model, we conduct a sensitivity analysis on the ABIDE dataset, shown in Fig. 7. We observe that the incorporation of fewer heads in our BrainGSLs-SRL generally leads to a better graph classification performance, and there is a general trend that our model performs worse as the number of heads increasing. The results show that it is critical to capture the temporal representations from BOLD signals with a suitable number of heads for graph embedding learning. Nevertheless, the model with more heads causes the overfitting in training, slightly worsening performance.

To explore the influence of the number of single BrainGSL in the ensemble on the classification performance, we compare different number of BrainGSL component in Fig. 8. From Fig. 8, we can see that the performance of the ensemble model is better than the single model. We also can find that as the number of single BrainGSL increasing, the classification performance increases and stabilizes, which is as expected. The result verifies that introducing the ensemble learning framework allows improving the self-supervised learning and the brain network classification.

B. Repeatability Using Different Brain Atlases

Different brain atlases divide the brain into different numbers of brain regions. To evaluate whether these observations were dependent on the choice of the atlas, we applied the same methodology on different atlas, including Automated Anatomical Labeling (AAL), Harvard-Oxford (HO), Eickhoff-Zilles (EZ), Talariach Daemon (TT) and Dosenbach 160-region atlas (DOS160) [50]. The performance of our model on multiple atlases is presented in Fig. 9. We observe that our method achieves the best results with the CC200 brain atlas, which proves that the finer brain region delineation can provide sufficient information. Moreover, it can be seen that our model achieves better classification performance than TA-encoder on different atlas, which indicates that the generalizability of our model.



(a) Masking ratios (b) Pre-training epochs (c) Different thresholds

Fig. 6: Performance of BrainGSL with different hyperparameters. (a) is the variation of classification performance with different masking ratios in fine-tuning mode. (b) is the effect of pre-training epochs for the classification. (c) is the effect of different thresholds on the model during the binarization of the association matrix.

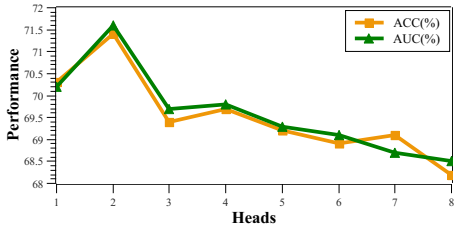


Fig. 7: The performance of BrainGSLs-SRL with varying number of heads in signal representation learning module.

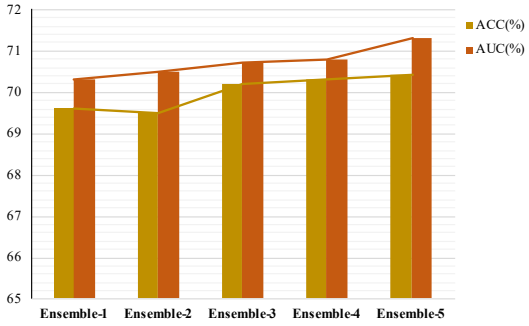


Fig. 8: The performance of BrainGSLs ensemble with varying number of the individual BrainGSL.

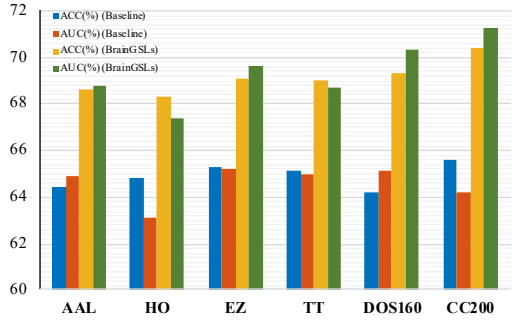
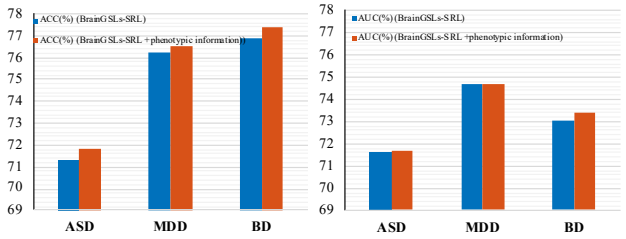


Fig. 9: Comparison of multiple atlases, including AAL, HO, EZ, TT, DOS160 and CC200 on ABIDE dataset. The baseline method is supervised TA-encoder.

C. Incorporation of Phenotypic Information

To facilitate better diagnosis performance, we attempt to fuse the learned graph embeddings with the corresponding phenotypic information (i.e. gender and age). The classification performance of BrainGSLs-SRL and BrainGSLs-SRL with the phenotypic information on the three disease classification tasks is illustrated in terms of ACC in Fig. 10(a) and AUC in Fig. 10(b).We can see that with the phenotypic information, the model can achieve better classification results on all three disease diagnosis tasks. The results suggest that the incorporation of extra prior knowledge is critical for improving the diagnosis performance.



(a) Comparison of ACC (b) Comparison of AUC

Fig. 10: Comparison between BrainGSLs-SRL and BrainGSLs-SRL with the phenotypic information in terms of (a) ACC and (b) AUC, respectively.

D. Association of Brain Diseases

An amount of works have demonstrated there exists associations between different psychiatric disorders [51], [52], i.e., patients with one psychiatric disorder are more susceptible to other psychiatric disorders. We design several experiments, in which we pre-train the model with one disorder dataset and fine-tuning on the other dataset, to explore the correlation among ASD, MDD and BD in Table. V. We find some interesting observations below:

1. By comparing the pre-trained models with the classification model without pre-training, we find that the pre-trained models can achieve better classification performance on the three tasks. The results again demonstrate that self-supervised learning can effectively alleviate the problem of insufficient learning of supervised models due to limited data. Furthermore, it also suggests that an appropriate transfer learning for learned knowledge from the auxiliary datasets can help solve the target task.

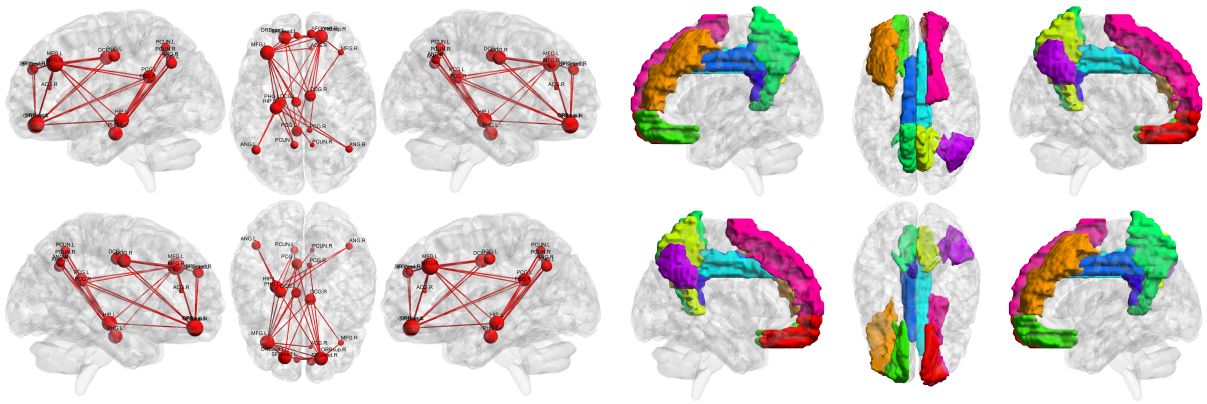
2. We observe that the model pre-trained from the BD classification achieves the second best in the HC vs. ASD classification task. In addition, the model using ASD data for pre-training achieves the best performance in the HC vs. BD classification task. It indicates that the two diseases are correlated, and the patients with ASD and BD may exhibit more similar characteristic in the functional brain network. However, the learned knowledge through transferring is different for ASD $\rightarrow$ BD and BD $\rightarrow$ ASD according to the observation that the pre-training from the task of ASD provides a higher improvement for the task of BD classification. The reason is that a larger amount

TABLE V: Comparison of different pre-train data for the three brain disease identification.

Source task \ Target task	HC vs. ASD		HC vs. MDD		HC vs. BD	
	ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)
-	64.4	64.9	70.6	69.2	68.6	68.3
HC vs. ASD	**68.3**	**68.4**	73.4	72.4	**77.6**	**75.8**
HC vs. MDD	65.4	65.7	**75.5**	**74.4**	74.8	72.2
HC vs. BD	66.2	66.6	74.8	72.2	75.3	71.8

TABLE VI: The top 30 discriminative connections for ASD diagnosis.

Datasets	Top 1-15			Top 16-30		
	Rank	Brain region pairs	Anatomical region pairs	Rank	Brain region pairs	Anatomical region pairs
ABIDE	1	Frontal_Sup_Orb_R-Frontal_Mid_L	Prefontal-Prefontal	16	Frontal_Sup_Orb_R-Frontal_Mid_R	Prefontal-Prefontal
	2	Cingulum_Mid_R-Frontal_Mid_L	Frontal-Prefontal	17	ParaHippocampal_L-Frontal_Mid_L	Temporal-Prefontal
	3	Cingulum_Mid_R-Frontal_Sup_Orb_R	Frontal-Prefontal	18	Cingulum_Post_L-Frontal_Sup_Orb_L	Parietal-Prefontal
	4	Cingulum_Post_L-Frontal_Mid_L	Parietal-Prefontal	19	Cingulum_Post_R-Frontal_Mid_L	Parietal-Prefontal
	5	ParaHippocampal_L-Angular_L	Temporal-Parietal	20	Frontal_Sup_Medial_L-Frontal_Sup_Orb_R	Prefontal-Prefontal
	6	Cingulum_Post_L-Frontal_Sup_Orb_R	Parietal-Prefontal	21	Frontal_Sup_Medial_R-Frontal_Sup_Orb_R	Prefontal-Prefontal
	7	Hippocampus_L-Frontal_Mid_L	Temporal-Prefontal	22	Hippocampus_L-Precuneus_L	Temporal-Parietal
	8	Cingulum_Mid_L-Frontal_Mid_L	Frontal-Prefontal	23	Frontal_Sup_Medial_R-Frontal_Mid_L	Prefontal-Prefontal
	9	Cingulum_Mid_L-Frontal_Sup_Orb_R	Frontal-Prefontal	24	Cingulum_Mid_R-Frontal_Mid_R	Frontal-Prefontal
	10	ParaHippocampal_L-Precuneus_L	Temporal-Parietal	25	ParaHippocampal_L-Precuneus_R	Temporal-Parietal
	11	Cingulum_Mid_R-Frontal_Sup_Orb_L	Frontal-Prefontal	26	Hippocampus_L-Angular_R	Temporal-Parietal
	12	ParaHippocampal_L-Angular_R	Temporal-Parietal	27	Cingulum_Post_R-Frontal_Sup_Orb_R	Parietal-Prefontal
	13	Frontal_Sup_Orb_L-Frontal_Mid_L	Prefontal-Prefontal	28	Cingulum_Ant_R-Frontal_Mid_L	Prefontal-Prefontal
	14	Hippocampus_L-Frontal_Sup_Orb_R	Temporal-Prefontal	29	Olfactory_L-Frontal_Mid_L	Prefontal-Prefontal
	15	Hippocampus_L-Angular_L	Temporal-Parietal	30	Frontal_Mid_L-Frontal_Sup_Orb_L	Prefontal-Prefontal



(a) Top-30 connections among brain regions

(b) Top-10 brain regions

Fig. 11: Illustration of discriminative brain regions and connections for ASD diagnosis. The left figure is the top-30 connections among brain regions. The right is the top-10 brain regions that are crucial for diagnosis.

of ASD data provides more sufficient information. Although the amount of the BD dataset is limited, the pre-trained model also helps increase the performance from 64.4% to 66.2% of ACC in the ASD classification task.

3. Interestingly, we find that the model pre-trained from the BD classification task performs better than from the ASD classification task in HC vs. MDD classification task, although BD dataset contains few instances. The results show that the correlation between MDD and BD is stronger than the one between MDD and ASD.

Previous work has investigated the association between MDD, ASD and BD from different perspectives. Radonji [52] computed the Pearson correlation between their sMRI phenotypics for each pair of disorders, including MDD, ASD, BD. For the three diseases of MDD, ASD and BD, the results indicates that the strongest association was the one between MDD and BD, followed by the one between ASD and BD. The weakest is the correlation between ASD and MDD. Notability, these diseases were positively associated. Despite the work and our study focus on the different data modality, their conclusions are consistent with ours, revealing an intrinsic association of diseases. In addition, BD is one of the most severe psychiatric disorders that can be found in comorbidity with ASD [53]. According to the observations of previous studies from the clinical presentation, the presence of autism spectrum conditions in the population seems to be associated with the early onset of BD. Specifically, the 3%-27% of patients with ASD suffer from BD, and 2%-30% of patients with BD suffer from ASD [54]. Recently, the study [55] reports that the percentage of participants with BD who also suffer from ASD is 42.7%. The results indicate that our model provides a new perspective for the study of disease associations through transfer learning on the brain network analysis.

E. Discriminative Brain Regions for Diagnosis

In this work, we construct a graph self-supervised model applicable to brain network learning and explore the interpretability of disorders

through the spatial attention mechanism in our model. The previous work indicates that patients with autism typically have abnormal functional connectivity [56].

Evidence of abnormal functional connectivity patterns in ASD has still been fully inconsistent. However, most studies summarize partially consistent findings. We show the top 30 connections of brain region pairs according to scores representing the contribution to the classification in Tab. VI. In addition, we also show the associations of brain regions with anatomical regions. From the Tab. VI, we find 23 out of 30 brain region pairs are associated with prefrontal region, which demonstrates that the prefrontal region is crucial for ASD diagnosis. Moreover, 9 brain region pairs occurred within the prefrontal. Therefore, we conclude that brain region interactions within the prefrontal region have a key role in the development of autism. We also find the temporal region is an important region for ASD diagnosis because 10 out of 30 brain region pairs are associated with it in our results. Fig. 11a is the illustration of top-30 connections. The results are consistent with the conclusion of [57]–[59]. Moverover, we performed a global analysis of the attention weight matrix and summed the weights in nodes to obtain the importance ranking of the brain regions, which is illustrated in Fig. 11b. The top 10 brain regions include superior frontal gyrus, middle prefrontal gyrus, precuneus, median cingulate and paracingulate gyri, posterior cingulate gyrus, and angular gyrus (no distinction between right and left brain), in which the posterior cingulate gyrus, the medial prefrontal cortex and the precuneus are the hub nodes of the DMN. Several works [57], [60] prove that the DMN is associated with the development of autism. Moreover, the random masking strategy also allows us to explore the discriminative brain regions. We trained 100 BrainGSL with different random masking and obtained the top 5 commonly visible nodes of the models with better classification performance. The commonly visible nodes include superior frontal gyrus, middle occipital gyrus, hippocampus, inferior temporal gyrus and middle prefrontal gyrus, in which the middle prefrontal gyrus [60], the superior frontal gyrus [61] and inferior temporal gyrus [62] are associated with autism. These findings are consistent with the current interpretation of the pathological pathways of ASD.

VI. Conclusion

Although deep supervised learning methods can achieve advanced performance in brain disease identification, it suffers from several challenges, including limited data and insufficient learning. In this work, we focus on self-supervised algorithms to compensate for the lack of supervision and propose a masking graph self-supervised framework called BrainGSL for brain network analysis. We conduct extensive experiments on the public ABIDE dataset and the center NMU dataset, which indicates that learning the graph representations with self-supervised training has led to remarkable improvement and our final BrainGSLs-SRL achieves promising performance compared with the state-of-the-art methods. In addition, we explore the correlations among ASD, MDD and BD, which provides a new perspective to study the association of multiple psychiatric disorders. We also discuss the interpretability of our model and find the discriminative brain regions and correlations for diagnosis. This work provides a researching direction to propose the self-supervision methods to facilitate the deep learning method training on the brain networks.

References

[1] N. Wang, D. Yao, L. Ma, M. Liu, Multi-site clustering and nested feature extraction for identifying autism spectrum disorder with resting-state fmri, Medical image analysis 75 (2022) 102279.

[2] P. H. Kassani, A. Gossmann, Y.-P. Wang, Multimodal sparse classifier for adolescent brain age prediction, IEEE Journal of Biomedical and Health Informatics 24 (2) (2019) 336–344.

[3] J. Zhuang, N. C. Dvornek, X. Li, P. Ventola, J. S. Duncan, Invertible network for classification and biomarker selection for asd, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 700–708.

[4] Z. Zhang, J. Ding, J. Xu, J. Tang, F. Guo, Multi-scale time-series kernel-based learning method for brain disease diagnosis, IEEE Journal of Biomedical and Health Informatics 25 (1) (2020) 209–217.

[5] Z. Wang, B. Jie, C. Feng, T. Wang, W. Bian, T. X. Ding, W. Zhou, M. Liu, Distribution-guided network thresholding for functional connectivity analysis in fmri-based brain disorder identification, IEEE journal of biomedical and health informatics (2021).

[6] M. Wang, J. Huang, M. Liu, D. Zhang, Functional connectivity network analysis with discriminative hub detection for brain disease identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 1198–1205.

[7] V. Kumar, R. Garg, Resting state functional connectivity alterations in individuals with autism spectrum disorders: A systematic review, Frontiers in Psychiatry (2021) 1–55.

[8] M. Khosla, K. Jamison, G. H. Ngo, A. Kuceyeski, M. R. Sabuncu, Machine learning in resting-state fmri analysis, Magnetic resonance imaging 64 (2019) 101–121.

[9] H. Guo, F. Zhang, J. Chen, Y. Xu, J. Xiang, Machine learning classification combining multiple features of a hyper-network of fmri data in alzheimer's disease, Frontiers in neuroscience 11 (2017) 615.

[10] S. I. Ktena, S. Parisot, E. Ferrante, M. Rajchl, M. Lee, B. Glocker, D. Rueckert, Metric learning with spectral graph convolutions on brain connectivity networks, NeuroImage 169 (2018) 431–442.

[11] X. Li, N. C. Dvornek, J. Zhuang, P. Ventola, J. S. Duncan, Brain biomarker interpretation in asd using deep learning and fmri, in: International conference on medical image computing and computer-assisted intervention, Springer, 2018, pp. 206–214.

[12] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. G. Moreno, B. Glocker, D. Rueckert, Spectral graph convolutions for population-based disease prediction, in: International conference on medical image computing and computer-assisted intervention, Springer, 2017, pp. 177–185.

[13] X. Li, Y. Zhou, N. Dvornek, M. Zhang, S. Gao, J. Zhuang, D. Scheinost, L. H. Staib, P. Ventola, J. S. Duncan, Braingnn: Interpretable brain graph neural network for fmri analysis, Medical Image Analysis 74 (2021) 102233.

[14] H. Jiang, P. Cao, M. Xu, J. Yang, O. Zaiane, Hi-gcn: A hierarchical graph convolution network for graph embedding learning of brain network and brain disorders prediction, Computers in Biology and Medicine 127 (2020) 104096.

[15] G. Wen, P. Cao, H. Bao, W. Yang, T. Zheng, O. Zaiane, Mvs-gcn: A prior brain structure learning-guided multi-view graph convolution network for autism spectrum disorder diagnosis, Computers in Biology and Medicine (2022) 105239.

[16] Y. Xie, Z. Xu, J. Zhang, Z. Wang, S. Ji, Self-supervised learning of graph neural networks: A unified review, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

[17] K. Yan, J. Cai, D. Jin, S. Miao, D. Guo, A. P. Harrison, Y. Tang, J. Xiao, J. Lu, L. Lu, Sam: Self-supervised learning of pixel-wise anatomical embeddings in radiological images, IEEE Transactions on Medical Imaging (2022).

[18] G. Li, R. Togo, T. Ogawa, M. Haseyama, Tribyol: Triplet byol for self-supervised representation learning, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 3458–3462.

[19] S. Zhang, H. Chen, H. Yang, X. Sun, P. S. Yu, G. Xu, Graph masked autoencoders with transformers, arXiv e-prints (2022) arXiv–2202.

[20] G. Liu, L. Shi, J. Qiu, W. Lu, Two neuroanatomical subtypes of males with autism spectrum disorder revealed using semi-supervised machine learning, Molecular autism 13 (1) (2022) 1–14.

[21] K. Ounap, Silver-russell syndrome and beckwith-wiedemann syndrome: opposite phenotypes with heterogeneous molecular etiology, Molecular Syndromology 7 (3) (2016) 110–121.

[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, arXiv preprint arXiv:2111.06377 (2021).

[23] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, J. Wang, Context autoencoder for self-supervised representation learning, arXiv preprint arXiv:2202.03026 (2022).

[24] B. Eckart, W. Yuan, C. Liu, J. Kautz, Self-supervised learning on 3d point clouds by learning discrete generative models, in: Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8248–8257.

[25] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, B. Gong, Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, Advances in Neural Information Processing Systems 34 (2021).

[26] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, J. Tang, Self-supervised learning: Generative or contrastive, IEEE Transactions on Knowledge and Data Engineering (2021).

[27] L. Jing, Y. Tian, Self-supervised visual feature learning with deep neural networks: A survey, IEEE transactions on pattern analysis and machine intelligence 43 (11) (2020) 4037–4058.

[28] F. J. Martinez-Murcia, A. Ortiz, J.-M. Gorriz, J. Ramirez, D. Castillo-Barnes, Studying the manifold structure of alzheimer's disease: a deep learning approach using convolutional autoencoders, IEEE journal of biomedical and health informatics 24 (1) (2019) 17–26.

[29] M. Wang, C. Lian, D. Yao, D. Zhang, M. Liu, D. Shen, Spatial-temporal dependency modeling and network hub detection for functional mri analysis via convolutional-recurrent network, IEEE Transactions on Biomedical Engineering 67 (8) (2019) 2241–2252.

[30] S. Gadgil, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, E. Adeli, K. M. Pohl, Spatio-temporal graph convolution for resting-state fmri analysis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 528–538.

[31] Y. Tang, L. Zhang, H. Wu, J. He, A. Song, Dual-branch interactive networks on multichannel time series for human activity recognition, IEEE Journal of Biomedical and Health Informatics 26 (10) (2022) 5223–5234.

[32] J. Ji, X. Xing, Y. Yao, J. Li, X. Zhang, Convolutional kernels with an element-wise weighting mechanism for identifying abnormal brain connectivity patterns, Pattern Recognition 109 (2021) 107570.

[33] J. F. A. Ronicko, J. Thomas, P. Thangavel, V. Koneru, G. Langs, J. Dauwels, Diagnostic classification of autism using resting-state fmri data improves with full correlation functional brain connectivity compared to partial correlation, Journal of Neuroscience Methods 345 (2020) 108884.

[34] R. M. Thomas, S. Gallo, L. Cerliani, P. Zhutovsky, A. El-Gazzar, G. Van Wingen, Classifying autism spectrum disorder using the temporal statistics of resting-state functional mri data with 3d convolutional neural networks, Frontiers in psychiatry 11 (2020) 440.

[35] D. Yao, M. Liu, M. Wang, C. Lian, J. Wei, L. Sun, J. Sui, D. Shen, Triplet graph convolutional network for multi-scale analysis of functional connectivity using functional mri, in: International Workshop on Graph Learning in Medical Imaging, Springer, 2019, pp. 70–78.

[36] L. Li, H. Jiang, G. Wen, P. Cao, M. Xu, X. Liu, J. Yang, O. Zaiane, Te-higcn: An ensemble of transfer hierarchical graph convolutional networks for disorder diagnosis, Neuroinformatics (2021) 1–23.

[37] W. Yang, G. Wen, P. Cao, J. Yang, O. R. Zaiane, Collaborative learning of graph generation, clustering and classification for brain networks diagnosis, Computer Methods and Programs in Biomedicine 219 (2022) 106772.

[38] J. Park, M. Lee, H. J. Chang, K. Lee, J. Y. Choi, Symmetric graph convolutional autoencoder for unsupervised graph representation learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6519–6528.

[39] Z. Hu, Y. Dong, K. Wang, K.-W. Chang, Y. Sun, Gpt-gnn: Generative pre-training of graph neural networks, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1857–1867.

[40] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, International Conference on Learning Representations (ICLR) (2016) 1–11.

[41] J. Kawahara, C. J. Brown, S. P. Miller, B. G. Booth, V. Chau, R. E. Grunau, J. G. Zwicker, G. Hamarneh, Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment, NeuroImage 146 (2017) 1038–1049.

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

[43] C. Craddock, Y. Benhajali, C. Chu, F. Chouinard, A. Evans, A. Jakab, B. S. Khundrakpam, J. D. Lewis, Q. Li, M. Milham, et al., The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives, Frontiers in Neuroinformatics 7 (2013).

[44] T. Xu, Z. Yang, L. Jiang, X.-X. Xing, X.-N. Zuo, A connectome computation system for discovery science of brain, Science Bulletin 60 (1) (2015) 86–95.

[45] C.-G. Yan, X.-D. Wang, X.-N. Zuo, Y.-F. Zang, Dpabi: data processing & analysis for (resting-state) brain imaging, Neuroinformatics 14 (3) (2016) 339–351.

[46] I. Mhiri, I. Rekik, Joint functional brain network atlas estimation and feature selection for neurological disorder diagnosis with application to autism, Medical image analysis 60 (2020) 101596.

[47] Y. Yan, J. Zhu, M. Duda, E. Solarz, C. Sripada, D. Koutra, Groupinn: Grouping-based interpretable neural network for classification of limited, noisy brain data, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 772–782.

[48] N. C. Dvornek, P. Ventola, K. A. Pelphrey, J. S. Duncan, Identifying autism from resting-state fmri using long short-term memory networks, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2017, pp. 362–370.

[49] T. Eslami, V. Mirjalili, A. Fong, A. R. Laird, F. Saeed, Asd-diagnet: a hybrid learning approach for detection of autism spectrum disorder using fmri data, Frontiers in neuroinformatics 13 (2019) 70.

[50] Z. Yao, B. Hu, Y. Xie, P. Moore, J. Zheng, A review of structural and functional brain networks: small world and atlas, Brain informatics 2 (1) (2015) 45–52.

[51] H. Zhao, D. R. Nyholt, Gene-based analyses reveal novel genetic overlap and allelic heterogeneity across five major psychiatric disorders, Human genetics 136 (2) (2017) 263–274.

[52] N. V. Radonjić, J. L. Hess, P. Rovira, O. Andreassen, J. K. Buitelaar, C. R. Ching, B. Franke, M. Hoogman, N. Jahanshad, C. McDonald, et al., Structural brain imaging studies offer clues about the effects of the shared genetic etiology among neuropsychiatric disorders, Molecular psychiatry 26 (6) (2021) 2101–2110.

[53] L. Dell'Osso, I. M. Cremone, G. Amatori, A. Cappelli, A. Cuomo, S. Barlati, G. Massimetti, A. Vita, A. Fagiolini, C. Carmassi, et al., Investigating the relationship between autistic traits, ruminative thinking, and suicidality in a clinical sample of subjects with bipolar disorder and borderline personality disorder, Brain Sciences 11 (5) (2021) 621.

[54] G. Joshi, J. Biederman, C. Petty, R. L. Goldin, S. L. Furtak, J. Wozniak, Examining the comorbidity of bipolar disorder and autism spectrum disorders: a large controlled analysis of phenotypic and familial correlates in a referred population of youth with bipolar i disorder with and without autism spectrum disorders, The Journal of clinical psychiatry 74 (6) (2013) 6865.

[55] L. Dell'Osso, B. Carpita, C. A. Bertelloni, E. Diadema, F. M. Barberi, C. Gesi, C. Carmassi, Subthreshold autism spectrum in bipolar disorder: prevalence and clinical correlates, Psychiatry research 281 (2019) 112605.

[56] J. V. Hull, L. B. Dokovna, Z. J. Jacokes, C. M. Torgerson, A. Irimia, J. D. Van Horn, Resting-state functional connectivity in autism spectrum disorders: a review, Frontiers in psychiatry 7 (2017) 205.

[57] S. Chen, Y. Xing, J. Kang, Latent and abnormal functional connectivity circuits in autism spectrum disorder, Frontiers in neuroscience 11 (2017) 125.

[58] L. Rabany, S. Brocke, V. D. Calhoun, B. Pittman, S. Corbera, B. E. Wexler, M. D. Bell, K. Pelphrey, G. D. Pearlson, M. Assaf, Dynamic functional connectivity in schizophrenia and autism spectrum disorder: Convergence, divergence and classification, NeuroImage: Clinical 24 (2019) 101966.

[59] J. M. Tyszka, D. P. Kennedy, L. K. Paul, R. Adolphs, Largely typical patterns of resting-state functional connectivity in high-functioning adults with autism, Cerebral cortex 24 (7) (2014) 1894–1905.

[60] Y. Jiang, M. Duan, X. Chen, X. Chang, H. He, Y. Li, C. Luo, D. Yao, Common and distinct dysfunctional patterns contribute to triple network model in schizophrenia and depression: a preliminary study, Progress in Neuro-Psychopharmacology and Biological Psychiatry 79 (2017) 302–310.

[61] I. I. Goldberg, M. Harel, R. Malach, When the brain loses its self: prefrontal inactivation during sensorimotor processing, Neuron 50 (2) (2006) 329–339.

[62] S. Ha, I.-J. Sohn, N. Kim, H. J. Sim, K.-A. Cheon, Characteristics of brains in autism spectrum disorder: structure, function and connectivity across the lifespan, Experimental neurobiology 24 (4) (2015) 273.