Information Bottleneck-Guided Adaptive Hypergraph Transformer for Brain Disease Diagnosis

Jingxi Feng[1] **Xudong Chen**[1] **Yifan Zhang**[1] **Heming Xu**[1]
Hongcheng Han[1] **Xijing Wang**[1] **Dong Zhang**[1] **Shaoyi Du**[1*]
[1]State Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
Institute of Artificial Intelligence and Robotics
Xi'an Jiaotong University, Xi'an 710049, China
jingxifeng8@gmail.com, chenxudongandying@gmail.com
dxny2nd@gmail.com, xuheming@stu.xjtu.edu.cn
hanhc@stu.xjtu.edu.cn, xjwang.2008@xjtu.edu.cn
dongzhang@xjtu.edu.cn, dushaoyi@xjtu.edu.cn

Abstract

Exploring high-order correlations and long-range dependencies in brain networks holds significant value for both neuroscience research and clinical diagnosis. However, previous studies have lacked a unified integration of high-order and long-range dependency information in brain networks, and there is substantial redundancy behind various types of information. These issues limit their effectiveness in the diagnosis of brain diseases. To address this, we propose an **I**nformation **B**ottleneck-Guided **A**daptive **H**yper**G**raph **T**ransformer (IBAHGT). By incorporating the information bottleneck (IB) principle, this approach enables adaptive learning of high-order correlations and both short- and long-range dependencies within a unified framework for brain network analysis, achieving high-precision brain disease diagnosis. IBAHGT consists of three key components: an information bottleneck-guided adaptive hypergraph convolution, which introduces a novel hypergraph information bottleneck (HIB) principle to adaptively learn hypergraph message-passing weights between nodes and hyperedges, optimizes information flow and captures high-order information in brain networks that is maximally informative and minimally redundant (MIMR). The Transformer encoder captures global information within brain networks through the attention mechanism, specifically modeling short- and long-range dependencies. An information bottleneck-guided node-level adaptive fusion employs the IB principle to learn independent weights for each node, facilitating the fine-grained integration of high-order information and global information to obtain an efficient representation for downstream tasks. Extensive experiments demonstrate that the proposed method outperforms current state-of-the-art methods and can identify biomarkers for clinical applications.

1 Introduction

The brain network describes the interactions and information processing patterns between regions of interest (ROIs) [1]. Brain network analysis, which explores interactions between ROIs, reveals the brain's functional organization and is crucial for understanding the pathogenesis of neurological diseases and their early intervention [2, 3]. As shown in Fig.1, many studies have shown the presence of extensive high-order correlations [4], along with short- and long-range dependencies, within brain networks [5, 6]. High-order correlations reflect the joint activation of multiple

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

ROIs during functional activities such as cognitive tasks, and are crucial for understanding the activity patterns of the brain's neural system [7]. Furthermore, during functional activities, the brain selectively integrates information through the coordinated work of high-order neural circuits to improve task execution efficiency and accuracy [8]. Short-range dependencies reflect the interactions between ROIs in the neighboring space [9], while long-range dependencies reflect long-distance communication between ROIs and play a critical role in understanding brain functions [10] and communication [11]. Therefore, it is necessary to simultaneously capture the high-order correlations and long-range dependencies in the brain network for unified analysis.

Many studies use graph learning methods for brain network analysis [12, 13], modeling the brain network as a graph and capturing pairwise dependencies in feature interactions through Graph Neural Networks (GNNs). However, these methods overlook the crucial high-order correlations and fail to capture the collaborative interactions among multiple ROIs involved in brain functions. Hypergraphs, a mathematical model that can express high-order correlations, have been increasingly used in brain network analysis [14]. For example, FC-HAT effectively captures high-order information in brain networks through hypergraph attention networks [15], thereby improving the accuracy of brain disease diagnosis.

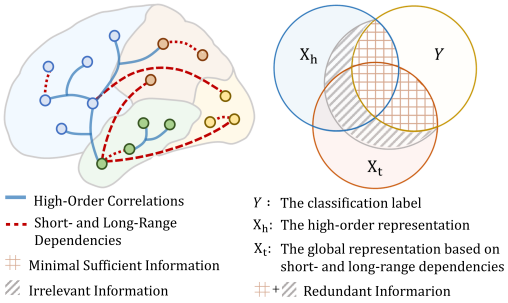


Figure 1: The brain network contains extensive high-order correlations as well as short-range and long-range dependencies. Multi-source information fusion, guided by the information bottleneck, aims to optimize the fused representation by capturing the minimal sufficient information from high-order representation $\mathbf{X}_h$ and global representation $\mathbf{X}_t$ for predicting the classification label Y.

GNNs and Hypergraph Neural Networks (HGNNs) [16, 17], which aggregate neighborhood information through message passing, struggle to capture long-range dependencies in brain networks. Unlike the local information aggregation mechanisms of GNNs and HGNNs, the attention mechanism in Transformer can capture long-range dependencies between nodes [18]. Brain network Transformer [19] has introduced Transformer into brain network analysis to learn long-range dependencies between ROIs, achieving promising performance in brain disease diagnosis.

However, these methods fail to integrate the high-order correlations and long-range dependencies according to the specific characteristics of each ROI. Furthermore, as shown in Fig.1, both the high-order information captured by general hypergraph convolution and the information captured based on different dependencies contain substantial redundancy, which dilutes the disease-related discriminative information. Thus, it is crucial to condense the information during the extraction of high-order features and the fusion of multi-source information to achieve efficient representations.

To address these issues, we propose an information bottleneck-guided adaptive hypergraph Transformer method that effectively captures the MIMR high-order correlations as well as short- and long-range dependencies in the brain network. Specifically, first, an information bottleneck-guided adaptive hypergraph convolution introduces an innovative hypergraph information bottleneck (HIB) principle to optimize the message passing process, achieving the goal of capturing MIMR high-order information. Secondly, the Transformer encoder captures short- and long-range dependencies between ROIs through the self-attention mechanism. Finally, an information bottleneck-guided node-level adaptive fusion method considers the unique roles and needs of each ROI, and adaptively learns the weights for each node (i.e., ROI) to fuse representations that contain high-order information and global information. In summary, this paper makes three main contributions:

1) We propose an information bottleneck-guided adaptive hypergraph Transformer method that captures high-order correlations as well as short- and long-range dependencies in the brain network within a unified framework, achieving high-precision brain disease diagnosis.

2) We introduce an information bottleneck-guided adaptive hypergraph convolution method. By incorporating the HIB principle, it adaptively learns the message passing weights between nodes and hyperedges, capturing MIMR high-order information in the brain network.

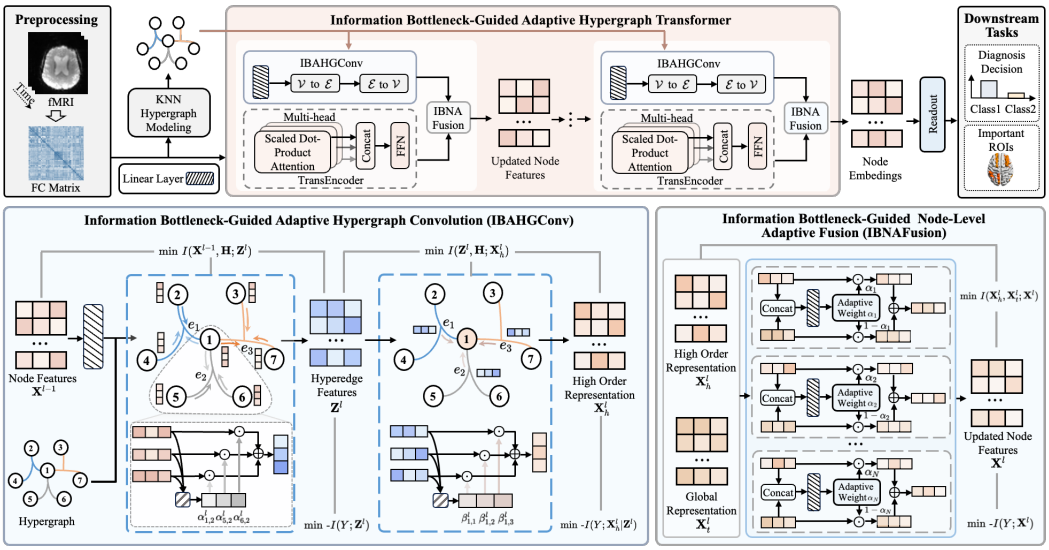


Figure 2: The overall framework of the proposed IBAHGT.

3) We propose an information bottleneck-guided node-level adaptive fusion method that fine-grained integrates high-order and global information in the brain network. Further optimization through the IB principle effectively eliminates redundancy during multi-source information fusion, resulting in efficient representations.

2 Related work

2.1 Deep Learning-Based Brain Network Analysis

GNNs are commonly used deep learning methods for graph-structured brain networks learning [20, 21]. BrainGNN [22] introduces a novel ROI-aware graph convolution (Ra-GConv) layer for functional brain network analysis. However, graph-based methods overlook the high-order correlations that are widespread in brain networks. Recent studies have incorporated hypergraph learning methods into brain network analysis [15]. For example, dwHGCN [23] learns high-order correlations in brain networks through dynamic hypergraph learning with learnable hyperedge weights. Graph- or hypergraph-based methods are unable to effectively capture long-range dependencies between ROIs. BrainNetTF [19] introduces a Transformer encoder to learn long-range dependencies and has demonstrated outstanding performance in brain disease diagnosis tasks. However, the above methods fail to provide a unified analysis that incorporates both the high-order correlations as well as short- and long-range dependencies present in brain networks.

2.2 Information Bottleneck

The Information Bottleneck (IB) [24] is a data compression technique aimed at encouraging representations derived from raw data to capture the maximum relevant information about the target, while excluding extraneous information unrelated to the prediction task. Deep learning methods based on the IB principle have been widely applied in various fields, such as computer vision [25, 26] and natural language processing [27]. Recently, BrainIB [28] has incorporated the IB principle into brain network analysis, effectively identifying disease-specific prominent brain network connections, thus enabling interpretable brain disease diagnosis.

3 Method

In brain network analysis, the ROIs are treated as nodes, and the functional connectivity (FC) matrix $\mathbf{X}^0 \in \mathbb{R}^{N\times N}$ is obtained by calculating Pearson correlation between ROIs, serving as the initial node features. Our goal is to comprehensively capture various types of correlations within brain networks

and distill this multi-source information into efficient feature representations for predicting the target Y, (i.e., the ground truth class label). The overall framework of IBAHGT is shown in Fig.2, with its core consisting of L layers of information bottleneck-guided adaptive hypergraph Transformer layers, each comprising three main components: information bottleneck-guided adaptive hypergraph convolution, an adaptive message passing process guided by the HIB principle, which produces high-order representation $\mathbf{X}_h^l$. Transformer encoder captures short- and long-range dependencies to obtain global representation $\mathbf{X}_t^l$. Information bottleneck-guided adaptive node-level fusion learns the fusion weights for each node, enabling fine-grained integration of MIMR high-order and global information to obtain updated node features $\mathbf{X}^l$. After L layers, the final node embeddings $\mathbf{X}^L$ is flattened and fed into a multilayer perceptron (MLP) to predict class label.

3.1 Information Bottleneck-Guided Adaptive Hypergraph Convolution

Hypergraph-based methods can model high-order correlations among ROIs. However, general hypergraph convolution employs equal-weight message passing when capturing high-order information, which introduces substantial redundant information. To address this, the information bottleneck-guided adaptive hypergraph convolution adaptively learns weights during the message passing and introduces the HIB principle to optimize this process, capturing the MIMR high-order information.

3.1.1 Adaptive Hypergraph Convolution

The hypergraph is constructed using the K Nearest Neighbor (KNN) method to model high-order correlations in the brain network. Specifically, ROIs are represented as a set of nodes $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$. The hypergraph structure is represented by an incidence matrix $\mathbf{H} \in \mathbb{R}^{N \times E}$, reflecting the node-hyperedge relation. The proximity between nodes is measured by the inverse of the Euclidean distance based on the node features $\mathbf{X}^0$.

Based on the constructed hypergraph structure, the adaptive hypergraph convolution performs a two-stage message passing process: adaptive gathering of node features to hyperedges and adaptive aggregating of hyperedge features to nodes. In the first stage, each node adaptively learns the weight and performs weighted aggregation to obtain the MIMR hyperedge feature $\mathbf{Z}^l$, with the HIB principle introduced to optimize the weight learning process. In the l-th layer, for a given hyperedge e_j, the process of generating the hyperedge feature $\mathbf{Z}_j^l$ based on the features of the nodes in its associated node set $\mathcal{N}_v(e_j)$ can be expressed as follows:

$$\mathbf{Z}_j^l = \sigma\left(\sum_{v_i \in \mathcal{N}_v(e_j)} \alpha_{i,j}^l \mathbf{W}_1^l \mathbf{X}_i^{l-1}\right), \quad \alpha_{i,j}^l = \frac{\exp\left(\mathbf{a}_1^l \mathbf{W}_1^l \mathbf{X}_i^{l-1}\right)}{\sum_{v_k \in \mathcal{N}_v(e_j)} \exp\left(\mathbf{a}_1^l \mathbf{W}_1^l \mathbf{X}_k^{l-1}\right)}, \tag{1}$$

where $\mathbf{a}_1^l$ is a learnable vector, $\mathbf{W}_1^l$ is a learnable weight matrix, σ represents the nonlinear activation function. $\alpha_{i,j}^l$ represents the weight when the feature of node v_i is aggregated into the hyperedge e_j.

In the second stage, each hyperedge adaptively learns the weight for the information transfer back to the nodes and performs a weighted aggregation to obtain the high-order representation $\mathbf{X}_h^l$. This process is also optimized through the HIB principle. For node v_i, the process of weighted aggregation of the hyperedge features from its connected hyperedge set $\mathcal{N}_e(v_i)$ to obtain the high-order representation $\mathbf{X}_{h,i}^l$ is formalized as follows:

$$\mathbf{X}_{h,i}^l = \sigma\left(\sum_{e_j \in \mathcal{N}_e(v_i)} \beta_{i,j}^l \mathbf{W}_2^l \mathbf{Z}_j^l\right), \quad \beta_{i,j}^l = \frac{\exp\left(\mathbf{a}_2^l \mathbf{W}_2^l \mathbf{Z}_j^l\right)}{\sum_{e_p \in \mathcal{N}_e(v_i)} \exp\left(\mathbf{a}_2^l \mathbf{W}_2^l \mathbf{Z}_p^l\right)}, \tag{2}$$

where $\mathbf{a}_2^l$ is a learnable vector, $\mathbf{W}_2^l$ is a learnable weight matrix, and β_{ij}^l represents the weight when the feature of hyperedge e_j is transmitted back to node v_i.

3.1.2 Optimization Based on the HIB Principle

There is redundancy in the high-order information within the brain captured by adaptive hypergraph convolution, for this reason, we propose the HIB principle to optimize the information flow process

to capture the MIMR high-order information. In the l-th layer of adaptive hypergraph convolution, the optimization objective based on the HIB principle is:

$$\min_{\mathbb{P}(\mathbf{Z}^l,\mathbf{X}^l|\mathbf{H},\mathbf{X}^{l-1})\in\Omega} \text{HIB}_{\gamma_1}(\mathbf{H},\mathbf{X}^{l-1},Y;\mathbf{Z}^l,\mathbf{X}_h^l) \triangleq -I(Y;\mathbf{Z}^l,\mathbf{X}_h^l)+\gamma_1 I(\mathbf{X}^{l-1},\mathbf{H};\mathbf{Z}^l,\mathbf{X}_h^l)\,. \quad (3)$$

where $\mathbf{X}^{l-1}\in\mathbb{R}^{N\times d}$ is the input node representation of layer l, $\mathbf{H}\in\{0,1\}^{N\times E}$ is the hypergraph incidence matrix with N ROIs and E hyperedges (here $E=N=90$ as each ROI centers one hyperedge via KNN), $\mathbf{Z}^l\in\mathbb{R}^{E\times d}$ is the intermediate hyperedge representation, $\mathbf{X}_h^l\in\mathbb{R}^{N\times d}$ is the output high-order node representation, Y is the diagnosis label, $I(\cdot;\cdot)$ denotes mutual information, and γ_1 balances prediction performance and information compression. Minimizing $-I(Y;\mathbf{Z}^l,\mathbf{X}_h^l)$ ensures that the hypergraph message passing process focuses on task-relevant information, while minimizing $I(\mathbf{X}^{l-1},\mathbf{H};\mathbf{Z}^l,\mathbf{X}_h^l)$ effectively removes redundant information. γ_1 is a hyperparameter that balances information and redundancy. However, accurately calculating mutual information terms in Eq.(3) is intractable. Therefore, we introduce the lower bound of $I(Y;\mathbf{Z}^l,\mathbf{X}_h^l)$ and the upper bound of $I(\mathbf{X}_h^l,\mathbf{Z}^l;\mathbf{X}^{l-1},\mathbf{H})$ for subsequent estimation.

Proposition 1. For any distribution $\mathbb{Q}_1(Y|\mathbf{Z}^l)$ and $\mathbb{Q}_2(Y)$, we have

$$I(Y;\mathbf{Z}^l,\mathbf{X}_h^l)=I(Y;\mathbf{Z}^l)+I(Y;\mathbf{X}_h^l\mid\mathbf{Z}^l)\,. \quad (4)$$

To represent the lower bounds of two terms on the right-hand side of the equation, in the adaptive hypergraph convolution layer, $\mathbf{Z}^l$ and $\mathbf{X}_h^l$ are read out and then mapped through the learnable parameter matrix to the predicted labels $\hat{y}_z^l$ and $\hat{y}_x^l$, respectively. Therefore, the two terms on the right-hand side of the Eq.(4) can reduce to the cross-entropy loss without constants as follows:

$$I(Y;\mathbf{Z}^l)\to-\mathcal{L}_{CE}(\hat{y}_z^l,Y)\quad,\quad I(Y;\mathbf{X}_h^l\mid\mathbf{Z}^l)\to-\mathcal{L}_{CE}(\hat{y}_x^l,Y)\,. \quad (5)$$

The proof of the above process is given in Appendix B. Subsequently, we derive the upper bound of the second term in Eq.(3).

Proposition 2. For any distributions $\mathbb{Q}(\mathbf{Z}^l)$ and $\mathbb{Q}(\mathbf{X}_h^l)$, we have

$$I(\mathbf{X}^{l-1},\mathbf{H};\mathbf{X}_h^l,\mathbf{Z}^l)\le I(\mathbf{X}^{l-1},\mathbf{H};\mathbf{Z}^l)+I(\mathbf{Z}^l,\mathbf{H};\mathbf{X}_h^l)\le \text{ZIB}^l+\text{XIB}^l\,, \quad (6)$$

$$\text{ZIB}^l=D_{KL}\left(\mathbb{P}(\mathbf{Z}^l\mid\mathbf{X}^{l-1},\mathbf{H})\parallel\mathbb{Q}(\mathbf{Z}^l)\right),\ \text{XIB}^l=D_{KL}\left(\mathbb{P}(\mathbf{X}_h^l\mid\mathbf{Z}^l,\mathbf{H})\parallel\mathbb{Q}(\mathbf{X}_h^l)\right). \quad (7)$$

The proof of the proposition can be found in Appendix B. To estimate ZIB^l and XIB^l, we set both $\mathbb{Q}(\mathbf{Z}^l)$ and $\mathbb{Q}(\mathbf{X}_h^l)$ as a mixture of Gaussians with learnable parameters [29]. Concretely, set $\mathbb{Q}(\mathbf{Z}^l)\sim\sum_{i=1}^m w_i\,\text{Gaussian}(\mu_{0,i},\sigma_{0,i}^2)$ and $\mathbb{Q}(\mathbf{X}_h^l)\sim\sum_{i=1}^m \hat{w}_i\,\text{Gaussian}(\hat{\mu}_{0,i},\hat{\sigma}_{0,i}^2)$, where $w_i,\mu_{0,i},\sigma_{0,i}$ and $\hat{w}_i,\hat{\mu}_{0,i},\hat{\sigma}_{0,i}$ are learnable parameters. Based on the aggregated $\mathbf{Z}^l$ and $\mathbf{X}^l$, the parameters are calculated and samples are drawn from a Gaussian distribution to obtain $\mathbb{P}(\mathbf{Z}^l\mid\mathbf{X}^{l-1},\mathbf{H})\sim\text{Gaussian}(\mu_l,\sigma_l^2)$ and $\mathbb{P}(\mathbf{X}_h^l\mid\mathbf{Z}^l,\mathbf{H})\sim\text{Gaussian}(\hat{\mu}_l,\hat{\sigma}_l^2)$. The estimation of ZIB^l and XIB^l can be written as:

$$\widehat{\text{ZIB}}^l=\log(\Phi(\mathbf{Z}^l;\mu_l,\sigma_l^2))-\log(\sum_{i=1}^m w_i\Phi(\mathbf{Z}^l;\mu_{0,i},\sigma_{0,i}^2))\,, \quad (8)$$

$$\widehat{\text{XIB}}^l=\log(\Phi(\mathbf{X}_h^l;\hat{\mu}_l,\hat{\sigma}_l^2))-\log(\sum_{i=1}^m \hat{w}_i\Phi(\mathbf{X}_h^l;\hat{\mu}_{0,i},\hat{\sigma}_{0,i}^2))\,. \quad (9)$$

Plugging the estimations of both mutual information terms in Eq.(3), the loss function for the information bottleneck-guided adaptive hypergraph convolution in the l-th layer can be obtained:

$$\mathcal{L}_H^l=\mathcal{L}_{CE}(\hat{y}_z^l,Y)+\mathcal{L}_{CE}(\hat{y}_x^l,Y)+\gamma_1[\widehat{\text{ZIB}}^l+\widehat{\text{XIB}}^l]\,. \quad (10)$$

3.2 Transformer Encoder

Within the IBAHGT framework, this branch complements the hypergraph convolution module by capturing full-range pairwise dependencies across ROIs. We briefly recap the standard multi-head self-attention formulation below, and focus on its fusion with high-order representations in Section

3.3. Hypergraph-based learning methods struggle to capture long-range dependencies, whereas the self-attention mechanism in the Transformer encoder can effectively model both short- and long-range dependencies between ROIs in the brain network. Formally, in the l-th layer, a multi-head self-attention mechanism is used to learn global representation $\mathbf{X}_t^l$:

$$\mathbf{X}_t^l = \left(\|_{m=1}^{M} \mathbf{h}^{l,m}\right) \mathbf{W}_O^l\,, \quad \mathbf{h}^{l,m} = \text{softmax}\left(\frac{\mathbf{W}_Q^{l,m}\mathbf{X}^{l-1}\left(\mathbf{W}_K^{l,m}\mathbf{X}^{l-1}\right)^\top}{\sqrt{d_K^{l,m}}}\right)\mathbf{W}_V^{l,m}\mathbf{X}^{l-1}\,, \quad (11)$$

where $\|$ denotes the concatenation operator, M is the number of heads, $\mathbf{W}_O^l$, $\mathbf{W}_Q^{l,m}$, $\mathbf{W}_K^{l,m}$, and $\mathbf{W}_V^{l,m}$ are learnable model parameters, and $d_K^{l,m}$ is the first dimension of $\mathbf{W}_K^{l,m}$.

3.3 Information Bottleneck-Guided Node-Level Adaptive Fusion

Considering the varying roles of each ROI in the brain network and their different tendencies in capturing various dependencies [30], the information bottleneck-guided node-level adaptive fusion learns an independent weight for each node to balance the high-order information from the information bottleneck-guided adaptive hypergraph convolution and the short- and long-range dependencies information from the Transformer encoder.

3.3.1 Node-Level Adaptive Fusion

In the l-th layer of the information bottleneck-guided adaptive hypergraph Transformer, node-level adaptive fusion integrates the high-order representation $\mathbf{X}_{h,i}^l$ and global representation $\mathbf{X}_{t,i}^l$ for each node v_i, resulting in the updated node features $\mathbf{X}_i^l$, which can be formulated as follows:

$$\mathbf{X}_i^l = \theta_i^l \mathbf{X}_{h,i}^l + (1-\theta_i^l)\mathbf{X}_{t,i}^l\,, \quad (12)$$

where θ_i^l denotes the learnable fusion weight of node v_i.

3.3.2 Optimization Based on the IB Principle

There is redundancy between the high-order representation of the brain network $\mathbf{X}_h^l$ and the global representation of the brain network $\mathbf{X}_t^l$ that contains short- and long-range dependencies. To reduce redundant information while preserving effective information in the updated node features $\mathbf{X}^l$, we apply the IB principle to optimize the fusion process. The optimization objective is:

$$\min_{\mathbb{P}(\mathbf{X}^l,\mathbf{X}_h^l,\mathbf{X}_t^l)\in\Omega} -I(Y;\mathbf{X}^l) + \gamma_2 I(\mathbf{X}_h^l,\mathbf{X}_t^l;\mathbf{X}^l)\,. \quad (13)$$

Minimizing $-I(Y;\mathbf{X}^l)$ ensures that the fused features effectively retain task-relevant information, while minimizing $I(\mathbf{X}_h^l,\mathbf{X}_t^l;\mathbf{X}^l)$ removes redundant information. The trade-off hyperparameter γ_2 is to balance the weights of two items. In the l-th layer, $\mathbf{X}^l$ is read out and mapped to the predicted label $\hat{y}_f^l$ through a learnable parameter matrix. Similar to Section 3.1.2, the minimization of the first term in Eq.(13) can be replaced by the cross-entropy loss, i.e.,

$$I(Y;\mathbf{X}^l) \to -\mathcal{L}_{CE}(\hat{y}_f^l, Y)\,. \quad (14)$$

The second term in Eq.(13) has an optimizable upper bound, which is given as follows:

Proposition 3. For any distribution $\mathbb{Q}(\mathbf{X}^l)$, we have

$$I(\mathbf{X}_h^l,\mathbf{X}_t^l;\mathbf{X}^l) \le D_{KL}\left(\mathbb{P}(\mathbf{X}^l \mid \mathbf{X}_h^l,\mathbf{X}_t^l) \,\|\, \mathbb{Q}(\mathbf{X}^l)\right)\,.$$

The proof of the proposition can be found in Appendix B. To specify the upper bound for $I(\mathbf{X}_h^l,\mathbf{X}_t^l;\mathbf{X}^l)$, we set $\mathbb{Q}(\mathbf{X}^l)$ as a mixture of Gaussians with learnable parameters: $\mathbb{Q}(\mathbf{X}^l) \sim \sum_{i=1}^m \tilde{w}_i\,\text{Gaussian}(\tilde{\mu}_{0,i},\tilde{\sigma}_{0,i}^2)$. The parameters are computed based on the fused representation $\mathbf{X}^l$, and sampling is performed from a Gaussian distribution to obtain $\mathbb{P}(\mathbf{X}^l \mid \mathbf{X}_h^l,\mathbf{X}_t^l) \sim \text{Gaussian}(\tilde{\mu}_l,\tilde{\sigma}_l^2)$. Thus, the estimation of $I(\mathbf{X}_h^l,\mathbf{X}_t^l;\mathbf{X}^l)$ is written as:

$$I(\mathbf{X}_h^l,\mathbf{X}_t^l;\mathbf{X}^l) \to \log(\Phi(\mathbf{X}^l;\tilde{\mu}_l,\tilde{\sigma}_l^2)) - \log(\sum_{i=1}^m \tilde{w}_i \Phi(\mathbf{X}^l;\tilde{\mu}_{0,i},\tilde{\sigma}_{0,i}^2))\,. \quad (15)$$

By plugging the estimations of both mutual information terms into Eq.(13), the loss function for the node-level adaptive fusion process guided by IB principle in the l-th layer can be obtained:

$$\mathcal{L}_F^l = \mathcal{L}_{CE}(\hat{y}_f^l, Y) + \gamma_2(\log(\Phi(\mathbf{X}^l; \tilde{\mu}_l, \tilde{\sigma}_l^2)) - \log(\sum_{i=1}^{m} \tilde{w}_i \Phi(\mathbf{X}^l; \tilde{\mu}_{0,i}, \tilde{\sigma}_{0,i}^2))) \,. \tag{16}$$

3.4 Information Bottleneck-Guided Adaptive Hypergraph Transformer Layer

The information bottleneck-guided adaptive hypergraph Transformer layer is composed of parallel information bottleneck-guided adaptive hypergraph convolution (IBAHGConv) layer and Transformer encoder (TransEncoder) layer with information bottleneck-guided node-level adaptive fusion (IBNAFusion) module. In the l-th layer, the high order representation $\mathbf{X}_h^l$ from the $\mathrm{IBAHGConv}^l$ layer and the global representation $\mathbf{X}_t^l$ from the $\mathrm{TransEncoder}^l$ layer are adaptively aggregated by $\mathrm{IBNAFusion}^l$ to obtain updated node features $\mathbf{X}^l$, as shown in the following equation:

$$\mathbf{X}_h^l = \mathrm{IBAHGConv}^l\left(\mathbf{X}^{l-1}, \mathbf{H}\right) , \tag{17}$$

$$\mathbf{X}_t^l = \mathrm{TransEcoder}^l\left(\mathbf{X}^{l-1}\right) , \tag{18}$$

$$\mathbf{X}^l = \mathrm{IBNAFusion}^l\left(\mathbf{X}_h^l, \mathbf{X}_t^l\right) . \tag{19}$$

After L layers, the final node embeddings $\mathbf{X}^L$ effectively capture information from various dependencies in the brain network.

3.5 Loss Function

By concatenating the node embeddings $\mathbf{X}^L$, the graph-level representation $\hat{\mathbf{x}}$ is obtained, which is then fed into an MLP to predict the label $\hat{y}$. During training, the overall loss function $\mathcal{L}$ is composed of the cross-entropy loss function $\mathcal{L}_{CE}$ for supervised disease prediction, the loss $\mathcal{L}_H$ for the adaptive hypergraph convolution part, and the loss $\mathcal{L}_F$ for the node-level adaptive fusion part:

$$\mathcal{L} = \mathcal{L}_{CE}(\hat{y}, Y) + \sum_{l=1}^{L}(\lambda_1 \mathcal{L}_H^l + \lambda_2 \mathcal{L}_F^l) \,, \tag{20}$$

where L denotes the number of layers in the information bottleneck-guided adaptive hypergraph Transformer, and λ_1 and λ_2 are trade-off hyperparameters balancing different losses.

4 Experiments

4.1 Experiments Settings

Datasets and Preprocessing. The proposed method is evaluated on two brain network analysis related fMRI datasets, ABIDE [31] and ADNI [32] . The ABIDE dataset includes brain imaging data from 871 subjects collected across 17 international sites, with 403 subjects diagnosed with Autism Spectrum Disorder (ASD). A subset of the ADNI dataset is used in this study, consisting of 66 patients with Alzheimer's Disease (AD) and 139 Normal Controls (NCs).

We preprocess the fMRI data using the standard steps provided by the Data Processing Assistant for Resting-State Function (DPARSF) toolkit [33], then use the Anatomical Automatic Labeling (AAL) brain template [34] to divide the brain space into ROIs and extract the corresponding BOLD signals. The FC matrix is obtained by calculating the Pearson correlation coefficients between the ROIs.

Metrics. Four machine learning and medical diagnostic-specific metrics are used: accuracy (ACC), sensitivity (SEN), specificity (SPE) and area under the receiver operating characteristic curve (AUC). We record the mean and standard deviation across 10 random runs on the test dataset.

Implementation Details. In the proposed method, we set the value of k to 3 in the KNN hypergraph modeling. The number of layers L in the adaptive hypergraph Transformer and the number of heads M in the multi-head attention module are set to 2 and 4, respectively. For all datasets, we randomly divide the training set, evaluation set, and test set by the ratio of $7:1:2$. During training, we use the Adam optimizer [35], with an initial learning rate set to 1×10^{-4}. The number of

Table 1: Experimental results of the comparison methods.

Datasets (Tasks)	ABIDE (ASD vs. NC)				ADNI (AD vs. NC)			
Method	ACC (%)	SPE (%)	SEN (%)	AUC (%)	ACC (%)	SPE (%)	SEN (%)	AUC (%)
BrainNetCNN	66.2±1.9	74.3±4.1	56.8±3.1	68.3±2.6	73.2±2.7	79.3±4.2	60.0±5.8	66.3±4.7
ContrastPool	63.0±2.5	71.1±8.3	53.5±4.6	68.1±7.7	70.7±4.1	72.1±5.7	67.7±10.2	69.5±3.9
BrainGNN	64.5±2.7	72.6±7.7	55.0±5.5	69.7±7.2	73.2±4.1	74.3±4.2	70.8±12.3	71.3±6.1
BrainIB	68.3±1.9	76.4±5.4	58.8±5.0	70.2±1.9	74.1±2.5	78.6±5.1	64.6±11.5	69.0±6.5
HGNN$^+$	68.9±2.1	77.0±4.4	59.3±5.9	71.0±1.9	74.6±2.9	81.4±4.2	60.0±5.8	69.2±6.8
FC-HAT	70.7±3.5	75.5±5.0	65.0±6.5	72.0±4.8	77.1±3.7	84.3±3.6	61.5±4.9	69.1±2.0
BrainNetTF	71.4±1.1	76.8±5.0	65.0±7.2	72.2±1.4	78.0±2.7	85.7±2.3	61.5±4.9	70.6±1.8
Com-BrainTF	72.1±1.5	74.3±3.3	69.5±6.0	73.3±2.4	79.0±3.3	86.4±2.7	63.1±5.8	71.4±2.5
ALTER	74.4±1.2	78.5±1.8	69.5±3.0	75.3±1.9	81.5±2.5	87.9±1.8	67.7±5.8	71.9±2.8
IBAHGT	**77.6±0.8**	**80.4±3.1**	**74.3±3.8**	**80.5±2.0**	**86.3±1.2**	**88.6±4.7**	**81.5±7.9**	**82.8±5.2**

epochs is set to 300. All our experiments are implemented in PyTorch and trained on one NVIDIA 4090. For interpretability analyses, all statistical tests use subject-level out-of-fold representations from 5 repeated nested cross-validation to avoid data leakage. We apply two-sided Welch's t-tests with Benjamini-Hochberg FDR correction ($q < 0.05$) per feature family, and conduct site-held-out validation to confirm cross-site generalizability.

4.2 Comparison Experiments

Baselines. The selected baselines correspond to three categories. The first category includes CNN-based method BrainNetCNN [36]. The second category includes graph or hypergraph-based learning methods, such as ContrastPool [37], BrainGNN [22], BrainIB [28], HGNN$^+$ [38], and FC-HAT [15]. The third category consists of Transformer-based methods, including BrainNetTF [19], Com-BrainTF [39] and ALTER [40]. The codes are reproduced based on the released codes.

Results. Table 1 shows the comparison results between the proposed and baseline methods. On all datasets, IBAHGT significantly outperforms the baseline methods. The hypergraph-based methods achieved higher accuracy in most cases compared to graph-based methods, indicating that hypergraphs can capture high-order correlations that are beneficial for classification. Compared to these methods, our proposed approach further improves accuracy on the ABIDE dataset (ASD vs. NC: 77.6%) and the ADNI dataset (AD vs. NC: 86.3%). The performance improvement is attributed to our method, which captures the MIMR high-order correlations and long-range dependencies in the brain network, and condenses the information from both at the ROI level to obtain an efficient representation.

4.3 Ablation Study

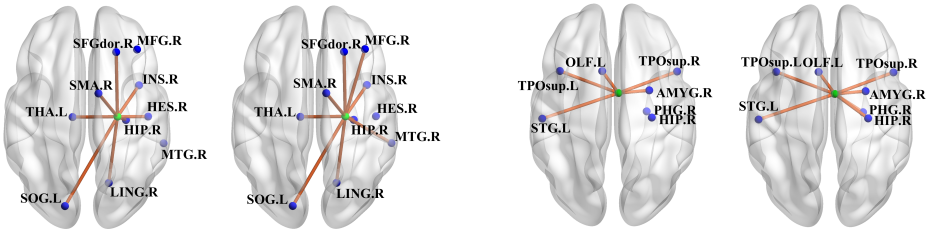
To validate the effectiveness of information bottleneck-guided adaptive hypergraph convolution (IBAHGConv) and information bottleneck-guided adaptive node-level fusion (IBNAFusion), ablation study are performed on two datasets, with results shown in Table 2. Using only IBAHGConv or the Transformer encoder (TransEncoder) performs worse than the combined method, indicating that integrating high-order information with long-range dependency information for comprehensive analysis enables a more thorough capture of abnormal changes in brain networks. On the ABIDE dataset, replacing IBAHGConv with HGNN$^+$ and AHGConv (without the HIB principle) led to a decrease in accuracy by 3.5% and 2.5%, respectively. This demonstrates that IBAHGConv optimizes the message-passing process and is more effective at capturing MIMR high-order information. Furthermore, diagnostic accuracy on the ABIDE dataset decreases when directly average fusion or NAFusion without IB principle, compared to IBNAFusion, which demonstrates the effectiveness of information concise during fusion at the node level guided by the IB principle.

4.4 Interpretation Analysis

Discriminative Hyperedge Analysis

Table 2: Experimental results of the ablation study.

Datasets (Tasks)	ABIDE (ASD vs. NC)				ADNI (AD vs. NC)			
Method	ACC (%)	SPE (%)	SEN (%)	AUC (%)	ACC (%)	SPE (%)	SEN (%)	AUC (%)
TransEnoder	69.3±0.9	77.9±4.5	59.3±4.2	72.7±5.3	76.1±2.8	77.9±6.9	72.3±10.4	75.6±5.8
IBAHGconv	70.9±1.3	81.7±3.3	58.3±2.9	73.4±5.1	78.0±2.2	82.9±5.3	67.7±10.2	77.3±3.8
IBAHGT w/o IBAHGConv	74.1±1.4	76.6±4.9	71.3±6.4	77.8±3.4	81.9±2.5	84.3±3.6	76.9±8.4	80.4±2.7
IBAHGT w/o HIB	75.1±1.5	78.1±4.1	71.5±6.5	79.2±2.3	83.4±2.4	84.3±3.6	81.5±7.9	81.0±5.1
IBAHGT w/o IBNAFusion	73.0±1.3	77.5±4.6	67.8±6.4	77.4±3.5	81.0±1.8	85.7±3.9	70.8±5.8	80.3±3.9
IBAHGT w/o IB	74.8±1.2	76.8±4.7	72.5±7.3	78.9±2.3	83.9±2.0	85.7±3.2	80.0±6.2	81.8±4.6
IBAHGT	**77.6±0.8**	**80.4±3.1**	**74.3±3.8**	**80.5±2.0**	**86.3±1.2**	**88.6±4.7**	**81.5±7.9**	**82.8±5.2**



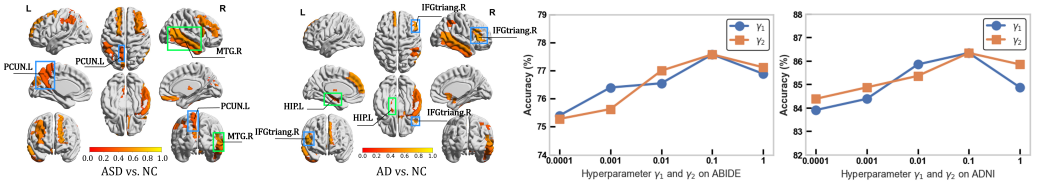
(a) ASD vs .NC (b) AD vs .NC

Figure 3: Visualization of discriminative hyperedges for each classification. In each figure, the blue balls are the ROI contained in the hyperedge and the green balls are the center of the involved ROIs.

We analyze the discriminative connectivity patterns of brain networks between patient and control groups from a high-order correlation perspective. After Benjamini-Hochberg FDR correction for multiple comparisons, the core discriminative hyperedges and ROIs reported in the main text remain statistically significant ($q < 0.05$). We visualize these significant patterns in Fig. 3.

For ASD vs. NC, ASD patients show missing connections with right middle frontal gyrus (MFG.R) as well as right middle temporal gyrus (MTG.R), which foreshadows defects in individual external goal-related behavior and self-consciousness, further validating the conclusions of the previous study [41]. For AD vs. NC, it can be observed that the high-order correlarions represented by the hyperedge in AD patients show missing connections with PHG.R and HIP.R. Previous studies have indicated that functional network impairments in AD patients often occur in the PHG and HIP, which affect memory [42].

Analysis of Fusion Weights for Important ROIs To evaluate the influence of ROIs on the classification tasks, we measure the significance of the nodes using the t-test method and visualize the top 10 most important ROIs with the smallest p-values, displaying the fusion weights of each ROI through color differences. For ASD vs. NC, the identified important ROIs include the left precuneus (PCUN.L), right rectus (REC.R), right middle temporal gyrus (MTG.R), superior frontal gyrus, medial (SFGmed.L) and IPL.R, etc. As mentioned in [43], the activations in PCUN and REC will be reduced when ASD patients inference mental states themselves or others. In addition, we found that the fusion weight of the MTG.R is larger, indicating a greater inclination towards high-order correlations. Previous studies have shown that MTG is critical for high-order cognitive functions, especially in language comprehension through collaboration with multiple other ROIs [44]. The fusion weight of the PCUN.L is relatively small, indicating that PCUN is more inclined to capture long-range dependencies. For AD vs. NC, the identified important ROIs include the HIP.R, left superior frontal gyrus, medial (SFGmed.L), right inferior frontal gyrus, triangular part (IFGtriang.R), right amygdala (AMYG.R), and left caudate nucleus (CAU.L), etc. The fusion weight of the HIP is relatively small, indicating that it relies more on long-range dependencies. Previous research has reported that low-frequency activity in the hippocampus can drive functional connectivity between the cortices of the two hemispheres [45], highlighting the HIP's key role in long-distance communication.



(a) Fusion weights for important ROIs on two tasks.

(b) Influence of two key hyperparameters for model performance.

Figure 4: Visualization of important ROIs with their fusion weights (Color bar represents the magnitude of the fusion weights) and the hyperparameters influence.

4.5 Hyperparameter Sensitivity Analysis

To analyze how the trade-off hyperparameters γ_1 and γ_2 affect model performance, we perform a hyperparameter search and plot the optimal model performance on the ABIDE and ADNI datasets for each case where $\gamma_1 \in \{0.0001, 0.001, 0.01, 0.1, 1\}$ and $\gamma_2 \in \{0.0001, 0.001, 0.01, 0.1, 1\}$, as shown in Fig.4(b). We observe that as the magnitude of γ_1 increases, the accuracy on both datasets initially rises and then decreases, reaching a peak at $\gamma_1 = 0.1$. A similar trend is observed with γ_2, where the IBAHGT achieves the highest classification accuracy when $\gamma_2 = 0.1$.

5 Conclusion

In this paper, the proposed IBAHGT learns the MIMR high-order correlations and short- and long-range dependencies in brain networks within a unified framework. It suppresses redundancy while enabling fine-grained fusion of multi-source information to obtain efficient representations for brain disease diagnosis. The information bottleneck-guided adaptive hypergraph convolution introduces the HIB principle, adaptively learning message-passing weights to capture MIMR high-order information. Information bottleneck-guided node-level fusion assigns weights based on each node's information preference, effectively integrating high-order and global information. The proposed method achieves state-of-the-art performance on two datasets, offering new perspectives for understanding ROIs' interactions and brain information processing patterns.

Acknowledgments and Disclosure of Funding

This work was supported by the National Key Research and Development Program of China under Grant No. 2025ZD0217300, the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant No. JYB2025XDXM504 and Guangdong Major Project of Basic and Applied Basic Research under Grant No. 2023B0303000009.

References

[1] Andrea Avena-Koenigsberger, Bratislav Misic, and Olaf Sporns. Communication dynamics in complex brain networks. *Nature reviews neuroscience*, 19(1):17–33, 2018.

[2] Jin Liu, Min Li, Yi Pan, Wei Lan, Ruiqing Zheng, Fang-Xiang Wu, and Jianxin Wang. Complex brain network analysis and its applications to brain disorders: a survey. *Complexity*, 2017(1):8362741, 2017.

[3] Kaustubh Supekar, Vinod Menon, Daniel Rubin, Mark Musen, and Michael D Greicius. Network analysis of intrinsic functional brain connectivity in alzheimer's disease. *PLoS computational biology*, 4(6):e1000100, 2008.

[4] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[5] Hae-Jeong Park and Karl Friston. Structural and functional brain networks: from connections to cognition. *Science*, 342(6158):1238411, 2013.

[6] Pablo Barttfeld, Bruno Wicker, Sebastián Cukier, Silvana Navarta, Sergio Lew, and Mariano Sigman. A big-world network in asd: dynamical connectivity analysis reflects a deficit in long-range connections and an excess of short-range connections. *Neuropsychologia*, 49(2):254–263, 2011.

[7] Christoph Von Der Malsburg. The correlation theory of brain function. In *Models of neural networks: Temporal aspects of coding and information processing in biological systems*, pages 95–119. Springer, 1994.

[8] Beatriz Luna and John A Sweeney. The emergence of collaborative brain function: Fmri studies of the development of response inhibition. *Annals of the New York Academy of Sciences*, 1021(1):296–309, 2004.

[9] Dina R Dajani and Lucina Q Uddin. Local brain connectivity across development in autism spectrum disorder: A cross-sectional investigation. *Autism Research*, 9(1):43–54, 2016.

[10] Gustavo Deco, Yonathan Sanz Perl, Peter Vuust, Enzo Tagliazucchi, Henry Kennedy, and Morten L Kringelbach. Rare long-range cortical connections enhance human information processing. *Current Biology*, 31(20):4436–4448, 2021.

[11] Caio Seguin, Olaf Sporns, and Andrew Zalesky. Brain network communication: concepts, models and applications. *Nature reviews neuroscience*, 24(9):557–574, 2023.

[12] Alaa Bessadok, Mohamed Ali Mahjoub, and Islem Rekik. Graph neural networks in network neuroscience. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5833–5848, 2022.

[13] Gang Qu, Wenxing Hu, Li Xiao, Junqi Wang, Yuntong Bai, Beenish Patel, Kun Zhang, and Yu-Ping Wang. Brain functional connectivity analysis via graphical deep learning. *IEEE Transactions on Biomedical Engineering*, 69(5):1696–1706, 2021.

[14] Ju Niu and Yuhui Du. Applications of hypergraph-based methods in classifying and subtyping psychiatric disorders: a survey. *Radiology Science*, 2(01):83–95, 2023.

[15] Junzhong Ji, Yating Ren, and Minglong Lei. Fc–hat: Hypergraph attention network for functional brain network classification. *Information Sciences*, 608:1301–1316, 2022.

[16] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3558–3565, 2019.

[17] Song Bai, Feihu Zhang, and Philip HS Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[19] Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. Brain network transformer. *Advances in Neural Information Processing Systems*, 35:25586–25599, 2022.

[20] Yanwu Yang, Chenfei Ye, Xutao Guo, Tao Wu, Yang Xiang, and Ting Ma. Mapping multi-modal brain connectome for brain disorder diagnosis via cross-modal mutual learning. *IEEE Transactions on Medical Imaging*, 43(1):108–121, 2023.

[21] Hongting Ye, Yalu Zheng, Yueying Li, Ke Zhang, Youyong Kong, and Yonggui Yuan. Rh-brainfs: regional heterogeneous multimodal brain networks fusion strategy. *Advances in Neural Information Processing Systems*, 36:59286–59303, 2023.

[22] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 74:102233, 2021.

[23] Junqi Wang, Hailong Li, Gang Qu, Kim M Cecil, Jonathan R Dillman, Nehal A Parikh, and Lili He. Dynamic weighted hypergraph convolutional network for brain functional connectome analysis. *Medical image analysis*, 87:102828, 2023.

[24] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[25] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6778–6787, 2019.

[26] Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. *arXiv preprint arXiv:1810.00821*, 2018.

[27] Rundong Wang, Xu He, Runsheng Yu, Wei Qiu, Bo An, and Zinovi Rabinovich. Learning efficient multi-agent communication: An information bottleneck approach. In *International conference on machine learning*, pages 9908–9918. PMLR, 2020.

[28] Kaizhong Zheng, Shujian Yu, Baojuan Li, Robert Jenssen, and Badong Chen. Brainib: Interpretable brain network-based psychiatric diagnosis with graph information bottleneck. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[29] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[30] James M Shine, Patrick G Bissett, Peter T Bell, Oluwasanmi Koyejo, Joshua H Balsters, Krzysztof J Gorgolewski, Craig A Moodie, and Russell A Poldrack. The dynamics of functional brain networks: integrated network states during cognitive task performance. *Neuron*, 92(2):544–554, 2016.

[31] Anibal Sólon Heinsfeld, Alexandre Rosa Franco, R Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi. Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage: Clinical*, 17:16–23, 2018.

[32] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[33] Xiao-Wei Song, Zhang-Ye Dong, Xiang-Yu Long, Su-Fang Li, Xi-Nian Zuo, Chao-Zhe Zhu, Yong He, Chao-Gan Yan, and Yu-Feng Zang. Rest: a toolkit for resting-state functional magnetic resonance imaging data processing. *PloS one*, 6(9):e25031, 2011.

[34] Edmund T Rolls, Chu-Chung Huang, Ching-Po Lin, Jianfeng Feng, and Marc Joliot. Automated anatomical labelling atlas 3. *Neuroimage*, 206:116189, 2020.

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[36] Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

[37] Jiaxing Xu, Qingtian Bian, Xinhang Li, Aihu Zhang, Yiping Ke, Miao Qiao, Wei Zhang, Wei Khang Jeremy Sim, and Balázs Gulyás. Contrastive graph pooling for explainable classification of brain networks. *IEEE Transactions on Medical Imaging*, 2024.

[38] Yue Gao, Yifan Feng, Shuyi Ji, and Rongrong Ji. Hgnn+: General hypergraph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3181–3199, 2022.

[39] Anushree Bannadabhavi, Soojin Lee, Wenlong Deng, Rex Ying, and Xiaoxiao Li. Community-aware transformer for autism prediction in fmri connectome. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 287–297. Springer, 2023.

[40] Shuo Yu, Shan Jin, Ming Li, Tabinda Sarwar, and Feng Xia. Long-range brain graph transformer. *Advances in Neural Information Processing Systems*, 37:24472–24495, 2024.

[41] Anne Pankow, Teresa Katthagen, Sarah Diner, Lorenz Deserno, Rebecca Boehme, Nobert Kathmann, Tobias Gleich, Michael Gaebler, Henrik Walter, Andreas Heinz, et al. Aberrant salience is related to dysfunctional self-referential processing in psychosis. *Schizophrenia bulletin*, 42(1):67–76, 2016.

[42] Saul L Miller, Kim Celone, Kristina DePeau, Eli Diamond, Bradford C Dickerson, Dorene Rentz, Maija Pihlajamäki, and Reisa A Sperling. Age-related memory impairment associated with loss of parietal deactivation but preserved hippocampal activation. *Proceedings of the National Academy of Sciences*, 105(6):2181–2186, 2008.

[43] Rajesh K Kana, Jose O Maximo, Diane L Williams, Timothy A Keller, Sarah E Schipul, Vladimir L Cherkassky, Nancy J Minshew, and Marcel Adam Just. Aberrant functioning of the theory-of-mind network in children and adolescents with autism. *Molecular autism*, 6:1–12, 2015.

[44] Karalyn Patterson, Peter J Nestor, and Timothy T Rogers. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature reviews neuroscience*, 8(12):976–987, 2007.

[45] Russell W Chan, Alex TL Leong, Leon C Ho, Patrick P Gao, Eddie C Wong, Celia M Dong, Xunda Wang, Jufang He, Ying-Shing Chan, Lee Wei Lim, et al. Low-frequency hippocampal–cortical activity drives brain-wide resting-state functional mri connectivity. *Proceedings of the National Academy of sciences*, 114(33):E6972–E6981, 2017.

[46] Iva Ilioska, Marianne Oldehinkel, Alberto Llera, Sidhant Chopra, Tristan Looden, Roselyne Chauvin, Daan Van Rooij, Dorothea L Floris, Julian Tillmann, Carolin Moessnang, et al. Connectome-wide mega-analysis reveals robust patterns of atypical functional connectivity in autism. *Biological psychiatry*, 94(1):29–39, 2023.

[47] Elysa J Marco, Leighton BN Hinkley, Susanna S Hill, and Srikantan S Nagarajan. Sensory processing in autism: a review of neurophysiologic findings. *Pediatric research*, 69(8):48–54, 2011.

[48] Nicholas Myers, Lorenzo Pasquini, Jens Göttler, Timo Grimmer, Kathrin Koch, Marion Ortner, Julia Neitzel, Mark Mühlau, Stefan Förster, Alexander Kurz, et al. Within-patient correspondence of amyloid-β and intrinsic network connectivity in alzheimer's disease. *Brain*, 137(7):2052–2064, 2014.

[49] Yigal Agam, Robert M Joseph, Jason JS Barton, and Dara S Manoach. Reduced cognitive control of response inhibition by the anterior cingulate cortex in autism spectrum disorders. *Neuroimage*, 52(1):336–347, 2010.

[50] Qinghua Zhao, Hong Lu, Hichem Metmer, Will XY Li, and Jianfeng Lu. Evaluating functional connectivity of executive control network and frontoparietal network in alzheimer's disease. *Brain research*, 1678:262–272, 2018.

[51] ADHD-200 consortium. The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience*, 6:62, 2012.

A Algorithm

The IBAHGT algorithm is described in Algorithm 1. To clearly present the process of our algorithm, the IBAHGT algorithm is described in the form of Algorithm 1. In the implementation process, we express the information-bottleneck-guided adaptive hypergraph convolution and the information-bottleneck-guided node-level adaptive fusion in matrix form to accelerate computation.

Algorithm 1: Information-Bottleneck-Guided Adaptive Hypergraph Transformer

Input : Initial node features $\mathbf{X}^0$; Incidence matrix $\mathbf{H}$
Initialize : The learnable weight matrix $\mathbf{W}_1^l$, $\mathbf{W}_2^l$; The learnable vector $\mathbf{a}_1^l$, $\mathbf{a}_2^l$, θ^l
Output : The predicted class label $\hat{y}$

for *Layers* $l = 1, 2, \ldots, L$ **do**
  Stage 1: Information-Bottleneck-Guided Adaptive Hypergraph Convolution
    Phase 1:Adaptive Gathering of Node Features to Hyperedges
      for $e_j \in \mathcal{E}$ **do**
        for $v_i \in \mathcal{V}$ **do**
          Construct Node Set $\mathcal{N}_v(e_j) = \{v_i \in \mathcal{V} \mid \mathbf{H}_{i,j} \neq 0\}$;
          $\alpha_{i,j}^l = \frac{\exp\left(\mathbf{a}_1^l \mathbf{W}_1^l \mathbf{X}_i^{l-1}\right)}{\sum_{v_k \in \mathcal{N}_v(e_j)} \exp\left(\mathbf{a}_1^l \mathbf{W}_1^l \mathbf{X}_k^{l-1}\right)}$;
        end
        $\mathbf{Z}_j^l = \sigma\left(\sum_{v_i \in \mathcal{N}_v(e_j)} \alpha_{i,j}^l \mathbf{W}_1^l \mathbf{X}_i^{l-1}\right)$;
      end
      $\mu_l \leftarrow \mathbf{Z}^l[0 : f_z^l]$;
      $\sigma_l^2 \leftarrow \text{softplus}(\mathbf{Z}^l[f_z^l : 2f_z^l])$;
      $\mathbb{P}(\mathbf{Z}^l \mid \mathbf{X}^{l-1}, \mathbf{H}) \sim \text{Gaussian}(\mu_l, \sigma_l^2)$;
    Phase 2:Adaptive Aggregating of Hyperedge Features to Nodes
      for $v_i \in \mathcal{V}$ **do**
        for $e_j \in \mathcal{E}$ **do**
          Construct Hyperedge Set $\mathcal{N}_e(v_i) = \{e_j \in \mathcal{E} \mid \mathbf{H}_{i,j} \neq 0\}$;
          $\beta_{i,j}^l = \frac{\exp\left(\mathbf{a}_2^l \mathbf{W}_2^l \mathbf{Z}_j^l\right)}{\sum_{e_p \in \mathcal{N}_e(v_i)} \exp\left(\mathbf{a}_2^l \mathbf{W}_2^l \mathbf{Z}_p^l\right)}$;
        end
        $\mathbf{X}_{h,i}^l = \sigma\left(\sum_{e_j \in \mathcal{N}_e(v_i)} \beta_{i,j}^l \mathbf{W}_2^l \mathbf{Z}_j^l\right)$;
      end
      $\hat{\mu}_l \leftarrow \mathbf{X}_h^l[0 : f_h^l]$;
      $\hat{\sigma}_l^2 \leftarrow \text{softplus}(\mathbf{X}_h^l[f_h^l : 2f_h^l])$;
      $\mathbb{P}(\mathbf{X}_h^l \mid \mathbf{Z}^l, \mathbf{H}) \sim \text{Gaussian}(\hat{\mu}_l, \hat{\sigma}_l^2)$;
  Stage 2: Transformer Encoder
    $\mathbf{X}_t^l \leftarrow \text{TransEncoder}(\mathbf{X}^{l-1})$
  Stage 3: Information-Bottleneck-Guided Node-Level Adaptive Fusion
    for $v_i \in \mathcal{V}$ **do**
      $\mathbf{X}_i^l = \theta_i^l \mathbf{X}_{h,i}^l + (1 - \theta_i^l)\mathbf{X}_{t,i}^l$;
    end
    $\tilde{\mu}_l \leftarrow \mathbf{X}^l[0 : f^l]$;
    $\tilde{\sigma}_l^2 \leftarrow \text{softplus}(\mathbf{X}^l[f^l : 2f^l])$;
    $\mathbb{P}(\mathbf{X}^l \mid \mathbf{X}_h^l, \mathbf{X}_t^l) \sim \text{Gaussian}(\tilde{\mu}_l, \tilde{\sigma}_l^2)$;
end
$\hat{\mathbf{x}} = \text{Readout}(\mathbf{X}^L))$;
$\hat{y} = \text{MLP}(\hat{\mathbf{x}})$;
return $\hat{y}$;

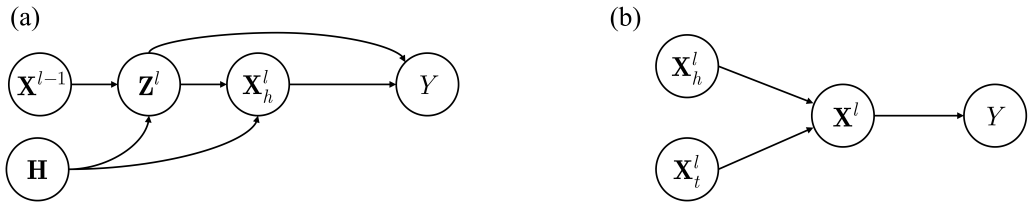


Figure 5: (a) The Markovian chain of information bottleneck-guided adaptive hypergraph convolution. (b) The Markovian chain of information bottleneck-guided node-level adaptive fusion.

B Proof

Proof for Proposition 1 We restate Proposition 1:

$$I(Y;\mathbf{Z}^l,\mathbf{X}_h^l) = I(Y;\mathbf{Z}^l) + I(Y;\mathbf{X}_h^l \mid \mathbf{Z}^l)\,. \tag{21}$$

For any probabilistic distribution functions $\mathbb{Q}_1(Y \mid \mathbf{Z}^l)$, $\mathbb{Q}_1(Y \mid \mathbf{Z}^l,\mathbf{X}_h^l)$, and $\mathbb{Q}_2(Y)$, we can express $I(Y;\mathbf{Z}^l)$ and $I(Y;\mathbf{X}_h^l \mid \mathbf{Z}^l)$ as follows:

$$I(Y;\mathbf{Z}^l) \geq 1 + \mathbb{E}\left[\log\frac{\mathbb{Q}_1(Y \mid \mathbf{Z}^l)}{\mathbb{Q}_2(Y)}\right] + \mathbb{E}_{\mathbb{P}(Y)\mathbb{P}(\mathbf{Z}^l)}\left[\frac{\mathbb{Q}_1(Y \mid \mathbf{Z}^l)}{\mathbb{Q}_2(Y)}\right], \tag{22}$$

$$I(Y;\mathbf{X}_h^l \mid \mathbf{Z}^l) \geq 1+\mathbb{E}_{\mathbb{P}(Y,\mathbf{X}_h^l\mid\mathbf{Z}^l)}\left[\log\frac{\mathbb{Q}_1(Y \mid \mathbf{Z}^l,\mathbf{X}_h^l)}{\mathbb{Q}_2(Y)}\right]+\mathbb{E}_{\mathbb{P}(Y)\mathbb{P}(\mathbf{X}_h^l\mid\mathbf{Z}^l)}\left[\frac{\mathbb{Q}_1(Y \mid \mathbf{Z}^l,\mathbf{X}_h^l)}{\mathbb{Q}_2(Y)}\right] \tag{23}$$

The Nguyen, Wainright & Jordan's bound is used here:

Lemma 1 For any two random variables $\mathbf{X}_1$ and $\mathbf{X}_2$ and any function g: $g(\mathbf{X}_1,\mathbf{X}_2) \in \mathbb{R}$, we have

$$I(X_1,X_2) \geq \mathbb{E}[g(X_1,X_2)] - \mathbb{E}_{P(X_1)P(X_2)}\left[\exp(g(X_1,X_2)-1)\right]. \tag{24}$$

The above lemma is used to $(Y;\mathbf{Z}^l)$. Plugging in $1+\log\frac{\text{Cat}(\hat{y}_z^l)}{\mathbb{P}(Y)}$, the right hand side of Eq.(24) is substituted by the cross-entropy loss, i.e.,

$$I(Y;\mathbf{Z}^l) \to -\mathcal{L}_{CE}(\hat{y}_z^l, Y)\,. \tag{25}$$

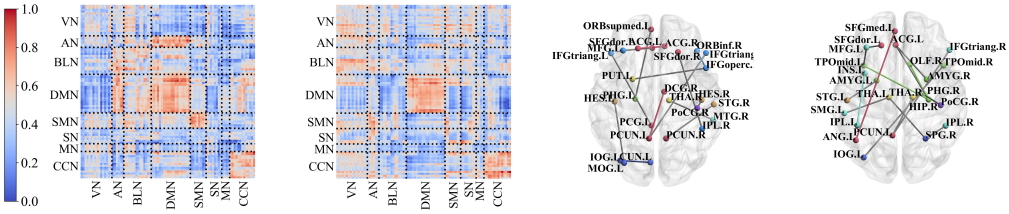
The above lemma is also used to $(Y;\mathbf{X}_h^l \mid \mathbf{Z}^l)$. Plugging in $1+\log\frac{\text{Cat}(\hat{y}_x^l)}{\mathbb{P}(Y)}$, it can be transformed into the following cross-entropy loss, i.e.,

$$I(Y;\mathbf{X}_h^l \mid \mathbf{Z}^l) \to -\mathcal{L}_{CE}(\hat{y}_x^l, Y)\,. \tag{26}$$

Proof for Proposition 2 According to the Markovian chain in Fig. 5(a), we restate Proposition 2: For any distributions $\mathbb{Q}(\mathbf{Z}^l)$ and $\mathbb{Q}(\mathbf{X}_h^l)$, we have

$$\begin{aligned}
I(\mathbf{X}^{l-1},\mathbf{H};\mathbf{X}_h^l,\mathbf{Z}^l) &\leq I(\mathbf{X}^{l-1},\mathbf{H};\mathbf{Z}^l) + I(\mathbf{Z}^l,\mathbf{H};\mathbf{X}_h^l)\\
&= \mathbb{E}\left(\log\frac{\mathbb{P}(\mathbf{Z}^l \mid \mathbf{X}^{l-1},\mathbf{H})}{\mathbb{Q}(\mathbf{Z}^l)}\right) - D_{KL}\left(\mathbb{P}(\mathbf{Z}^l)\|\mathbb{Q}(\mathbf{Z}^l)\right)\\
&+ \mathbb{E}\left(\log\frac{\mathbb{P}(\mathbf{X}_h^l \mid \mathbf{Z}^l,\mathbf{H})}{\mathbb{Q}(\mathbf{X}_h^l)}\right) - D_{KL}\left(\mathbb{P}(\mathbf{X}_h^l)\|\mathbb{Q}(\mathbf{X}_h^l)\right)\\
&\leq \mathbb{E}\left(\log\frac{\mathbb{P}(\mathbf{Z}^l \mid \mathbf{X}^{l-1},\mathbf{H})}{\mathbb{Q}(\mathbf{Z}^l)}\right) + \mathbb{E}\left(\log\frac{\mathbb{P}(\mathbf{X}_h^l \mid \mathbf{Z}^l,\mathbf{H})}{\mathbb{Q}(\mathbf{X}_h^l)}\right)\\
&= \text{ZIB}^l + \text{XIB}^l
\end{aligned} \tag{27}$$

$$\text{ZIB}^l = D_{KL}\left(\mathbb{P}(\mathbf{Z}^l \mid \mathbf{X}^{l-1},\mathbf{H}) \parallel \mathbb{Q}(\mathbf{Z}^l)\right), \text{XIB}^l = D_{KL}\left(\mathbb{P}(\mathbf{X}_h^l \mid \mathbf{Z}^l,\mathbf{H}) \parallel \mathbb{Q}(\mathbf{X}_h^l)\right). \tag{28}$$



(a) ASD vs. NC (b) AD vs. NC (c) ASD vs. NC (d) AD vs. NC

Figure 6: Visualization of attention maps and discriminative connections. connections within the same neural system (VN, AN, BLN, DMN, SMN, SN, MN, CCN) are colored accordingly, while connections across different systems are colored gray.

Proof for Proposition 3 According to the Markovian chain in Fig. 5(b), we restate Proposition 3: For any distribution $\mathbb{Q}(\mathbf{X}^l)$, we have

$$\begin{aligned} I(\mathbf{X}_h^l, \mathbf{X}_t^l; \mathbf{X}^l) &= \mathbb{E}(\log \frac{\mathbb{P}(\mathbf{X}^l \mid \mathbf{X}_h^l, \mathbf{X}_t^l)}{\mathbb{Q}(\mathbf{X}^l)}) - D_{KL}(\mathbb{P}(\mathbf{X}^l) \| \mathbb{Q}(\mathbf{X}^l)) \\ &\leq \mathbb{E}(\log \frac{\mathbb{P}(\mathbf{X}^l \mid \mathbf{X}_h^l, \mathbf{X}_t^l)}{\mathbb{Q}(\mathbf{X}^l)}) \\ &= D_{KL}(\mathbb{P}(\mathbf{X}^l \mid \mathbf{X}_h^l, \mathbf{X}_t^l) \parallel \mathbb{Q}(\mathbf{X}^l)\,. \end{aligned} \tag{29}$$

C Additional Experiments.

C.1 Analysis of Attention Maps and Discriminative Connections

To evaluate the contribution of the self-attention mechanism in the Transformer module, we visualize the average attention maps along the group level, as shown in Fig.6. Additionally, to clearly display the variations within and between sub-networks, we divide the entire brain network into eight sub-networks: visual network (VN), auditory network (AN), basal ganglia network (BLN), default mode network (DMN), sensorimotor network (SMN), salience network (SN), motor network (MN), and cognitive control network (CCN). In the ASD vs. NC classification task, the attention scores within the SMN are relatively high, indicating that the inter-ROI connectivity within the SMN plays a significant role in ASD prediction. This is consistent with previous studies that have shown reduced functional connectivity within the SMN [46], reflecting alterations in sensory and motor processing [47]. In the AD vs. NC classification task, the positions with higher attention scores are primarily located within the DMN. Previous studies have indicated that the substantial accumulation of β-amyoid protein in the DMN of AD patients leads to reduced functional connectivity in this region, which is associated with cognitive impairments in AD patients [48].

We perform a t-test on attention maps to identify discriminative connections. For clarity, only connections with significant differences (p-value < 0.01) are shown in Fig. 6. In the ASD vs. NC classification task, the identified discriminative connections include those between the left anterior cingulate and paracingulate gyri (ACG.L) and right Inferior parietal (IPL.R), as well as between the right superior frontal gyrus, dorsolateral (SFGdor.R). This validates previous research, which suggests that changes in these connections are often associated with the repeat stereotypical behaviors commonly observed in individuals with ASD [49]. In the AD vs. NC classification task, the discriminative connections include the connection between the left middle frontal gyrus (MFG.L) and the left superior frontal gyrus, dorsolateral (SFGdor.L). This is consistent with previous findings [50], which suggest that changes in functional connectivity between the right frontal lobe and the superior frontal gyrus in AD patients are associated with cognitive decline.

C.2 Hyperparameter Analysis

To evaluate the influence of hyperparameters on classification performance, we conduct controlled variable experiments on two main hyperparameters: the value of k for KNN-based hypergraph

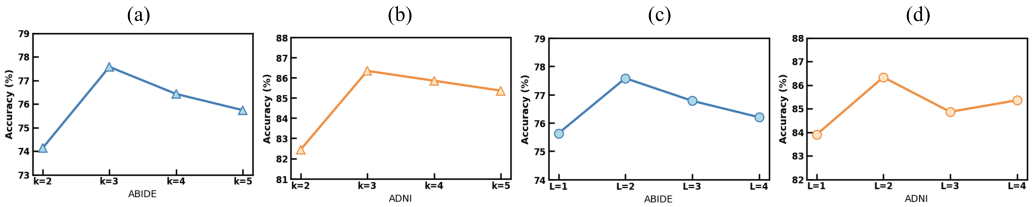


Figure 7: The effect of varying hyperparameters. (a) and (c) are experiments on ABIDE dataset. (b) and (d) are experiments on ADNI dataset.

Table 3: Running time with different methods.

Method	Running Time on ABIDE(s)	Running Time on ADNI(s)
BrainNetCNN	39.75 ± 1.03	8.04 ± 0.98
ContrastPool	334.89 ± 4.23	83.62 ± 2.50
BrainGNN	294.01 ± 1.11	74.97 ± 2.08
BrainIB	212.43 ± 1.32	51.66 ± 1.58
HGNN$^+$	36.55 ± 1.22	7.63 ± 3.22
FC-HAT	82.56 ± 0.69	22.76 ± 1.08
BrainNetTF	49.72 ± 1.02	10.82 ± 0.34
Com-BrainTF	61.26 ± 2.31	14.32 ± 0.74
ALTER	54.87 ± 0.98	11.35 ± 2.62
IBAHGT	72.43 ± 0.88	15.97 ± 0.49

construction and the number of layers L in the information bottleneck-guided hypergraph Transformer. Fig.7(a) and Fig.7(b) illustrate the impact of different values of k on two datasets. The accuracy is highest when $k = 3$ for both classification tasks. The accuracy is highest when $k = 3$ for both classification tasks. When $k = 2$, the hypergraph degenerates into a graph, causing a loss of high-order information in the brain network. When $k > 3$, the information aggregation range increases, and the representations of different nodes become more homogeneous, resulting in a drop in accuracy.

Fig.7(c) and Fig.7(d) show the variation in accuracy with respect to the number of layers L on two datasets. We find that when the number of layers $L = 2$, the model achieved the highest accuracy in both classification tasks. When $L < 2$, the model's depth was insufficient to effectively capture the complex dependencies within the brain network. On the other hand, when $L > 2$, an excessive number of layers led to overfitting, resulting in a decline in performance.

C.3 Running Time

Table 3 shows the running time of IBAHGT compared to other methods on ABIDE and ADNI datasets. BrainGNN and ContrastPool are significantly slower than the other methods, primarily due to their pooling processes, which consume more time. BrainIB dynamically samples subgraphs during training and exhibits slower convergence, resulting in a longer training time. The hypergraph-based method FC-HAT is slower than both our IBAHGT and HGNN$^+$, mainly because FC-HAT dynamically updates the hypergraph structure at each layer using KNN and k-means methods, which adds additional time cost. The adaptive hypergraph convolution and node-level fusion in the proposed IBAHGT are formulated in matrix form during implementation to accelerate training on GPU. Although the training time is slightly longer compared to Transformer-based methods, this increase is acceptable given the significant performance improvement.

C.4 Baselines and Additional Comparative Experiments on ADHD-200 Dataet

The selected baselines mainly include three categories. The first category consists of convolutional neural network (CNN)-based method, such as BrainNetCNN [36]. BrainNetCNN employs new feature extraction structures based on the topological locality characteristics of brain networks. The second category includes graph or hypergraph-based methods, such as ContrastPool [37], BrainGNN

Table 4: Comparison experiments results on the ADHD-200 dataset.

Datasets (Tasks)	ADHD-200 (ADHD vs. NC)			
Method	ACC (%)	SPE (%)	SEN (%)	AUC (%)
BrainNetCNN	62.3±1.7	67.6±2.7	53.5±1.5	61.8±1.4
ContrastPool	61.9±1.6	69.0±4.0	50.4±4.2	60.6±3.5
BrainGNN	62.5±1.9	67.3±3.3	54.7±4.5	62.3±2.1
BrainIB	63.5±2.0	68.7±2.8	55.2±5.0	63.4±2.2
HGNN$^+$	64.4±1.0	68.9±2.5	57.2±5.8	64.1±1.6
FC-HAT	66.5±1.1	70.1±2.0	60.6±4.5	66.3±1.2
BrainNetTF	67.2±1.7	70.4±5.7	62.0±5.0	66.6±1.1
Com-BrainTF	67.7±1.7	69.7±1.7	64.5±3.9	67.6±2.5
ALTER	68.8±2.1	**74.1±4.6**	60.3±7.6	68.0±2.4
IBAHGT	**72.3±1.3**	73.7±1.9	**69.9±4.0**	**72.2±1.2**

[22], BrainIB [28], HGNN$^+$ [38], and FC-HAT [15]. ContrastPool generates contrast graph through a dual-attention module to guide differentiable graph pooling, which is used to produce effective brain network representations for disease classification. BrainIB introduces the IB principle into brain network analysis to effectively identify disease-specific prominent brain network connections. HGNN$^+$ is a general hypergraph convolution method. FC-HAT combines KNN and k-means to dynamically construct hypergraphs and captures information within the hypergraph during the hypergraph aggregation attention phase. The third category consists of Transformer-based methods, including BrainNetTF [19], Com-BrainTF [39], and ALTER [40]. BrainNetTF employs self-attention to learn the long-range dependencies between ROIs and utilizes orthogonal clustering for graph-level representation. Com-BrainTF is a hierarchical local-global transformer architecture that learns intra- and inter-community aware node embeddings. ALTER proposes a brain graph transformer to capture long-range dependencies between ROIs utilizing biased random walk. To ensure a fair comparison, we conduct experiments using the open-source code provided in the papers and the optimal parameters specified therein.

To further validate the performance of IBAHGT, the ADHD-200 dataset [51] is also used for comparative experiments. The ADHD-200 dataset is a multi-site neuroimaging resource for attention deficit hyperactivity disorder (ADHD) research, consisting of 357 ADHD patients and 573 normal controls (NCs) from eight international imaging centers. As shown in Table 4, the experimental results demonstrate that IBAHGT outperforms the baselines on the majority of metrics, further validating the superiority of the proposed method in brain disease diagnosis.

D Further Discussion.

D.1 Limitations.

This study primarily focuses on functional brain networks constructed from fMRI data, providing a unified analysis of high-order relationships and both short- and long-range dependencies in brain network. However, the influence of structural brain networks constructed from DTI data on the diverse dependencies among ROIs in functional brain networks are also worthy of exploration. In future work, we will delve into the effects of structural connectivity on multiple types of dependencies in functional brain network and propose methods to integrate functional and structural brain networks for more comprehensive analyses of high-order correlations and both short- and long-range dependencies among ROIs.

D.2 Possible Societal Impacts.

This research involves brain disease diagnosis, and it is therefore necessary to acknowledge its potential societal impact. The proposed method contributes to the discovery of potential biomarkers, offering positive impacts for neuroscience research and computer-aided diagnosis. However, the use of artificial intelligence in diagnostic assistance may lead to misdiagnoses, which are difficult

to completely avoid and could have negative impacts for both patients and society. In real-world clinical settings, AI-based methods should serve solely as decision-support tools, with final diagnostic decisions remaining the responsibility of medical professionals.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction accurately reflect the paper's contributions and scope, namely, the implementation of a unified framework for modeling and learning high-order relationships, as well as both short- and long-range dependencies in brain networks. This framework further condenses a minimal sufficient representation from the information captured across different relationships for brain disease diagnosis.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of IBAHGT in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For each theoretical result, we provide the full set of assumptions and a complete (and correct) proof in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In the paper, we provide comprehensive details of all information necessary for reproducing the experimental results, including the datasets and preprocessing procedures, baselines, evaluation metrics, and experimental settings, thereby ensuring the reproducibility of the research.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the code in the supplementary material and will release it publicly upon the paper's acceptance. The experiments are conducted on three publicly available datasets, which can be directly accessed and downloaded from the internet.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We elaborate on the dataset splitting strategy, hyperparameters selection, type of optimizer, and other relevant implementation details in Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All results are the average of 10 random runs on test sets with the standard deviation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details on computer resources can be found in the experimental settings subsection of Section 4.1, while the time of execution is reported in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research is conducted in compliance with the NeurIPS Code of Ethics. Our experiments do not involve human subjects and potential harmful consequences for society.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Our study involves brain disease diagnosis, and we outline both potential positive and negative societal impacts of its practical application in Appendix D.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We explicitly cite the original papers for all the code used in this study.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: In this paper, the core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.