Causal Invariance-aware Augmentation for Brain Graph Contrastive Learning

Minqi Yu [1] **Jinduo Liu** [† 1] **Junzhong Ji** [1]

Abstract

Deep models are increasingly used to analyze brain graphs to diagnose and understand brain diseases. However, due to the multi-site data aggregation and individual differences, brain graph datasets exhibit widespread distribution shifts, which impair the model's generalization ability to the test set, thereby limiting the performance of existing methods. To address these issues, we propose a Causally Invariance-aware Augmentation for brain Graph Contrastive Learning, called CIA-GCL. This method first generates a brain graph by extracting node features based on the topological structure. Then, a learnable brain invariant subgraph is identified based on a causal decoupling approach to capture the maximum label-related invariant information with invariant learning. Around this invariant subgraph, we design a novel invariance-aware augmentation strategy to generate meaningful augmented samples for graph contrast learning. Finally, the extracted invariant subgraph is utilized for brain disease classification, effectively mitigating distribution shifts while also identifying critical local graph structures, enhancing the model's interpretability. Experiments on three real-world brain disease datasets demonstrate that our method achieves state-of-the-art performance, effectively generalizes to multi-site brain datasets, and provides certain interpretability. The code is available at https://github.com/qinsheng1900/CIA-GCL.

1. Introduction

Brain connectomes have consistently been regarded as a rich source of information in neuroscience and neuroinformatics (Sporns et al., 2005), and their value has become more apparent in recent years (Liu et al., 2022; Bazinet et al., 2023; Liu et al., 2024a). Functional magnetic resonance imaging (fMRI) provides a non-invasive solution to capture abnormal interactions between regions of interest (ROIs) in the brain by tracking blood oxygen level-dependent (BOLD) signals that naturally fluctuate over time (Chen et al., 2017) (Chong et al., 2019). Based on fMRI data, the brain functional connectivity (BFC), which captures statistical dependencies between ROIs, is modeled as an undirected graph. BFC abnormalities are often linked to cerebral diseases, making brain graph analysis valuable for computer-aided diagnosis of many brain diseases (Liu et al., 2024b; Taspinar & Ozkurt, 2024; Ji et al., 2024).

In the field of brain graph analysis, the rapid development of deep learning techniques has been notable in recent years. Cui et al. proposed a pipeline of node feature construction, message passing, and graph pooling for brain graph analysis (Cui et al., 2023). Zheng et al. introduced a Granger causality-inspired network that identifies causally relevant subgraphs for disease diagnosis (Zheng et al., 2024a). However, these existing methods fail to handle the characteristics of brain graph data, resulting in poor performance:

- **Local distinguishing biomarkers resist capture.** The brain graph exhibits "small-world" properties with a degree distribution following a power law, concentrating key information in modular structures (Liang et al., 2010). However, the low signal-to-noise ratio (SNR) and high dimensionality of fMRI data (Vizioli et al., 2021) make it challenging for models to focus on critical features, limiting the extraction of discriminative neuroimaging biomarkers for disease detection.
- **Data shift hinders generalization.** Due to multi-site collection protocols and individual differences (Eslami et al., 2021; Mueller et al., 2013), the brain graph data often suffers from significant distribution shifts. These shifts hinder the generalization of models, leading to degraded performance on test sets. Models may rely on spurious correlations, such as site-specific information or extracting fragile features from specific samples that lack robustness across test distributions.

Graph contrastive learning (GCL) (Sun et al., 2021) has

[1]Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, College of Computer Science, Beijing University of Technology, Beijing, China. Correspondence to: Jinduo Liu <jinduo@bjut.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

emerged as an effective approach to cope with fMRI data due to its robustness in capturing distinguishing features across different graph augmentations. Peng et al. used sliding windows to build positive samples from fMRI images and applied semi-supervised learning for classification (Peng et al., 2023). Xu et al. introduce ContrastPool with a contrastive dual-attention block to improve graph neural network (GNN) based model (Xu et al., 2024a). Although these methods have achieved promising results, their random truncating of BOLD signals or changing edges to construct augmented samples may destroy the local structure with rich saliency information. These operations may disrupt the semantic feature in brain graph data and thus cannot cope with the data shift well (Soon et al., 2021).

Meanwhile, the invariant learning (Zhu et al., 2024; Sui et al., 2024) emerges as a prevalent strategy for tackling the challenge of generalization to out-of-distribution (OOD) data which aims to exploit the invariant relationships with labels across different distributions (Chang et al., 2020; Ahuja et al., 2021). Therefore, if the invariant properties of the brain graph can be used to guide the augmented sample generation, it is expected to identify the distinguishing markers and mitigate the impact of the data distribution shift.

In this paper, we propose a Causally Invariance-aware Augmentation for brain Graph Contrastive Learning, called CIA-GCL. We begin by constructing the brain graph using the complete BOLD data, then analyze the brain graph data from a causal perspective. Under the causal assumption, we enable the model to progressively learn how to extract invariant subgraphs during optimization. And through the constraints of two loss functions, the invariant subgraph maintains two properties of *Causal Relationship with Label* and *Invariance Property*. Subsequently, we design a novel invariance-aware augmentation strategy based on the invariant subgraph to generate an augmented sample set. This strategy ensures that the augmented samples *retain the local structures, label preservation, and provide diversity*. In GCL, we use the invariant subgraph as the anchor graph and the augmented sample set as the positive samples to enhance the model's ability to capture discriminative features. The principal contributions can be summarized as follows:

- We propose a novel CIA-GCL framework for brain graph analysis, designed to address the challenges of data shift in multi-site brain data.
- We propose a brain invariant subgraph extraction method based on causal disentanglement and invariant learning to better capture the discriminative local structures in brain graphs.
- We design a novel invariance-aware augmentation strategy to generate augmented samples with diverse distribution shifts, facilitating invariant subgraph learning.
- Systematic experiments conducted on three real disease datasets demonstrate that the proposed method outperforms several state-of-the-art methods.

2. Related Works

2.1. Brain Graph Analysis

Medical research links brain diseases to abnormal brain network patterns (Rudie et al., 2013), and brain graph analysis identifies connectivity changes to aid in diagnosis and understanding. The GCL-based methods, with their robustness in capturing discriminative features from graph augmentations, show potential for brain graph analysis.

Wang et al. calculated correlations of truncated BOLD signals to build positive samples for contrastive learning to create node features (Wang et al., 2022). Yang et al. proposed a GNN pretraining framework that leverages contrastive learning for brain graph analysis (Yang et al., 2023). A-GCL used adversarial contrastive learning to extract sparse graph-level features for analysis (Zhang et al., 2023). CMV-CGCN leveraged GCL to integrate functional and high-order functional connectivities along with phenotypic information for disease diagnosis (Zhu et al., 2023). Zong et al. (2024) employs GCL to optimize and constrain the learning process of brain connections, leading to the reconstruction of the brain network. Xu et al. (2024b) propose a novel contrastive brain network transformer, Contrasformer, which aligns brain region features across subjects via contrastive learning. Compared to the above methods, we combine invariant learning with GCL, and specifically design an invariance-aware augmentation strategy tailored to brain graph data. This strategy ensures semantic consistency while enhancing sample diversity and serves as a bridge between the GCL and the invariant learning. Detailed comparisons can be found in Appendix B.

2.2. Graph Invariant learning

Graph invariant learning is an approach to OOD generalization that aims to exploit the stable relationships between features and labels across different distributions. According to (Li et al., 2022a), such methods can be broadly categorized into two groups. The first category is invariance optimization, which assumes part of the input captures invariant label relationships across environments, enabling OOD generalization (Arjovsky et al., 2019). The second category is explicit representation alignment, which improves generalization by explicitly aligning graph representations across multiple environments.

CIGA leveraged causal models to capture graph invariance, ensuring OOD generalization by focusing on the most informative subgraphs (Chen et al., 2022). GIL automatically inferred environment labels and learned the invariant sub-

graph in a mixture of latent environments (Li et al., 2022b). Jia et al. propose a graph invariant learning method with a mixup strategy to generate diverse environments for enhancing OOD generalization (Jia et al., 2024). Wu et al. introduced a causal inference-based approach to train GNNs, enhancing generalization without environment label knowledge (Wu et al., 2024). BrainOOD enhances GNNs' OOD generalization in brain graph by leveraging an improved graph information bottleneck (Xu et al., 2025). Unlike previous methods that rely on only a single graph invariance learning strategy, our approach combines both invariance optimization and explicit representation alignment through jointly optimizing two invariance-driven objectives to more effectively address the OOD problem.

3. Methodology

We first analyze multi-site brain data from a causal view, then provide definitions for brain invariant subgraph and good brain augmented samples, and then provide a concise introduction to the implementation process of the CIA-GCL.

3.1. A Causal View on Multi-site Brain Dataset

To address the two problems with brain graph data mentioned in the introduction, we attempt to analyze them from a causal perspective and propose the following assumptions.

Definition 3.1. (Brain Graph) A brain graph $G = (V, E)$: $V = \{v_i\}_{i=1}^{N}$ is the node set indicating brain regions and N represents the number of ROIs. $E = \{e_{ij}\}_{i,j=1}^{N}$ is the edge set describing the connection relationship between ROIs.

In this paper, the k-th subject's brain functional connection network is described using G_k to describe. The generation process of the G_k from the fMRI time series form is detailed in Appendix E2. In other literature (Said et al., 2023), brain graphs can also be referred to as brain functional networks.

Assumption 3.2. The brain graph G consists of G^c, which has a causal relationship with the label Y by f_{inv}, and G^s, which has no causal relationship with Y. G^s is generated by the external environment factors $\mathcal{E}$ through f_{env}.

$$G = G^c \cup G^s, Y = f_{inv}(G^c), G^s = f_{env}(\mathcal{E}).$$

The G^c refers to the localized brain structures that exhibit differences between patients and the typically developing control (TC). For example, (Doyle-Thomas et al., 2015) found that consistent abnormalities in the default mode network (DMN) were found across ASD patients. These findings support the assumption that a causally relevant subgraph (e.g., within DMN) exists and is shared across subjects, even under distribution shifts.

The environment $\mathcal{E}$ consists of various factors, such as site differences, individual differences, and the noise inherent in the brain graph itself. Under different environmental factors, the spurious variable G^s exhibits data with different distributional variations. Since the graph G is composed of G^c and G^s, and there exists a statistical dependency between G and the label Y, the mixture of G^c and G^s may lead to interference:

$$P(Y \mid G) = P(Y \mid G^c) + \Delta G^s, \quad \Delta G^s \neq 0.$$

where G^s can be regarded as an interfering factor. The statistical dependency of G^s through G interferes with the relationship between G^c and Y.

The distribution shift problem refers to the phenomenon where subjects sharing the same underlying invariant substructure G^c may still appear statistically different due to the influence of G^s. Therefore, separating G^c and G^s in the brain graph and only using G^c for disease diagnosis can reduce brain graph complexity, for discriminative local feature extraction, and mitigate data distribution shifts for boosting model generalization.

Definition 3.3. (Brain Invariant Subgraph) Given a brain graph $G = (V, E)$ and under the Assumption 3.2, the brain invariant subgraph $G^{inv} = (V^{inv}, E^{inv})$, where $V^{inv} \subseteq V$, $E^{inv} \subseteq E$ satisfying:

a. Causal Relationship with Label Y: The invariant subgraph G^{inv} has a causal relationship with the label Y

Theorem 3.4. *If the G^{inv} obtains the causal property, it shares the maximum information with invariant subgraphs from subjects with the same label:*

$$G_k^{inv} = \underset{m \in \{i \mid Y_i = Y_k, i \neq k\}}{\arg\max} I\left(G_k^{inv}; G_m^{inv}\right),$$

where m represents a subject selected from the set of other subjects with the same label as k-th subject.

b. Invariance property: The G^{inv} captures the invariant relations with Y across different environment $\boldsymbol{e}$:

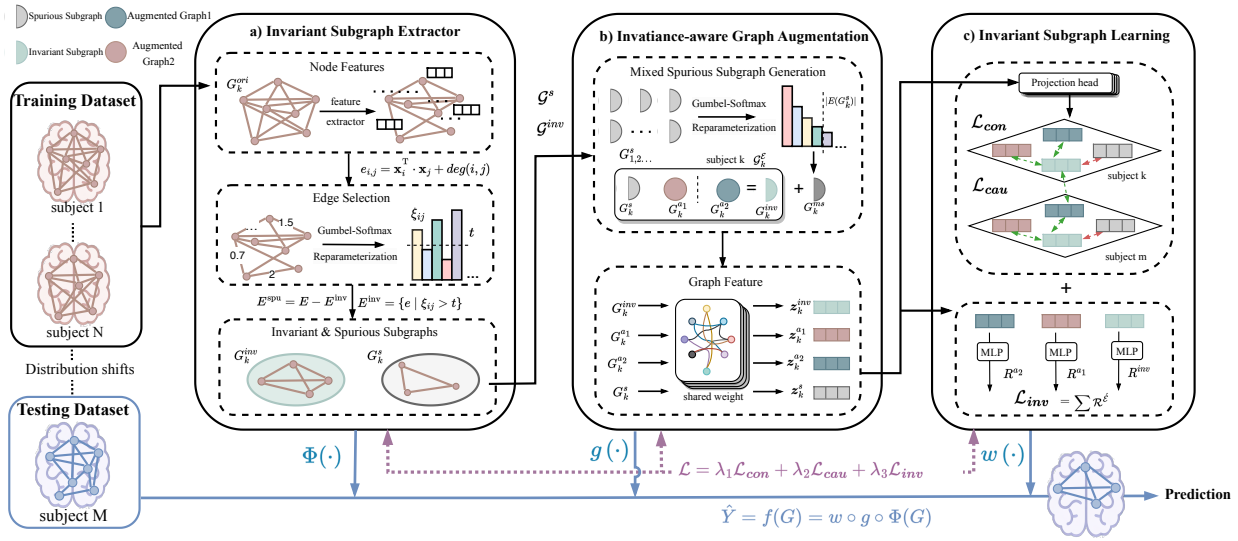
$$\forall \boldsymbol{e}, \boldsymbol{e}' \in \text{supp}(\mathcal{E}), P^{\boldsymbol{e}}(Y \mid G^{inv}) = P^{\boldsymbol{e}'}(Y \mid G^{inv}),$$

$\mathcal{E}$ is the set of possible environments in the study.

Theorem 3.5. *If the G^{inv} satisfies the invariant property, it follows that the G^{inv} maximizes the expected mutual information with the Y across the environment set $\mathcal{E}$:*

$$G^{inv} \in \arg\max \mathbb{E}_{\boldsymbol{e} \in \mathcal{E}}\left[I(G^{inv}; Y \mid \boldsymbol{e})\right].$$

Detailed proofs are given in Appendix D. According to our definition, under the Assumption 3.2, there exists a G^{inv} with the semantic information of the G that not only holds a causal relationship with the Y but also maintains its predictive consistency across various environments, i.e., $G^{inv} = G^c$. A detailed description of the G^{inv} extraction method, including its formulation and implementation, is provided in Section 3.3.



. *Figure 1.* The overall framework of the CIA-GCL. (a) an invariant subgraph extractor $\Phi(\cdot)$ to partition the entire graph G^{ori} into invariant subgraph G^{inv} and spurious subgraph G^s (b) displays the invariance-aware graph augmentation process by mixed spurious subgraph generation and using $g(\cdot)$ to extract graph-level feature. (c) The loss functions of CIA-GCL, $w(\cdot)$, represent the predictor. The brown graph structure represents the training process, blue represents the testing process, and purple stands for the iterative feedback process. The whole process of CIA-GCL can be summarized as $\hat{Y} = f(G) = w \circ g \circ \Phi(G)$.

3.2. Good Augmentation for Brain GCL

The literature (Wu et al., 2023) suggests that if the anchor graph retains essential information from the input graph, it ensures the stability of key information during contrastive learning, avoiding random disruptions and loss of critical features. Therefore, we apply G^{inv} as the anchor graph and generate brain graph augmentations. We name this method as the invariance-aware graph augmentation strategy.

Definition 3.6. (Good Brain Augmented Samples) Given a brain invariant subgraph G^{inv} as the anchor view, we define a good brain augmented sample for GCL can be obtained by $G^a = \zeta(G^{inv} \oplus \Delta G)$, where ΔG representing additional different information, $\oplus$ denotes a subgraph concatenation operation, and the ζ function is a learnable operation during the optimization process. Based on this definition, the augmented view has the following three properties:

a. Retain local structures of the brain graph: The augmented sample leverages the complete BOLD signals and uses subgraphs as modules for augmentation to preserve the local topological structure of the brain graph.

b. Label-perserving: The augmented view G^a maintains the relationship with Y, i.e., $P(Y \mid G^a) = P(Y \mid G^{inv})$.

c. Provide Diversity: The augmented view introduces diverse information that complements the anchor view, encouraging the model to learn a richer representation, i.e., $H(G^a) > H(G^{inv})$, $H(G^{a_1}) \neq H(G^{a_2})$.

According to Definition 3.3, a good brain augmented sample preserves the local topological structure without randomly changing nodes or edges and retains the critical relationship with the Y. Furthermore, redundant information between augmented samples should be minimized, while ensuring they carry complementary and diverse information to enhance the robustness of graph contrastive learning.

3.3. Causal Decoupling for Invariant Subgraph Extraction

Brain Node Features Representation. Before extracting the invariant subgraph, we first process the brain graph data to obtain the node features. Following the Edge-to-Edge (E2E) and Edge-to-Graph (E2N) components proposed by (Kawahara et al., 2017), we utilize the topological locality of the structural brain graph to update node features. The node features extraction module consists of three layers, the first layer L1 is an E2E layer, which contains a cross-shape filter with 16 output channels to update edge-level features e_{ij} from combining the weights of edges that share nodes i and j together, resulting in $\mathbb{R}^{16\times n\times n}$. L2 and L3 are both E2N layers, which contain 1 $\times$ n or n $\times$ 1 convolution filter with 32 channels. The node features are extracted from the output of L1 in both horizontal and vertical directions. Now, the graph is renewed to $G = (V, E, X, \dot{A}), X \in \mathbb{R}^{n\times d}$, where d refers to the feature vectors. $\dot{A}$ is the sparse adjacency matrix formed by selecting the top 5% values from E. Because according to (Said et al., 2023), the setting of sparse graphs enhances the model performance.

Learnable Invariant Subgraph Extractor. We obtain the G^{inv} by performing edge deletion operations, as shown in Eq. 1. Specifically, we define an edge mask matrix $\mathcal{M}^{inv}$,

which follows a Bernoulli distribution, $\mathcal{M}^{inv} \in \{0,1\}^{n\times n}$. If $\mathcal{M}_{ij}^{inv} = 1$, the corresponding edge e_{ij} is retained in the invariant subgraph. The $\circ$ represents the element-wise product. By applying $\mathcal{M}^{inv}$ to the sparse adjacency matrix $\dot{A}$, we can obtain the G^{inv}:

$$G^{inv} = \big(V, \dot{A} \circ \mathcal{M}^{inv}, X\big),\ \mathcal{M}^{inv} \sim \mathrm{Bern}(\xi_{ij}), \quad (1)$$

$$\xi_{ij} = \sigma\big(\mathrm{MLPs}(\mathbf{x}_i^{\mathrm{T}} \cdot \mathbf{x}_j) + \deg(i,j)\big). \quad (2)$$

Considering the importance of node degrees in brain graphs, we divide the edge features into two components. The first component is $\mathbf{x}_i^{\mathrm{T}} \cdot \mathbf{x}_j$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the node features of the endpoints of the edge e_{ij}, followed by a linear layer. The second component, $deg(i,j) = Scale_{[0,1]}(deg(i)) + Scale_{[0,1]}(deg(j))$, is obtained by normalizing the degrees of the endpoints using Min-Max scaling. The edge features are processed through the $\sigma(\cdot)$ sigmoid function, resulting in ξ_{ij}, which represents the probability of e_{ij} being retained in the brain invariant subgraph.

However, since Bernoulli sampling is discrete and non-differentiable, directly sampling from $\mathrm{Bern}(\xi_{ij})$ blocks gradient backpropagation, hindering the model in learning to prune edges for the invariant subgraph during optimization. To address this, we adopt the Gumbel-Softmax technology to approximate the generation of the binary behavior of $\mathcal{M}^{inv}$, which can be formulated as:

$$p_{ij} = \mathrm{GumbelSoftmax}(\xi_{ij}), \quad (3)$$

$$t_k = \mathrm{Quantile}\big(\{p_{ij} \mid i,j = 1,\dots,n\}, r\big), \quad (4)$$

$$\mathcal{M}_k^{inv} = \mathbb{1}(p_{ij} > t_k). \quad (5)$$

We first employ Gumbel-Softmax (Jang et al., 2022) to transform the probability ξ_{ij} into a binary concrete distribution p_{ij} to approximate discrete selections, then determine the binary matrix $\mathcal{M}_k^{inv}$ through threshold selection.

To address the first problem in the introduction, we set a maximum edge proportion r for the invariant subgraph to ensure that the extracted subgraph remains compact. According to Eq. 4, we calculate the minimum threshold t_k. In Eq. 5, $\mathbb{1}(\cdot)$ denotes an indicator function, where an edge is retained if $p_{ij} > t_k$. Similarly, the spurious subgraph $G^s = (V, \dot{A} \circ (1 - \mathcal{M}^{inv}), X)$, utilizes the unselected edges.

Decoupling Causal and Spurious Subgraph. We have divided each subject's brain graph G into an invariant subgraph G^{inv} and a spurious subgraph G^s. According to Assumption 3.2, the G^{inv} has a causal relationship with the label Y, while the G^s acts as a confounding factor, disrupting the relationship between G^{inv} and Y. To mitigate the interference of G^s , we propose the following causal learning objective to disentangle the G^{inv} and G^s:

$$\max\big(I(G^{inv}, Y) - I(G^{inv}, G^s)\big). \quad (6)$$

The first term, $\max I(G^{inv}, Y)$, ensures that G^{inv} contains as much causal information as possible, strengthening the correlation between G^{inv} and Y. The second term, $\min I(G^{inv}, G^s)$, reduces the correlation between the invariant subgraph and the spurious subgraph. This prevents G^s from interfering with G^{inv}, preserving the causal relationship with the label Y. By optimizing this causal learning objective, the model can better capture the discriminative subgraph G^{inv} from confounded brain graph data, effectively addressing problem 1 in the introduction.

3.4. Invariance-Aware Graph Augmentations

Mixed Spurious Subgraph Generation. To create good augmented samples providing more diverse information for the anchor graph G^{inv}, we generate mixed spurious subgraphs G^{ms}. These subgraphs exhibit significant differences from G^{inv}, ensuring that the brain augmented samples obtained by combining G^{inv} and G^{ms} satisfies the conditions outlined in Definition 3.6.

Accounting for individual differences, we analyze each subject separately. For G_k, we first collect the G^s from the same batch with different labels. Because compared to G_k^{inv}, spurious subgraphs with different labels contain more diverse and complex information. Then, the p values by Eq.3 of all edges in the collected set $E^s_{Y_k}$ are used as parameters for Gumbel-Softmax sampling to obtain q_{ij}, which represents the probability of the edge e_{ij} being retained in G_k^{ms}:

$$E^s_{Y_k} = \big\{e \mid e \in E(G_i^s),\ G_i^s \in \{G_i^s \mid Y_i \neq Y_k\}\big\}, \quad (7)$$

$$q_{ij} = \mathrm{GumbelSoftmax}(p_{ij}), \quad \forall e_{ij} \in E^s_{Y_k} \quad (8)$$

Then, we can construct the G_k^{ms} :

$$\begin{aligned} G_k^{ms} &= (V_k^{ms}, E_k^{ms}) \\ &= \big(V_k^{ms}, \{q_{ij} \mid i,j = 1,\dots,n\}^{\downarrow}_{|E(G_k^s)|}\big), \end{aligned} \quad (9)$$

we sort all q_{ij} values in descending order and select the top $|E(G_k^s)|$ edges, $|E(G_k^s)|$ equals the number of edges in G_k^s.

Graph Augmentations. For k-th subject, we use the invariance-aware augmentation strategy to generate a positive sample set $\mathcal{G}_k^{pos}$. Each sample is constructed by centering around G_k^{inv} and augmenting it with additional information-rich ΔG. Because combining different ΔG, $\mathcal{G}^{\mathcal{E}}$ is a set of distribution-shifted graph views which are used to minimize both the $\mathcal{L}_{con}$ and $\mathcal{L}_{inv}$ to perform two types of graph invariant learning for OOD generalization.

Thus, the invariant subgraph extraction method and the augmentation strategy are tightly coupled: the augmentation relies on the quality of the extracted G^{inv}, and the learning objectives under the augmentation reinforce the extraction of robust, invariant substructures. And both subgraphs are generated using Gumbel-Softmax sampling, enabling the

learned subgraphs to be optimized end-to-end:

$$\begin{aligned}\mathcal{G}_k^{pos} &= \{G_k^{a_1}, \quad G_k^{a_2}\} \\ &= \{G_k^{inv} \oplus G_k^s, \quad G_k^{inv} \oplus G_k^{ms}\}. \end{aligned} \tag{10}$$

Notably, the original brain graph G, which is the $G_k^{a_1}$ already satisfies the conditions for a good augmented sample in Definition 3.6. Through the causal learning objective in Eq. 6, G^{inv} and G^s have been decoupled, ensuring that G^s contains information distinct from G^{inv}.

So far, every subject has four types of graphs, namely invariant subgraph G^{inv}, spurious subgraph G^s, and two augmented samples $G_k^{a_1}$ (which corresponds to the original graph G^{ori}) and $G_k^{a_2}$. We define each subject's graph structure as $\mathcal{G}^{\mathcal{E}}$, where $\{\mathcal{G}^{\mathcal{E}}, \mathcal{E} \in (inv, s, a_1, a_2)\}$. $\mathcal{G}^{\mathcal{E}}$, which includes invariant subgraphs that have deterministic and truly predictive relations with the labels and other environmental graphs. Next, we will use $\mathcal{G}^{\mathcal{E}}$ to strengthen the invariant subgraph G^{inv} learning.

3.5. Loss function definition

The proposed model employs a composite loss function consisting of causal loss, contrastive loss, and invariant loss, optimized based on two invariance-driven objectives. The first loss, $\mathcal{L}_{cau}$, employs invariance optimization, a method that assumes the existence of invariant substructures (G^{inv}) in graph G to guide the model in internally discovering features that are stable across samples. The losses $\mathcal{L}_{con}$ and $\mathcal{L}_{inv}$ implement explicit representation alignment, enforcing the model to align features from different environments both in terms of predictive risk and feature space. This dual approach allows us to both discover and enforce invariance from different perspectives.

Causal Loss. We use causal loss $\mathcal{L}_{cau}$ to achieve the causal decoupling object of Eq. 6, ensuring that G^{inv} maintains its causal relationship with label Y. According to Theorem 3.4, we jointly achieve maximizing $I(G^{inv}, Y)$ by $\max I\left(G_k^{inv}; G_m^{inv}\right)$ and minimizing the cross entropy loss on G^{inv}. We treat G^{inv} with the same label as positive samples: $\boldsymbol{z}^{inv} = g(G^{inv})$, where $g(\cdot)$ represents the encoder for learning the graph-level features. And minimize $I(G^{inv}, G^s)$ using an empirical distribution-based mutual information estimation method:

$$\begin{aligned}\mathcal{L}_{cau} &: \max I(G_k^{inv}; G_m^{inv}) + \min I(G^{inv}, G^s) \\ &= \mathbb{E}_{k \in I} \frac{-1}{|P(k)|} \sum_{p \in P(k)} \log \frac{\exp\left(\boldsymbol{z}_k^{inv} \cdot \boldsymbol{z}_p^{inv} / \tau\right)}{\sum_{a \in A(k)} \exp\left(\boldsymbol{z}_k^{inv} \cdot \boldsymbol{z}_a^{inv} / \tau\right)} \\ &\quad + \sum_{i,j} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i) P(y_j)}, \end{aligned} \tag{11}$$

where $A(k) = S \setminus \{k\}$ include samples other than k in the same batch, $P(k) = \{p \in A(k) : Y_p = Y_k\}$ is the set of indices of all positives in the batch and $\cdot$ represents cosine similarity. The $P(x_i, y_j)$ represents the joint probability distribution of G^{inv} and G^s. $P(x_i)$ and $P(y_j)$ are the marginal probability distributions of G^{inv} and G^s , respectively.

Invariant Loss. For k-th subject, we treat G_k^{inv} ,$G_k^{a_1}$, and $G_k^{a_2}$ as graph data generated under different environments. Because they mixed ΔG with different data distributions. While there are significant distribution differences between these graphs, a certain proportion of the data, G_k^{inv}, captures the relationship with the Y across different environments. We then achieve the invariant learning by minimizing the risk of the predictor $w(\cdot)$, ensuring it makes the least risky predictions across these three environments:

$$\begin{aligned}\mathcal{L}_{inv} &: \max\ \mathbb{E}_{\boldsymbol{e} \in \mathcal{E}}\left[I(G^{inv}; Y \mid \boldsymbol{e})\right] \\ &= \mathbb{E}_{G_k \in \mathcal{G}} \sum_{\boldsymbol{e} \in \acute{\mathcal{E}}} \mathcal{R}(f \mid \boldsymbol{e}), f = w \circ g \circ \Phi(G_k) \end{aligned} \tag{12}$$

where $\acute{\mathcal{E}} \in (inv, a_1, a_2)$ and $\mathcal{R}(f \mid \boldsymbol{e})$ represents cross-entropy loss. $w(\cdot)$ is the predictor with two fully connected layers, $g(\cdot)$ represents the encoder that learns graph-level features, while $\Phi(\cdot)$ is the extractor for invariant subgraphs.

Contrastive Loss. In GCL, we use G^{inv} as the anchor graph, and the graph in $\mathcal{G}^{pos}$ as the positive sample:

$$\begin{aligned}\mathcal{L}_{\text{con}} &: \max_g I\left(g\left(\mathcal{G}^{pos}\right), g\left(G^{inv}\right)\right) \\ &= \mathbb{E}_{k \in I}\, \mathbb{E}_{+ \in (a_1, a_2)} \log \frac{\exp\left(\boldsymbol{z}_k^{inv} \cdot \boldsymbol{z}_k^{+} / \tau\right)}{\sum_{b \in B(i)} \exp\left(\boldsymbol{z}_k^{inv} \cdot \boldsymbol{z}_k^{b} / \tau\right)} \\ &\quad + \|\boldsymbol{z}^{a_1\top} \cdot \boldsymbol{z}^{a_2}\|_F^2, \end{aligned} \tag{13}$$

here, $B(i) = \mathcal{E} \setminus \{inv\}$ include samples other than G^{inv}. And $B(i)$ includes the G^s, which can be considered as a negative sample of G^{inv}. We further $\min I(G^{inv}, G^s)$ by maximizing the distance between $\boldsymbol{z}_k^{inv}$ and $\boldsymbol{z}_k^s$, ensuring the disentanglement of G^{inv} and G^s again. Let $\|\cdot\|_F$ represent the Frobenius norm. We impose an orthogonality loss to reduce redundancy between two augmented samples, enhancing GCL's feature diversity and informativeness.

Finally, the model completes iterative training by $\mathcal{L} = \lambda_1 \mathcal{L}_{con} + \lambda_2 \mathcal{L}_{cau} + \mathcal{L}_{inv}$, combining the above three losses. The pseudocode description is detailed in Appendix A, and a comprehensive summary of key notations and their explanations can be found in Appendix C.

4. Experiments

We primarily validate the method through three main aspects in our experiments. **Q1**: Does the method effectively perform on multi-site brain datasets, and does it help mitigate the OOD problem? **Q2**: Do the proposed brain invariant subgraph and invariance-aware augmentation strategies lead to improved model performance? **Q3**: Does the method cap-

Dataset	Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	bAcc (%)	AUC (%)	Avg (%)
ABIDE I	RF	62.02±4.72	63.28±5.33	63.02±3.79	63.01±3.53	62.00±4.81	66.60±5.96	62.79±0.48
	SVM	65.12±2.93	65.49±4.15	60.81±5.01	62.93±3.44	65.03±2.93	70.83±4.79	62.67±6.75
	BrainNetCNN	68.88±3.23	70.21±5.29	71.13±13.8	69.48±5.37	68.85±3.22	72.39±4.09	70.16±1.39
	CIGA	66.96±1.64	69.55±2.42	57.84±6.20	62.92±3.43	66.75±1.69	67.53±2.81	65.26±4.22
	GATE	71.67±2.95	71.93±4.92	74.68±6.08	72.99±2.55	71.42±3.03	73.04±4.62	72.62±1.21
	Com-BrainTF	70.14±4.38	70.28±4.32	72.83±4.15	70.01±4.49	70.00±4.65	71.67±6.16	70.82±1.17
	BrainGB	71.07±4.92	72.79±8.02	72.90±6.20	70.80±4.47	70.93 ±4.79	74.93±5.10	72.24±1.62
	METAFormer	70.31±2.86	71.80±3.20	74.38±6.64	72.85±3.29	69.91±2.80	72.29±3.54	71.18±2.91
	Contrasformer	68.90±2.33	67.68±3.96	70.91±6.04	68.70±2.74	68.86±2.23	70.68±2.64	69.29±1.24
	BrainIB	69.97±2.82	71.07±4.19	70.70±4.61	70.74±2.90	69.63±3.15	73.44±4.35	70.92±1.34
	CI-GNN	71.89±2.91	70.49±2.98	73.37±4.80	70.58±2.21	71.44±2.75	73.32±3.62	71.85±1.27
	CIA-GCL	**73.42±2.75**	**73.82±3.23**	**74.91±5.63**	**74.21±3.03**	**73.38±2.76**	**76.28±3.74**	**74.34±1.10**
ABIDE II	RF	62.56±2.67	62.71±4.29	48.87±2.97	54.83±2.55	61.68±2.56	66.60±5.96	59.54±6.47
	SVM	64.78±3.75	63.79±5.30	56.53±6.83	59.75±4.90	64.25±3.88	69.13±3.62	63.04±4.36
	BrainNetCNN	67.75±2.72	69.74±5.70	56.92±9.87	61.87±5.57	67.21±2.72	69.59±3.89	65.51±5.08
	CIGA	66.40±3.22	63.01±3.16	68.32±13.39	64.85±7.92	66.54±3.72	66.63±3.94	65.53±2.81
	GATE	69.51±2.76	71.21±2.97	65.64±6.88	68.83±2.94	70.33±5.25	70.87±4.55	69.37±2.02
	Com-BrainTF	68.91±2.29	69.12±1.95	62.92±9.42	68.34±2.87	68.80±6.50	70.40±4.72	68.08±2.62
	BrainGB	69.23±3.33	67.37±4.61	65.97±6.45	66.34±3.89	69.01±3.28	71.33±4.12	68.21±2.03
	METAFormer	69.10±3.17	68.34±3.91	64.00±7.66	65.89±4.41	69.04±3.21	70.88±4.36	67.88±2.49
	Contrasformer	69.79±2.99	67.39±8.16	57.86±8.87	61.78±6.25	67.94±3.50	70.02±3.67	65.78±4.92
	BrainIB	68.92±3.60	66.73±6.70	68.13±10.74	66.65±4.19	68.58±3.22	69.83±4.61	68.29±1.96
	CI-GNN	70.04±2.62	70.22±3.59	**69.56±5.28**	68.85±3.60	69.93±3.39	69.13±4.82	69.62±0.54
	CIA-GCL	**72.09±2.47**	**72.67±3.49**	64.39±9.14	**68.94±4.74**	**71.45±2.71**	**72.58±2.41**	**70.51±3.25**
ADHD200	RF	63.88±3.28	54.52±10.37	21.67±9.99	30.14±11.37	55.53±4.19	60.72±7.90	47.74±17.46
	SVM	66.48±3.99	62.90±12.19	28.25±7.83	38.49±8.51	58.91±4.37	67.58±7.91	53.77±16.41
	BrainNetCNN	70.71±4.15	61.69±5.90	59.30±12.29	59.90±7.94	68.47±5.21	70.02±6.20	65.01±5.27
	CIGA	67.73±2.44	61.95±6.04	43.25±17.31	48.40±12.75	62.75±4.97	64.01±2.95	58.02±9.79
	GATE	72.02±3.55	71.27±6.09	45.18±17.04	52.82±14.87	66.67±6.01	69.31±9.42	62.88±11.17
	Com-BrainTF	71.78±4.50	70.38±4.33	56.71±17.17	68.28±7.69	68.77±6.89	69.75±6.96	67.61±5.48
	BrainGB	71.91±2.89	71.51±9.69	45.56±9.96	54.62±6.65	66.01±3.49	71.63±6.30	63.54±11.02
	METAFormer	70.27±2.54	73.00±7.87	44.09±10.01	54.66±9.45	62.63±4.83	69.13±4.65	62.29±11.09
	Contrasformer	72.19±2.65	62.68±5.76	**59.53±12.73**	60.26±7.32	68.54±4.06	72.63±3.41	65.97±5.91
	BrainIB	70.07±1.56	60.69±2.54	56.47±8.48	58.10±4.11	67.10±2.15	67.28±2.72	63.29±5.60
	CI-GNN	71.03±3.05	61.32±5.08	53.44±10.84	66.93±4.26	67.93±4.26	71.95±2.97	65.43±6.97
	CIA-GCL	**75.21±2.57**	**79.40±10.59**	58.75±11.73	**69.80±8.87**	**69.93±4.01**	**73.51±4.01**	**71.10±7.03**

Table 1. Classification results for ABIDE I, ABIDE II, ADHD200 on AAL atlas comparing with different methods (mean ± std%). **Bold** denotes the best performance while Underline represents the second best performance. Avg (%) is the average of these 6 metrics.

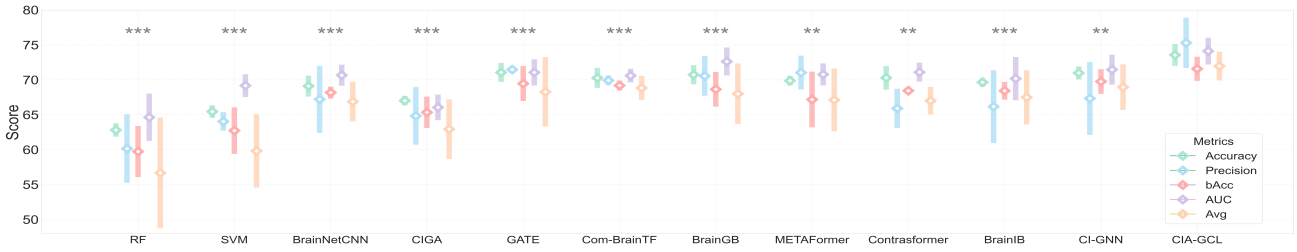


Figure 2. Average classification performance of three datasets on AAL atlas.

ture local biomarkers and provide interpretability? Details of the experimental configurations are in Appendix E.

4.1. Comparison with State-of-the-Art Methods (Q1)

4.1.1. Overall Performance

We present the comparison results of three brain datasets in Table 1. The results show that our method outperforms the comparison method in seven metrics, highlighting the superiority of CIA-GCL's performance. All experimental results are obtained based on 10-fold cross-validation. On three datasets, our method outperforms other methods in most metrics. The causal analysis-based CI-GNN and the contrastive learning-based methods both performed well in three datasets, demonstrating the feasibility of causal analysis and contrastive learning in brain graph analysis. Our method performs poorly on the Recall index, indicating that it tends to make conservative decisions. We also calculated the average results for all methods across 3 datasets and performed t-tests in Figure 2. It is evident that our method outperforms the other methods in most of the metrics. And the overall performance across the 3 datasets is statistically significantly different from others, with significance levels of $p < 0.01(**)$, and $p < 0.001(***)$.

4.1.2. Generalization on Multi-Site Dataset

In section 4.1.1, we tested by mixing data from all sites. In this summary, to verify whether the proposed method can alleviate the OOD phenomenon, we followed the testing

Dataset	Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	bAcc (%)	AUC (%)	Avg (%)
ABIDE I	RF	63.58±4.83	63.06±5.81	59.95±3.03	61.26±1.57	63.23±3.87	67.80±3.65	63.15±2.67
	SVM	64.92±3.95	63.58±6.03	64.05±3.73	63.62±2.67	64.66±3.15	68.00±2.88	64.81±1.66
	BrainNetCNN	67.97±0.86	69.35±6.55	60.35±12.57	63.78±6.81	67.36±1.57	70.24±2.19	66.51±3.75
	CIGA	65.51±1.97	64.32±4.19	63.17±2.06	63.65±1.54	65.35±1.78	66.74±3.40	64.79±1.33
	GATE	67.76±1.74	64.60±3.55	66.86±4.93	65.63±3.53	66.93±2.75	69.16±1.74	66.82±1.59
	Com-BrainTF	69.29±0.81	70.87±5.29	67.66±6.88	69.15±0.61	69.26±0.81	71.54±0.63	69.63±1.38
	BrainGB	69.91±1.17	68.11±3.02	68.16±10.27	68.01±6.70	69.13±1.78	69.68±1.96	68.83±0.85
	METAFormer	70.13±3.40	69.60±1.56	**72.26±8.13**	67.88±4.34	69.47±3.78	71.13±5.15	70.08±1.50
	Contrasformer	70.50±2.83	73.14±6.80	67.19±5.60	70.73±4.41	70.78±2.95	69.54±4.11	70.31±1.93
	BrainIB	64.92±1.34	61.27±1.42	66.06±9.45	63.89±8.98	63.32±2.88	61.59±1.17	63.51±1.86
	CI-GNN	69.94±1.98	72.97±2.62	64.96±2.57	69.93±0.95	70.06±2.93	71.31±2.86	69.86±2.68
	CIA-GCL	**72.03±0.90**	**74.12±3.57**	69.92±3.31	**71.86±1.51**	**72.08±1.65**	**74.26±3.43**	**72.38±1.62**
ABIDE II	RF	60.44±5.27	68.10±12.47	56.18±11.66	60.20±3.71	62.42±1.30	66.28±1.77	62.27±4.35
	SVM	62.58±4.21	68.49±15.29	67.07±15.39	65.47±3.12	65.20±1.50	67.60±4.71	66.07±2.12
	BrainNetCNN	67.20±4.18	67.36±8.87	73.31±11.96	70.22±10.28	63.65±2.18	68.19±3.01	68.32±3.24
	CIGA	68.11±3.20	71.61±9.12	62.60±13.97	67.36±10.32	66.45±2.81	67.72±3.62	67.31±2.90
	GATE	67.83±2.15	60.60±9.88	66.53±14.11	62.26±6.10	63.26±3.52	65.70±2.21	64.36±2.76
	Com-BrainTF	69.34±3.24	67.09±1.67	69.60±19.07	64.94±4.02	64.79±4.40	67.10±0.85	67.14±2.06
	BrainGB	69.68±4.45	71.89±9.45	70.53±12.90	70.93±10.19	**67.36±0.92**	68.39±1.79	69.79±1.68
	METAFormer	66.18±2.33	64.32±6.68	73.60±11.12	67.97±2.23	66.48±2.38	67.01±0.18	67.59±3.17
	Contrasformer	69.81±2.05	70.18±3.36	68.98±10.25	69.25±6.43	65.47±4.54	66.99±0.95	68.45±1.82
	BrainIB	68.46±3.13	**72.14±8.44**	64.89±16.37	68.02±12.17	65.99±2.01	65.60±2.45	67.50±2.67
	CI-GNN	68.14±2.35	66.28±2.43	62.63±8.61	63.89±3.68	66.81±2.50	67.32±2.82	65.85±2.13
	CIA-GCL	**71.51±2.94**	70.83±4.63	**77.87±15.94**	**73.97±9.73**	66.52±5.24	**68.81±5.72**	**71.59±3.98**
ADHD200	RF	60.35±6.83	41.03±15.54	7.02±4.80	11.34±7.46	49.44±3.71	49.58±7.05	36.46±22.04
	SVM	63.00±4.95	55.93±19.57	8.53±3.90	14.56±6.37	51.90±2.64	62.03±2.74	42.66±24.51
	BrainNetCNN	67.96±2.22	61.77±7.54	38.75±24.33	44.62±16.99	61.51±5.86	63.79±4.45	56.40±11.77
	CIGA	67.27±1.75	69.78±4.47	21.24±14.06	30.71±17.30	57.70±4.55	59.60±7.98	51.05±20.16
	GATE	67.50±1.58	62.36±14.01	30.3±17.12	42.93±14.14	60.66±3.41	62.86±5.44	54.44±14.54
	Com-BrainTF	69.02±0.32	69.04±0.94	31.51±15.03	**60.11±5.45**	61.51±5.86	65.28±3.11	59.41±14.16
	BrainGB	68.29±1.23	66.31±4.17	32.52±13.95	45.84±10.84	61.98±4.94	63.92±1.14	56.48±14.20
	METAFormer	67.22±2.01	**71.16±12.90**	39.90±0.22	43.76±10.57	59.21±0.98	63.58±3.88	57.38±13.02
	Contrasformer	68.05±1.35	60.37±5.54	43.10±16.63	48.33±11.32	62.75±3.87	68.15±4.18	58.46±10.45
	BrainIB	68.43±1.93	58.76±4.27	33.44±9.67	42.27±9.18	60.18±3.80	64.03±3.76	54.52±13.62
	CI-GNN	68.57±1.25	58.58±5.50	49.33±1.27	50.23±3.75	64.08±1.23	64.57±2.48	59.23±7.98
	CIA-GCL	**71.32±1.05**	64.77±3.28	**51.17±9.23**	56.09±5.89	**66.63±5.92**	**68.97±2.35**	**63.16±7.85**

Table 2. Classification results for different methods on the three target domains from three brain datasets on AAL atlas (mean ± std%). **Bold** denotes the best performance while Underline represents the second best performance. Avg (%) is the average of these 6 metrics.

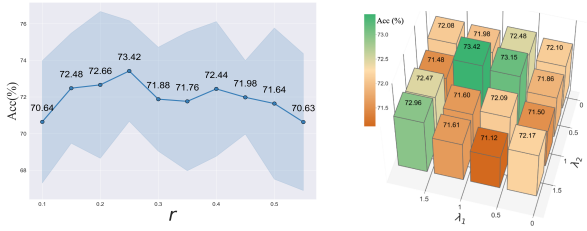


Figure 3. Accuracy sensitivity of hyper-parameters r, λ_1, λ_2.

approach used in (Qiu et al., 2024) The test and train sets consist of data from different sites, similar to the validation methods in domain generalization problems. We randomly selected one site's data as the test set (target domain) and used the remaining sites as the train set (source domain). For each dataset, we randomly selected three target domains for experimentation and averaged the results for comparison. Details on the site data selection can be found in Appendix E4. We present the comparison results in Table 2. Under this experimental design, the performance of most methods degrades, indicating the heterogeneity of multi-site brain data. But on three datasets, our method still performed the best on most metrics. In the ADHD200 dataset, due to the imbalance in the number of patients and the TC group, the Recall index is relatively low.

4.2. Ablation Studies and Parameter Analysis (Q2)

In the ablation setting where only ERM loss is used, we remove the augmentation strategy in module $g(\cdot)$, and the invariant subgraph extraction method does not constrain the extraction process to satisfy the theoretical properties. This model is optimized solely via the cross-entropy between the invariant subgraph features and labels. By adding $\mathcal{L}_{cau}$ and $\mathcal{L}_{inv}$ one by one, we introduce constraints that force the extracted subgraph to satisfy the properties defined in our framework, allowing us to validate the effectiveness of the brain invariant subgraph extraction method through two invariance-driven objectives. When adding $\mathcal{L}_{con}$, we incorporate the augmentation strategy in $g(\cdot)$, which enables us to assess the impact of the proposed invariance-aware augmentation strategy. The results based on 10-fold cross-validation are presented in Table 3. It proves that, through the optimization of the three losses, the model can effectively learn the invariant subgraphs under various data distribution shifts.

We conduct experiments on ABIDE I to examine the sensitivity to hyperparameters. We select three critical parameters of the model, including the maximum edge ratio r, the coefficient of losses λ_1, and λ_2. The results are shown in Fig. 3. For r, a too-small r may result in performance degra-

Dataset	ERM	$\mathcal{L}_{cau}$	$\mathcal{L}_{con}$	$\mathcal{L}_{inv}$	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)	bAcc(%)	AUC(%)	Avg(%)
ABIDE I	✓				71.13±4.01	71.72±4.01	73.20±10.93	71.89±5.34	71.14±3.95	72.20±4.01	71.88±0.77
	✓	✓			72.17±4.10	73.42±6.66	73.77±11.10	72.44±5.10	72.06±4.07	72.41±4.71	72.71±0.70
	✓	✓	✓		72.63±2.66	**73.91±4.05**	72.9±10.02	72.89±4.53	72.63±2.55	73.02±2.83	73.01±0.47
	✓	✓	✓	✓	**73.42±2.75**	73.82±3.23	**74.91±5.63**	**74.21±3.03**	**73.38±2.76**	**76.28±3.74**	**74.34±1.10**
ABIDE II	✓				70.07±2.21	68.89±5.37	67.12±10.15	67.32±3.71	69.91±2.18	70.38±4.62	68.95±1.43
	✓	✓			69.87±3.00	70.49±5.40	62.32±13.42	65.06±6.78	69.29±3.42	69.14±4.01	67.70±3.25
	✓	✓	✓		71.41±2.48	69.47±4.29	**70.99±10.44**	**69.15±4.54**	71.32±2.78	71.83±5.21	69.96±1.11
	✓	✓	✓	✓	**72.09±2.47**	**72.67±3.49**	64.39±9.14	68.94±4.74	**71.45±2.71**	**72.58±2.41**	**70.51±3.25**
ADHD200	✓				72.96±2.71	76.07±11.03	44.41±13.21	54.26±9.26	68.07±3.74	70.54±6.89	64.39±12.35
	✓	✓			74.02±4.17	69.54±7.44	57.14±11.59	61.93±7.84	70.67±5.14	72.26±6.55	67.59±6.59
	✓	✓	✓		74.26±2.36	68.16±5.50	**61.76±11.82**	63.88±5.82	**71.80±3.53**	73.33±4.53	68.87±5.16
	✓	✓	✓	✓	**75.21±2.57**	**79.40±10.59**	58.75±11.73	**69.80±8.87**	69.93±4.01	**73.51±4.01**	**71.10±7.03**

Table 3. Ablation experiment of three loss functions for three brain datasets on AAL atlas (mean ± std%). **Bold** denotes the best performance while Underline represents the second best performance. Avg(%) is the average of these 6 metrics.

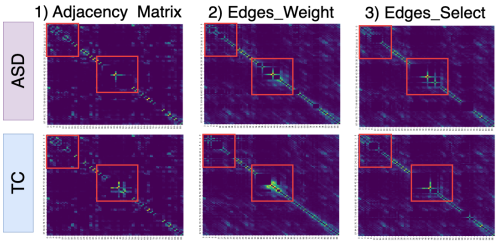


Figure 4. Illustration of the original adjacency matrix, learned edge weight, and invariant subgraph of patients and TC groups.

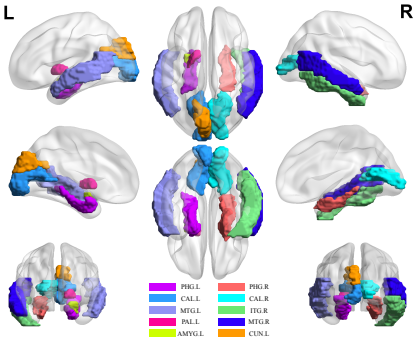


Figure 5. Top 10 important ROIs of ABIDE I on AAL atlas.

dation due to the extracted G^{inv} being too small and having too little available information. Meanwhile, if r is too large, G^{inv} may contain a lot of redundant information, leading to a decrease in performance. The optimal hyperparameters were fixed across three datasets to ensure consistency.

4.3. Capability in capturing local brain biomarkers (Q3)

We conducted an interpretability analysis on the invariant subgraphs and important ROIs. The extracted G^{inv} are considered as the local biomarkers identified by the model. 1) shows the original adjacency matrix in the Figure. 4. The elements in 2) are p_{ij} in $\dot{A}$, which represents the possibility of edge selection in G^{inv} obtained through Eq. 3. 3) is the adjacency matrix of the G^{inv} after edge selection. They only selected the top 5% values from the matrix to present clearly. We can see that the 2) learned after optimization can effectively capture the obvious features in the original adjacency matrix, while also exploring other new features. After selection, the 3) edges_select retains the most discriminative features, making it clearer to distinguish the patients. The red box in the middle belongs to the occipital region. It can be seen that the intensity of ASD in this area is higher than TC, which is consistent with (Keehn et al., 2008).

In the analysis of important brain regions, we obtain the ROIs by the invariant subgraph. The detailed calculation process can be found in Appendix F1 and the ROIs of the AAL atlas are shown in Figure 5. On the AAL atlas, the top 10 brain regions mainly focus on temporal and occipital regions. Among them, Temporal_Mid (MTG.R, MTG.L) belongs to the middle temporal gyrus, which is pointed out as one of the key regions in the "social brain" network, and they are more vulnerable in people with ASD than in TC (Xu et al., 2020). The study found that lower ParaHippocampal (PHG.R, PHG.L) activity in individuals with ASD, compared to TD individuals, likely reflects their proficiency in scene recognition and spatial navigation, requiring fewer cognitive resources (Mouga et al., 2022). For the visualization and analysis of important brain regions on the other two datasets, please refer to Appendix F2.

5. Conclusions

In this article, to address the two issues in handling brain graph data, we combine invariant learning in the field of out-of-distribution generalization with contrastive learning. We propose the CIA-GCL with a novel invariance-aware augmentation strategy that embeds a causal disentanglement method to find brain invariant subgraphs for disease diagnosis. The experiments on real-world brain disease datasets show the effectiveness and generalization of our method and demonstrate the potential of out-of-distribution generalization techniques for processing brain datasets with data distribution shifts. However, this method still has shortcomings. In the generation of ΔG in the augmented samples, we regard different labels as the changing environment, with data distribution differences being too simple and rough. In the future, we aim to introduce more techniques from the field of OOD generalization into brain graph analysis.

Impact Statement

This study delves into brain graph analysis, which is an important tool for revealing the complex structure and function of complex neural networks. As an AI application with significant societal benefits, our model aids in the early detection of neurological disorders and advances neuroscience research. By drawing and analyzing the interconnections between different brain regions, this analysis helps with the initial diagnosis of brain diseases and the localization of biomarkers in brain diseases, providing new avenues for personalized medical strategies and interventions. However, there are also issues of inaccurate diagnosis and a lack of universality and generalizability in cross-populations and cross-sites. This type of research needs to be combined with the professional knowledge and experience of healthcare professionals, who can provide valuable insights into the clinical context and ensure that the model is suitable for different patient populations.

Acknowledgements

Thanks for the professional and constructive suggestions from the four reviewers. This work was sponsored by Beijing Nova Program (20240484635), partly supported by National Natural Science Foundation of China Research Program (62106009, 62276010, 62171297), and in part by R&D Program of Beijing Municipal Education Commission (KM202210005030, KZ202210005009).

References

Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

Ahuja, K., Caballero, E., Zhang, D., Gagnon-Audet, J.-C., Bengio, Y., Mitliagkas, I., and Rish, I. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

and others. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bannadabhavi, A., Lee, S., Deng, W., Ying, R., and Li, X. Community-aware transformer for autism prediction in fmri connectome. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 287–297. Springer, 2023.

Bazinet, V., Hansen, J. Y., and Misic, B. Towards a biologically annotated brain connectome. *Nature reviews neuroscience*, 24(12):747–760, 2023.

Chang, S., Zhang, Y., Yu, M., and Jaakkola, T. Invariant rationalization. In *International Conference on Machine Learning*, pp. 1448–1458. PMLR, 2020.

Chen, X., Zhang, H., Zhang, L., Shen, C., Lee, S.-W., and Shen, D. Extraction of dynamic functional connectivity from brain grey matter and white matter for mci classification. *Human brain mapping*, 38(10):5019–5034, 2017.

Chen, Y., Zhang, Y., Bian, Y., Yang, H., Kaili, M., Xie, B., Liu, T., Han, B., and Cheng, J. Learning causally invariant representations for out-of-distribution generalization on graphs. *Advances in Neural Information Processing Systems*, 35:22131–22148, 2022.

Chong, C. D., Schwedt, T. J., and Hougaard, A. Brain functional connectivity in headache disorders: a narrative review of mri investigations. *Journal of Cerebral Blood Flow & Metabolism*, 39(4):650–669, 2019.

Cui, H., Dai, W., Zhu, Y., Kan, X., Gu, A., Lukemire, J., Zhan, L., He, L., Guo, Y., and Yang, C. Braingb: A benchmark for brain network analysis with graph neural networks. *IEEE Transactions on Medical Imaging*, 42(2): 493–506, 2023.

Doyle-Thomas, K. A., Lee, W., Foster, N. E., Tryfon, A., Ouimet, T., Hyde, K. L., Evans, A. C., Lewis, J., Zwaigenbaum, L., Anagnostou, E., et al. Atypical functional brain connectivity during rest in autism spectrum disorders. *Annals of neurology*, 77(5):866–876, 2015.

Eslami, T., Almuqhim, F., Raiker, J. S., and Saeed, F. Machine learning methods for diagnosing autism spectrum disorder and attention-deficit/hyperactivity disorder using functional and structural mri: a survey. *Frontiers in neuroinformatics*, 14:575999, 2021.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2022.

Ji, J., Wang, F., Han, L., and Liu, J. Causal learning and knowledge fusion mechanism for brain functional network classification. *IEEE Transactions on Signal and Information Processing over Networks*, 2024.

Jia, T., Li, H., Yang, C., Tao, T., and Shi, C. Graph invariant learning with subgraph co-mixup for out-of-distribution generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8562–8570, 2024.

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

Keehn, B., Brenner, L., Palmer, E., Lincoln, A. J., and Mueller, R.-A. Functional brain organization for visual search in asd. *Journal of the International Neuropsychological Society*, 14(6):990–1003, 2008.

Li, H., Wang, X., Zhang, Z., and Zhu, W. Out-of-distribution generalization on graphs: A survey. *arXiv preprint arXiv:2202.07987*, 2022a.

Li, H., Zhang, Z., Wang, X., and Zhu, W. Learning invariant graph representations for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:11828–11841, 2022b.

Liang, X., Wang, J., and He, Y. Human connectome: Structural and functional brain networks. *Chinese Science Bulletin*, 55(16):1565–1583, 2010.

Liu, J., Ji, J., Xun, G., and Zhang, A. Inferring effective connectivity networks from fmri time series with a temporal entropy-score. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10):5993–6006, 2022.

Liu, J., Han, L., and Ji, J. Mcan: Multimodal causal adversarial networks for dynamic effective connectivity learning from fmri and eeg data. *IEEE transactions on medical imaging*, 43(8):2913–2923, 2024a.

Liu, J., Wang, F., and Ji, J. Concept-level causal explanation method for brain function network classification. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 3087–3096, 2024b.

Mahler, L., Wang, Q., Steiglechner, J., Birk, F., Heczko, S., Scheffler, K., and Lohmann, G. Pretraining is all you need: A multi-atlas enhanced transformer framework for autism spectrum disorder classification. In *International Workshop on Machine Learning in Clinical Neuroimaging*, pp. 123–132. Springer, 2023.

Mouga, S., Duarte, I. C., Café, C., Sousa, D., Duque, F., Oliveira, G., and Castelo-Branco, M. Parahippocampal deactivation and hyperactivation of central executive, saliency and social cognition networks in autism spectrum disorder. *Journal of Neurodevelopmental Disorders*, 14(1):9, 2022.

Mueller, S., Wang, D., Fox, M. D., Yeo, B. T., Sepulcre, J., Sabuncu, M. R., Shafee, R., Lu, J., and Liu, H. Individual variability in functional connectivity architecture of the human brain. *Neuron*, 77(3):586–595, 2013.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Peng, L., Wang, N., Xu, J., Zhu, X., and Li, X. Gate: Graph cca for temporal self-supervised learning for label-efficient fmri analysis. *IEEE Transactions on Medical Imaging*, 42(2):391–402, 2023.

Qiu, X., Wang, F., Sun, Y., Lian, C., and Ma, J. Towards graph neural networks with domain-generalizable explainability for fmri-based brain disorder diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 454–464. Springer, 2024.

Rudie, J. D., Brown, J., Beck-Pancer, D., Hernandez, L., Dennis, E., Thompson, P., Bookheimer, S., and Dapretto, M. Altered functional and structural brain network organization in autism. *NeuroImage: clinical*, 2:79–94, 2013.

Said, A., Bayrak, R., Derr, T., Shabbir, M., Moyer, D., Chang, C., and Koutsoukos, X. Neurograph: Benchmarks for graph machine learning in brain connectomics. *Advances in Neural Information Processing Systems*, 36: 6509–6531, 2023.

Soon, C. S., Vinogradova, K., Ong, J. L., Calhoun, V. D., Liu, T., Zhou, J. H., Ng, K. K., and Chee, M. W. Respiratory, cardiac, eeg, bold signals and functional connectivity over multiple microsleep episodes. *NeuroImage*, 237: 118129, 2021.

Sporns, O., Tononi, G., and Kötter, R. The human connectome: a structural description of the human brain. *PLoS computational biology*, 1(4):e42, 2005.

Sui, Y., Tang, C., Chu, Z., Fang, J., Gao, Y., Cui, Q., Li, L., Zhou, J., and Wang, X. Invariant graph learning for causal effect estimation. In *Proceedings of the ACM on Web Conference 2024*, pp. 2552–2562, 2024.

Sun, L., Yu, K., and Batmanghelich, K. Context matters: Graph-based self-supervised representation learning for medical images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4874–4882, 2021.

Taspinar, G. and Ozkurt, N. A review of adhd detection studies with machine learning methods using rsfmri data. *NMR in Biomedicine*, pp. e5138–e5138, 2024.

Vizioli, L., Moeller, S., Dowdle, L., Akçakaya, M., De Martino, F., Yacoub, E., and Uğurbil, K. Lowering the thermal noise barrier in functional brain mapping with magnetic

resonance imaging. *Nature communications*, 12(1):5181, 2021.

Wang, X., Yao, L., Rekik, I., and Zhang, Y. Contrastive functional connectivity graph learning for population-based fmri classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 221–230. Springer, 2022.

Wu, J., Chen, X., Shi, B., Li, S., and Xu, K. Sega: Structural entropy guided anchor view for graph contrastive learning. In *International Conference on Machine Learning*, pp. 37293–37312. PMLR, 2023.

Wu, Q., Nie, F., Yang, C., Bao, T., and Yan, J. Graph out-of-distribution generalization via causal intervention. In *Proceedings of the ACM on Web Conference 2024*, pp. 850–860, 2024.

Xu, J., Wang, C., Xu, Z., Li, T., Chen, F., Chen, K., Gao, J., Wang, J., and Hu, Q. Specific functional connectivity patterns of middle temporal gyrus subregions in children and adults with autism spectrum disorder. *Autism Research*, 13(3):410–422, 2020.

Xu, J., Bian, Q., Li, X., Zhang, A., Ke, Y., Qiao, M., Zhang, W., Sim, W. K. J., and Gulyás, B. Contrastive graph pooling for explainable classification of brain networks. *IEEE Transactions on Medical Imaging*, 2024a.

Xu, J., He, K., Lan, M., Bian, Q., Li, W., Li, T., Ke, Y., and Qiao, M. Contrasformer: a brain network contrastive transformer for neurodegenerative condition identification. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2671–2681, 2024b.

Xu, J., Chen, Y., Dong, X., Lan, M., Huang, T., Bian, Q., Cheng, J., and Ke, Y. Brainood: Out-of-distribution generalizable brain network analysis. *arXiv preprint arXiv:2502.01688*, 2025.

Yang, Y., Cui, H., and Yang, C. Ptgb: Pre-train graph neural networks for brain network analysis. In *Conference on Health, Inference, and Learning*, pp. 526–544. PMLR, 2023.

Zhang, S., Chen, X., Shen, X., Ren, B., Yu, Z., Yang, H., Jiang, X., Shen, D., Zhou, Y., and Zhang, X.-Y. A-gcl: Adversarial graph contrastive learning for fmri analysis to diagnose neurodevelopmental disorders. *Medical Image Analysis*, 90:102932, 2023.

Zheng, K., Yu, S., and Chen, B. Ci-gnn: A granger causality-inspired graph neural network for interpretable brain network-based psychiatric diagnosis. *Neural Networks*, 172:106147, 2024a.

Zheng, K., Yu, S., Li, B., Jenssen, R., and Chen, B. Brainib: Interpretable brain network-based psychiatric diagnosis with graph information bottleneck. *IEEE Transactions on Neural Networks and Learning Systems*, 2024b.

Zhu, H., Wang, J., Zhao, Y.-P., Lu, M., and Shi, J. Contrastive multi-view composite graph convolutional networks based on contribution learning for autism spectrum disorder classification. *IEEE transactions on bio-medical engineering*, 70(6):1943–1954, 2023.

Zhu, Y., Feng, L., Deng, Z., Chen, Y., Amor, R., and Witbrock, M. Robust node classification on graph data with graph and label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17220–17227, 2024.

Zong, Y., Zuo, Q., Ng, M. K.-P., Lei, B., and Wang, S. A new brain network construction paradigm for brain disorder via diffusion-based graph contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10389–10403, 2024.

A. Algorithm Pseudocode

Algorithm 1 CIA-GCL for Brain Graph analysis

Input: brain graph G, labels Y
Output: brain invariant graph G^{inv}

for number of training iterations **do**
  Sample mini training batch
  Invariant Subgraphs Extract:
  Calculate the probabilities of edges in G^{inv} by Eq. (3)
  Select the reserved edge mask $\mathcal{M}^{inv}$ based on the threshold t by Eq.(4)(5)
  Obtain the invariant graph G^{inv} and the spurious graph G^{s} by Eq.(1)
  for each subject G_k **do**
    Augmented Samples Generation:
    Select G_i^s with $Y_i \neq Y_k$
    Obtain the mixed spurious subgraph G^{ms_k} by Eqs.(8)(9)
    Combine G_k^{inv} and ΔG to get $\mathcal{G}_k^{pos}$ by Eq.(10)
  end for
  Gradient-based optimization via backpropagation:
  Calculate the causal loss $\mathcal{L}_{cau}$ by Eq.(11)
  Calculate the invariant loss $\mathcal{L}_{inv}$ supervised loss by Eq.(12)
  Calculate the contrastive loss $\mathcal{L}_{con}$ supervised loss by Eq.(13)
  Update the model by minimizing the combination of the above three losses
end for
return brain invariant graph G^{inv}

B. Graph Augmentation Methods Comparison

Since some models do not have names, we use the first author's last name and the year as substitutes. For example, Wang2022 refers to (Wang et al., 2022), Xu2024 refers to (Xu et al., 2024a), and Song2024 refers to (Zong et al., 2024). We present the comparison results in Table 4.

Property	Wang2022	GATE	CMV-CGCN	A-GCL	PTGB	Xu2024	Contrasformer	Zong2024	CIA-GCL
Label-preserve	✓	✓	✓			✓	✓	✓	✓
All time series			✓	✓	✓	✓	✓	✓	✓
Learnble			✓	✓	✓			✓	✓
Diversity					✓				✓
Invariant-aware									✓

Table 4. Graph augmentation method comparison for brain graph analysis based on graph contrastive learning.

The "property" in the table refers to the conditions satisfied during the generation of augmented samples. "Label preserve" indicates whether the effect of labels was considered during the contrastive process, whether it was between groups in contrastive learning, or through simultaneous cross-entropy loss optimization. "All time series" highlights the use of the complete time-length BOLD signal in the augmentation process, as opposed to generating samples through techniques such as sliding windows. "Learnable" indicates that the augmented samples are automatically refined during optimization, with the model learning to modify nodes and edges. "Diversity" indicates whether multiple augmented samples were constructed for the same anchor graph for contrastive optimization in graph contrastive learning. "Invariant-aware" refers to whether important local information, specifically the invariant semantic details, is intentionally preserved in the brain graph during the augmentation process, avoiding arbitrary node and edge modifications.

C. Summary of Notations

The key symbols used in this paper and their corresponding descriptions are summarized in Table 5.

Symbol	Meaning
$\mathcal{G}$	Brain graph data set
Y	Subject's label
$\mathcal{E}, \acute{\mathcal{E}}$	Environment set
V	The set of nodes in the graph
E	The set of edges in the graph
G, G^{ori}	Original brain graph data
G^c, G^{inv}	Invariant subgraph
G^s, G^{ms}	Spurious subgraph, mixed spurious subgraph
G^a, G^{a_1}, G^{a_2}	Brain Augmented graph
n	Number of brain regions
d	Node feature dimensions in brain graph
p, q	The probability of an edge being selected
$\mathcal{M}$	Brain graph mask
k	The k-th subject
$\boldsymbol{z}$	Graph-level features
A	Adjacency matrix of a graph
$\boldsymbol{e}$	Environment
e	An edge in the graph
r	Maximum edge ratio to select
λ_1, λ_2	Hyperparameters in loss functions
$I(,)$	Invariant subgraph extractor
$\Phi(\cdot)$	Trade-off hyper-parameters
$g(\cdot)$	Graph Encoder
$w(\cdot)$	Predictor

Table 5. Notations and their descriptions.

D. Detailed Proofs

D.1. Proof of Theorem 3.4

Under the Definition 3.3, Theorem 3.4 indicates that if the G^{inv} obtains the $G_{inv} \rightarrow Y$ property, it shares the maximum information with invariant subgraphs from subjects with the same label:

$$G_k^{inv} \in \underset{m \in \{i | Y_i = Y_k, i \neq k\}}{\arg\max} I\left(G_k^{inv}; G_m^{inv}\right), \tag{14}$$

where m represents a subject selected from the set of other subjects with the same label. Theorem 3.4 essentially shows that the estimated graph G^{inv} obtained through Eq. 14. can produce the most informative and stable predictions about Y. We can get the optimal G^{inv} through comparative learning between groups in Eq. 11, namely, the mutual information between G^{inv} obtained from each training environment sample is minimized:

Proof.

$$
\begin{aligned}
&\max I\left(G_k^{inv}; G_m^{inv}\right), \quad m \in \{i \mid Y_i = Y_k, i \neq k\} \\
&\quad = I(G_k^{inv}, E = e_k; G_m^{inv}, E = e_m \mid Y) \\
&\quad = H(G_k^{inv}, E = e_k \mid Y) - H(G_k^{inv}, E = e_k \mid G_k^{inv}, E = e_m, Y) \\
&\quad = H(G_k^{inv} \cup G_k^{inv} \mid E = e_k, Y) - H(G_k^{inv}, G_k^{inv}, E = e_k \mid G_m^{inv}, E = e_m) \\
&\quad = H(G_k^{inv^g} \mid E = e_k, Y) + H(G_k^{inv^s} \mid G_k^{inv}, E = e_k, Y) \\
&\qquad - \Big\{ H(G_k^{inv^s} \mid G_m^{inv}, E = e_m, Y, G_k^{inv^s}, E = e_k) \\
&\qquad + H(G_k^{inv^c}, E = e_k \mid G_m^{inv}, E = e_m, Y) \Big\}
\end{aligned} \tag{15}
$$

where $G_k = G_k^{inv_c} \cup G_k^{inv_s}$, because there might be some subset $G_k^{inv_s} \subseteq G_k^s$ from the underlying G_k^s that entail the same information about label, i.e., $I(G_k^{inv_c} \cup G_s; Y) = I(G_k^{inv}; Y)$. And when G_k^{inv} = G_k^{inv*}, we have the entropy change as:

$$
\begin{aligned}
&\Delta I(G_k^{inv}, E = e_k; G_m^{inv}, E = e_m | Y) \\
&= \Delta H(G_k^{inv}, E = e_k | Y) - \Delta H(G_k^{inv}, E = e_k | G_m^{inv}, E = e_m, Y) \\
&= \left\{ H(G_k^{inv_s} | G_k^{inv_c}, E = e_k, Y) - H(G_k^{inv_l} | G_k^{inv_c}, E = e_k, Y) \right\} \\
&\quad - \left\{ H(G_k^{inv_l} | G_m^{inv_c}, G_m^{inv_s}, E = e_m, Y, G_k^{inv_c}, E = e_k) \right. \\
&\quad - \left. H(G_k^{inv_s} | G_m^{inv_c}, G_m^{inv_s}, E = e_m, Y, G_k^{inv_c}, E = e_k) \right\} \\
&= -H(G_k^{inv_l} | G_k^{inv_c}, E = e_k, Y) + H(G_k^{inv_l} | G_m^{inv_c}, G_m^{inv_s}, E = e_m, G_k^{inv_c}, Y, E = e_k)
\end{aligned} \tag{16}
$$

where $G_k^{inv_l} = G_k^{inv*} - G_k^{inv_c}$, since additionally conditioning on $G_k^{inv_s}$ in $H(H(G_k^{inv_l} | G_m^{inv_c}, G_m^{inv_s}, E = e_m, G_k^{inv_c}, Y, E = e_k)$ can not lead to new information about $G_k^{inv_c}$, we have:

$$
\begin{aligned}
H(G_k^{inv_l} | G_m^{inv_c}, G_m^{inv_s}, E = e_m, G_k^{inv_c}, Y, E = e_k) &= H(G_k^{inv_l} | G_m^{inv_c}, E = e_m, G_k^{inv_c}, Y, E = e_k), G_s^p, Y, E = \hat{e}) \\
&< H(G_k^{inv_l} | G_k^{inv_c}, E = e_k, Y)
\end{aligned} \tag{17}
$$

which follows that:

$$
\Delta I(G_k^{inv}, E = e_k; G_m^{inv}, E = e_m | Y) < 0
$$

□

D.2. Proof of Theorem 3.5

Under the Definition 3.3, Theorem 3.5 indicates that If the G^{inv} satisfies the invariant property, it follows that the G^{inv} maximizes the expected mutual information with the Y across the environment set $\mathcal{E}$:

$$
G_k^{inv} \in \arg\max \mathbb{E}_{e \in \mathcal{E}} \left[I(G_k^{inv}; Y \mid e) \right], \tag{18}
$$

where $I(\cdot;\cdot)$ is the mutual information between the label and the generated subgraph. This is a sufficient condition for the subgraph to be considered an G^{inv}. We can obtain the optimal G^{inv} under distribution shifts, i.e., the solution to handle the first characteristic of brain graph data mentioned in the introduction:

$$
\min \mathbb{E}_{G_k \in \mathcal{G}} \sum_{e \in \acute{e}} \mathcal{R}(f \mid e), \quad f = w \circ g \circ \Phi(G_k) = G_k^{inv} \tag{19}
$$

Proof. Under the Assumption 3.2, the graph G is composed of spurious subgraph G^s and invariant subgraph G^{inv}, where G^s is the complement of G^{inv}. Further denote $\hat{f} = \arg\min_{w,g} w \circ g \circ \Phi^*(G)$. To show that the subgraph extracted according to $\hat{f}$ is the optimal brain invariant subgraph G^{inv}, our proof strategy is to show that $\forall e \in \text{supp}(\mathcal{E})$, for any possible subgraph $\acute{G}$, $R(\acute{G}|\mathcal{E}) \leq R(G^{inv}|\mathcal{E})$ and therefore $\sup_{e \in \text{supp}(\mathcal{E})} R(\acute{G}|\mathcal{E}) \leq \sup_{e \in \text{supp}(\mathcal{E})} R(G^{inv}|\mathcal{E})$. To show the inequality, we have:

$$
R(\acute{G}|\mathcal{E}) = \mathbb{E}_{G,Y}[\ell(f(G), Y)] = \sum_{G,Y} P^e(G, Y) \ell(f(G), Y)
$$

$$= \sum_{G^s} P^e(G^s) \sum_{\Phi(G),Y} P^e(\Phi(G))\ell(w^*(g^*(\Phi^*(G)))), Y)$$

$$= \sum_{\Phi(G),Y} P^e(\Phi(G), Y)\ell(w^*(g^*(\Phi^*(G)))), Y)$$

$$\leq \sum_{\Phi(G),Y} P^e(\Phi(G), Y)\ell(w(g(\Phi^*(G)))), Y)$$

$$= \sum_{G^s} P^e(G^s) \sum_{\Phi(G),Y} P^e(\Phi(G))\ell(w(g(\Phi^*(G)))), Y)$$

$$= \sum_{G^s,Y} P^e(G^s, Y)\ell(G^s) = \mathbb{E}_{G,Y}[\ell(f(G), Y)] = R(G^{inv}|\mathcal{E}). \tag{20}$$

□

E. Data Preparation

E.1. Data preprocessing

Then we utilized the Data Processing Assistant for Resting-State fMRI (DPARSF)[1] A program to process the raw fMRI data. Specifically, we first removed the first five time points for fMRI and performed slice timing corrections. Then, we completed head motion correction by eliminating data from subjects with head movement exceeding 2mm horizontally and head rotational movement exceeding 2 degrees. Finally, we conducted image registration, smoothing, and filtering in sequence. The next step is brain parcellation, which is the process of dividing the brain into smaller regions or packages guided by a specific brain atlas. In this paper, we select the AAL atlas (tzo, 2002) with 90 brain regions and 26 cerebellar brain regions. This step allows for the analysis of functional connectivity within and between these parcels. After brain parcellation, for each atlas, the mean time series, namely the blood oxygen level-dependent (BOLD) signal in each region, is calculated by averaging the time series of all the voxels. This gives a representative measure of the average neural activity in each specific brain region, allowing for subsequent analysis of connectivity.

E.2. Brain graph generation

Each person's fMRI image is segmented into multiple ROIs using a designated brain atlas. In each ROI, the mean time series is calculated, and every person has $n \times t$ dimensions of time series data, where n represents the number of ROIs and t represents the time length. We then temporally normalize all subjects' signals to zero mean and unit variance. In this paper, each subject's brain functional connection network uses a graph structure G to describe the interconnections between brain regions. And the e_{ij} is calculated by the Pearson correlation coefficient between time-series signals of i-th ROI and j-th ROI, and the value represents the connection strength.

E.3. Datasets

We validate our approach on three real-world brain disease datasets. The three rs-fMRI datasets are Autism Brain Imaging Data Exchange (ABIDE) I, ABIDE II, and ADHD200, which are publicly available MRI datasets collected from different international imaging sites. The demographic statistics of these datasets are provided in Table 5.

- **ABIDE I**: ABIDE I[2] (and others, 2014) is a common dataset to evaluate the effectiveness of the Autism Spectrum Disorder (ASD) brain network classification tasks, which anonymously collected and shared fMRI and phenotype data for a total of 1035 subjects from 17 different sites around the world, including 505 subjects with ASD and 530 typical controls (TC). We used two brain segmentation atlases to divide the brain into smaller regions and obtained two datasets of ABIDE I. The first atlas is AAL (tzo, 2002), which contains 90 brain regions and 26 cerebellar brain regions and is often used in brain disease analysis.

[1] http://rfmri.org/DPARSF/

[2] https://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html

- **ABIDE II**: ABIDE I[3] was created to advance scientific discovery regarding the brain connectome in autism spectrum disorder (ASD). To date, it has aggregated over 1000 additional datasets with enhanced phenotypic characterization, particularly focusing on core ASD symptoms and related features. Currently, ABIDE II includes 19 sites, consisting of ten charter institutions and seven new members, collectively contributing 1114 datasets from 521 individuals with ASD and 593 controls.

- **ADHD200**: ADHD200[4] dataset, collected from 8 independent imaging sites, consists of 491 datasets from typically developing individuals and 285 from children and adolescents with ADHD (ages 7-21 years). The accompanying phenotypic data includes diagnostic status, ADHD symptom dimensions, age, sex, intelligence quotient (IQ), and lifetime medication history.

Dataset	Subgroup	Number	Gender (M/F)	Age (mean±std.)
ABIDE I	ASD	505	443/62	17.06±8.52
	TC	530	435/95	16.78±7.43
ABIDE II	ASD	483	410/73	15.09±9.24
	TC	556	382/174	15.27±9.51
ADHD200	ADHD	319	253/66	11.98±3.01
	TC	528	280/248	12.45±3.41

Table 6. The demographic statistics of the three brain datasets.

E.4. Selection of target domain data in Section 4.1.2.

Dataset	ABIDE I			ABIDE II			ADHD200		
information	Site	ASD	TC	Site	ASD	TC	Site	ADHD	TC
Target1	MAX_MUN	75	100	NYU_1NYU_2	73	30	Peking_1	48	88
Target2	UCLA_1,UCLA_2	54	44	BNI_1,IP_1	51	62	OHSUPeking_2	41	60
Target3	SBL,SDSU,STANFORD	49	57	TCD_1,UCD_1,UCLA_1	55	51	KKI,NeurolMAGE	60	106

Table 7. The demographic statistics of the datasets used in Section 4.1.2.

The data information of the target domain for testing the generalization ability of the model is in Section 4.1.2. summary is shown in Table 6. The target corresponds to the test set in the generalization verification experiment, and then we use the data of all the remaining sites as the training set. Since the data of a single site is relatively small, we also chose data from 2 or 3 sites as the target domain. Then we take the average of the experimental results of these three target domains. The data shown in the ASD (ADHD) and TC columns in the table are the number of samples in the test set.

E.5. Model Setup

The model uses Adam optimizer with lr=4e-5, batchsize=32, and max epochs=300. In the subgraph selection section, r=0.25. In the summary of contrastive loss, τ_1 and τ_2 are both set to 0.01. In $\mathcal{L} = \lambda_1\mathcal{L}_{con} + \lambda_2\mathcal{L}_{cau} + \mathcal{L}_{inv}$ formula, the two types of losses:λ_1,λ_2 = $\{1,0.5\}$. We employ ten-fold cross-validation to get a dependable and stable model, and the ratio of the training set, validation set, and test set is 8:1:1. Finally, we take the average of the ten-fold results for model comparison. All the experiments are conducted on a server equipped with NVIDIA GeForce RTX 3090 alongside the computational prowess of an AMD Ryzen 9 5950X 16-Core Processor CPU.

E.6. Baselines

We provide detailed descriptions of baselines in our experiments as follows:

- Traditional machine learning methods: The traditional methods include support vector machine (SVM) classifier and a random forest (RF) classifier, which were all implemented using the scikit-learn library (Pedregosa et al., 2011). We

[3] https://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html
[4] https://fcon_1000.projects.nitrc.org/indi/adhd200/

directly calculate the Pearson correlation coefficient on the time series to obtain the FC network, and then flatten it and concatenate it into a vector, which is fed to RF or SVM.

- BrainCNN (Kawahara et al., 2017): This method utilizes innovative graph-based convolutional filters to predict neurodevelopmental outcomes from brain networks. The learning rate is set to 1×10^{-3}, and the batch size is 32.
- GATE (Peng et al., 2023): This approach used sliding windows to build positive samples from fMRI images and applied semi-supervised learning for node classification. All hyperparameters are configured as per the original implementation.
- CIGA (Chen et al., 2022): We use the CIGAv2 model. Because CIGAv2 performs better than CIGAv1 in the experiment. We set the coefficient of $I(\widehat{G}_s; Y)$ to 1. The learning rate is set to 1×10^{-3}, and the batch size is 64.
- BrainGB (Cui et al., 2023): This approach proposed a standard pipeline including node feature construction, message passing, and graph pooling for brain graph analysis. The learning rate is set to 1×10^{-3}, and the batch size is 32.
- Com-BrainTF (Bannadabhavi et al., 2023): This hierarchical transformer model incorporates intra- and inter-community features of brain ROIs. For the ABIDE I, ϕ is set to 0.03, and the learning rate is 1×10^{-3}. For the ABIDE II atlas, ϕ is set to 0.0005. For the ADHD200 atlas, ϕ is set to 0.0005. and the learning rate is 1×10^{-4}. The pretraining is conducted over 50 epochs with a batch size of 32.
- METAFormer (Mahler et al., 2023): This approach is a transformer-based framework that integrates multi-atlas functional connectivity representations and self-supervised pretraining via masked input reconstruction. We only used the AAL atlas data for training.
- Contrasformer (Xu et al., 2024b): Contrasformer applies contrastive learning to brain networks by aligning ROI-level representations across subjects using identity-aware cross-attention. All other hyperparameters are consistent with the original implementation.
- BrainIB (Zheng et al., 2024b): BrainIB is a graph neural network framework that utilizes the information bottleneck principle to analyze fMRI data, aiming to resolve issues related to overfitting and generalization. All other hyperparameters are consistent with the original implementation.
- CI-GNN (Zheng et al., 2024a): This approach proposed a Granger causality-inspired network that identifies causally relevant subgraphs for disease diagnosis. The learning rate is set to 1×10^{-3}, and the batch size is 32.

E.7. Evaluation Indicators

We used six common metrics to evaluate the performance of the model, including accuracy (Acc), recall (Rec), precision (Pre), F1-score (F1), and balanced accuracy (bAcc). They can be calculated by:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}, \tag{21}$$

$$Rec = \frac{TP}{TP + FN}, \tag{22}$$

$$Pre = \frac{TP}{TP + FP}, \tag{23}$$

$$F1 = 2 \times \frac{Pre \times Rec}{Pre + Rec}, \tag{24}$$

$$bAcc = \frac{1}{2} \times (\frac{TP}{TP + FN} + \frac{TN}{FP + TN}), \tag{25}$$

where TP, FP, TN, and FN are the number of true positive subjects, false positive subjects, true negative subjects, and false negative subjects, respectively. In addition, we also utilize the area under curve (AUC) to evaluate the expected performance. It refers specifically to the area under the ROC curve in this paper, which takes the false positive rate (FPR) as abscissa and the true positive rate (TPR) as ordinate. They can be calculated by:

$$FPR = \frac{FP}{FP + TN}, TPR = \frac{TP}{TP + FN}. \tag{26}$$

F. Extra Experiments

F.1. Calculation process and ROI's scores of ABIDE I dataset

In the analysis of important brain regions, we first obtain the invariant subgraph of each person. Based on this, we acquire the important brain regions of each person, namely the points in the invariant subgraphs. Specifically, we use the adjacency matrix of the invariant subgraph and add it separately by row and column. Then we take the average of these two results as the importance score for each ROI. Sort these scores in descending order to obtain the top 10 important brain regions. The detailed information on the important brain regions of the ABIDE I dataset on the AAL atlas is shown in Table 7.

No.	Label	ROI	Abbreviation	Score
1	44	Calcarine_R	CAL.R	2.0311
2	43	Calcarine_L	CAL.L	1.9218
3	86	Temporal_Mid_R	MTG.R	1.885
4	75	Pallidum_L	PAL.L	1.8457
5	41	Amygdala_L	AMYG.L	1.8336
6	40	ParaHippocampal_R	PHG.R	1.7985
7	39	ParaHippocampal_L	PHG.L	1.772
8	90	Temporal_Inf_R	ITG.R	1.7571
9	85	Temporal_Mid_L	MTG.L	1.6722
10	45	Cuneus_L	CUN.L	1.6647

Table 8. Details on top 10 important ROIs of ABIDE I dataset on AAL atlas.

"No." represents the descending sorting order, "Label" is the default order of ROI in the AAL atlas, and "Score" indicates scientific counting after the calculation process in A4 (Calculation on the Important ROIs).

F.2. More Visualization Results

F.2.1. Important ROI analysis of ABIDE II dataset

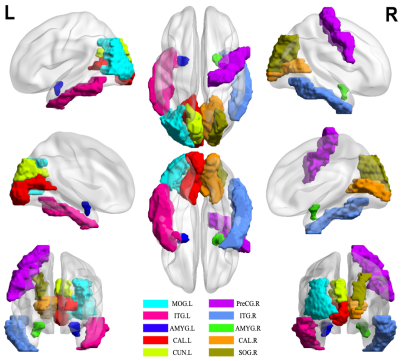


Figure 6. Top 10 important ROIs of ABIDE II on AAL atlas.

No.	Label	ROI	Abbreviation	Score
1	43	Calcarine_L	CAL.L	2.9844
2	44	Calcarine_R	CAL.R	2.9699
3	45	Cuneus_L	CUN.L	2.9299
4	42	Amygdala_R	AMYG.R	2.5436
5	50	Occipital_Sup_R	SOG.R	2.3723
6	51	Occipital_Mid_L	MOG.L	2.2975
7	41	Amygdala_L	AMYG.L	2.2292
8	90	Temporal_Inf_R	ITG.R	2.1846
9	2	Precentral_R	PreCG.R	2.1511
10	89	Temporal_Inf_L	ITG.L	2.1249

Table 9. Details on top 10 important ROIs of ABIDE II on AAL atlas

F.2.2. Important ROI analysis of ADHD200 dataset

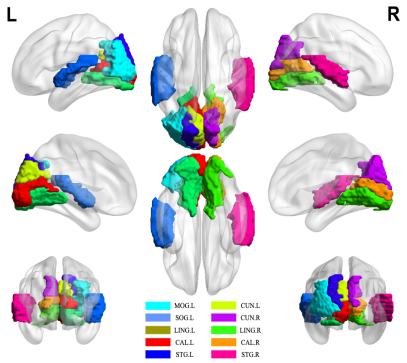


Figure 7. Top 10 important ROIs of ADHD200 on AAL atlas.

No.	Label	ROI	Abbreviation	Score
1	48	Lingual_R	LING.R	2.7251
2	47	Lingual_L	LING.L	2.6500
3	81	Temporal_Sup_L	STG.L	2.6270
4	44	Calcarine_R	CAL.R	2.6033
5	43	Calcarine_L	CAL.L	2.601
6	49	Occipital_Sup_L	SOG.L	2.5678
7	46	Cuneus_R	CUN.R	2.5666
8	82	Temporal_Sup_R	STG.R	2.4937
9	51	Occipital_Mid_L	MOG.L	2.4241
10	45	Cuneus_L	CUN.L	2.3938

Table 10. Details on top 10 important ROIs of ADHD200 on AAL atlas.