

# BrainHGT: A Hierarchical Graph Transformer for Interpretable Brain Network Analysis

Jiajun Ma, Yongchao Zhang, Chao Zhang, Zhao Lv, Shengbing Pei\*

Anhui Province Key Laboratory of Multimodal Cognitive Computation,  
School of Computer Science and Technology, Anhui University

e23301163@stu.ahu.edu.cn, e24301188@stu.ahu.edu.cn, iiphci\_ahu@163.com, kjlz@ahu.edu.cn, shengbingpei@ahu.edu.cn

## Abstract

Graph Transformer shows remarkable potential in brain network analysis due to its ability to model graph structures and complex node relationships. Most existing methods typically model the brain as a flat network, ignoring its modular structure, and their attention mechanisms treat all brain region connections equally, ignoring distance-related node connection patterns. However, brain information processing is a hierarchical process that involves local and long-range interactions between brain regions, interactions between regions and sub-functional modules, and interactions among functional modules themselves. This hierarchical interaction mechanism enables the brain to efficiently integrate local computations and global information flow, supporting the execution of complex cognitive functions. To address this issue, we propose BrainHGT, a hierarchical Graph Transformer that simulates the brain’s natural information processing from local regions to global communities. Specifically, we design a novel long-short range attention encoder that utilizes parallel pathways to handle dense local interactions and sparse long-range connections, thereby effectively alleviating the over-globalizing issue. To further capture the brain’s modular architecture, we design a prior-guided clustering module that utilizes a cross-attention mechanism to group brain regions into functional communities and leverage neuroanatomical prior to guide the clustering process, thereby improving the biological plausibility and interpretability. Experimental results indicate that our proposed method significantly improves performance of disease identification, and can reliably capture the sub-functional modules of the brain, demonstrating its interpretability.

**Code** — <https://github.com/null-cks/BrainHGT>

## Introduction

The diagnosis of neurological and psychiatric disorders, such as autism spectrum disorder (ASD) and Alzheimer’s disease (AD), faces a significant challenge due to the lack of objective and reliable biomarkers (Lord et al. 2020; Zhang et al. 2024c). Resting-state functional magnetic resonance imaging (fMRI) has emerged as an important neuroimaging tool to address this problem. The non-invasive technique

\*Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

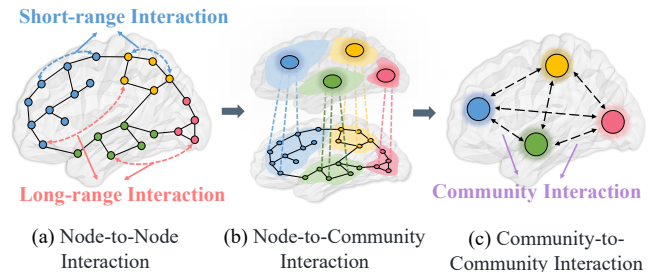


Figure 1: The hierarchical model of brain information processing, showing three levels of interaction.

measures blood-oxygen-level-dependent (BOLD) signals to construct the brain’s functional connectivity network, providing a unique insight into the brain’s intrinsic organization and, more critically, helping to identify biomarkers that are crucial for the diagnosis and characterization of these diseases (Li et al. 2019; Xu et al. 2024a; Kong et al. 2024).

For the brain network analysis, graph neural networks (GNNs) are a commonly used class of methods (Zhang and Hong 2024; Luo et al. 2024). By modeling brain regions of interest (ROIs) as nodes and their connections as edges, GNNs automatically learn discriminative features directly from complex brain network topologies, bypassing extensive and often subjective manual feature engineering (Zhang et al. 2022). Despite this advantage, GNNs are inherently limited by a receptive field confined to a local  $k$ -hop neighborhood, this make them ineffective at capturing the long-range dependencies that are crucial to the brain’s large-scale functional organization. In contrast, Graph Transformer (Cai, Gao, and Liu 2023; Song et al. 2024; Qu et al. 2024) can model interactions between any pair of ROIs via self-attention mechanism, naturally capturing the global long-range connectivity patterns inherent in brain networks.

However, the global self-attention of Graph Transformer brings new challenges. Recent studies reveal that standard Graph Transformer exhibits a tendency towards over-globalizing, where attention is indiscriminately distributed across all nodes (Xing et al. 2024). This mechanism is inconsistent with the small-world network properties of the brain, where connectivity is heavily biased towards local, dense connections. And the probability of connection decays

exponentially with distance, global communication is supplemented only by sparse yet critical long-range pathways (Bassett and Bullmore 2017; Bullmore and Sporns 2012). Consequently, the over-globalizing effect in Graph Transformer can dilute functionally relevant signals from nearby neighbors with irrelevant information from distant nodes, weakening the model’s representational capacity. To address this, we process local and global information in parallel, using a learnable decay mask to strengthen local interactions, this effectively mitigates the over-globalizing problem and produces richer ROI-level spatial representations.

A more critical issue is that prevailing methods treat the brain as a flat, non-hierarchical graph structure, neglecting its inherent modular properties. In reality, the functional organization of brain is not confined to singular inter-regional interactions, but relies on higher-level coordination between functional modules to achieve whole-brain functions (Sporns and Betzel 2016; Betzel and Bassett 2017). The brain’s information processing follows a natural hierarchical pattern. As shown in Figure 1, it begins with local and long-range interactions between brain regions, then organizes into large-scale networks (communities) with distinct functions, and finally achieves complex cognitive and behavioral functions through the dynamic coordination of these communities. Due to the lack of effective modeling for this hierarchical process, the decision-making process of current methods is often not comprehensive enough. Furthermore, the learned features typically lack clear neuroscientific meaning, severely hinders the method’s potential for clinical translation and scientific discovery. To capture this hierarchical structure, we propose a prior-guided clustering module that uses neuroanatomical knowledge to guide the integration of information from microscopic nodes to macroscopic communities, and subsequently models community-level interactions to improve the biological plausibility and interpretability of the learning process.

Here, we propose BrainHGT, a novel Graph Transformer framework that simulates hierarchical brain processing to enhance interpretability. The main contributions include:

- A long-short range attention (LSRA) module is designed to balance strengthening functionally relevant local interactions and capturing sparse critical global dependencies, thereby generating richer node representations.
- A prior-guided clustering module is designed to enhance model interpretability by using anatomical prior knowledge and cross-attention to dynamically group individual ROIs into functional communities, thereby effectively modeling the brain’s hierarchical interactions.
- Our method enables identification of disease biomarkers across multiple scales, including ROI-level connections, ROI-to-community affiliations, and inter-community coordination. This capability enhances diagnostic accuracy while ensuring a transparent decision-making process.

## Related Work

### ROI-Level Brain Network Analysis

ROI-level studies focus on the interaction patterns among whole-brain ROIs, which is a common paradigm in brain

network analysis. Cui et al. (Cui et al. 2023) introduced BrainGB, a comprehensive benchmark for GNN-based brain network analysis that evaluates the effectiveness of different node features, message-passing mechanisms, and pooling strategies. Zheng et al. (Zheng et al. 2024) developed BrianIB, which uses the information bottleneck principle to identify disease-related subgraphs, improving diagnostic accuracy and providing interpretable biomarkers. However, the local message-passing mechanism of GNNs limits their ability to capture long-range connections. In contrast, Graph Transformer can alleviate this issue by virtue of its global attention mechanism. Peng et al. (Peng et al. 2024) proposed GBT, which learns representative graph features by applying a low-rank approximation to the attention matrix. Yu et al. (Yu et al. 2024) proposed ALTER, which adopts random walk to better model long-range dependencies. However, the global attention in these methods often ignores the network’s topological structure, leading to an over-globalizing problem and making it difficult to balance the local and global characteristics of brain information processing.

### Modular-Aware Brain Network Analysis

Modular-aware studies focus on integrating the community structure of brain function, which corresponds to the sub-functional activities of the brain. This can be broadly classified into two categories. The first category of methods relies on predefined network groups to study functional community interactions. Shehzad et al. developed BrainGT and BrainDGT, which are based on voxel overlap to assign ROIs to functional networks (Shehzad et al. 2024, 2025). Similarly, Bannadabhavi et al. (Bannadabhavi et al. 2023) designed Com-BrainTF, which rearranges the functional connectivity matrix based on predefined community labels to learn connection patterns within and between communities. The main limitation of these methods lies in their reliance on predefined groups, which are typically derived from general anatomical or functional divisions. Such groups may fail to align accurately with the true functional communities specific to a given task or disease, thereby constraining the model’s flexibility. The second category of methods is dedicated to adaptively learning functional communities directly from data. Li et al. (Li et al. 2021) proposed BrainGNN, which utilizes a special graph convolution layer to learn soft membership scores for each brain region, building communities that are relevant to the classification task. Zhu et al. (Zhu et al. 2022) introduced M2CDCA, which utilizes spectral clustering and structural networks to guide community detection in functional networks. Kan et al. (Kan et al. 2022) proposed BRAINNETTF, which draws on the idea of deep embedding clustering, using a self-supervised soft clustering process to group functionally similar brain region nodes. Zhang et al. (Zhang et al. 2024a) proposed HFBN, which features a unique hierarchical node fusion method that adaptively merges fine-grained nodes into coarse-grained nodes layer by layer, thereby constructing a hierarchical brain network. Peng et al. (Peng et al. 2025) developed BioBGT, which uses the classic Louvain algorithm for unsupervised community detection and introduces a community contrastive learning strategy to optimize the

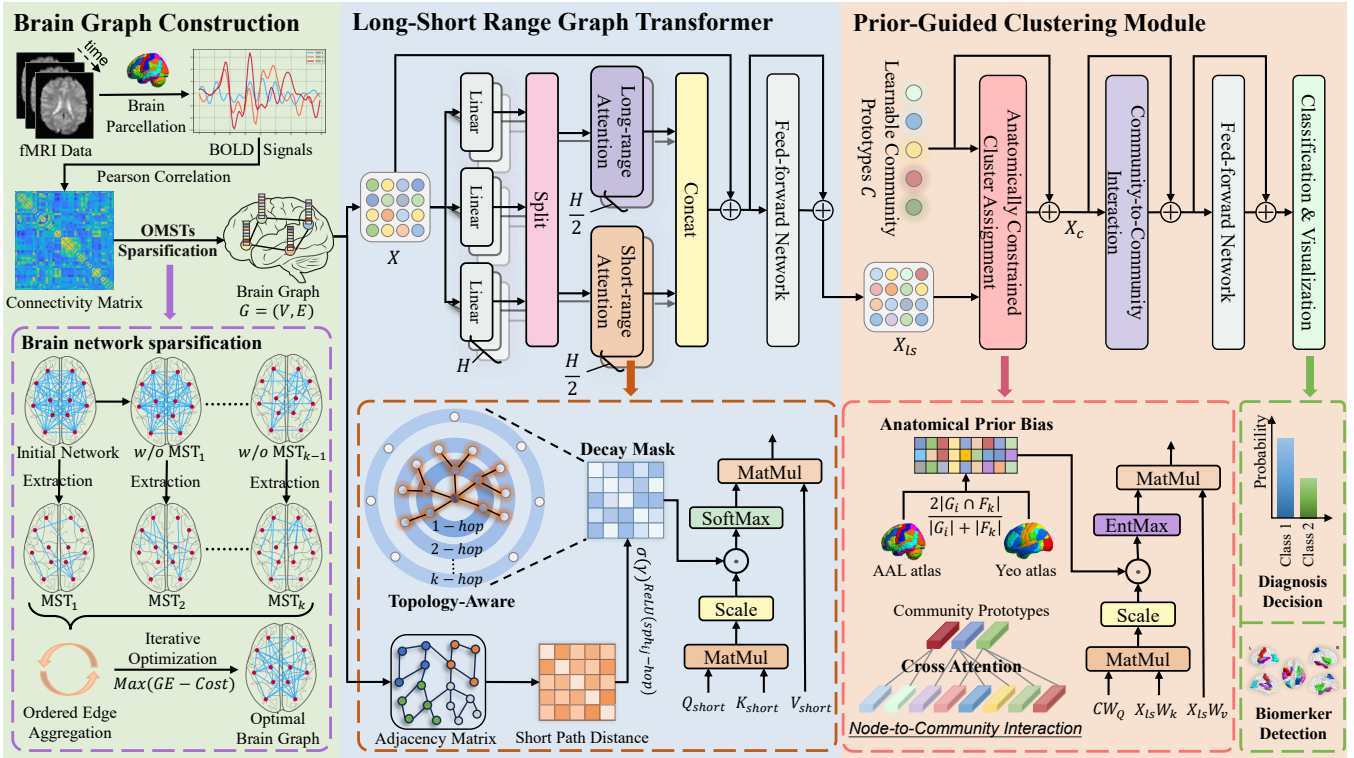


Figure 2: The overall framework of the proposed BrainHGT method. The model first learns multi-scale features of brain regions via a long-short range graph Transformer, and then aggregates these features into biologically plausible functional communities using a prior-guided clustering module for the final classification task.

partition by enhancing intra-module integration and inter-module segregation. Although these adaptive methods are more flexible, certain studies lack sufficient guidance from biological priors. This can lead to the learned community structures being unstable or their functional meaning being ambiguous, which poses a challenge to their interpretability in the field of neuroscience.

## Method

The overall framework of the proposed BrainHGT is presented in Figure 2. First, we construct a sparse and topologically efficient brain graph from fMRI data. Then, a long-short range attention encoder is designed to learn multi-scale node representations that align with brain topology. Finally, an interpretable prior-guided clustering module that aggregates node information into biologically meaningful functional communities to facilitate classification.

### Brain Graph Construction

For each subject, the preprocessed fMRI time series from  $N$  Regions of Interest (ROIs) are used to compute a Pearson correlation matrix  $\mathbf{R} \in \mathbb{R}^{N \times N}$ , representing the initial fully-connected functional connectivity network. To preserve essential functional connections while discarding redundant information, we employ the Orthogonal Minimal Spanning Trees algorithm (Dimitriadis et al. 2017) for network sparsification. This constructs a sparse and topo-

logically optimized brain graph that represents the backbone of the brain functional network. The process begins by converting the correlation matrix  $\mathbf{R}$  into an inverse weighted graph  $\mathbf{D}$ , where  $d_{ij} = 1/|r_{ij}|$  (for  $r_{ij} \neq 0$ ), to ensure that stronger functional connections correspond to shorter topological distances. Subsequently, a sequence of orthogonal minimal spanning trees (MSTs), namely  $\{\text{MST}_1, \text{MST}_2, \dots, \text{MST}_k\}$ , is iteratively extracted from  $\mathbf{D}$  using Kruskal’s algorithm. In each iteration  $i$ , the  $\text{MST}_i$  with the minimum total edge weight is identified, and its constituent edges are then excluded from the graph for all subsequent iterations to ensure orthogonality. Finally, the edges from this sequence of MSTs are aggregated one-by-one in their extraction order, forming a set of candidate networks. The final adjacency matrix  $\tilde{\mathbf{A}}$  is the specific candidate network  $\mathbf{A}$  that maximizes a global cost-efficiency metric:

$$\tilde{\mathbf{A}} = \arg \max_{\mathbf{A}} (\text{GE}(\mathbf{A}) - \text{Cost}(\mathbf{A})). \quad (1)$$

Here, Global Efficiency (GE) measures the network’s information transfer capability and is defined as the average inverse shortest path length:

$$\text{GE}(\mathbf{A}) = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{L_{ij}}, \quad (2)$$

where  $L_{ij}$  is the shortest path length between nodes  $i$  and  $j$  in the graph  $\mathbf{A}$ . Cost represents the wiring cost of the net-

work, defined as the ratio of the total weight of  $\mathbf{A}$  normalized by the total weights in the original dense graph:

$$\text{Cost}(\mathbf{A}) = \frac{\sum_{(i,j) \in E_k} |r_{ij}|}{\sum_{i < j} |r_{ij}|}, \quad (3)$$

where  $E_k$  is the set of edges in  $\mathbf{A}$ . This procedure yields a topologically optimized brain network that preserves core functional pathways while balancing efficiency and cost. As a result, we construct a brain graph  $G(V, E)$ , where the adjacency matrix is denoted as  $\tilde{\mathbf{A}} \in \mathbb{R}^{N \times N}$ , and the node features are represented by the correlation vectors from the original matrix  $\mathbf{R}$ .

### Long-Short Range Graph Transformer

The brain’s multi-scale organization features dense local processing within modules and sparse long-range connections for global integration. To learn node representations sensitive to this structure, we introduce the long-short range attention mechanism, designed to overcome the topology-agnostic limitations of the standard Transformer. LSRA strategically divides the  $H$  attention heads into two equal sets:  $H_{short} = H/2$  heads dedicated to modeling short-range, topologically-informed interactions, while  $H_{long} = H/2$  heads for capturing long-range, global dependencies.

Let  $\mathbf{X} \in \mathbb{R}^{N \times d}$  denote the input node feature matrix, where  $d$  is the feature dimension. Specifically,  $\mathbf{X}$  is derived from a linear mapping of  $\mathbf{R}$ . The input is projected into query ( $\mathbf{Q}$ ), key ( $\mathbf{K}$ ), and value ( $\mathbf{V}$ ) representations, each of which has a dimension of  $\mathbb{R}^{N \times d}$ . The query, key, and value vectors for the  $i$ -th node within head  $h$  are denoted as  $\mathbf{q}_i^{(h)}$ ,  $\mathbf{k}_i^{(h)}$ ,  $\mathbf{v}_i^{(h)}$  respectively, corresponding to the  $i$ -th rows of these matrices.

**Short-Range Attention:** To focus on interactions between topologically proximal brain regions, the attention scores are modulated by a Topological Decay Mask, derived from the shortest path length (SPL) matrix  $\mathbf{S} \in \mathbb{R}^{N \times N}$ . For each short-range head  $h_s = 1, \dots, H_{short}$ , we introduce learnable parameters, a hop threshold  $\text{hop}^{(h_s)}$  and a decay factor  $\gamma^{(h_s)}$ . The raw attention energy  $e_{ij}^{(h_s)}$  between node  $i$  and node  $j$  for head  $h_s$  is first computed:

$$e_{ij}^{(h_s)} = \frac{(\mathbf{q}_i^{(h_s)})^T \mathbf{k}_j^{(h_s)}}{\sqrt{d_{head}}}, \quad (4)$$

where  $d_{head}$  is the dimension of each attention head. This energy is then modulated by the topological decay:

$$\hat{e}_{ij}^{(h_s)} = e_{ij}^{(h_s)} \odot \left( \sigma(\gamma^{(h_s)}) \right)^{\text{ReLU}(S_{ij} - \text{hop}^{(h_s)})}, \quad (5)$$

where  $\sigma(\cdot)$  is the Sigmoid function constraining  $\gamma^{(h_s)}$  to  $(0, 1)$ ,  $S_{ij}$  is the SPL between node  $i$  and  $j$ ,  $\odot$  denotes the Hadamard product. This formulation ensures that attention to nodes beyond the  $\text{hop}^{(h_s)}$  distance decays exponentially, focusing the head on local information processing. The final output for head  $h_s$  is:

$$\mathbf{o}_i^{(h_s)} = \sum_{j=1}^N \text{Softmax}_j(\hat{e}_{ij}^{(h_s)}) \mathbf{v}_j^{(h_s)}. \quad (6)$$

**Long-Range Attention:** To capture dependencies between any pair of brain regions irrespective of distance, the long-range heads operate as standard scaled dot-product attention. This facilitates the modeling of global information integration. For each long-range head  $h_l = 1, \dots, H_{long}$ :

$$\mathbf{o}_i^{(h_l)} = \sum_{j=1}^N \text{Softmax}_j \left( \frac{(\mathbf{q}_i^{(h_l)})^T \mathbf{k}_j^{(h_l)}}{\sqrt{d_{head}}} \right) \mathbf{v}_j^{(h_l)}. \quad (7)$$

The outputs from all  $H$  heads, namely  $\{\mathbf{o}_i^{(h_s)}\}_{s=1}^{H_{short}}$  and  $\{\mathbf{o}_i^{(h_l)}\}_{l=1}^{H_{long}}$ , are concatenated and passed through a final linear projection  $\mathbf{W}_O \in \mathbb{R}^{d \times d}$  to produce the output of the LSRA layer:

$$\mathbf{X}_{ls}(i) = \mathbf{W}_O \left( \text{Concat}(\mathbf{o}_i^{(h_1)}, \dots, \mathbf{o}_i^{(h_H)}) \right). \quad (8)$$

This dual-branch design enables the encoder to concurrently process information at distinct spatial granularities, effectively capturing the multi-scale functional architecture inherent in brain networks.

### Prior-Guided Brain Region Clustering

To bridge the transition from node-level interactions to community-level insights and simulate brain’s hierarchical information processing, we introduce the prior-guided clustering module. This module groups the  $N$  brain regions into  $K$  latent functional communities in a data-driven yet biologically-informed manner. To ensure these communities are neuroscientifically meaningful, this process is guided by an explicit anatomical prior.

This process begins by predefining a set of learnable community prototypes  $\mathbf{C} \in \mathbb{R}^{K \times d}$ , where each vector  $\mathbf{c}_k$  represents the initial embedding of a functional community. Inspired by Kan et al (Kan et al. 2022), the community prototypes are initialized using the Xavier uniform method (Glorot and Bengio 2010) and then orthogonalized with the Gram-Schmidt process to enhance the spatial separability of node embeddings from different functional modules. To ground the clustering process in established neuroanatomy, we construct a Dice similarity prior matrix,  $\mathbf{D}_{\text{prior}} \in \mathbb{R}^{N \times K}$ , based on the spatial overlap between input  $N$  ROIs and the  $K$  canonical large-scale functional networks.

Let  $G_i$  be the set of voxels corresponding to the  $i$ -th ROI in atlas, and  $F_k$  be the set of voxels for the  $k$ -th functional network. The Dice coefficient between them is:

$$\mathbf{D}_{\text{prior}}(i, k) = \frac{2|G_i \cap F_k|}{|G_i| + |F_k|} \quad (9)$$

where  $|\cdot|$  denotes the number of voxels in a set. Each element  $(\mathbf{D}_{\text{prior}})_{ik}$  thus represents a similarity score between ROI  $i$  and functional network  $k$ . This matrix injects valuable biological knowledge, guiding the model to form communities that align with known functional systems.

**Anatomically Constrained Cluster Assignment:** The core of the module is a cross-attention mechanism that computes the soft assignment of nodes to communities. Here, the community prototypes  $\mathbf{C}$  act as queries, while the node

Type	Method	Dataset: ABIDE				Dataset: ADNI			
		ACC(%) $\uparrow$	AUC(%) $\uparrow$	SEN(%) $\uparrow$	SPE(%) $\uparrow$	ACC(%) $\uparrow$	AUC(%) $\uparrow$	SEN(%) $\uparrow$	SPE(%) $\uparrow$
ML Methods	SVM	60.0 $\pm$ 1.78	60.09 $\pm$ 1.88	63.15 $\pm$ 4.56	57.03 $\pm$ 5.49	67.72 $\pm$ 5.43	56.76 $\pm$ 3.88	85.16 $\pm$ 3.25	18.37 $\pm$ 8.75
	Random Forest	55.25 $\pm$ 2.71	55.39 $\pm$ 2.80	59.55 $\pm$ 5.74	51.23 $\pm$ 7.57	64.23 $\pm$ 2.41	53.74 $\pm$ 1.87	80.59 $\pm$ 5.46	16.89 $\pm$ 7.31
Graph Neural Networks	GAT	60.41 $\pm$ 2.50	59.40 $\pm$ 2.48	57.96 $\pm$ 11.73	59.43 $\pm$ 12.31	65.26 $\pm$ 3.66	64.22 $\pm$ 3.76	80.30 $\pm$ 9.92	49.16 $\pm$ 15.96
	BrainGNN	57.74 $\pm$ 4.40	62.10 $\pm$ 3.38	39.96 $\pm$ 24.04	67.23 $\pm$ 24.29	62.14 $\pm$ 3.09	65.12 $\pm$ 3.97	71.96 $\pm$ 16.05	41.24 $\pm$ 14.17
	IBGNN	59.85 $\pm$ 2.78	60.67 $\pm$ 3.27	64.56 $\pm$ 13.88	53.99 $\pm$ 16.25	57.56 $\pm$ 7.49	57.12 $\pm$ 4.93	64.91 $\pm$ 21.02	46.03 $\pm$ 21.21
	BrainGB	63.12 $\pm$ 3.24	66.32 $\pm$ 2.97	65.10 $\pm$ 12.01	60.40 $\pm$ 8.55	67.50 $\pm$ 3.10	70.81 $\pm$ 3.70	82.07 $\pm$ 9.21	43.40 $\pm$ 11.10
	BrainUSL	62.36 $\pm$ 2.73	62.80 $\pm$ 2.40	59.21 $\pm$ 10.03	66.34 $\pm$ 8.82	60.08 $\pm$ 3.32	49.64 $\pm$ 3.23	79.07 $\pm$ 9.34	20.20 $\pm$ 10.08
	BrainIB	59.47 $\pm$ 2.52	62.18 $\pm$ 2.38	64.98 $\pm$ 7.86	53.26 $\pm$ 6.64	65.73 $\pm$ 2.69	65.63 $\pm$ 4.15	75.37 $\pm$ 4.68	47.14 $\pm$ 6.54
Graph Transformer	VanillaTF	64.33 $\pm$ 2.70	72.97 $\pm$ 2.02	62.60 $\pm$ 11.30	66.78 $\pm$ 15.80	68.46 $\pm$ 2.54	72.39 $\pm$ 3.64	80.11 $\pm$ 14.26	<b>51.85<math>\pm</math>19.61</b>
	BrainNetTF	67.98 $\pm$ 3.63	75.63 $\pm$ 2.74	<u>70.31<math>\pm</math>15.69</u>	61.07 $\pm$ 20.43	71.46 $\pm$ 3.75	78.02 $\pm$ 2.92	83.15 $\pm$ 10.72	48.93 $\pm$ 16.91
	GBT	67.34 $\pm$ 2.37	75.67 $\pm$ 3.37	65.47 $\pm$ 13.83	<u>69.86<math>\pm</math>14.29</u>	72.20 $\pm$ 2.03	77.64 $\pm$ 1.50	85.19 $\pm$ 4.90	37.50 $\pm$ 9.62
	RGTNet	63.35 $\pm$ 3.06	68.42 $\pm$ 3.50	63.83 $\pm$ 10.0	62.88 $\pm$ 7.89	66.59 $\pm$ 2.63	64.96 $\pm$ 4.65	76.48 $\pm$ 3.41	47.50 $\pm$ 8.60
	Contrasformer	61.70 $\pm$ 3.18	66.75 $\pm$ 3.30	54.95 $\pm$ 10.37	68.11 $\pm$ 8.20	67.80 $\pm$ 1.90	70.60 $\pm$ 2.97	70.50 $\pm$ 11.60	39.42 $\pm$ 12.85
	ALTER	<u>68.87<math>\pm</math>1.89</u>	<u>75.80<math>\pm</math>2.63</u>	69.98 $\pm$ 6.56	67.40 $\pm$ 6.81	<u>73.29<math>\pm</math>2.41</u>	<u>79.56<math>\pm</math>2.72</u>	82.78 $\pm$ 9.37	45.20 $\pm$ 21.01
	BioBGT	63.97 $\pm$ 1.74	62.60 $\pm$ 1.77	60.91 $\pm$ 6.01	65.71 $\pm$ 5.49	68.90 $\pm$ 1.84	59.48 $\pm$ 4.50	78.70 $\pm$ 15.62	38.21 $\pm$ 14.23
Ours	BrainHGT	<b>71.33<math>\pm</math>1.91</b>	<b>76.65<math>\pm</math>1.45</b>	<b>72.40<math>\pm</math>3.96</b>	<b>69.93<math>\pm</math>2.45</b>	<b>74.27<math>\pm</math>1.76</b>	<b>80.84<math>\pm</math>1.38</b>	<b>87.22<math>\pm</math>5.52</b>	<u>49.29<math>\pm</math>8.42</u>

Table 1: Performance comparison with existing methods on two datasets (Mean  $\pm$  standard deviation). **Bold** indicates the best results and underlining signifies second outcomes.

features  $\mathbf{X}_{ls} \in \mathbb{R}^{N \times d}$  from the LSRA encoder serve as keys and values. The attention energy between community  $k$  and ROI  $i$  is modulated by the Dice prior.

$$\hat{e}_{ki} = \frac{(\mathbf{c}_k \mathbf{W}_Q)((\mathbf{x}_{ls})_i \mathbf{W}_K)^T}{\sqrt{d_{head}}} \odot (\mathbf{D}_{\text{prior}})_{ik} \quad (10)$$

where  $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d \times d}$  are learnable projection matrices and  $d_{head}$  is the dimension of each attention head. This step ensures that a high feature similarity ( $\hat{e}_{ki}$ ) is amplified if it corresponds to a high spatial overlap ( $(\mathbf{D}_{\text{prior}})_{ik}$ ), and suppressed otherwise. To obtain a sparse and more definitive assignment distribution, we apply the Entmax function transformation to the biased scores. The resulting attention weights form the soft assignment matrix  $\mathbf{P} \in \mathbb{R}^{N \times K}$ :

$$\mathbf{P}[i, :] = \text{Entmax}((\hat{\mathbf{e}}[:, i])^T) \quad (11)$$

The updated representation for each community,  $\hat{\mathbf{c}}_k$ , is then computed by taking the weighted sum of node value vectors:  $\hat{\mathbf{c}}_k = \sum_{i=1}^N \mathbf{P}_{ik}((\mathbf{x}_{ls})_i \mathbf{W}_V)$ , where  $\mathbf{W}_V \in \mathbb{R}^{d \times d}$  is learnable projection matrices.

**Refinement of Community Interaction.** To further model community-level coordination pattern, these aggregated community representations  $\mathbf{X}_c = \{\hat{\mathbf{c}}_k\}_{k=1}^K \in \mathbb{R}^{K \times d}$  are then refined through a self-attention layer followed by a feed-forward network. This crucial step models inter-community dynamics, allowing the framework to capture complex relationships between the identified functional networks. These hierarchical-aware community representations are then aggregated via mean pooling and passed through a 256-32-2 multilayer perceptron for final classification.

By utilizing a hierarchical architecture that first learns multi-scale node representations and then aggregates them into meaningful functional communities. This multi-stage process produces a final representation that is optimized for classification while remaining highly interpretable.

## Experiments and Results

### Materials

We evaluate our proposed method on two commonly used public fMRI datasets.

**ABIDE Dataset.** The ABIDE dataset is a database for ASD research, comprising data from 17 international sites (Craddock et al. 2013; Liu et al. 2024). In total, 1009 subjects are utilized, including 516 individuals diagnosed with ASD and 493 normal controls (NC). The brain is parcellated into 200 ROIs using the Craddock 200 (CC200) atlas (Craddock et al. 2012).

**ADNI Dataset.** The ADNI dataset is a longitudinal data study focused on discovering early biomarkers for AD (Petersen et al. 2010; Zhang et al. 2024b). A cohort of 410 subjects is utilized, including 263 patients with mild cognitive impairment (MCI) and 147 normal controls, this carefully matches for both age and sex ratio. The Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer et al. 2002) is employed to define 90 ROIs for the brain of each subject.

The preprocessing of fMRI data follows the standardized pipeline (Shehzad et al. 2015; Yan et al. 2016; Esteban et al. 2019). The anatomical prior serves as a key component of the Prior-Guided Brain Region Clustering module. Here, we utilize the Yeo 7-network atlas (Yeo et al. 2011) to cluster the ROIs into eight sub-networks as a prior, they are Default Mode Network (DMN), Frontoparietal Network (FPN), Limbic Network (LN), Ventral Attention Network (VAN), Dorsal Attention Network (DAN), Somatomotor Network (SMN), Visual Network (VN), and Cerebellum and Subcortical Structures (CB & SC).

### Experimental Settings

All experiments were conducted on a GeForce GTX 3080 Ti GPU, with implementation in PyTorch v1.12.1. For training, we employed the Adam optimizer with an initial learning rate of 1e-4 and a weight decay of 1e-4. The model was

Datasets	Component	ACC(%) $\uparrow$	AUC(%) $\uparrow$	SEN(%) $\uparrow$	SPE(%) $\uparrow$
ABIDE	w/o LSRA	69.99 $\pm$ 3.06	75.29 $\pm$ 3.17	71.64 $\pm$ 3.95	68.0 $\pm$ 5.01
	w/o $D_{\text{prior}}$	70.20 $\pm$ 1.70	75.78 $\pm$ 1.58	70.99 $\pm$ 4.66	69.16 $\pm$ 5.06
	w/o Entmax	70.0 $\pm$ 2.16	75.73 $\pm$ 1.93	71.95 $\pm$ 4.70	67.47 $\pm$ 4.83
	w/o Clustering	65.02 $\pm$ 1.69	69.85 $\pm$ 1.83	70.27 $\pm$ 11.6	58.45 $\pm$ 11.6
	BrainHGT	<b>71.33<math>\pm</math>1.91</b>	<b>76.65<math>\pm</math>1.45</b>	<b>72.40<math>\pm</math>3.96</b>	<b>69.93<math>\pm</math>2.45</b>
ADNI	w/o LSRA	72.80 $\pm$ 1.89	79.12 $\pm$ 1.85	81.74 $\pm$ 3.71	47.86 $\pm$ 3.98
	w/o $D_{\text{prior}}$	73.17 $\pm$ 2.44	80.21 $\pm$ 2.27	83.51 $\pm$ 5.33	53.21 $\pm$ 9.90
	w/o Entmax	72.20 $\pm$ 2.30	79.65 $\pm$ 2.37	80.81 $\pm$ 3.49	47.86 $\pm$ 5.80
	w/o Clustering	69.88 $\pm$ 3.41	73.55 $\pm$ 4.23	76.48 $\pm$ 4.27	<b>57.14<math>\pm</math>10.71</b>
	BrainHGT	<b>74.27<math>\pm</math>1.76</b>	<b>80.84<math>\pm</math>1.38</b>	<b>87.22<math>\pm</math>5.52</b>	49.29 $\pm$ 8.42

Table 2: Ablation Studies of different components on two disease classification tasks. **Bold** indicates the best results.

trained for 100 epochs with a batch size of 32, and the checkpoint that achieved the highest Area Under the Receiver Operating Characteristic Curve (AUC) on the validation set was selected for the final performance evaluation on the test set.

For each dataset, we followed a consistent evaluation protocol by randomly splitting the data into 70% for training, 10% for validation, and 20% for testing. This process was repeated 10 times, with the results averaged to ensure robustness and reliability.

### Comparison with Existing Methods

We compare our BrainHGT with three types of methods. 1) typical machine learning (ML) methods, including SVM (with a linear kernel) and Random Forest; 2) GNN-based methods, including GAT (Velickovic et al. 2017), BrainGNN (Li et al. 2021), IBGNN (Cui et al. 2022), BrainGB (Cui et al. 2023), BrainUSL (Zhang et al. 2023) and BrainIB (Zheng et al. 2024). 3) Graph Transformer-based methods, including vanillaTF (Kan et al. 2022), BrainNetTF (Kan et al. 2022), GBT (Peng et al. 2024), ALTER (Yu et al. 2024), RGTNet (Wang et al. 2024), Contrasformer (Xu et al. 2024b) and BioBGT (Peng et al. 2025). For these models, we used the official open-source code and its default configuration.

Table 1 summarizes the experimental results on the two datasets, using Accuracy (ACC), AUC, Sensitivity (SEN), and Specificity (SPE) as the primary metrics. As can be seen, our method achieves the best performance across most metrics on both datasets, which strongly validates that simulating the brain’s hierarchical functional structure is a more effective strategy for learning discriminative features for disease classification.

In addition, deep learning methods that model brain connectome topology consistently outperform traditional methods, highlighting the importance of learning features directly from the network structure. And Graph Transformer-based methods generally surpass those GNN-based methods by overcoming the latter’s reliance on local message passing to better capture global dependencies.

### Ablation Studies

We conduct ablation studies on two datasets to validate the effectiveness of the designed components in BrainHGT.

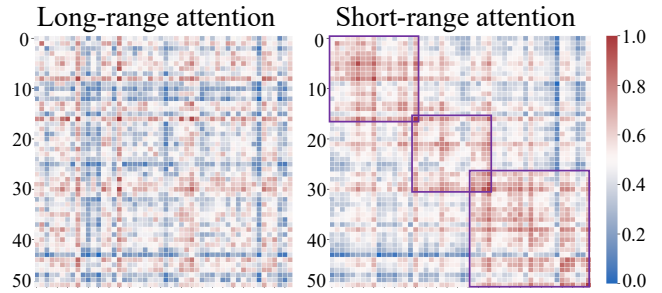


Figure 3: Visualizing attention score from the long-short range attention module.

As shown in Table 2, replacing our long-short range graph Transformer with a standard Transformer (w/o LSRA) degrades performance, demonstrating the crucial role of LSRA in balancing local and global information processing. Removing the anatomical constraint (w/o  $D_{\text{prior}}$ ) or replacing Entmax with Softmax for generating soft cluster assignments (w/o Entmax) also leads to a decline in performance, which confirms the importance of biological guidance and sparse assignment strategies. Notably, the most significant performance drop occurs when the entire hierarchical clustering module is removed (w/o Clustering), strongly indicating that modeling the brain’s hierarchical functional structure is central to our model’s superior performance.

### Interpretability and Visualization

To validate the interpretability of our method, we analyze correctly classified samples from test set to identify potential biomarkers and visualize its internal decision-making process. Without loss of generality, we use the ABIDE dataset for illustration.

**Analysis of ROI-level Interaction Patterns.** Figure 3 visualizes the long-range and short-range attention from the LSRA module. The long-range matrix shows a diffuse pattern to capture global dependencies, while the short-range matrix has a clear block or banded structure focused on local interactions. This indicates that the short-range attention is constrained by the topological decay mask, whereas the long-range attention operates free from spatial restrictions. This observation verifies the necessity of LSRA in segregating information processing across different spatial scales.

**Visualization of Learned Functional Communities.** To evaluate the biological plausibility of the learned functional communities, we first extracted the soft assignment matrix from the prior-guided clustering module, then assigned each ROI to its highest-probability community for visualization. As shown in Figure 4 (a), the resulting functional communities demonstrate high spatial correspondence with the standard Yeo-7 atlas, intuitively confirming their biological plausibility. However, this correspondence is not a simple replication of the anatomical prior but performs a data-driven refinement. As shown in Figure 4 (b), the learned clustering assignment is similar to the Dice prior matrix while incorporating disease-relevant characteristics. The model trans-

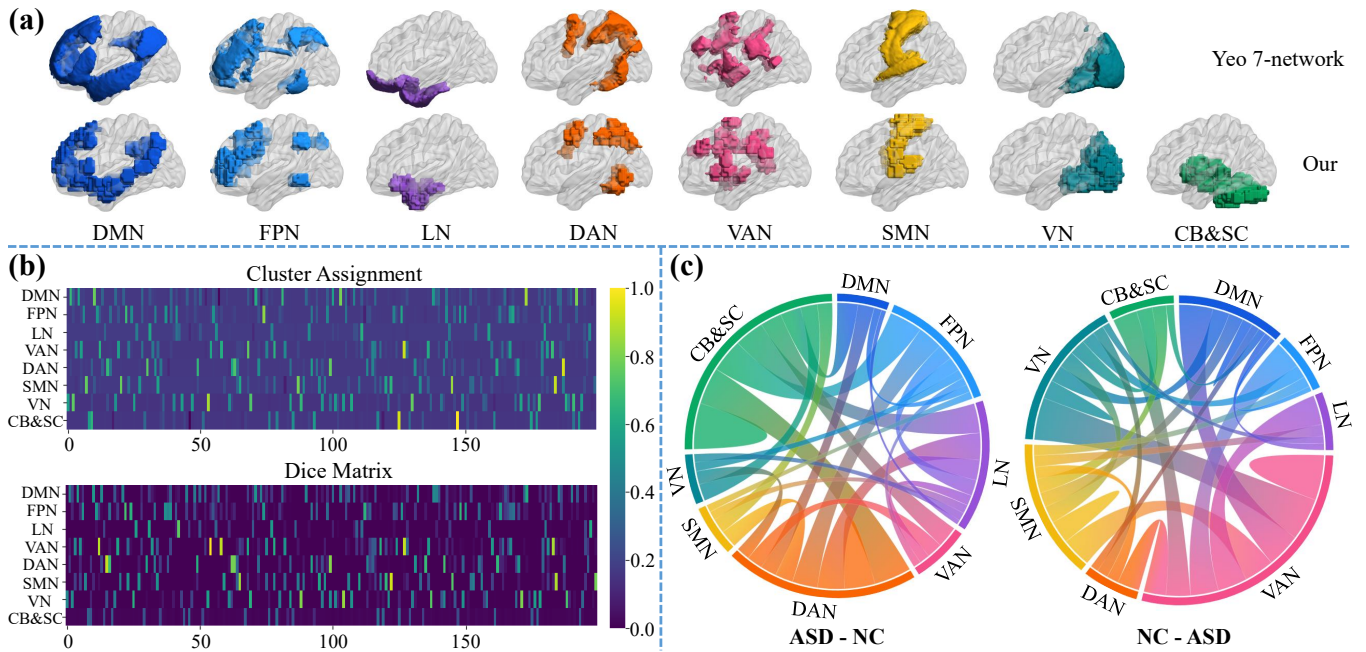


Figure 4: (a) Comparison of our learned functional communities with the standard Yeo 7-network. (b) Visualizing soft cluster assignment matrix and Dice prior matrix. (c) Differential community interactions highlight hyper-connectivity (left) and hypo-connectivity (right) in ASD.

Datasets	Atlas	ACC(%) $\uparrow$	AUC(%) $\uparrow$	SEN(%) $\uparrow$	SPE(%) $\uparrow$
ABIDE	Yeo 7-network	71.33 $\pm$ 1.91	76.65 $\pm$ 1.45	72.40 $\pm$ 3.96	69.93 $\pm$ 2.45
	Yeo 17-network	70.10 $\pm$ 1.38	75.54 $\pm$ 1.95	71.12 $\pm$ 2.57	68.73 $\pm$ 3.97
ADNI	Yeo 7-network	74.27 $\pm$ 1.76	<b>80.84<math>\pm</math>1.38</b>	87.22 $\pm$ 5.52	<b>49.29<math>\pm</math>8.42</b>
	Yeo 17-network	<b>74.76<math>\pm</math>1.54</b>	80.63 $\pm$ 2.51	<b>88.70<math>\pm</math>4.18</b>	47.86 $\pm$ 5.35

Table 3: Model performance comparison using different atlas as the anatomical prior. **Bold** indicates the best results.

forms the rigid anatomical constraints into a soft assignment relationship where each ROI can contribute to multiple communities to varying degrees. This flexibility is key to capturing the brain’s functional complexity and achieving an adaptive partitioning of the connectome.

**Identifying Disease-Relevant Communities.** To identify community-level biomarkers, we analyzed the differential interaction patterns between the ASD and NC groups by subtracting their average inter-community attention weights. As shown in Figure 4 (c), our model discovered patterns of aberrant connectivity that are highly consistent with the neuroscience literature. It revealed significant hyper-connectivity between the cerebellum and cortical association networks like the LN and DAN, consistent with extensive findings of abnormal cerebellar-cortical circuits in ASD (D’Mello and Stoodley 2015; Cerliani et al. 2015). Concurrently, the model identified prominent hypo-connectivity primarily on the VAN, with decreased interaction with VN and DMN, aligning with previously reported findings (Duan et al. 2017; Anteraper et al. 2020). This automated discov-

ery of clinically-relevant biomarkers highlights our model’s potential for understanding complex neurological disorders.

### Influence of Anatomical Prior Atlases

To evaluate the model’s robustness to the choice of anatomical prior, we evaluated its performance using both the Yeo 7-network and the more fine-grained Yeo 17-network atlases (Yeo et al. 2011). As shown in Table 3, both atlases yield improved performance, confirming the model’s robustness to atlas selection. In addition, the coarser 7-network atlas yields superior performance on the ABIDE dataset, whereas results are largely comparable on the ADNI dataset, this suggests the optimal prior granularity may be task-dependent.

### Conclusion

In this work, we propose a novel hierarchical Graph Transformer framework that enhances brain network analysis by simulating the brain’s natural information processing from local regions to global communities. The long-short range attention module mitigates the over-globalizing problem by processing dense local interactions and sparse global connections in parallel, yielding richer node representations. The prior-guided clustering module leverages anatomical prior knowledge to guide the dynamic partitioning of functional communities while modeling hierarchical interaction, significantly improving the model’s biological plausibility and interpretability. Our method not only achieves superior performance but also reliably identifies key biomarkers associated with neurological disorders.

## Acknowledgments

This work was supported in part by the Natural Science Research Project of Anhui Educational Committee (2023AH050081), the National Natural Science Foundation of China (No. 62476004), the Collaborative Innovation Project of Key Laboratory of Philosophy and Social Sciences in Anhui Province (HZ2302), the Scientific Research Project of Colleges and Universities in Anhui Province (2024AH040115) and the Anhui Province Science and Technology Innovation and Tackling Key Problems Project (202423k09020041).

## References

- Anteraper, S. A.; Guell, X.; Hollinshead, M. O.; D’Mello, A.; Whitfield-Gabrieli, S.; Biederman, J.; and Joshi, G. 2020. Functional alterations associated with structural abnormalities in adults with high-functioning autism spectrum disorder. *Brain Connectivity*, 10(7): 368–376.
- Bannadabhavi, A.; Lee, S.; Deng, W.; Ying, R.; and Li, X. 2023. Community-aware transformer for autism prediction in fmri connectome. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 287–297. Springer.
- Bassett, D. S.; and Bullmore, E. T. 2017. Small-world brain networks revisited. *The Neuroscientist*, 23(5): 499–516.
- Betzal, R. F.; and Bassett, D. S. 2017. Multi-scale brain networks. *Neuroimage*, 160: 73–83.
- Bullmore, E.; and Sporns, O. 2012. The economy of brain network organization. *Nature Reviews Neuroscience*, 13(5): 336–349.
- Cai, H.; Gao, Y.; and Liu, M. 2023. Graph Transformer Geometric Learning of Brain Networks Using Multimodal MR Images for Brain Age Estimation. *IEEE Transactions on Medical Imaging*, 42(2): 456–466.
- Cerliani, L.; Mennes, M.; Thomas, R. M.; Di Martino, A.; Thioux, M.; and Keyers, C. 2015. Increased functional connectivity between subcortical and cortical resting-state networks in autism spectrum disorder. *JAMA Psychiatry*, 72(8): 767–777.
- Craddock, C.; Benhajali, Y.; Chu, C.; Chouinard, F.; Evans, A.; Jakab, A.; Khundrakpam, B. S.; Lewis, J. D.; Li, Q.; Milham, M.; et al. 2013. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, 7(27): 5.
- Craddock, R. C.; James, G. A.; Holtzheimer III, P. E.; Hu, X. P.; and Mayberg, H. S. 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, 33(8): 1914–1928.
- Cui, H.; Dai, W.; Zhu, Y.; Kan, X.; Gu, A. A. C.; Lukemire, J.; Zhan, L.; He, L.; Guo, Y.; and Yang, C. 2023. BrainGB: A Benchmark for Brain Network Analysis With Graph Neural Networks. *IEEE Transactions on Medical Imaging*, 42(2): 493–506.
- Cui, H.; Dai, W.; Zhu, Y.; Li, X.; He, L.; and Yang, C. 2022. Interpretable graph neural networks for connectome-based brain disorder analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 375–385. Springer.
- Dimitriadis, S. I.; Salis, C.; Tarnanas, I.; and Linden, D. E. 2017. Topological filtering of dynamic functional brain networks unfolds informative chronnectomics: a novel data-driven thresholding scheme based on orthogonal minimal spanning trees (OMSTs). *Frontiers in Neuroinformatics*, 11: 28.
- D’Mello, A. M.; and Stoodley, C. J. 2015. Cerebro-cerebellar circuits in autism spectrum disorder. *Frontiers in neuroscience*, 9: 408.
- Duan, X.; Chen, H.; He, C.; Long, Z.; Guo, X.; Zhou, Y.; Uddin, L. Q.; and Chen, H. 2017. Resting-state functional under-connectivity within and between large-scale cortical networks across three low-frequency bands in adolescents with autism. *Progress in Neuro-psychopharmacology and Biological Psychiatry*, 79: 434–441.
- Esteban, O.; Markiewicz, C. J.; Blair, R. W.; Moodie, C. A.; Isik, A. I.; Erramuzpe, A.; Kent, J. D.; Goncalves, M.; DuPre, E.; Snyder, M.; et al. 2019. fMRIPrep: a robust pre-processing pipeline for functional MRI. *Nature methods*, 16(1): 111–116.
- Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256. JMLR Workshop and Conference Proceedings.
- Kan, X.; Dai, W.; Cui, H.; Zhang, Z.; Guo, Y.; and Yang, C. 2022. Brain network transformer. *Advances in Neural Information Processing Systems*, 35: 25586–25599.
- Kong, Y.; Zhang, X.; Wang, W.; Zhou, Y.; Li, Y.; and Yuan, Y. 2024. Multi-Scale Spatial-Temporal Attention Networks for Functional Connectome Classification. *IEEE Transactions on Medical Imaging*, 1–1.
- Li, X.; Dvornek, N. C.; Zhou, Y.; Zhuang, J.; Ventola, P.; and Duncan, J. S. 2019. Graph neural network for interpreting task-fMRI biomarkers. In *International conference on medical image computing and computer-assisted intervention*, 485–493. Springer.
- Li, X.; Zhou, Y.; Dvornek, N.; Zhang, M.; Gao, S.; Zhuang, J.; Scheinost, D.; Staib, L. H.; Ventola, P.; and Duncan, J. S. 2021. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 74: 102233.
- Liu, M.; Zhang, H.; Liu, M.; Chen, D.; Zhuang, Z.; Wang, X.; Zhang, L.; Peng, D.; and Wang, Q. 2024. Randomizing human brain function representation for brain disease diagnosis. *IEEE Transactions on Medical Imaging*, 43(7): 2537–2546.
- Lord, C.; Brugha, T. S.; Charman, T.; Cusack, J.; Dumas, G.; Frazier, T.; Jones, E. J.; Jones, R. M.; Pickles, A.; State, M. W.; et al. 2020. Autism spectrum disorder. *Nature reviews Disease primers*, 6(1): 1–23.
- Luo, X.; Wu, J.; Yang, J.; Xue, S.; Beheshti, A.; Sheng, Q. Z.; McAlpine, D.; Sowman, P.; Giral, A.; and Yu, P. S. 2024. Graph Neural Networks for Brain Graph Learning: A Survey. In *Proceedings of the Thirty-Third International*

- Joint Conference on Artificial Intelligence, IJCAI-24, 8170–8178.
- Peng, C.; Huang, Y.; Dong, Q.; Yu, S.; Xia, F.; Zhang, C.; and Jin, Y. 2025. Biologically plausible brain graph transformer. *arXiv preprint arXiv:2502.08958*.
- Peng, Z.; He, Z.; Jiang, Y.; Wang, P.; and Yuan, Y. 2024. GBT: Geometric-Oriented Brain Transformer for Autism Diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 142–152. Springer.
- Petersen, R. C.; Aisen, P. S.; Beckett, L. A.; Donohue, M. C.; Gamst, A. C.; Harvey, D. J.; Jack Jr, C.; Jagust, W. J.; Shaw, L. M.; Toga, A. W.; et al. 2010. Alzheimer’s disease Neuroimaging Initiative (ADNI) clinical characterization. *Neurology*, 74(3): 201–209.
- Qu, G.; Orlichenko, A.; Wang, J.; Zhang, G.; Xiao, L.; Zhang, K.; Wilson, T. W.; Stephen, J. M.; Calhoun, V. D.; and Wang, Y.-P. 2024. Interpretable Cognitive Ability Prediction: A Comprehensive Gated Graph Transformer Framework for Analyzing Functional Brain Networks. *IEEE Transactions on Medical Imaging*, 43(4): 1568–1578.
- Shehzad, A.; Yu, S.; Zhang, D.; Abid, S.; Cheng, X.; Zhou, J.; and Xia, F. 2024. BrainGT: Multifunctional Brain Graph Transformer for Brain Disorder Diagnosis. *medRxiv*, 2024–08.
- Shehzad, A.; Zhang, D.; Yu, S.; Abid, S.; and Xia, F. 2025. Dynamic Graph Transformer for Brain Disorder Diagnosis. *IEEE Journal of Biomedical and Health Informatics*.
- Shehzad, Z.; Giavasis, S.; Li, Q.; Benhajali, Y.; Yan, C.; Yang, Z.; Milham, M.; Bellec, P.; and Craddock, C. 2015. The Preprocessed Connectomes Project Quality Assessment Protocol—a resource for measuring the quality of MRI data. *Frontiers in Neuroscience*, 9(10.3389).
- Song, X.; Shu, K.; Yang, P.; Zhao, C.; Zhou, F.; Frangi, A. F.; Xiao, X.; Dong, L.; Wang, T.; Wang, S.; et al. 2024. Knowledge-aware Multisite Adaptive Graph Transformer for Brain Disorder Diagnosis. *IEEE transactions on medical imaging*.
- Sporns, O.; and Betzel, R. F. 2016. Modular brain networks. *Annual Review of Psychology*, 67(1): 613–640.
- Tzourio-Mazoyer, N.; Landeau, B.; Papathanassiou, D.; Crivello, F.; Etard, O.; Delcroix, N.; Mazoyer, B.; and Joliot, M. 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1): 273–289.
- Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *stat*, 1050(20): 10–48550.
- Wang, Y.; Long, H.; Bo, T.; and Zheng, J. 2024. Residual graph transformer for autism spectrum disorder prediction. *Computer Methods and Programs in Biomedicine*, 247: 108065.
- Xing, Y.; Wang, X.; Li, Y.; Huang, H.; and Shi, C. 2024. Less is more: on the over-globalizing problem in graph transformers. *arXiv preprint arXiv:2405.01102*.
- Xu, J.; Bian, Q.; Li, X.; Zhang, A.; Ke, Y.; Qiao, M.; Zhang, W.; Sim, W. K. J.; and Gulyás, B. 2024a. Contrastive graph pooling for explainable classification of brain networks. *IEEE Transactions on Medical Imaging*, 43(9): 3292–3305.
- Xu, J.; He, K.; Lan, M.; Bian, Q.; Li, W.; Li, T.; Ke, Y.; and Qiao, M. 2024b. Contrasformer: A Brain Network Contrastive Transformer for Neurodegenerative Condition Identification. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2671–2681.
- Yan, C.-G.; Wang, X.-D.; Zuo, X.-N.; and Zang, Y.-F. 2016. DPABI: data processing & analysis for (resting-state) brain imaging. *Neuroinformatics*, 14: 339–351.
- Yeo, B. T.; Krienen, F. M.; Sepulcre, J.; Sabuncu, M. R.; Lashkari, D.; Hollinshead, M.; Roffman, J. L.; Smoller, J. W.; Zöllei, L.; Polimeni, J. R.; et al. 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*.
- Yu, S.; Jin, S.; Li, M.; Sarwar, T.; and Xia, F. 2024. Long-range Brain Graph Transformer. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Zhang, H.; Song, R.; Wang, L.; Zhang, L.; Wang, D.; Wang, C.; and Zhang, W. 2022. Classification of brain disorders in rs-fMRI via local-to-global graph neural networks. *IEEE transactions on medical imaging*, 42(2): 444–455.
- Zhang, J.; Guo, Y.; Zhou, L.; Wang, L.; Wu, W.; and Shen, D. 2024a. Constructing hierarchical attentive functional brain networks for early AD diagnosis. *Medical Image Analysis*, 94: 103137.
- Zhang, J.; Wu, X.; Tang, X.; Zhou, L.; Wang, L.; Wu, W.; and Shen, D. 2024b. Asynchronous functional brain network construction with spatiotemporal transformer for mci classification. *IEEE Transactions on Medical Imaging*.
- Zhang, P.; Wen, G.; Cao, P.; Yang, J.; Zhang, J.; Zhang, X.; Zhu, X.; Zaiane, O. R.; and Wang, F. 2023. BrainUSL: Unsupervised Graph Structure Learning for Functional Brain Network Analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 205–214. Springer.
- Zhang, W.; and Hong, Q. 2024. Modeling brain functional networks using graph neural networks: A review and clinical application. *ICCK Transactions on Intelligent Systematics*, 1(2): 58–68.
- Zhang, Z.; Liu, X.; Zhang, S.; Song, Z.; Lu, K.; and Yang, W. 2024c. A review and analysis of key biomarkers in Alzheimer’s disease. *Frontiers in Neuroscience*, 18: 1358998.
- Zheng, K.; Yu, S.; Li, B.; Jenssen, R.; and Chen, B. 2024. BrainIB: Interpretable Brain Network-Based Psychiatric Diagnosis With Graph Information Bottleneck. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.
- Zhu, Q.; Yang, J.; Wang, S.; Zhang, D.; and Zhang, Z. 2022. Multi-modal non-euclidean brain network analysis with community detection and convolutional autoencoder. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 7(2): 436–446.