BRAIN DISEASE DETECTION BASED ON BOTTOM-UP HIERARCHICAL GRAPH REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Current research on brain disorders focuses on the functional connectivity networks of regions of interest (ROIs), but less attention has been paid to explicit brain functional cluster recognition, which neuroscience has identified as playing a crucial role in brain disease detection. In this paper, we propose a bottom-up hierarchical graph representation learning method for brain disease detection. Bottom-level brain graph representation with ROI node importance is designed to suppress noise for effective recognition of top-level brain functional clusters. Top-level brain graph representation with clustering consistency constraints explicitly characterizes functional clusters and interaction patterns. We demonstrate through theoretical analysis that bottom-level gating conditionally bounds perturbations in spectral alignment, while top-level clustering consistency constraints yield a conditional upper bound on the continuous normalized cut objective. Our model is validated on the publicly available neuroimaging datasets ADHD-200, OpenNeuro-ds002424, and OpenNeuro-ds000030, demonstrating significant advantages in brain disease detection. Code is available at `https://anonymous.4open.science/r/Bottom-Up-Hierarchical-Graph-Representation-Learning-0670`.

1 INTRODUCTION

Functional MRI (fMRI) has emerged as a promising source of objective biomarkers by capturing BOLD fluctuations across brain regions, enabling construction of functional connectivity networks reflecting coordinated activity patterns underlying cognition and behavior [26, 27]. A graph neural network (GNN) captures the relational structure inherent in brain networks by modeling ROIs as graph nodes and functional correlations as edges, enabling end-to-end learning directly on the connectivity graph [1, 2], spawning a line of brain-specific GNN methods. BrainGNN [3] introduced ROI-aware pooling to identify disease-related biomarkers from task-critical regions. BrainNetTF [8] leveraged self-attention for long-range inter-regional dependencies to improve disorder discrimination. BrainGB [4] implemented node-edge concatenation to preserve representations from the previous step during updating. GroupBNA [12] incorporated group-adaptive network enhancement to strengthen collective perception.

Despite these advances, existing methods extract brain functional clusters implicitly by encoding graph structure with learned weight matrices or attention scores of ROI nodes. Yet neurological research has revealed that brain regions differ substantially in functional importance [22, 23] and operate in a cluster manner [20, 24, 25] in brain disease detection. The absence of explicit cluster representations leaves functional cluster recognition opaque, limiting both classification performance and neurological interpretability.

To address these limitations, we propose a bottom-up hierarchical graph representation learning method that explicitly characterizes functional clusters and interaction patterns. Figure 1 compares our model to BrainGNN with implicit community clustering. The main contributions are as follows:

1. Bottom-level brain graph representation with ROI node importance is designed to suppress noise for robust recognition of top-level brain functional clusters. Top-level brain graph representation with clustering consistency constraints explicitly captures functional clusters and interaction patterns.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

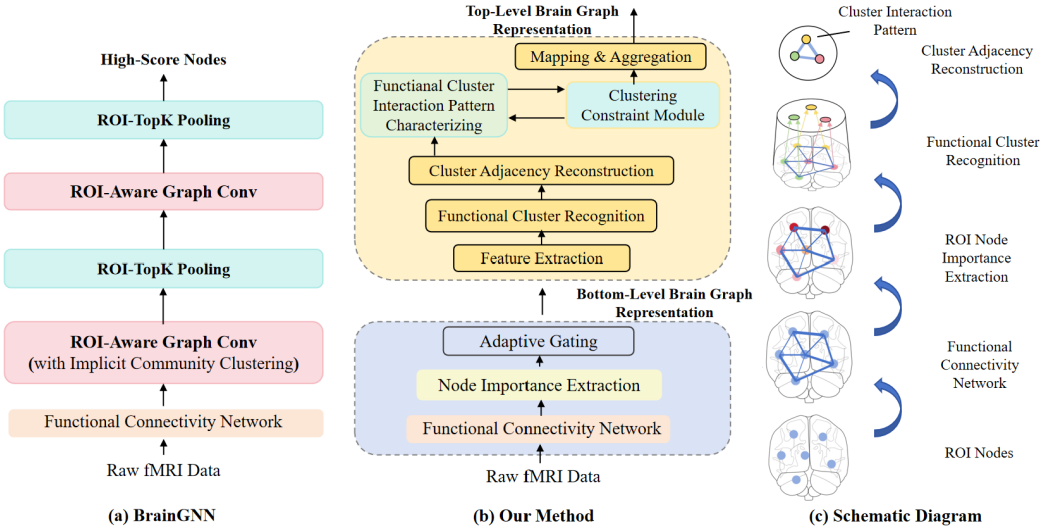


Figure 1: Comparison of the proposed method with BrainGNN. (a) Architecture of BrainGNN. (b) Architecture of the proposed method. (c) Schematic diagram of a bottom-up brain hierarchical graph representation.

2. We demonstrate through theoretical analysis that bottom-level gating conditionally bounds perturbations in spectral alignment, while top-level clustering consistency constraints yield a conditional upper bound on the continuous normalized cut objective.
3. Extensive experiments on the ADHD-200, OpenNeuro-ds002424, and OpenNeuro-ds000030 datasets demonstrate state-of-the-art performance over existing baselines.

2 Related Work

2.1 Graph Neural Networks for Brain Disease Detection

Ktena et al. [2] pioneered the use of Graph Convolutional Networks (GCNs) with spectral filters for brain connectivity classification, demonstrating the efficacy of graph-based representations over traditional vectorized features. Li et al. [3] proposed BrainGNN, an interpretable architecture incorporating ROI-aware graph convolution and pooling regularization to identify disease-related biomarkers. Cui et al. [4] established BrainGB, a benchmark for brain network analysis with GNNs standardizing evaluation protocols across multiple neuroimaging datasets. Despite these advances, existing brain GNNs predominantly rely on local message passing with limited receptive fields, capturing the modular organization of brain regions only implicitly through stacked neighborhood aggregations rather than explicit cluster representations.

2.2 Graph Transformer for Brain Disease Detection

The Transformer architecture [21] has been successfully adapted to graph data through various formulations. Veličković et al. [5] introduced Graph Attention Networks (GAT), employing masked self-attentional layers to weight neighbor contributions. However, GAT's attention is restricted to local neighborhoods, limiting global information aggregation. To overcome this, Dwivedi and Bresson [6] proposed a general Graph Transformer incorporating Laplacian eigenvectors as positional encodings. Kreuzer et al. [7] advanced this direction with Spectral Attention Networks (SAN), utilizing the full Laplacian spectrum to enable fully-connected attention without over-squashing. In neuroimaging, Kan et al. [8] introduced BrainNetTF, applying Transformer architectures to fMRI time series analysis. However, these approaches typically treat node features as sequences and pay insufficient attention to node importance for noise suppression.

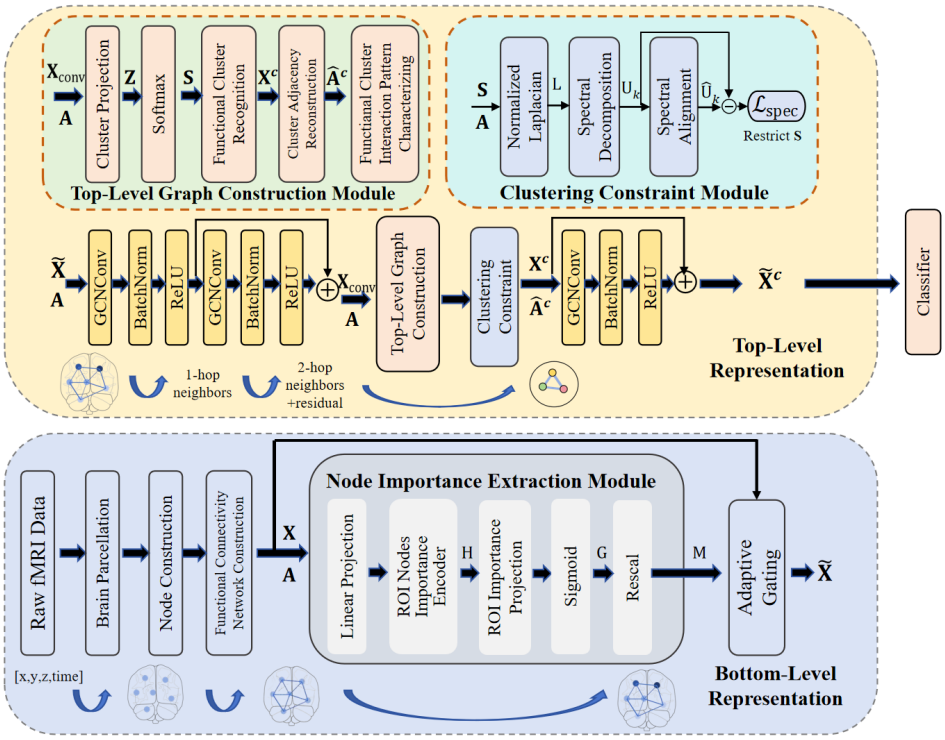


Figure 2: Overview of the proposed bottom-up hierarchical graph representation learning for brain disease detection.

3 Bottom-Up Hierarchical Brain Graph Representation

3.1 Motivation

Neuroscientific evidence demonstrates that brain functional networks are organized into modules that exhibit distinct roles in cognitive function and neuropsychiatric disorders [25], implying that explicit functional cluster recognition for brain ROIs is a biologically grounded structural property. However, existing graph-based methods primarily use implicit clustering and overlook explicit functional cluster recognition. In this paper, we aim to build an explicit cluster representation that accurately reflects functional clusters and interaction patterns of the brain. On the one hand, clustering quality must be ensured: a good functional clustering should minimize inter-cluster connectivity relative to intra-cluster connectivity, formally evaluated by the normalized cut (NCut) criterion [18]. On the other hand, clustering stability must be ensured: raw ROI features extracted from fMRI data are inherently noisy due to physiological artifacts and individual variability, and feeding noisy node features directly into the clustering module degrades convergence and reduces representational compactness.

As illustrated in Figure 2, our bottom-up hierarchical framework addresses these two challenges. At the bottom level, the node importance extraction module (NIEM) adaptively weights ROI features, aiming to suppress irrelevant information and noise for subsequent clustering. At the top level, the top-level graph construction module (TLGCM) explicitly models functional cluster and interaction patterns. The clustering constraint module (CCM) promotes spectral alignment and, under stated rank and graph assumptions, yields a theoretical upper bound on the continuous NCut objective.

3.2 Bottom-Level Brain Graph Representation with ROI Node Importance

We introduce the node importance extraction module (NIEM), which leverages the self-attention mechanism of the Transformer architecture to learn global inter-node dependencies, generating

adaptive gating signals that enhance task-critical features while suppressing noise from physiological artifacts, head motion, and scanner drift. The theoretical analysis is provided in Section 4.1.

Specifically, for an input node feature matrix $\mathbf{X} \in \mathbb{R}^{N\times F}$, a linear projection yields $\mathbf{X}' \in \mathbb{R}^{N\times H_L}$, which is fed into a Transformer encoder to produce $\mathbf{H}$. A gating matrix $\mathbf{G} \in \mathbb{R}^{N\times F}$ is then produced via a linear layer and sigmoid activation:

$$\mathbf{G} = \sigma(\mathbf{H}\mathbf{W}_g + \mathbf{b}_g). \tag{1}$$

To prevent information loss from excessively small gating values, $\mathbf{G}$ is rescaled to obtain $\mathbf{M}$. The importance-weighted ROI-level representation is obtained via element-wise multiplication:

$$\tilde{\mathbf{X}} = \mathbf{X} \odot \mathbf{M}, \tag{2}$$

where $\tilde{\mathbf{X}} \in \mathbb{R}^{N\times F}$ constitutes the bottom-level representation, serving as input to the top-level network.

3.3 Top-Level Brain Graph Representation with Clustering Consistency Constraints

Taking $\tilde{\mathbf{X}}$ from the bottom level as input, the top-level network first employs a GCN to extract bottom-level features and aggregate structural neighborhood information, then applies TLGCM to produce an explicit functional cluster assignment for brain regions, and finally constrains the clustering via CCM. The theoretical analysis is provided in Section 4.2.

Specifically, $\tilde{\mathbf{X}}$ is passed through the first GCN layer, followed by Batch Normalization (BN) and ReLU activation, yielding intermediate embeddings $\mathbf{H}^{(1)}$. The second GCN layer further refines these representations with a residual connection to mitigate the vanishing gradient problem:

$$\mathbf{X}_{\text{conv}} = \text{ReLU}\big(\text{BN}(\text{GCN}(\mathbf{H}^{(1)}))\big) + \mathbf{H}^{(1)}. \tag{3}$$

Top-Level Graph Construction Module (TLGCM). To recognize functional clusters, TLGCM learns a mapping from the original N nodes to K functional clusters. We map $\mathbf{X}_{\text{conv}}$ to clustering logits $\mathbf{Z} \in \mathbb{R}^{N\times K}$ via a linear projection, then convert them into a soft assignment matrix $\mathbf{S} \in \mathbb{R}^{N\times K}$ via temperature-scaled softmax:

$$S_{ij} = \frac{\exp(Z_{ij}/\tau)}{\sum_{m=1}^{K}\exp(Z_{im}/\tau)}, \tag{4}$$

where τ controls the smoothness of the assignment distribution. Based on the recognized functional clusters, we reconstruct the functional cluster features and coarsened adjacency matrix as:

$$\mathbf{X}^c = \mathbf{S}^\top \mathbf{X}_{\text{conv}}, \qquad \mathbf{A}^c = \mathbf{S}^\top \mathbf{A}\mathbf{S}. \tag{5}$$

A graph convolution layer is applied to learn interaction patterns among the K functional clusters:

$$\mathbf{H}^c = \hat{\mathbf{D}}^{c^{-\frac{1}{2}}} \hat{\mathbf{A}}^c \hat{\mathbf{D}}^{c^{-\frac{1}{2}}} \mathbf{X}^c \mathbf{W}^c, \tag{6}$$

where $\hat{\mathbf{A}}^c = \mathbf{A}^c + \mathbf{I} \in \mathbb{R}^{K\times K}$, $\hat{\mathbf{D}}^c$ is the corresponding degree matrix, and $\mathbf{W}^c$ is the learnable weight matrix.

Clustering Constraint Module (CCM). CCM constrains $\mathbf{S}$ to preserve key topological information encoded in the low-frequency eigenvectors of the graph Laplacian, encouraging spectral-structural consistency without guaranteeing full column rank. We compute the symmetric normalized Laplacian $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ and extract eigenvectors $\mathbf{U}_k \in \mathbb{R}^{N\times k}$ corresponding to the k smallest eigenvalues. For full-column-rank $\mathbf{S}$, the eigenvectors reconstructed via orthogonal projection onto $\text{span}(\mathbf{S})$ are:

$$\hat{\mathbf{U}}_k = \mathbf{P}_S \mathbf{U}_k, \qquad \mathbf{P}_S = \mathbf{S}(\mathbf{S}^\top\mathbf{S})^{-1}\mathbf{S}^\top. \tag{7}$$

The CCM loss is defined as the MSE between $\mathbf{U}_k$ and $\hat{\mathbf{U}}_k$:

$$\mathcal{L}_{\text{spec}} = \frac{1}{Nk}\sum_{i=1}^{N}\sum_{j=1}^{k}(U_{ij} - \hat{U}_{ij})^2. \tag{8}$$

After mapping and aggregation with a GCN, the top-level brain graph representation $\tilde{\mathbf{X}}^c \in \mathbb{R}^{K \times d_h}$ is obtained by applying BN, ReLU and the residual connection to $\mathbf{H}^c$. Together, $\mathbf{S}$ and $\tilde{\mathbf{X}}^c$ constitute the top-level representation, providing an explicit functional parcellation of the brain for the downstream classification module.

4 Theoretical Analysis

Under the implemented setting $K = k$ and assuming full-column-rank assignment matrices, the proposed framework has two theoretical components. The top-level CCM constrains the cluster representation $\mathbf{S}$ to preserve spectral structure, yielding a conditional upper bound on the normalized cut. The bottom-level NIEM controls feature perturbations from residual noise and gating error, which provides a conditional upper bound on the CCM loss and supports spectral alignment. These results do not claim universal denoising, optimality, or classification guarantees for arbitrary graphs.

4.1 ROI-Importance Bottom-Level Brain Graph Suppresses Noise for Robust Functional Clustering

Under the assumptions in Appendix C, NIEM controls how residual feature noise and gating approximation error propagate to the CCM loss through a fixed downstream encoder:

$$\mathcal{L}_{\text{spec}}(\mathbf{S}) \leq \frac{2C_P^2 L_F^2}{Nk} \Big(\|\mathbf{X}^*\|_F \, \epsilon_m + \delta\rho + \delta\epsilon_m \Big)^2 + 2\epsilon_0, \tag{9}$$

where δ bounds noise, ϵ_m bounds gating error, $\rho = \|\mathbf{M}^*\|_{\max} < 1$, ϵ_0 bounds reference alignment error, and C_P, L_F are stability constants. With bounded constants, the bound vanishes as $\delta, \epsilon_m, \epsilon_0 \to 0$. Combined with the regular-graph theorem above, this yields convergence of the continuous NCut objective to its spectral minimum. The complete derivation and sufficient conditions for a tighter bound relative to ungated inference are provided in Appendix C.

4.2 Clustering-Consistency Top-Level Brain Graph Maintains Clustering Quality and Captures Interactions

CCM preserves selected low-frequency spectral components in the cluster subspace, which under regularity yields an upper bound on the continuous NCut objective.

Theorem 1 (Spectral Information Preservation for Top-Level Representation)**.** *For orthonormal* $\mathbf{U}_k$ *and full-column-rank* $\mathbf{S}$*, minimizing the CCM loss*

$$\mathcal{L}_{\text{spec}} = \frac{1}{Nk} \|\mathbf{U}_k - \mathbf{P}_S \mathbf{U}_k\|_F^2 \tag{10}$$

encourages preservation of the k *lowest-frequency Laplacian eigenvectors in* $\text{span}(\mathbf{S})$*. Specifically, the retained-energy factor for the* i*-th eigenvector is:*

$$\alpha_i = 1 - \frac{\|\mathbf{u}_i - \mathbf{P}_S \mathbf{u}_i\|^2}{\|\mathbf{u}_i\|^2}, \tag{11}$$

and $\mathcal{L}_{\text{spec}} = \frac{1}{Nk} \sum_{i=1}^{k} (1 - \alpha_i)$.

Proof. Since $\mathbf{P}_S$ is an orthogonal projector, α_i measures the fraction of $\mathbf{u}_i$'s energy retained in $\text{span}(\mathbf{S})$. Summing the reconstruction errors of the unit eigenvectors yields $\mathcal{L}_{\text{spec}} = \frac{1}{Nk} \sum_{i=1}^{k} (1 - \alpha_i)$. The complete derivation is provided in Appendix A. □

Remark 1. This result concerns spectral reconstruction within the assignment subspace, rather than biological validity or preservation of the entire graph spectrum. We consider $k = K$, matching the cluster dimension to the number of preserved spectral components.

We next quantify how CCM controls an upper bound on the continuous normalized cut (NCut) objective.

Definition 1 (Continuous Normalized Cut)**.** For symmetric nonnegative adjacency $\mathbf{A}$, positive degree matrix $\mathbf{D}$, and full-column-rank soft assignments $\mathbf{S} \in \mathbb{R}^{N\times K}$, define the continuous NCut extension [18]:

$$\mathrm{NCut}(\mathbf{S}) = K - \mathrm{tr}\left((\mathbf{S}^\top\mathbf{D}\mathbf{S})^{-1/2}\mathbf{S}^\top\mathbf{A}\mathbf{S}(\mathbf{S}^\top\mathbf{D}\mathbf{S})^{-1/2}\right). \tag{12}$$

Lemma 1 (Projection Quality)**.** *Let $\mathbf{U}_k \in \mathbb{R}^{N\times k}$ contain orthonormal eigenvectors for the k smallest normalized Laplacian eigenvalues, and let $\mathbf{S} \in \mathbb{R}^{N\times k}$ have full column rank, with $\mathbf{P}_S$ the orthogonal projector onto* $\mathrm{span}(\mathbf{S})$*. If*

$$\mathcal{L}_{\mathrm{spec}} = \frac{1}{Nk}\|\mathbf{U}_k - \mathbf{P}_S\mathbf{U}_k\|_F^2 = \epsilon, \tag{13}$$

then there exists an orthonormal basis $\mathbf{V}$ for $\mathrm{span}(\mathbf{S})$ *such that*

$$\|\mathbf{V} - \mathbf{U}_k\mathbf{R}\|_F \le 2\sqrt{Nk\epsilon} \tag{14}$$

for some orthogonal matrix $\mathbf{R} \in \mathbb{R}^{k\times k}$.

Lemma 2 (Trace Perturbation)**.** *Let $\mathbf{L}$ be a symmetric positive semidefinite matrix with $\|\mathbf{L}\|_2 \le \lambda_{\max}$, and let $\mathbf{V}, \tilde{\mathbf{V}} \in \mathbb{R}^{N\times k}$ have orthonormal columns and satisfy*

$$\|\mathbf{V} - \tilde{\mathbf{V}}\|_F \le \delta. \tag{15}$$

Then

$$\left|\mathrm{tr}(\mathbf{V}^\top\mathbf{L}\mathbf{V}) - \mathrm{tr}(\tilde{\mathbf{V}}^\top\mathbf{L}\tilde{\mathbf{V}})\right| \le 2\lambda_{\max}\sqrt{k}\,\delta. \tag{16}$$

Classical spectral clustering theory [19] gives $\sum_{i=1}^K \lambda_i$ as the minimum of the unrestricted orthonormal trace relaxation, rather than the general discrete NCut optimum. The following theorem bounds the continuous objective evaluated at $\mathbf{S}$; its proof appears in Appendix A.

Theorem 2 (Restricted NCut Bound under Regularity)**.** *Assume $K = k$, full-column-rank $\mathbf{S} \in \mathbb{R}^{N\times K}$, and a symmetric nonnegative adjacency matrix with $\mathbf{D} = d\mathbf{I}$ for some $d > 0$. Let*

$$\epsilon = \mathcal{L}_{\mathrm{spec}}(\mathbf{S}) = \frac{1}{NK}\left\|(\mathbf{I} - \mathbf{P}_S)\mathbf{U}_K\right\|_F^2. \tag{17}$$

Then

$$\mathrm{NCut}(\mathbf{S}) \le \sum_{i=1}^K \lambda_i + 4\lambda_{\max}K\sqrt{N\epsilon}. \tag{18}$$

For a fixed graph, as $\epsilon \to 0$, this continuous objective converges to the spectral minimum $\sum_{i=1}^K \lambda_i$.

Remark 2 (Scope and Implementation)**.** Setting $K = k$ matches the dimensions; full rank and graph regularity remain separate assumptions. Under these conditions, the spectral lower bound gives

$$\sum_{i=1}^k \lambda_i \le \mathrm{NCut}(\mathbf{S}) \le \sum_{i=1}^k \lambda_i + 4\lambda_{\max}k\sqrt{N\epsilon}. \tag{19}$$

Reducing ϵ tightens this upper bound without guaranteeing monotonic improvement or a bound after hard rounding. Appendix A derives an additional degree-heterogeneity term for approximately regular graphs.

5 Experimental Results

5.1 Datasets and Experimental Settings

After completing the representation, we construct a main classifier and an auxiliary classifier to ensure classification accuracy while promoting the emergence of patterns in the cluster construction process that are both class-discriminative and structurally meaningful. We evaluate our model on three neuroimaging datasets: ADHD-200 [9], OpenNeuro-ds002424 (OpenNeuro) [13], and OpenNeuro-ds000030 [28], which offer fMRI data for neuropsychiatric disorders. The ADHD-200 includes data from 357 children with ADHD and 582 controls. We clean the dataset by removing samples with missing labels. OpenNeuro-ds002424 includes multimodal data from 232 ADHD

Table 1: Performance comparison (Mean ± Std) and p-values on three public datasets: OpenNeuro-ds002424, ADHD-200, and OpenNeuro-ds000030.

Method		Accuracy (%)		AUC (%)		F1-Score (%)	
		Mean ± Std	**p-value**	**Mean ± Std**	**p-value**	**Mean ± Std**	**p-value**
OpenNeuro-ds002424							
Machine Learning	Naive Bayes	61.01 ± 2.50	3.43×10^{-7}	63.07 ± 2.61	1.10×10^{-7}	61.00 ± 2.39	2.31×10^{-7}
	Decision Tree	60.73 ± 2.59	5.66×10^{-7}	59.83 ± 2.90	3.15×10^{-7}	60.65 ± 2.71	8.83×10^{-7}
Graph Transformer	SAN (NeurIPS'21)	55.82 ± 1.83	3.41×10^{-6}	56.65 ± 2.47	1.34×10^{-5}	32.59 ± 9.97	3.69×10^{-4}
	BioBGT (ICLR'25)	56.45 ± 4.39	1.21×10^{-3}	53.80 ± 3.49	7.35×10^{-3}	38.83 ± 13.64	2.92×10^{-2}
Graph Neural Networks	BrainGNN (MIA'21)	64.22 ± 5.82	2.92×10^{-5}	75.06 ± 4.19	1.49×10^{-8}	63.42 ± 6.96	1.80×10^{-4}
	BrainGB (TMI'23)	54.73 ± 2.90	5.97×10^{-4}	51.77 ± 2.83	7.93×10^{-2}	34.31 ± 7.71	1.20×10^{-4}
	GroupBNA (ICLR'24)	54.36 ± 8.36	1.34×10^{-1}	51.30 ± 6.02	5.12×10^{-1}	54.36 ± 8.36	1.34×10^{-1}
	BrainOOD (ICLR'25)	63.91 ± 2.82	2.15×10^{-7}	69.73 ± 2.56	1.53×10^{-9}	45.99 ± 2.96	2.06×10^{-3}
	EBIGNN (JBHI'25)	54.64 ± 6.24	4.32×10^{-2}	52.62 ± 7.51	2.98×10^{-1}	33.56 ± 15.99	1.01×10^{-2}
	Ours	$\mathbf{78.62 \pm 3.65}$	2.15×10^{-9}	$\mathbf{86.41 \pm 2.37}$	5.43×10^{-12}	$\mathbf{78.42 \pm 3.65}$	2.32×10^{-9}
ADHD-200							
Machine Learning	Naive Bayes	57.59 ± 3.90	2.47×10^{-4}	58.19 ± 3.75	1.05×10^{-4}	56.60 ± 4.02	7.97×10^{-4}
	Decision Tree	58.13 ± 2.36	2.68×10^{-6}	56.15 ± 2.22	3.15×10^{-5}	58.37 ± 2.50	8.83×10^{-6}
Graph Transformer	SAN (NeurIPS'21)	60.85 ± 1.94	2.68×10^{-8}	55.72 ± 3.47	5.55×10^{-4}	14.08 ± 12.53	8.05×10^{-6}
	BioBGT (ICLR'25)	55.77 ± 4.08	1.55×10^{-3}	51.53 ± 3.19	1.64×10^{-1}	31.49 ± 17.92	9.74×10^{-3}
Graph Neural Networks	BrainGNN (MIA'21)	50.63 ± 2.92	5.12×10^{-1}	50.63 ± 2.92	5.12×10^{-1}	46.05 ± 6.82	1.00×10^{-1}
	BrainGB (TMI'23)	59.15 ± 2.45	8.82×10^{-7}	54.46 ± 3.27	1.95×10^{-3}	27.72 ± 13.20	4.70×10^{-4}
	GroupBNA (ICLR'24)	50.05 ± 10.66	9.88×10^{-1}	49.62 ± 1.00	2.60×10^{-1}	50.05 ± 10.66	9.88×10^{-1}
	BrainOOD (ICLR'25)	61.12 ± 2.71	4.76×10^{-7}	55.50 ± 2.65	9.26×10^{-5}	37.32 ± 9.87	2.85×10^{-3}
	EBIGNN (JBHI'25)	60.78 ± 2.34	1.60×10^{-5}	46.94 ± 2.75	2.19×10^{-2}	5.59 ± 8.02	1.10×10^{-7}
	Ours	$\mathbf{63.05 \pm 3.15}$	5.71×10^{-7}	$\mathbf{61.58 \pm 2.64}$	3.53×10^{-7}	$\mathbf{61.23 \pm 2.95}$	1.17×10^{-6}
OpenNeuro-ds000030							
Machine Learning	Naive Bayes	60.00 ± 8.94	8.45×10^{-3}	46.83 ± 10.84	4.04×10^{-1}	58.86 ± 8.00	8.92×10^{-3}
	Decision Tree	63.82 ± 6.45	1.21×10^{-4}	53.83 ± 9.56	2.60×10^{-1}	63.25 ± 7.81	6.58×10^{-4}
Graph Transformer	SAN (NeurIPS'21)	70.29 ± 2.91	3.84×10^{-9}	47.76 ± 7.84	3.90×10^{-1}	4.69 ± 7.25	1.01×10^{-8}
	BioBGT (ICLR'25)	56.57 ± 10.67	8.33×10^{-2}	47.70 ± 6.62	3.00×10^{-1}	23.57 ± 12.85	1.11×10^{-4}
Graph Neural Networks	BrainGNN (MIA'21)	66.47 ± 4.84	1.90×10^{-6}	52.21 ± 6.09	2.80×10^{-1}	58.62 ± 6.74	3.00×10^{-3}
	BrainGB (TMI'23)	60.86 ± 8.29	2.51×10^{-3}	47.32 ± 4.99	1.24×10^{-1}	27.78 ± 4.20	4.36×10^{-8}
	GroupBNA (ICLR'24)	51.71 ± 19.85	7.91×10^{-1}	51.20 ± 2.04	9.58×10^{-2}	51.71 ± 19.85	7.91×10^{-1}
	BrainOOD (ICLR'25)	67.14 ± 6.90	2.30×10^{-5}	50.60 ± 6.80	7.86×10^{-1}	14.85 ± 17.28	1.30×10^{-4}
	EBIGNN (JBHI'25)	62.29 ± 6.98	3.54×10^{-4}	59.08 ± 9.52	1.47×10^{-2}	37.67 ± 13.23	1.61×10^{-2}
	Ours	$\mathbf{73.24 \pm 6.37}$	1.68×10^{-6}	$\mathbf{62.03 \pm 6.54}$	3.72×10^{-4}	$\mathbf{68.73 \pm 7.41}$	3.40×10^{-5}

Table 2: Combined ablation study results on three datasets. BL: Bottom-Level Brain Graph Network, TL: Top-Level Brain Graph Network.

Dataset	BL	TL	Accuracy (%)		AUC (%)		F1-Score (%)	
			Mean ± Std	p-value	Mean ± Std	p-value	Mean ± Std	p-value
OpenNeuro-ds002424	×	×	$\mathbf{72.20 \pm 3.21}$	6.49×10^{-9}	$\mathbf{79.54 \pm 3.13}$	4.20×10^{-10}	$\mathbf{72.21 \pm 3.22}$	6.73×10^{-9}
	✓	×	$\mathbf{76.24 \pm 3.29}$	1.88×10^{-9}	$\mathbf{84.17 \pm 3.28}$	1.70×10^{-10}	$\mathbf{76.01 \pm 3.42}$	2.84×10^{-9}
	×	✓	$\mathbf{71.10 \pm 4.77}$	3.26×10^{-7}	$\mathbf{79.19 \pm 3.99}$	4.01×10^{-9}	$\mathbf{70.98 \pm 4.86}$	4.00×10^{-7}
	✓	✓	$\mathbf{78.62 \pm 3.65}$	2.15×10^{-9}	$\mathbf{86.41 \pm 2.37}$	5.43×10^{-12}	$\mathbf{78.42 \pm 3.65}$	2.32×10^{-9}
ADHD-200	×	×	$\mathbf{60.70 \pm 1.92}$	4.34×10^{-8}	$\mathbf{57.27 \pm 3.23}$	8.37×10^{-5}	$\mathbf{56.46 \pm 2.87}$	8.31×10^{-5}
	✓	×	$\mathbf{61.98 \pm 2.62}$	2.49×10^{-7}	$\mathbf{60.72 \pm 3.05}$	2.33×10^{-6}	$\mathbf{60.03 \pm 3.51}$	1.27×10^{-5}
	×	✓	$\mathbf{61.82 \pm 3.17}$	1.38×10^{-6}	$\mathbf{60.12 \pm 2.79}$	1.76×10^{-6}	$\mathbf{58.28 \pm 4.10}$	1.87×10^{-4}
	✓	✓	$\mathbf{63.05 \pm 3.15}$	5.71×10^{-7}	$\mathbf{61.58 \pm 2.64}$	3.53×10^{-7}	$\mathbf{61.23 \pm 2.95}$	1.17×10^{-6}
OpenNeuro-ds000030	×	×	$\mathbf{70.88 \pm 5.95}$	2.32×10^{-6}	$\mathbf{65.47 \pm 6.67}$	6.64×10^{-5}	$\mathbf{68.14 \pm 9.86}$	3.71×10^{-4}
	✓	×	$\mathbf{69.41 \pm 6.47}$	8.54×10^{-6}	$\mathbf{57.11 \pm 9.97}$	6.11×10^{-2}	$\mathbf{63.81 \pm 7.18}$	2.68×10^{-4}
	×	✓	$\mathbf{70.00 \pm 7.87}$	3.25×10^{-5}	$\mathbf{63.77 \pm 13.88}$	1.55×10^{-2}	$\mathbf{65.23 \pm 10.26}$	1.59×10^{-3}
	✓	✓	$\mathbf{73.24 \pm 6.37}$	1.68×10^{-6}	$\mathbf{62.03 \pm 6.54}$	3.72×10^{-4}	$\mathbf{68.73 \pm 7.41}$	3.40×10^{-5}

samples and 314 healthy control samples. OpenNeuro-ds000030 includes multimodal data from 58 schizophrenia samples and 138 control samples. We use an NVIDIA RTX 6000 Ada Generation GPU for training. The data are split into training (60%), validation (20%), and test (20%) sets. Further configuration details are provided in Appendix E. To assess statistical significance, we conduct 10 independent runs with different random seeds and use a one-sample t-test against the null hypothesis that each metric has a mean of 0.5. The t-test results are presented in Table 1.

Table 3: Discussion on the number of clusters (K) in the top-level brain graph network across three datasets.

Dataset	K	Accuracy (%)		AUC (%)		F1-Score (%)	
		Mean ± Std	p-value	Mean ± Std	p-value	Mean ± Std	p-value
OpenNeuro-ds002424	6	75.69 ± 4.25	2.15×10^{-8}	84.51 ± 3.14	1.08×10^{-10}	75.26 ± 4.71	6.08×10^{-8}
	7	75.69 ± 3.90	1.00×10^{-8}	82.46 ± 5.63	3.26×10^{-8}	75.02 ± 4.57	5.13×10^{-8}
	8	76.15 ± 4.18	1.61×10^{-8}	83.92 ± 5.21	1.11×10^{-8}	75.68 ± 4.72	5.41×10^{-8}
	9	78.62 ± 3.65	2.15×10^{-9}	$\mathbf{86.41 \pm 2.37}$	$\mathbf{5.43 \times 10^{-12}}$	78.42 ± 3.65	2.32×10^{-9}
	10	$\mathbf{78.72 \pm 3.91}$	$\mathbf{3.86 \times 10^{-9}}$	86.30 ± 3.76	3.41×10^{-10}	$\mathbf{78.50 \pm 3.94}$	$\mathbf{4.40 \times 10^{-9}}$
ADHD-200	6	62.14 ± 2.08	2.95×10^{-8}	61.50 ± 2.92	8.86×10^{-7}	59.10 ± 2.30	8.56×10^{-7}
	7	60.70 ± 2.80	1.15×10^{-6}	59.40 ± 3.84	4.40×10^{-5}	57.90 ± 3.40	6.60×10^{-5}
	8	$\mathbf{63.05 \pm 3.15}$	$\mathbf{5.71 \times 10^{-7}}$	$\mathbf{61.58 \pm 2.64}$	$\mathbf{3.53 \times 10^{-7}}$	$\mathbf{61.23 \pm 2.95}$	$\mathbf{1.17 \times 10^{-6}}$
	9	62.09 ± 3.14	1.07×10^{-6}	61.09 ± 2.72	6.54×10^{-7}	60.30 ± 2.29	2.86×10^{-7}
	10	62.57 ± 3.63	2.64×10^{-6}	59.68 ± 3.34	1.13×10^{-5}	59.53 ± 4.79	2.12×10^{-4}
OpenNeuro-ds000030	6	70.59 ± 9.11	8.11×10^{-5}	63.72 ± 12.05	7.67×10^{-3}	68.51 ± 8.42	$\mathbf{9.97 \times 10^{-5}}$
	7	$\mathbf{73.24 \pm 6.37}$	1.68×10^{-6}	62.03 ± 6.54	3.72×10^{-4}	$\mathbf{68.73 \pm 7.41}$	3.40×10^{-5}
	8	67.06 ± 3.90	3.60×10^{-7}	$\mathbf{64.25 \pm 10.22}$	$\mathbf{2.37 \times 10^{-3}}$	65.03 ± 6.44	6.29×10^{-5}
	9	69.12 ± 7.58	3.46×10^{-5}	61.59 ± 9.72	5.98×10^{-3}	66.17 ± 8.32	2.51×10^{-4}
	10	66.76 ± 5.10	4.04×10^{-6}	54.68 ± 11.70	2.60×10^{-1}	63.32 ± 6.74	2.21×10^{-4}

5.2 Overall Performance

We compare our model with state-of-the-art methods: (1) typical machine learning (ML) methods, including Naive Bayes [31] and Decision Tree [32]; (2) graph transformer models, including SAN [7], and BioBGT [11]; (3) graph neural networks for brain graph analysis, including BrainGNN [3], BrainGB [4], GroupBNA [12], BrainOOD [33], and EBIGNN [34].

Table 1 reports the performance comparison on OpenNeuro-ds002424, ADHD-200, and OpenNeuro-ds000030. Overall, **Ours** achieves the best average performance on almost all dataset–metric combinations. In particular, on OpenNeuro-ds002424 and ADHD-200, **Ours** consistently outperforms the compared methods in Accuracy, AUC, and F1-Score. On OpenNeuro-ds000030, **Ours** also obtains the best Accuracy and F1-Score, while its AUC is slightly lower than that of Polynormer. The reported p-values test departures from 0.5, not improvements over competing methods. These results demonstrate the effectiveness of the proposed method across the three public datasets.

5.3 Ablation Studies

The ablation results in Table 2 demonstrate the complementary contributions of the bottom-level (BL) and top-level (TL) networks. BL enhances node-level feature discriminability, suppressing noise and highlighting disease-relevant brain regions; introducing BL alone substantially improves performance on most datasets. TL maps variable-sized brain graphs into a fixed-dimensional cluster space via spectral structure alignment; however, without BL's denoising, it tends to aggregate noisy features and often falls below the baseline. When combined, BL's fine-grained feature refinement stabilizes TL's hierarchical cluster construction, while TL's spectral regularization guides BL toward globally meaningful structures. This joint optimization achieves the best Accuracy and F1-Score across all three datasets, although the baseline retains higher AUC on OpenNeuro-ds000030, indicating metric-dependent benefits from combining BL and TL.

5.4 Analysis of Clustering Effectiveness

To assess the connectivity structure and inference-time stability of the learned functional clusters, we conduct a connectivity consistency analysis on OpenNeuro-ds002424, using a fixed pretrained model and 109 test graphs, with details in Appendix D. Connectivity, measured by absolute correlations between ROI feature profiles, is slightly higher within than between clusters, though the mean stabilized ratio is sensitive to near-zero between-cluster connectivity and should not be read as a consistent advantage (Fig. 3). Near-zero modularity also indicates limited community separation, despite a positive modularity–connectivity-ratio correlation (Fig. 4). Moreover, the model is trained for classification with spectral regularization, not for maximizing modularity in this feature-

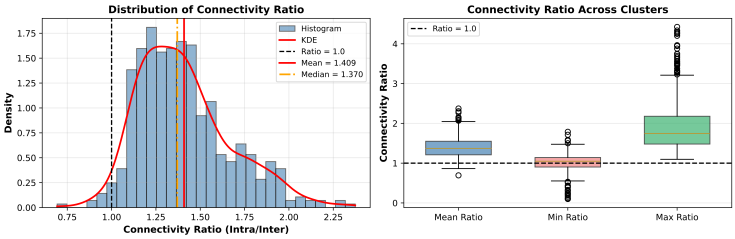


Figure 3: Distribution of connectivity ratios across samples. (Left) Histogram of mean intra/inter-cluster connectivity ratio. (Right) Box plots of connectivity ratios across clusters within samples.

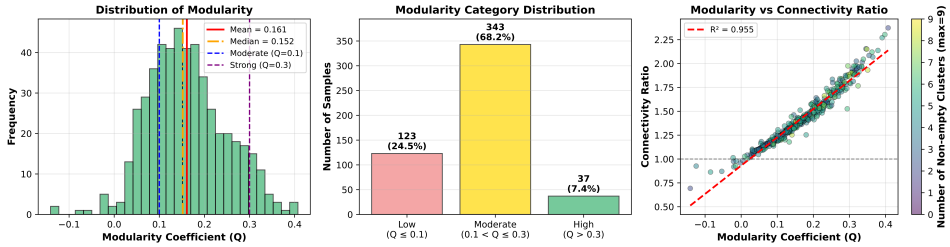


Figure 4: Modularity analysis of learned cluster patterns. (Left) Distribution of modularity coefficients Q. (Middle) Sample proportions across modularity categories. (Right) Positive correlation between modularity and connectivity ratio.

correlation matrix. Connectivity scores thus complement the perturbation results by characterizing partitions, not by directly measuring diagnostic utility.

Inference-only perturbation experiments in Appendix D provide more direct evidence of assignment stability. Across 10 trials, partitions retain high agreement under mild Gaussian noise and substantial agreement under moderate ROI subsampling, with ARI and NMI decreasing as perturbations strengthen. Overall, results support a limited claim: partitions show inference-time stability under fixed evaluation, while connectivity separation is modest. Without random-partition controls or cross-model comparisons, these results do not establish non-random functional communities or reproducibility across training runs.

5.5 Discussion on the Number of Clusters

The results in Table 3 demonstrate that our model is robust to the choice of cluster number K: performance remains competitive across all tested configurations, with the optimal K differing modestly between datasets. This small shift is likely attributable to differences in data source and disorder type, which affect the underlying functional network topology. Small K over-aggregates neurobiological communities, reducing discriminative capacity, while large K risks fragmenting coherent functional systems and overfitting to noise. The Accuracy range is 3.03, 2.35, and 6.48 percentage points on OpenNeuro-ds002424, ADHD-200, and OpenNeuro-ds000030, respectively, indicating dataset-dependent sensitivity.

6 Conclusion

We propose a hierarchical graph representation learning method for brain disease detection that explicitly models functional clusters and interaction patterns. Bottom-level brain graph representation with ROI node importance is designed to suppress noise for recognition of top-level brain functional clusters. Top-level brain graph representation with clustering consistency constraints characterizes functional clusters and interaction patterns. Theoretical analysis provides conditional bounds linking bottom-level gating, top-level spectral alignment, and the continuous normalized cut objective. Experiments on three datasets demonstrate state-of-the-art performance, outperforming existing baselines. Limitations and broader societal impacts are discussed in Appendix F and G.

AI USE STATEMENT

Generative AI tools were used solely to assist with LaTeX typesetting, the formatting and organization of mathematical expressions, and limited language polishing. The authors reviewed these edits and take full responsibility for the scientific content and final manuscript.

REFERENCES

[1] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[2] S. I. Ktena, S. Parisot, E. Ferrante, M. Rajchl, M. Lee, B. Glocker, and D. Rueckert, "Distance metric learning using graph convolutional networks: Application to functional brain networks," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2017, pp. 469–477.

[3] X. Li, Y. Zhou, N. Dvornek, M. Zhang, S. Gao, J. Zhuang, D. Scheinost, P. Ventola, and J. S. Duncan, "BrainGNN: Interpretable brain graph neural network for fMRI analysis," *Medical Image Analysis*, vol. 74, p. 102233, 2021.

[4] H. Cui, W. Dai, Y. Zhu, X. Li, L. He, A. B. Nielsen, W. Zhang, and C. Yang, "BrainGB: A benchmark for brain network analysis with graph neural networks," *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 493–506, 2023.

[5] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[6] V. P. Dwivedi and X. Bresson, "A generalization of transformer networks to graphs," in *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.

[7] D. Kreuzer, D. Beaini, W. Hamilton, V. Létourneau, and P. Tossou, "Rethinking graph transformers with spectral attention," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 21 618–21 629.

[8] X. Kan, H. Cui, V. T. Annavarapu, and C. Yang, "Brain network transformer," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 25 586–25 599.

[9] ADHD-200 Consortium, "The ADHD-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience," *Frontiers in Systems Neuroscience*, vol. 6, p. 62, 2012.

[10] C. Deng, Z. Yue, and Z. Zhang, "Polynormer: Polynomial-expressive graph transformer," in *International Conference on Learning Representations (ICLR)*, 2024.

[11] C. Peng, Y. Huang, Q. Dong, S. Yu, F. Xia, C. Zhang, and Y. Jin, "Biologically plausible brain graph transformer," in *International Conference on Learning Representations (ICLR)*, 2025.

[12] C. Peng, M. Liu, C. Meng, S. Yu, and F. Xia, "Adaptive brain network augmentation based on group-aware graph learning," in *Tiny Papers Track at ICLR*, 2024.

[13] J. R. Booth, G. E. Cooke, J. Gayda, R. Hammer, M. N. Lytle, M. A. Stein, and M. Tennekoon, "Working memory and reward in children with and without attention deficit hyperactivity disorder (ADHD)," *OpenNeuro*, 2021.

[14] X. Han, R. Xue, S. Du, and Y. Gao, "Inter-intra high-order brain network for ASD diagnosis via functional MRIs," in *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, ser. LNCS 15002, 2024, pp. 216–226.

[15] F. Wang, K. Pena-Pena, W. Qian, and G. R. Arce, "T-HyperGNNs: Hypergraph neural networks via tensor representations," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[16] Y. Feng, H. You, et al., "Hypergraph neural networks," in *AAAI*, 2019, pp. 3558–3565.

[17] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*. Academic Press, 1990.

[18] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[19] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[20] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, no. 3, pp. 186–198, 2009.

[21] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[22] O. Sporns, "Structure and function of complex brain networks," *Dialogues in Clinical Neuroscience*, vol. 8, no. 4, pp. 381–388, 2006.

[23] M. P. van den Heuvel and H. E. Hulshoff Pol, "Exploring the brain network: A review on resting-state fMRI functional connectivity," *European Neuropsychopharmacology*, vol. 20, no. 8, pp. 519–534, 2010.

[24] J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, and S. E. Petersen, "Functional network organization of the human brain," *Neuron*, vol. 72, no. 4, pp. 665–678, 2011.

[25] B. T. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, B. Fischl, H. Liu, and R. L. Buckner, "The organization of the human cerebral cortex estimated by intrinsic functional connectivity," *Journal of Neurophysiology*, vol. 106, no. 3, pp. 1125–1165, 2011.

[26] B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde, "Functional connectivity in the motor cortex of resting human brain using echo-planar MRI," *Magnetic Resonance in Medicine*, vol. 34, no. 4, pp. 537–541, 1995.

[27] M. D. Fox and M. E. Raichle, "Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging," *Nature Reviews Neuroscience*, vol. 8, no. 9, pp. 700–711, 2007.

[28] R. Bilder, R. Poldrack, T. Cannon, E. London, N. Freimer, E. Congdon, K. Karlsgodt, and F. Sabb, "UCLA Consortium for Neuropsychiatric Phenomics LA5c Study," OpenNeuro, 2018.

[29] A. Virmaux and K. Scaman, "Lipschitz regularity of deep neural networks: Analysis and efficient estimation," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, pp. 3839–3848, 2018.

[30] B. Gao and L. Pavel, "On the properties of the softmax function with application in game theory and reinforcement learning," *arXiv preprint arXiv:1704.00805*, 2017.

[31] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[32] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Wadsworth, 1984.

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

[34] S. Mamoon, Z. Xia, W. Jin, A. Alfakih, and J. Lu, "Joint dynamic brain network estimation and graph representation learning for the recognition of neurological disorders," *IEEE Journal of Biomedical and Health Informatics*, 2025.

A THEORETICAL PROOFS

A.1 THEORETICAL PROOF OF THE CLUSTERING CONSTRAINT MODULE (CCM)

We establish the theoretical foundation of the Clustering Constraint Module (CCM) by demonstrating that, under the setting $K = k$ and the full-column-rank assumption on $\mathbf{S}$, constraining the cluster assignment matrix to reconstruct low-frequency Laplacian eigenvectors preserves these spectral components in the learned cluster subspace.

Let $\mathcal{G} = (\mathbf{A}, \mathbf{X})$ denote a brain network graph with N nodes (ROIs), where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is a symmetric, nonnegative adjacency matrix and $\mathbf{X} \in \mathbb{R}^{N\times F}$ is the node feature matrix. Throughout this appendix, we assume strictly positive node degrees. The symmetric normalized graph Laplacian is defined as

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}, \tag{20}$$

where $\mathbf{D}$ is the degree matrix. Eigendecomposition yields

$$\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top, \tag{21}$$

with eigenvalues $0 = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_N \le 2$ and corresponding orthonormal eigenvectors $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_N]$.

The low-frequency eigenvectors $\mathbf{U}_k = [\mathbf{u}_1, \ldots, \mathbf{u}_k] \in \mathbb{R}^{N\times k}$ describe slowly varying graph signals and form the spectral subspace used in the analysis. For the soft cluster assignment matrix $\mathbf{S} \in \mathbb{R}^{N\times K}$ learned by the Top-Level Graph Construction Module (TLGCM), we focus on the regime $K = k$. We further assume that $\mathbf{S}$ has full column rank and that $\mathbf{S}^\top\mathbf{S}$ is invertible. We consider the orthogonal projection matrix

$$\mathbf{P}_S = \mathbf{S}(\mathbf{S}^\top\mathbf{S})^{-1}\mathbf{S}^\top \tag{22}$$

onto the column space of $\mathbf{S}$. The reconstructed spectral features are $\hat{\mathbf{U}}_k = \mathbf{P}_S\mathbf{U}_k$. The identities below concern this exact orthogonal projector; replacing the inverse by a ridge-regularized inverse generally changes its idempotence and requires a separate approximation analysis.

We now present the formal proof of Theorem 1, which characterizes the CCM loss in terms of spectral reconstruction error and establishes the connection between loss minimization and spectral preservation.

Proof of Theorem 1. **Step 1: Decomposing the CCM loss.**

By the definition of the Frobenius norm,

$$\begin{aligned}\|\mathbf{U}_k - \mathbf{P}_S\mathbf{U}_k\|_F^2 &= \|(\mathbf{I} - \mathbf{P}_S)\mathbf{U}_k\|_F^2 \\ &= \sum_{i=1}^{k}\|(\mathbf{I} - \mathbf{P}_S)\mathbf{u}_i\|^2 = \sum_{i=1}^{k}\|\mathbf{u}_i - \mathbf{P}_S\mathbf{u}_i\|^2.\end{aligned} \tag{23}$$

Therefore,

$$\mathcal{L}_{\text{spec}} = \frac{1}{Nk}\sum_{i=1}^{k}\|\mathbf{u}_i - \mathbf{P}_S\mathbf{u}_i\|^2. \tag{24}$$

Step 2: Expressing the reconstruction error via orthogonal projection.

Since $\mathbf{P}_S$ is an orthogonal projection, it satisfies $\mathbf{P}_S^2 = \mathbf{P}_S$ and $\mathbf{P}_S^\top = \mathbf{P}_S$. Hence $\mathbf{I} - \mathbf{P}_S$ is also an orthogonal projection. For each i,

$$\|\mathbf{u}_i - \mathbf{P}_S\mathbf{u}_i\|^2 = \mathbf{u}_i^\top(\mathbf{I} - \mathbf{P}_S)\mathbf{u}_i = \|\mathbf{u}_i\|^2 - \mathbf{u}_i^\top\mathbf{P}_S\mathbf{u}_i. \tag{25}$$

Moreover, because $\mathbf{P}_S$ is symmetric and idempotent,

$$\mathbf{u}_i^\top\mathbf{P}_S\mathbf{u}_i = \|\mathbf{P}_S\mathbf{u}_i\|^2. \tag{26}$$

Therefore,

$$\|\mathbf{u}_i - \mathbf{P}_S\mathbf{u}_i\|^2 = \|\mathbf{u}_i\|^2 - \|\mathbf{P}_S\mathbf{u}_i\|^2. \tag{27}$$

Step 3: Defining the retained-energy factor.

Each $\mathbf{u}_i$ is a unit eigenvector of the normalized Laplacian, so $\|\mathbf{u}_i\| = 1$. Define the retained-energy factor

$$\alpha_i = \frac{\|\mathbf{P}_S \mathbf{u}_i\|^2}{\|\mathbf{u}_i\|^2} = \|\mathbf{P}_S \mathbf{u}_i\|^2 \in [0, 1], \tag{28}$$

which measures the fraction of energy retained when $\mathbf{u}_i$ is projected onto $\mathrm{span}(\mathbf{S})$. Then

$$\|\mathbf{u}_i - \mathbf{P}_S \mathbf{u}_i\|^2 = 1 - \alpha_i. \tag{29}$$

Here α_i is an energy ratio, rather than a claim that $\mathbf{P}_S \mathbf{u}_i$ is a scalar multiple of $\mathbf{u}_i$.

Step 4: Deriving the final form of the CCM loss.

Substituting the above result into the CCM loss gives

$$\mathcal{L}_{\mathrm{spec}} = \frac{1}{Nk} \sum_{i=1}^{k} (1 - \alpha_i). \tag{30}$$

The normalization factor $\frac{1}{Nk}$ matches the mean squared reconstruction error over all $N \times k$ entries of $\mathbf{U}_k$. In particular, $0 \leq \mathcal{L}_{\mathrm{spec}} \leq 1/N$.

Step 5: Spectral preservation under loss minimization.

Since each $\alpha_i \in [0, 1]$, we have $\mathcal{L}_{\mathrm{spec}} \geq 0$, with equality if and only if $\alpha_i = 1$ for all $i = 1, \ldots, k$. This occurs if and only if

$$\mathbf{P}_S \mathbf{u}_i = \mathbf{u}_i, \qquad \forall i = 1, \ldots, k, \tag{31}$$

meaning that each selected low-frequency eigenvector lies entirely in $\mathrm{span}(\mathbf{S})$. Under $K = k$ and full column rank, this is equivalent to $\mathrm{span}(\mathbf{S}) = \mathrm{span}(\mathbf{U}_k)$. Therefore, minimizing $\mathcal{L}_{\mathrm{spec}}$ is equivalent to maximizing the total retained spectral energy $\sum_{i=1}^{k} \alpha_i$. This establishes that CCM loss minimization encourages the learned assignment subspace to preserve the selected low-frequency spectral structure. □

This spectral preservation property provides a mathematical motivation for CCM as a structural regularizer. It does not, by itself, establish that the learned clusters correspond to neurobiologically meaningful functional communities or that the coarsened graph preserves the entire spectrum of the original graph.

A.2 Spectral Clustering Quality Guarantees

While Theorem 1 shows that CCM preserves low-frequency spectral structure, the next step is to quantify how this preservation relates to clustering quality of the top-level brain graph. We connect CCM to the normalized cut (NCut) expression in Definition 1.

The Spectral-NCut Connection. Classical spectral clustering theory [19] relates discrete NCut minimization to a continuous spectral relaxation. The relaxed trace problem over all K-dimensional orthonormal matrices has minimum $\sum_{i=1}^{K} \lambda_i$, attained by the first K eigenvectors of the normalized Laplacian. This spectral minimum is a lower bound on the discrete NCut optimum and need not be attained by a discrete partition. For a hard assignment matrix with nonempty clusters, Definition 1 agrees with the usual discrete NCut; for a full-column-rank soft assignment matrix, we use the same expression as a continuous extension. The bounds below concern this soft trace objective and do not directly guarantee the NCut of labels obtained by rounding or taking $\arg\max$.

We focus on $K = k$ and assume that $\mathbf{S}$ has full column rank. Together with strictly positive degrees, this ensures that $\mathbf{S}^\top \mathbf{S}$ and $\mathbf{S}^\top \mathbf{D} \mathbf{S}$ are invertible. We now present the proofs of Lemma 1, Lemma 2, and Theorem 2. The proof strategy proceeds in three steps: (i) reformulating NCut as a Laplacian trace; (ii) relating the CCM reconstruction error to the distance from the low-frequency spectral subspace; and (iii) identifying the regular-graph bound and an explicit extension under approximate regularity.

Proof of Lemma 1. From the definition of CCM loss,

$$\|\mathbf{U}_k - \mathbf{P}_S \mathbf{U}_k\|_F^2 = Nk\epsilon. \tag{32}$$

Since $K = k$ and $\mathbf{S}$ has full column rank, $\mathbf{P}_S$ projects onto a k-dimensional subspace. Let $\mathbf{V}$ be any orthonormal basis of this subspace, so that $\mathbf{P}_S = \mathbf{V}\mathbf{V}^\top$.

Let $\mathbf{U}_k^\top \mathbf{V} = \mathbf{A}_0 \mathbf{\Sigma} \mathbf{B}_0^\top$ be a singular value decomposition, where the singular values satisfy $0 \le \sigma_i \le 1$. Choose the orthogonal matrix $\mathbf{R} = \mathbf{A}_0 \mathbf{B}_0^\top$. The orthogonal Procrustes identity [17] gives

$$\|\mathbf{V} - \mathbf{U}_k \mathbf{R}\|_F^2 = 2k - 2\sum_{i=1}^{k} \sigma_i. \tag{33}$$

Meanwhile,

$$\|(\mathbf{I} - \mathbf{P}_S)\mathbf{U}_k\|_F^2 = k - \|\mathbf{V}^\top \mathbf{U}_k\|_F^2 = \sum_{i=1}^{k} (1 - \sigma_i^2). \tag{34}$$

Since $1 - \sigma_i \le 1 - \sigma_i^2$ for $\sigma_i \in [0, 1]$,

$$\|\mathbf{V} - \mathbf{U}_k \mathbf{R}\|_F^2 \le 2\|(\mathbf{I} - \mathbf{P}_S)\mathbf{U}_k\|_F^2 = 2Nk\epsilon. \tag{35}$$

Taking square roots yields the slightly stronger bound $\|\mathbf{V} - \mathbf{U}_k \mathbf{R}\|_F \le \sqrt{2Nk\epsilon}$ and, in particular, the stated inequality

$$\|\mathbf{V} - \mathbf{U}_k \mathbf{R}\|_F \le 2\sqrt{Nk\epsilon}. \tag{36}$$

Thus, small CCM reconstruction error controls the distance between orthonormal bases after an appropriate rotation. □

Proof of Lemma 2. Let $\mathbf{L}$ be symmetric positive semidefinite with $\|\mathbf{L}\|_2 \le \lambda_{\max}$, and let $\mathbf{V}, \tilde{\mathbf{V}} \in \mathbb{R}^{N \times k}$ have orthonormal columns and satisfy $\|\mathbf{V} - \tilde{\mathbf{V}}\|_F \le \delta$. Then

$$\begin{aligned} &\operatorname{tr}(\mathbf{V}^\top \mathbf{L} \mathbf{V}) - \operatorname{tr}(\tilde{\mathbf{V}}^\top \mathbf{L} \tilde{\mathbf{V}}) \\ &\quad = \operatorname{tr}((\mathbf{V} - \tilde{\mathbf{V}})^\top \mathbf{L} \mathbf{V}) + \operatorname{tr}(\tilde{\mathbf{V}}^\top \mathbf{L} (\mathbf{V} - \tilde{\mathbf{V}})). \end{aligned} \tag{37}$$

By the Cauchy–Schwarz inequality for the Frobenius inner product,

$$\begin{aligned} |\operatorname{tr}((\mathbf{V} - \tilde{\mathbf{V}})^\top \mathbf{L} \mathbf{V})| &\le \|\mathbf{V} - \tilde{\mathbf{V}}\|_F \|\mathbf{L}\mathbf{V}\|_F \\ &\le \delta \|\mathbf{L}\|_2 \|\mathbf{V}\|_F \le \delta \lambda_{\max} \sqrt{k}. \end{aligned} \tag{38}$$

The second term satisfies the same bound because $\mathbf{L}$ is symmetric and $\|\tilde{\mathbf{V}}\|_F = \sqrt{k}$. Combining these estimates gives

$$\left| \operatorname{tr}(\mathbf{V}^\top \mathbf{L} \mathbf{V}) - \operatorname{tr}(\tilde{\mathbf{V}}^\top \mathbf{L} \tilde{\mathbf{V}}) \right| \le 2\lambda_{\max} \sqrt{k}\, \delta, \tag{39}$$

which establishes the trace perturbation bound. □

Proof of Theorem 2. **Step 1: Reformulating Normalized Cut.**

For a full-column-rank soft assignment matrix $\mathbf{S} \in \mathbb{R}^{N \times K}$ with nonnegative entries and unit row sums, the continuous NCut expression is

$$\operatorname{NCut}(\mathbf{S}) = K - \operatorname{tr}\left((\mathbf{S}^\top \mathbf{D} \mathbf{S})^{-1/2} \mathbf{S}^\top \mathbf{A} \mathbf{S} (\mathbf{S}^\top \mathbf{D} \mathbf{S})^{-1/2} \right). \tag{40}$$

Using $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$, we obtain

$$\mathbf{A} = \mathbf{D}^{1/2} (\mathbf{I} - \mathbf{L}) \mathbf{D}^{1/2}. \tag{41}$$

Define $\tilde{\mathbf{S}} = \mathbf{D}^{1/2} \mathbf{S}$. Then

$$\mathbf{S}^\top \mathbf{A} \mathbf{S} = \tilde{\mathbf{S}}^\top (\mathbf{I} - \mathbf{L}) \tilde{\mathbf{S}}, \qquad \mathbf{S}^\top \mathbf{D} \mathbf{S} = \tilde{\mathbf{S}}^\top \tilde{\mathbf{S}}. \tag{42}$$

Let

$$\mathbf{Q} = \tilde{\mathbf{S}} (\tilde{\mathbf{S}}^\top \tilde{\mathbf{S}})^{-1/2} = \mathbf{D}^{1/2} \mathbf{S} (\mathbf{S}^\top \mathbf{D} \mathbf{S})^{-1/2}. \tag{43}$$

Then $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_K$, and

$$\operatorname{NCut}(\mathbf{S}) = \operatorname{tr}(\mathbf{Q}^\top \mathbf{L} \mathbf{Q}). \tag{44}$$

The variational characterization of the sum of the smallest eigenvalues consequently gives

$$\mathrm{NCut}(\mathbf{S}) \geq \sum_{i=1}^{K} \lambda_i. \tag{45}$$

Step 2: Relating CCM loss to the optimal spectral subspace.

Under $K = k$, Lemma 1 gives an orthonormal basis $\mathbf{V}$ of $\mathrm{span}(\mathbf{S})$ and an orthogonal matrix $\mathbf{R}$ such that

$$\|\mathbf{V} - \mathbf{U}_k\mathbf{R}\|_F \leq 2\sqrt{Nk\epsilon}. \tag{46}$$

Applying Lemma 2 and using invariance of the trace under orthogonal changes of basis,

$$\begin{aligned} \mathrm{tr}(\mathbf{V}^\top \mathbf{L}\mathbf{V}) &\leq \mathrm{tr}((\mathbf{U}_k\mathbf{R})^\top \mathbf{L}(\mathbf{U}_k\mathbf{R})) + 4\lambda_{\max} k\sqrt{N\epsilon} \\ &= \sum_{i=1}^{k} \lambda_i + 4\lambda_{\max} k\sqrt{N\epsilon}. \end{aligned} \tag{47}$$

This controls the trace on $\mathrm{span}(\mathbf{S})$, whereas NCut uses $\mathrm{span}(\mathbf{D}^{1/2}\mathbf{S})$. The next step relates these two subspaces.

Step 3: Relating the subspaces under regularity.

Under the theorem's regular-graph assumption $\mathbf{D} = d\mathbf{I}$ with $d > 0$, the matrices $\mathbf{Q}$ and $\mathbf{V}$ span the same subspace. Therefore,

$$\mathrm{NCut}(\mathbf{S}) = \mathrm{tr}(\mathbf{V}^\top \mathbf{L}\mathbf{V}). \tag{48}$$

For completeness, we also derive a separate extension to approximately regular graphs. Suppose

$$\max_{i\in[N]} |d_i - \bar{d}| \leq \eta\bar{d}, \qquad 0 \leq \eta < 1, \tag{49}$$

where $\bar{d} > 0$ is the average degree. Define $\mathbf{B} = \mathbf{D}^{1/2}/\sqrt{\bar{d}}$. Then

$$\|\mathbf{B} - \mathbf{I}\|_2 \leq 1 - \sqrt{1-\eta} \leq \eta, \qquad \lambda_{\min}(\mathbf{B}^2) \geq 1 - \eta. \tag{50}$$

An orthonormal basis of $\mathrm{span}(\mathbf{D}^{1/2}\mathbf{S})$ is

$$\mathbf{W} = \mathbf{B}\mathbf{V}(\mathbf{V}^\top \mathbf{B}^2 \mathbf{V})^{-1/2}. \tag{51}$$

Since $\mathbf{P}_S = \mathbf{V}\mathbf{V}^\top$, we have

$$(\mathbf{I} - \mathbf{P}_S)\mathbf{W} = (\mathbf{I} - \mathbf{P}_S)(\mathbf{B} - \mathbf{I})\mathbf{V}(\mathbf{V}^\top \mathbf{B}^2 \mathbf{V})^{-1/2}. \tag{52}$$

It follows that

$$\|(\mathbf{I} - \mathbf{P}_S)\mathbf{W}\|_F \leq \frac{\sqrt{k}\eta}{\sqrt{1-\eta}}. \tag{53}$$

Applying the same Procrustes argument used in Lemma 1, there exists an orthogonal matrix $\mathbf{O}$ such that

$$\|\mathbf{W} - \mathbf{V}\mathbf{O}\|_F \leq 2\|(\mathbf{I} - \mathbf{P}_S)\mathbf{W}\|_F \leq \frac{2\sqrt{k}\eta}{\sqrt{1-\eta}}. \tag{54}$$

Because $\mathbf{W}$ and $\mathbf{Q}$ span the same subspace, Lemma 2 yields

$$\begin{aligned} \left|\mathrm{NCut}(\mathbf{S}) - \mathrm{tr}(\mathbf{V}^\top \mathbf{L}\mathbf{V})\right| &= \left|\mathrm{tr}(\mathbf{W}^\top \mathbf{L}\mathbf{W}) - \mathrm{tr}((\mathbf{V}\mathbf{O})^\top \mathbf{L}(\mathbf{V}\mathbf{O}))\right| \\ &\leq \frac{4\lambda_{\max} k\eta}{\sqrt{1-\eta}}. \end{aligned} \tag{55}$$

This explicitly bounds the additional discrepancy due to degree heterogeneity.

Step 4: Final bound.

For the regular graphs covered by Theorem 2, combining Steps 1–3 and using $K = k$ gives

$$\sum_{i=1}^{K} \lambda_i \leq \mathrm{NCut}(\mathbf{S}) \leq \sum_{i=1}^{K} \lambda_i + 4\lambda_{\max} K\sqrt{N\epsilon}, \tag{56}$$

which proves the stated theorem. For a fixed graph, the lower and upper bounds imply convergence of this continuous NCut objective to the spectral minimum as $\epsilon \to 0$.

The separate approximately regular extension is

$$\mathrm{NCut}(\mathbf{S}) \leq \sum_{i=1}^{k} \lambda_i + 4\lambda_{\max} k\sqrt{N\epsilon} + \frac{4\lambda_{\max} k\eta}{\sqrt{1-\eta}}, \qquad 0 \leq \eta < 1. \tag{57}$$

If $\eta \leq \eta_0 < 1$ for a fixed η_0, the last term is at most $C\eta$ with the explicit constant

$$C = \frac{4\lambda_{\max} k}{\sqrt{1-\eta_0}}. \tag{58}$$

Thus, reducing the CCM error tightens the spectral approximation term in the upper bound. This does not imply that the actual NCut decreases monotonically with CCM loss, and for fixed nonzero degree heterogeneity the additional term remains. □

Remark 3 (Implemented Regime)**.** Theorem 2 assumes $K = k$, full-column-rank $\mathbf{S}$, and a regular graph with positive degree. Setting $K = k$ matches the cluster and spectral dimensions, but does not itself ensure the rank or graph assumptions. Softmax assignments and orthogonality regularization do not automatically guarantee full column rank or a uniform lower bound on the smallest singular value. The applicability of the theorem therefore requires these conditions to be checked separately. Rank-deficient assignments and unequal cluster and spectral dimensions are outside the present proof.

Remark 4 (Approximate Regularity)**.** The main theorem concerns exactly regular graphs; the extension above assumes $0 \leq \eta < 1$ and gives an explicit degree-heterogeneity term. Graphs containing isolated nodes fall outside this extension, and the estimate becomes uninformative as η approaches one. The empirical relationships reported in Appendix B can provide descriptive evidence beyond the regular-graph setting, but do not establish that the theorem's assumptions hold or prove robustness of its guarantee outside the stated regime.

B Empirical Analysis of CCM Loss and the NCut Objective

To examine the empirical relationship underlying Theorem 2, we analyze CCM loss and the reported NCut scores on 546 functional brain graphs from OpenNeuro-ds002424 [13]. These experiments characterize associations between spectral reconstruction error and the clustering objective; they do not directly verify the theoretical upper bound.

Experimental Configuration. We use $K = 9$ clusters in the pooling layer and $k = 9$ spectral components in the CCM loss. This matches the dimensional condition $K = k$ in Theorem 2. However, this configuration alone does not establish full column rank or graph regularity. The theorem concerns the continuous NCut expression in Definition 1, rather than the NCut of hard labels obtained by rounding. Its direct application additionally requires the evaluated objective and numerical treatment of singular matrices to match the theoretical formulation. The retained results do not establish whether NCut was evaluated on soft assignments or rounded labels, or how isolated nodes and singular matrices were handled; these implementation details remain unverified, limiting reproducibility and direct comparison with the theorem.

Association between CCM Loss and NCut. Figure 5 shows a positive association between CCM reconstruction error and NCut across the 546 graphs. A linear fit using $\mathcal{L}_{\mathrm{spec}}$ gives Pearson's $\rho = 0.9055$ ($p = 1.11 \times 10^{-204}$, $R^2 = 0.8199$), while using $\sqrt{\mathcal{L}_{\mathrm{spec}}}$ gives $\rho = 0.9100$ ($p = 3.37 \times 10^{-210}$, $R^2 = 0.8282$). Thus, lower reconstruction error is associated with lower NCut in the evaluated collection, and the square-root transformation provides a slightly better fit. However, the theorem bounds NCut from above and does not prescribe an exact relationship between observed scores and reconstruction error. These fits therefore do not establish the tightness of the bound or show that reducing CCM loss necessarily decreases NCut.

Degree Heterogeneity. Figure 6 shows NCut against the degree-heterogeneity statistic $\eta = \max_i |d_i - \bar{d}|/\bar{d}$. The reported values range from 0.000 to 1.000, with mean 0.214 ± 0.410. The distribution is bimodal: 429 graphs (78.6%) are near-complete with approximately uniform degrees and $\eta \approx 0$, whereas 117 graphs (21.4%) contain isolated nodes and exhibit $\eta \approx 1$. The pooled

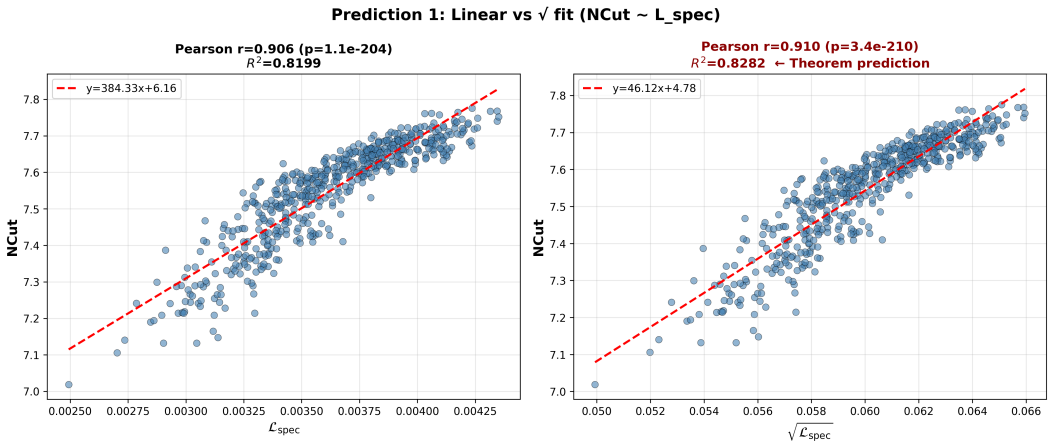


Figure 5: **Empirical Association between CCM Loss and NCut.** Fits of NCut against $\mathcal{L}_{\text{spec}}$ (left, $\rho = 0.9055$, $R^2 = 0.8199$) and $\sqrt{\mathcal{L}_{\text{spec}}}$ (right, $\rho = 0.9100$, $R^2 = 0.8282$) across 546 brain graphs. The square-root transformation provides a slightly better descriptive fit; neither fit directly tests the theoretical upper bound.

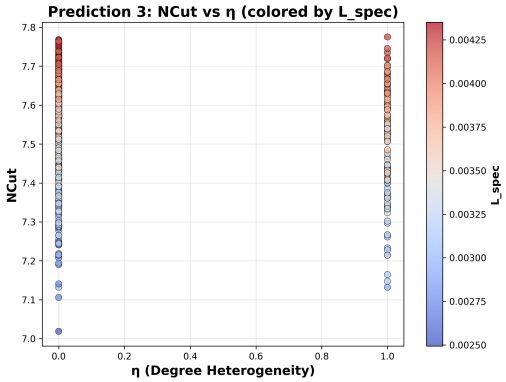


Figure 6: **NCut and Degree Heterogeneity.** NCut versus η, colored by $\mathcal{L}_{\text{spec}}$, across 546 graphs. The bimodal distribution distinguishes approximately regular graphs from graphs containing isolated nodes. The latter lie outside the assumptions of the theoretical bounds, so this plot describes empirical variation rather than validating the guarantee under arbitrary degree heterogeneity.

CCM–NCut correlation is strong ($\rho = 0.9055$), but this does not establish equally strong associations within each subgroup. Moreover, the isolated-node graphs fall outside both the positive-degree regular-graph theorem and the approximately regular extension in Appendix A. Their reported scores must therefore be interpreted as descriptive results under the evaluation implementation, rather than as instances covered by the guarantee.

Summary. These results show that CCM reconstruction error and NCut are positively associated across the evaluated graphs. The square-root transformation yields a modest improvement in fit, and the dataset includes graphs with substantially different degree structure. The observations provide empirical support for the structural relevance of CCM, while leaving verification of the numerical bound, its assumptions, and the behavior of hard partitions to separate analyses. The reported significance values are nominal graph-level statistics; repeated observations from the same participant may violate independence.

C Theoretical Analysis of Bottom-Level Brain Graph Representation

Here we establish a conditional theoretical connection between the bottom-level node importance extraction module (NIEM) and the top-level graph construction and clustering modules (TLGCM and CCM). We show how feature perturbations and gating approximation error control an upper bound on the CCM loss. A comparison with ungated inference requires additional assumptions on the shared encoding map, projection stability, and reference alignment errors. These results do not establish universal improvements in classification or optimization.

C.1 Problem Setup

Consider a brain functional graph $\mathcal{G} = (\mathbf{A}, \mathbf{X})$, where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the adjacency matrix and $\mathbf{X} \in \mathbb{R}^{N\times F}$ is the node feature matrix. We hold the graph and network parameters fixed throughout the analysis. Following a signal-plus-noise decomposition [20], we write

$$\mathbf{X} = \mathbf{X}^* + \mathbf{N}, \tag{59}$$

where $\mathbf{X}^*$ denotes a reference noise-free feature matrix and $\mathbf{N}$ denotes residual feature noise.

The bottom-level NIEM produces a gated feature matrix

$$\tilde{\mathbf{X}} = \mathbf{X} \odot \mathbf{M}, \qquad \mathbf{M} \in (0.5, 1)^{N\times F}. \tag{60}$$

Here $\odot$ denotes entrywise multiplication. The gate is generated by a Transformer encoder and a sigmoid-rescaled affine head:

$$\mathbf{M} = 0.5 + 0.5\,\sigma(\mathbf{H}\mathbf{W}_g + \mathbf{b}_g), \tag{61}$$

where the bias is broadcast over nodes and σ is applied entrywise.

The gated features are subsequently processed by two GCN layers with a residual connection. Letting $\mathbf{H}^{(1)} = \mathrm{ReLU}(\mathrm{BN}(\mathrm{GCN}_0(\tilde{\mathbf{X}}, \mathbf{A})))$, the node embeddings are

$$\mathbf{X}_{\mathrm{conv}} = \mathrm{ReLU}\big(\mathrm{BN}(\mathrm{GCN}_1(\mathbf{H}^{(1)}, \mathbf{A}))\big) + \mathbf{H}^{(1)}. \tag{62}$$

TLGCM generates the soft assignment matrix

$$\mathbf{S} = \mathrm{softmax}\left(\frac{\mathbf{X}_{\mathrm{conv}}\mathbf{W}_c + \mathbf{b}_c}{\tau}\right) \in \mathbb{R}^{N\times K}, \qquad \tau > 0. \tag{63}$$

All operations are considered in inference mode, with dropout disabled and BatchNorm statistics fixed.

We denote the downstream encoding pipeline, consisting of the GCN layers, residual connection, cluster projection, and softmax, by $\mathcal{F} : \mathbb{R}^{N\times F} \to \mathbb{R}^{N\times K}$. Thus $\mathbf{S} = \mathcal{F}(\tilde{\mathbf{X}})$. The gate is applied before $\mathcal{F}$ and is not included a second time in this map.

For a fixed symmetric normalized Laplacian $\mathbf{L}$, let $\mathbf{U}_k$ contain orthonormal eigenvectors corresponding to its k smallest eigenvalues. The CCM loss is

$$\mathcal{L}_{\mathrm{spec}}(\mathbf{S}) = \frac{1}{Nk}\|(\mathbf{I} - \mathbf{P}_S)\mathbf{U}_k\|_F^2, \qquad \mathbf{P}_S = \mathbf{S}(\mathbf{S}^\top\mathbf{S})^{-1}\mathbf{S}^\top. \tag{64}$$

As in Appendix A, the analysis uses the exact orthogonal projector and assumes full column rank. It does not directly cover changes to the graph Laplacian or replacement of this projector by a ridge-regularized approximation.

C.2 Assumptions

Assumption 1 (Bounded Noise)**.** *The noise matrix satisfies $\|\mathbf{N}\|_F \le \delta$ for some $\delta \ge 0$.*

Assumption 2 (Reference Assignment)**.** *Let $\mathbf{M}^* \in (0.5, 1)^{N\times F}$ be a reference gate, with reference features $\tilde{\mathbf{X}}^* = \mathbf{X}^* \odot \mathbf{M}^*$ and assignment $\mathbf{S}^* = \mathcal{F}(\tilde{\mathbf{X}}^*)$. Assume*

$$\mathcal{L}_{\mathrm{spec}}(\mathbf{S}^*) \le \epsilon_0, \qquad \|\mathbf{M} - \mathbf{M}^*\|_F \le \epsilon_m. \tag{65}$$

Define $\rho = \|\mathbf{M}^*\|_{\max} = \max_{i,f} |M^*_{if}| < 1$.

Remark 5. The reference error ϵ_0 is an assumption on the alignment induced by the reference gated features. A small value is not implied by the gate range, and the analysis does not prove that training drives either ϵ_0 or ϵ_m to zero. Allowing nonzero ϵ_0 makes the dependence on reference quality explicit.

Assumption 3 (Lipschitz Continuity of $\mathcal{F}$)**.** *The fixed downstream map is L_F-Lipschitz on a domain containing the compared feature matrices:*

$$\|\mathcal{F}(\mathbf{Z}_1) - \mathcal{F}(\mathbf{Z}_2)\|_F \le L_F \|\mathbf{Z}_1 - \mathbf{Z}_2\|_F. \tag{66}$$

For fixed graph propagation operators, finite network weights, positive temperature, and fixed inference-time BatchNorm statistics, the GCN layers, affine maps, ReLU, and softmax admit finite Lipschitz constants. Residual addition preserves this property, giving a finite constant for their composition.

Assumption 4 (Lipschitz Continuity of Projection)**.** *The compared assignments belong to a set $\mathcal{S}$ satisfying $\sigma_{\min}(\mathbf{S}) \ge \sigma_l > 0$ for every $\mathbf{S} \in \mathcal{S}$. In particular, $\mathbf{S}$ and $\mathbf{S}^*$ belong to this set. For a common constant C_P,*

$$\|\mathbf{P}_{S_1}\mathbf{U}_k - \mathbf{P}_{S_2}\mathbf{U}_k\|_F \le C_P \|\mathbf{S}_1 - \mathbf{S}_2\|_F, \qquad \mathbf{S}_1, \mathbf{S}_2 \in \mathcal{S}. \tag{67}$$

Appendix C.6 proves that $C_P = \sqrt{2}/\sigma_l$ is sufficient. Full column rank alone does not provide a uniform constant over matrices approaching rank deficiency.

C.3 Auxiliary Lemmas

Lemma 3 (Noise Propagation Bound)**.** *Under Assumptions 1–3, the actual assignment $\mathbf{S} = \mathcal{F}(\mathbf{X} \odot \mathbf{M})$ and reference assignment $\mathbf{S}^* = \mathcal{F}(\mathbf{X}^* \odot \mathbf{M}^*)$ satisfy*

$$\|\mathbf{S} - \mathbf{S}^*\|_F \le L_F\Big(\|\mathbf{X}^*\|_F \epsilon_m + \delta\rho + \delta\epsilon_m\Big). \tag{68}$$

Proof. Expanding the difference between the gated feature matrices gives

$$\begin{aligned}\tilde{\mathbf{X}} - \tilde{\mathbf{X}}^* &= (\mathbf{X}^* + \mathbf{N}) \odot \mathbf{M} - \mathbf{X}^* \odot \mathbf{M}^* \\ &= \mathbf{X}^* \odot (\mathbf{M} - \mathbf{M}^*) + \mathbf{N} \odot \mathbf{M}^* + \mathbf{N} \odot (\mathbf{M} - \mathbf{M}^*).\end{aligned} \tag{69}$$

The first and third terms can be bounded using $\|\mathbf{A} \odot \mathbf{B}\|_F \le \|\mathbf{A}\|_F \|\mathbf{B}\|_F$. For the second term, the sharper entrywise bound gives

$$\|\mathbf{N} \odot \mathbf{M}^*\|_F \le \|\mathbf{N}\|_F \|\mathbf{M}^*\|_{\max} \le \delta\rho. \tag{70}$$

Consequently,

$$\|\tilde{\mathbf{X}} - \tilde{\mathbf{X}}^*\|_F \le \|\mathbf{X}^*\|_F \epsilon_m + \delta\rho + \delta\epsilon_m. \tag{71}$$

Applying the Lipschitz property of $\mathcal{F}$ yields

$$\|\mathbf{S} - \mathbf{S}^*\|_F \le L_F \|\tilde{\mathbf{X}} - \tilde{\mathbf{X}}^*\|_F \le L_F\Big(\|\mathbf{X}^*\|_F \epsilon_m + \delta\rho + \delta\epsilon_m\Big). \tag{72}$$

The three terms correspond to gating error on the clean features, noise passed through the reference gate, and the interaction between noise and gating error. □

Lemma 4 (CCM Loss Sensitivity to Assignment Error)**.** *Under Assumptions 2 and 4, for $\mathbf{S}, \mathbf{S}^* \in \mathcal{S}$,*

$$\mathcal{L}_{\text{spec}}(\mathbf{S}) \le \frac{2C_P^2}{Nk} \|\mathbf{S} - \mathbf{S}^*\|_F^2 + 2\epsilon_0. \tag{73}$$

Proof. By the triangle inequality,

$$\|(\mathbf{I} - \mathbf{P}_S)\mathbf{U}_k\|_F \le \|(\mathbf{I} - \mathbf{P}_{S^*})\mathbf{U}_k\|_F + \|(\mathbf{P}_{S^*} - \mathbf{P}_S)\mathbf{U}_k\|_F. \tag{74}$$

Using the reference error and projection stability assumptions gives

$$\|(\mathbf{I} - \mathbf{P}_S)\mathbf{U}_k\|_F \le \sqrt{Nk\epsilon_0} + C_P \|\mathbf{S} - \mathbf{S}^*\|_F. \tag{75}$$

Squaring and applying $(a+b)^2 \le 2a^2 + 2b^2$ yields

$$\|(\mathbf{I} - \mathbf{P}_S)\mathbf{U}_k\|_F^2 \le 2Nk\epsilon_0 + 2C_P^2 \|\mathbf{S} - \mathbf{S}^*\|_F^2. \tag{76}$$

Dividing by Nk establishes the result. □

C.4 Main Theorem: Gating Controls the CCM Loss Upper Bound

Theorem 3 (CCM Loss Bound under Feature Gating)**.** *Under Assumptions 1–4, the actual assignment satisfies*

$$\mathcal{L}_{\text{spec}}(\mathbf{S}) \leq \frac{2C_P^2 L_F^2}{Nk}\Big(\|\mathbf{X}^*\|_F \epsilon_m + \delta\rho + \delta\epsilon_m\Big)^2 + 2\epsilon_0. \tag{77}$$

For fixed graph dimensions and uniformly bounded $\|\mathbf{X}^*\|_F$, L_F, and C_P, this upper bound tends to zero as $\epsilon_m, \delta, \epsilon_0 \to 0$. If the additional conditions of Theorem 2 hold, including $K = k$ and graph regularity, the continuous NCut objective then converges to its spectral minimum.

Proof. Combining Lemmas 3 and 4 gives

$$\begin{aligned}\mathcal{L}_{\text{spec}}(\mathbf{S}) &\leq \frac{2C_P^2}{Nk}\|\mathbf{S} - \mathbf{S}^*\|_F^2 + 2\epsilon_0 \\ &\leq \frac{2C_P^2 L_F^2}{Nk}\Big(\|\mathbf{X}^*\|_F \epsilon_m + \delta\rho + \delta\epsilon_m\Big)^2 + 2\epsilon_0.\end{aligned} \tag{78}$$

The limiting statement follows from uniform boundedness of the constants and nonnegativity of the loss. Under the additional regular-graph assumptions, Theorem 2 bounds the continuous NCut gap by a quantity proportional to $\sqrt{\mathcal{L}_{\text{spec}}}$, which also tends to zero. □

Remark 6 (Interpretation of the Bound)**.** The bound separates gating approximation error on the clean signal, $\|\mathbf{X}^*\|_F \epsilon_m$, reference-gated noise, $\delta\rho$, and their interaction, $\delta\epsilon_m$. The additive term $2\epsilon_0$ accounts for imperfect reference alignment. If only δ and ϵ_m vanish, the displayed upper bound approaches $2\epsilon_0$, not necessarily zero.

Remark 7 (Noise-Suppression Interpretation)**.** Because $\rho < 1$, the reference gate contracts the noise component in Frobenius norm. This fact alone does not establish an improved CCM loss: gating also changes the clean reference features and may alter their spectral alignment. A comparison of the resulting upper bounds therefore requires the additional conditions below.

C.5 Corollary: Gating as a Noise-Suppression Mechanism

Corollary 1 (Conditional Comparison with Ungated Inference)**.** *Consider gated and ungated inference through the same fixed downstream map $\mathcal{F}$ on the same graph. Define $\mathbf{S}_{\text{gate}} = \mathcal{F}(\mathbf{X} \odot \mathbf{M})$, $\mathbf{S}_{\text{no-gate}} = \mathcal{F}(\mathbf{X})$, and the clean ungated reference $\mathbf{S}_u^* = \mathcal{F}(\mathbf{X}^*)$. Let $\epsilon_u = \mathcal{L}_{\text{spec}}(\mathbf{S}_u^*)$. Assume common positive constants L_F and C_P apply to both comparisons, all four assignments belong to the projection-stable set, and $\epsilon_0 \leq \epsilon_u$. For $\delta > 0$, if*

$$\epsilon_m < \frac{\delta(1-\rho)}{\|\mathbf{X}^*\|_F + \delta}, \tag{79}$$

then the following valid upper bounds satisfy $\text{UB}_{\text{gate}} < \text{UB}_{\text{no-gate}}$:

$$\begin{aligned}\text{UB}_{\text{gate}} &= \frac{2C_P^2 L_F^2}{Nk}\Big(\|\mathbf{X}^*\|_F \epsilon_m + \delta\rho + \delta\epsilon_m\Big)^2 + 2\epsilon_0, \\ \text{UB}_{\text{no-gate}} &= \frac{2C_P^2 L_F^2}{Nk}\delta^2 + 2\epsilon_u.\end{aligned} \tag{80}$$

Proof. For ungated inference, the actual and reference feature matrices differ by $\mathbf{N}$, so

$$\|\mathbf{S}_{\text{no-gate}} - \mathbf{S}_u^*\|_F \leq L_F\|\mathbf{N}\|_F \leq L_F\delta. \tag{81}$$

Applying the argument of Lemma 4 with reference $\mathbf{S}_u^*$ gives

$$\mathcal{L}_{\text{spec}}(\mathbf{S}_{\text{no-gate}}) \leq \frac{2C_P^2 L_F^2}{Nk}\delta^2 + 2\epsilon_u. \tag{82}$$

For gated inference, Theorem 3 gives UB_{gate}. Condition equation 79 implies

$$(\|\mathbf{X}^*\|_F + \delta)\epsilon_m + \delta\rho < \delta. \tag{83}$$

All quantities are nonnegative, so the gated squared perturbation term is strictly smaller. Combining this with $\epsilon_0 \leq \epsilon_u$ and the common positive constants proves the stated comparison. □

Remark 8. Condition equation 79 requires the reduction in the noise bound, $\delta(1-\rho)$, to exceed the feature error introduced by imperfect gating. The reference-alignment condition additionally requires that the gated clean reference be no worse than the ungated clean reference. Neither condition follows solely from the gate range or from training.

Remark 9. The corollary compares specified upper bounds, not the actual gated and ungated losses. A smaller upper bound does not imply $\mathcal{L}_{\mathrm{spec}}(\mathbf{S}_{\mathrm{gate}}) < \mathcal{L}_{\mathrm{spec}}(\mathbf{S}_{\text{no-gate}})$. The result also does not directly compare independently trained ablation models, whose downstream maps, reference errors, and stability constants may differ.

C.6 Bound on the Lipschitz Constant C_P

We provide an explicit estimate for the projection Lipschitz constant in Assumption 4.

Proposition 1 (Explicit Bound on C_P)**.** *Let $\mathbf{S}_1, \mathbf{S}_2 \in \mathbb{R}^{N\times K}$ have minimum singular values at least $\sigma_l > 0$, and let $\mathbf{U}_k$ have orthonormal columns. Then*

$$\|(\mathbf{P}_{S_1} - \mathbf{P}_{S_2})\mathbf{U}_k\|_F \le \frac{\sqrt{2}}{\sigma_l}\|\mathbf{S}_1 - \mathbf{S}_2\|_F. \tag{84}$$

Thus $C_P = \sqrt{2}/\sigma_l$ is sufficient.

Proof. Write $\mathbf{P}_i = \mathbf{S}_i\mathbf{S}_i^\dagger$, where $\mathbf{S}_i^\dagger = (\mathbf{S}_i^\top\mathbf{S}_i)^{-1}\mathbf{S}_i^\top$ and $\|\mathbf{S}_i^\dagger\|_2 \le 1/\sigma_l$. Since $(\mathbf{I}-\mathbf{P}_2)\mathbf{S}_2 = 0$,

$$(\mathbf{I}-\mathbf{P}_2)\mathbf{P}_1 = (\mathbf{I}-\mathbf{P}_2)(\mathbf{S}_1 - \mathbf{S}_2)\mathbf{S}_1^\dagger. \tag{85}$$

Consequently,

$$\|(\mathbf{I}-\mathbf{P}_2)\mathbf{P}_1\|_F \le \frac{\|\mathbf{S}_1 - \mathbf{S}_2\|_F}{\sigma_l}. \tag{86}$$

Exchanging the two indices gives the same bound for $\|(\mathbf{I}-\mathbf{P}_1)\mathbf{P}_2\|_F$. For orthogonal projectors,

$$\|\mathbf{P}_1 - \mathbf{P}_2\|_F^2 = \|(\mathbf{I}-\mathbf{P}_2)\mathbf{P}_1\|_F^2 + \|(\mathbf{I}-\mathbf{P}_1)\mathbf{P}_2\|_F^2. \tag{87}$$

Both sides equal $\mathrm{tr}(\mathbf{P}_1) + \mathrm{tr}(\mathbf{P}_2) - 2\mathrm{tr}(\mathbf{P}_1\mathbf{P}_2)$. It follows that

$$\|\mathbf{P}_1 - \mathbf{P}_2\|_F \le \frac{\sqrt{2}}{\sigma_l}\|\mathbf{S}_1 - \mathbf{S}_2\|_F. \tag{88}$$

Finally, $\|\mathbf{U}_k\|_2 = 1$ implies

$$\|(\mathbf{P}_1 - \mathbf{P}_2)\mathbf{U}_k\|_F \le \|\mathbf{P}_1 - \mathbf{P}_2\|_F\|\mathbf{U}_k\|_2 \le \frac{\sqrt{2}}{\sigma_l}\|\mathbf{S}_1 - \mathbf{S}_2\|_F, \tag{89}$$

which proves the proposition. □

Remark 10. The orthogonality penalty encourages better-conditioned assignments but does not automatically guarantee a positive uniform lower bound on their singular values. A sufficient quantitative condition is $\|\mathbf{S}^\top\mathbf{S} - \mathbf{I}\|_2 \le a < 1$, which implies $\sigma_{\min}(\mathbf{S}) \ge \sqrt{1-a}$. Such a condition, or a direct singular-value bound, must hold throughout the compared set to justify a common C_P.

D Functional Connectivity Consistency Analysis of Top-Level Brain Graph Clustering

We evaluate the connectivity structure and inference-time stability of the ROI partitions learned by TLGCM using a fixed pretrained model and a fixed test subset from the OpenNeuro experiments. The processed dataset contains 546 graph samples. The checkpoint was obtained from the first training run with random seed 42, using a sample-level random split into 328 training, 109 validation, and 109 test graphs. We use the 109 test indices saved with this checkpoint throughout the analysis, without retraining. Each graph contains 182 ROIs with 182-dimensional input features, and the pooling layer uses $K = 9$ clusters. The evaluation unit is a graph sample, rather than necessarily a distinct participant.

D.1 Evaluation Protocol and Metric Definitions

Cluster Assignments and Connectivity Matrix. For each sample s, TLGCM produces a soft assignment matrix $\mathbf{S}^{(s)} \in \mathbb{R}^{N\times K}$, where K is the configured number of clusters. The hard label of ROI i is $c_i^{(s)} = \arg\max_k S_{ik}^{(s)}$. Connectivity metrics are computed from the original input feature matrix $\mathbf{X}^{(s)}$, rather than the GCN hidden representations or the sparse graph adjacency. Specifically, the implementation defines

$$A_{ij}^{(s)} = \begin{cases} \left|\mathrm{corr}(\mathbf{X}_{i,:}^{(s)}, \mathbf{X}_{j,:}^{(s)})\right|, & i \neq j, \\ 0, & i = j, \end{cases} \tag{90}$$

where corr is the Pearson correlation across feature dimensions; non-finite correlation values are replaced by zero. Thus, $\mathbf{A}^{(s)}$ measures unsigned similarity between ROI input feature profiles. If these profiles encode connectivity features, this quantity represents similarity between connectivity profiles and should not be interpreted as the original signed functional connectivity.

Within- and Between-Cluster Connectivity. Define the unordered ROI-pair sets $\mathcal{P}_{\mathrm{in}}^{(s)} = \{(i,j) : i < j,\ c_i^{(s)} = c_j^{(s)}\}$ and $\mathcal{P}_{\mathrm{out}}^{(s)} = \{(i,j) : i < j,\ c_i^{(s)} \neq c_j^{(s)}\}$. The sample-level connectivity means are

$$I_s = \frac{\sum_{(i,j)\in\mathcal{P}_{\mathrm{in}}^{(s)}} A_{ij}^{(s)}}{|\mathcal{P}_{\mathrm{in}}^{(s)}|}, \qquad O_s = \frac{\sum_{(i,j)\in\mathcal{P}_{\mathrm{out}}^{(s)}} A_{ij}^{(s)}}{|\mathcal{P}_{\mathrm{out}}^{(s)}|}. \tag{91}$$

These are averages over all eligible pairs within a sample, not equally weighted averages over clusters. Following the implementation, a mean is set to zero when its pair set is empty. Consequently, degenerate partitions, including single-cluster assignments, require caution when interpreting these summaries.

Connectivity Ratios and Low Inter-Cluster Fraction. With denominator floor $\tau = 10^{-3}$ and numerical constant $\epsilon = 10^{-10}$, the stabilized ratio and the diagnostic raw ratio are

$$R_s = \frac{I_s}{\max(O_s, \tau) + \epsilon}, \qquad R_s^{\mathrm{raw}} = \frac{I_s}{O_s + \epsilon}. \tag{92}$$

For M evaluation samples, we report $\bar{I} = M^{-1}\sum_s I_s$, $\bar{O} = M^{-1}\sum_s O_s$, and $\bar{R} = M^{-1}\sum_s R_s$. The low inter-cluster fraction is $f_{\mathrm{low}} = M^{-1}\sum_s \mathbb{I}[O_s < \tau]$, where $\mathbb{I}[\cdot]$ is the indicator function. This is a fraction of samples, not clusters or edges, and includes samples with no between-cluster pairs. The denominator floor limits numerical inflation but does not eliminate sensitivity to small O_s. In particular, $\bar{R}$ is generally different from $\bar{I}/\bar{O}$.

Weighted Modularity. Using the same nonnegative matrix $\mathbf{A}^{(s)}$, we compute Newman–Girvan modularity [20]:

$$Q_s = \frac{1}{2m_s}\sum_{i,j}\left(A_{ij}^{(s)} - \frac{d_i^{(s)} d_j^{(s)}}{2m_s}\right)\mathbb{I}[c_i^{(s)} = c_j^{(s)}], \quad d_i^{(s)} = \sum_j A_{ij}^{(s)}, \quad 2m_s = \sum_{i,j} A_{ij}^{(s)}. \tag{93}$$

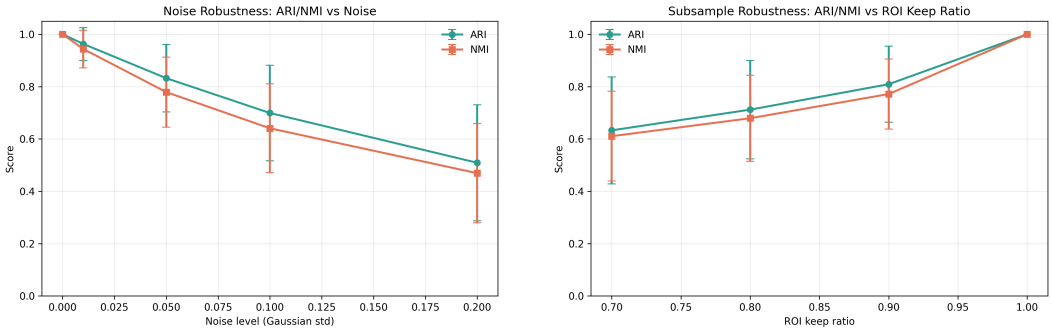
We set $Q_s = 0$ when $m_s = 0$ and report $\bar{Q} = M^{-1}\sum_s Q_s$. Modularity measures excess within-cluster weight relative to a degree-based reference. A value near zero indicates limited separation under this criterion; it is reported as a descriptive measure because the model is trained for classification rather than modularity maximization.

Inference-Time Perturbations. All assignment comparisons use single-graph inference in evaluation mode. For Gaussian perturbations, independent noise $\eta_{if} \sim \mathcal{N}(0, \sigma^2)$ is added to each input feature entry while graph edges and edge weights remain fixed. The tested standard deviations are $\sigma \in \{0, 0.01, 0.05, 0.10, 0.20\}$, expressed in input-feature units without additional variance normalization. For ROI subsampling, we uniformly retain $\max(2, \mathrm{round}(rN))$ nodes without replacement, with $r \in \{1.0, 0.9, 0.8, 0.7\}$, and construct the induced subgraph. Retained nodes preserve their original feature vectors and feature dimensions. Perturbed labels are compared with clean labels on exactly the same retained ROIs.

Partition Agreement and Aggregation. We use Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), both invariant to permutations of cluster labels. Let n_{ab} be the contingency

Table 4: Sample-averaged connectivity summaries on the fixed evaluation subset. Connectivity refers to absolute correlations between input feature profiles.

Metric	**Value**
Mean stabilized ratio $\bar{R}$	8.6740
Low inter-cluster fraction f_{low}	0.0183
Mean modularity $\bar{Q}$	0.0100
Mean within-cluster connectivity $\bar{I}$	0.6534
Mean between-cluster connectivity $\bar{O}$	0.6373



(a) Agreement under Gaussian feature noise.

(b) Agreement under ROI subsampling.

Figure 7: Inference-time partition agreement with the clean reference. Error bars show standard deviations across pooled sample–trial scores.

count between clean cluster a and perturbed cluster b, with marginals $u_a = \sum_b n_{ab}$, $v_b = \sum_a n_{ab}$, and $n = \sum_{a,b} n_{ab}$. Then

$$\text{ARI} = \frac{T - UV/B}{(U+V)/2 - UV/B}, \quad T = \sum_{a,b} \binom{n_{ab}}{2}, \quad U = \sum_a \binom{u_a}{2}, \quad V = \sum_b \binom{v_b}{2}, \quad B = \binom{n}{2}. \tag{94}$$

ARI equals one for identical partitions and is adjusted for chance agreement. Using empirical label probabilities, the implementation's arithmetic-normalized NMI is

$$\text{NMI}(C, C') = \frac{2\ \text{MI}(C, C')}{H(C) + H(C')}, \quad \text{MI}(C, C') = \sum_{a,b} p_{ab} \log \frac{p_{ab}}{p_a p'_b}, \quad H(C) = -\sum_a p_a \log p_a. \tag{95}$$

Here $p_{ab} = n_{ab}/n$, $p_a = u_a/n$, and $p'_b = v_b/n$, with zero-probability terms omitted. NMI ranges from zero to one; degenerate cases follow the scikit-learn conventions used in the code. Each perturbation level is evaluated over 10 random trials on the same 109 test graphs. The reported mean and population standard deviation are computed over the resulting 1,090 sample–trial agreement scores at each level. Because the same graphs are evaluated repeatedly, these scores are not independent observations. The standard deviations describe score dispersion rather than confidence intervals or variability across independently trained models.

D.2 Results and Interpretation

Table 4 shows slightly higher mean within-cluster than between-cluster connectivity, with an absolute difference of 0.0161. Their ratio is approximately 1.0253, whereas the mean of sample-level stabilized ratios is 8.6740. These statistics use different aggregation operations, and the latter can be strongly influenced by samples with small between-cluster connectivity. The low inter-cluster fraction of 0.0183 further motivates caution. Together with the near-zero mean modularity of 0.0100, these results provide limited evidence of strong community separation under the chosen connectivity measure.

Table 5: ARI and NMI under inference-time perturbations, reported as pooled mean ± standard deviation. The checkpoint and evaluation subset remain fixed.

Perturbation Level	**ARI**	**NMI**
Gaussian noise standard deviation σ		
0.000	1.0000 ± 0.0000	1.0000 ± 0.0000
0.010	0.9629 ± 0.0627	0.9435 ± 0.0722
0.050	0.8321 ± 0.1288	0.7789 ± 0.1341
0.100	0.6991 ± 0.1823	0.6406 ± 0.1695
0.200	0.5088 ± 0.2219	0.4690 ± 0.1895
ROI retention ratio r		
1.00	1.0000 ± 0.0000	1.0000 ± 0.0000
0.90	0.8091 ± 0.1454	0.7714 ± 0.1336
0.80	0.7119 ± 0.1881	0.6790 ± 0.1641
0.70	0.6328 ± 0.2044	0.6104 ± 0.1717

Figure 3: Cluster Assignment Heatmaps (Fixed Pretrained Model)

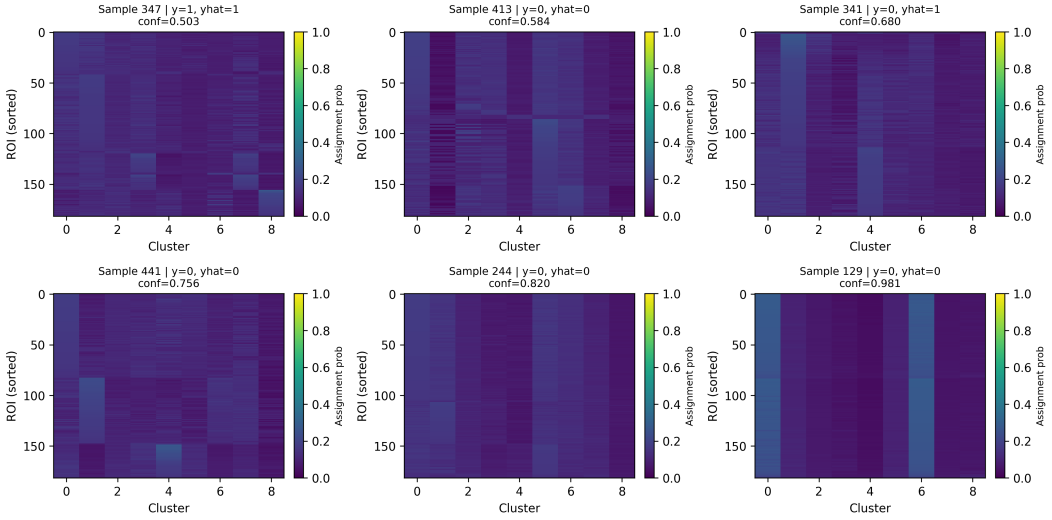


Figure 8: Example soft assignment matrices under the fixed pretrained model. Rows denote ROIs and columns denote learned clusters. These visualizations illustrate assignment patterns but do not independently establish biological validity or statistical significance.

Figure 7 and Table 5 show decreasing mean agreement as perturbations increase. At $\sigma = 0.01$, ARI and NMI remain high at 0.9629 and 0.9435, respectively; at $\sigma = 0.20$, they decrease to 0.5088 and 0.4690. Retaining 90% of ROIs yields ARI 0.8091 and NMI 0.7714, while retaining 70% yields 0.6328 and 0.6104. The unperturbed settings yield perfect agreement, as expected for repeated evaluation of the same input. These results support stability under mild perturbations while showing substantial sensitivity at stronger perturbation levels.

Figure 8 illustrates the soft assignments underlying the hard partitions, and Figure 9 complements the aggregate connectivity values with their sample-level distributions. Overall, the evidence supports a restricted conclusion: the fixed model produces ROI partitions that retain substantial agreement under mild inference-time perturbations, while connectivity separation is modest under the implemented metric. Without random-partition controls, independent training comparisons, or external biological validation, these analyses do not establish non-random community structure, reproducibility across training runs, or correspondence with canonical brain networks.

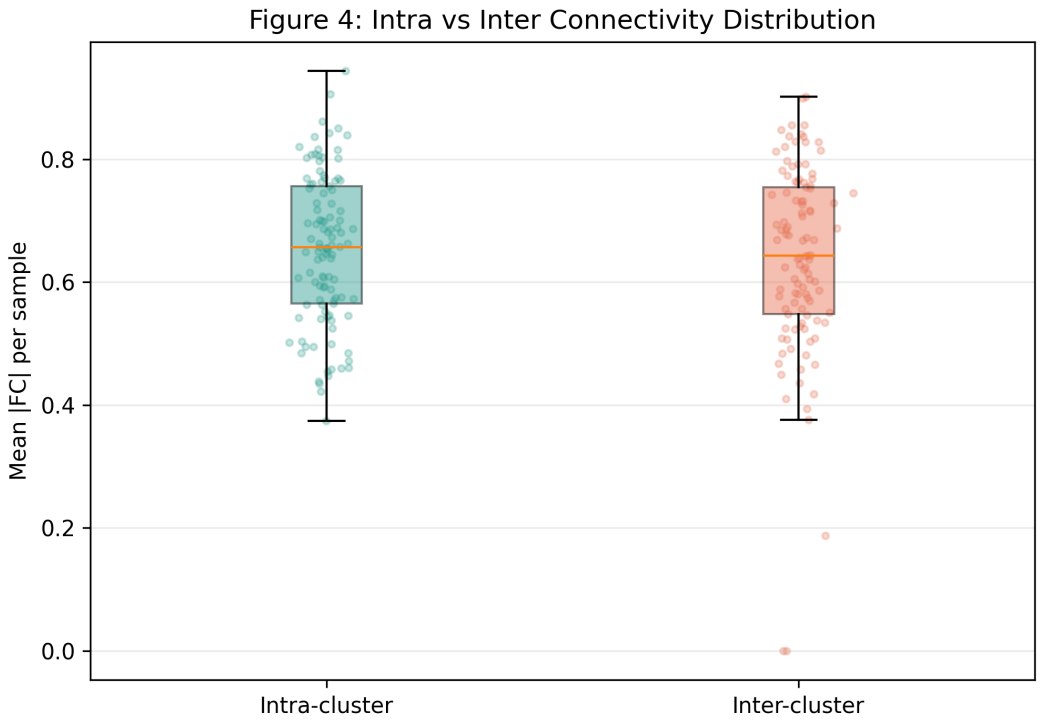


Figure 9: Distributions of sample-level within- and between-cluster connectivity, computed from absolute correlations between ROI input feature profiles. The comparison is descriptive; no significance test is reported.

E Complexity Analysis

E.1 Full Model Configuration

Table 6 summarizes the complete hyperparameter configuration of the full model (BL + TL), corresponding to the full model in the experimental codebase. The optimal number of clusters K is treated as an experimental variable and is validated separately (Table 3); all other hyperparameters are shared across datasets.

E.2 Time Complexity

We analyze the per-training-step complexity of the full model using the notation and configuration in Table 6. Let $N = Bn$ denote the total number of nodes in the batch (note: n and N here denote single-graph and batch-level node counts respectively, distinct from the single-graph context in the main text), E the total number of edges, and F the input feature dimension. Per-module costs are listed in Table 7.

Summing over all modules, the total time complexity of a single training step is

$$T = O\left(\underbrace{n_L N^2 H_L + n_L N H_L d_{\text{ff}}}_{\text{NIEM}} + \underbrace{BKn^2}_{\text{TLGCM}} + \underbrace{Bn^3}_{\text{CCM}} + NFd_h + Ed_h\right). \tag{96}$$

Dominant term. The self-attention in NIEM operates over all $N = Bn$ nodes in the batch as a single sequence, yielding an attention matrix of size $N \times N$ and a total cost of $O(n_L N^2 H_L) = O(n_L B^2 n^2 H_L)$; this dominates all other terms for typical values of B and n. The CCM spectral decomposition term $O(Bn^3)$, arising from computing the eigenvectors of the $n \times n$ graph Laplacian for each of the B graphs per training step, constitutes the secondary bottleneck. The TLGCM cluster coarsening term $O(BKn^2)$ is smaller by a factor of n/K and represents the tertiary cost.

Table 6: Complete hyperparameter configuration of the full model.

Component	Hyperparameter	Symbol	Value
NIEM	Transformer hidden dimension	H_L	96
	Transformer encoder layers	n_L	1
	Attention heads	n_h	2
	Dropout rate	p_T	0.5
GCN encoder (BL)	Hidden dimension	d_h	64
	Dropout rate	p_G	0.65
	Number of layers	n_{gcn}	2
TLGCM	Number of clusters	K	see Table 3
	Softmax temperature	τ	1.0
Loss weights	Classification loss	w_{cls}	1.0
	Clustering regularization	w_{cl}	0.008
	Spectral alignment loss	w_{spec}	0.001
	Orthogonality loss	w_{orth}	0.015
	Balance loss	w_{bal}	0.002
Optimizer	Optimizer	—	AdamW
	Learning rate	η_{lr}	8×10^{-4}
	Weight decay	w_d	2×10^{-5}
	Learning rate scheduler	—	StepLR
	Step size / decay factor	T_s / γ	20 / 0.5
Training	Batch size	B	32
	Training epochs	T_{ep}	50
	Edge dropout (augmentation)	p_e	0.2
	Feature masking (augmentation)	p_m	0.2
Evaluation	Independent runs	R	10

E.3 Space Complexity

Per-module activation memory costs are listed in Table 8.

The total activation memory required is

$$M = O\left(\underbrace{N^2}_{\text{attention map (NIEM)}} + \underbrace{Bn^2}_{\text{dense adj. (TLGCM)}} + Nd_h \right). \tag{97}$$

Dominant term. The global attention weight matrix stored by NIEM contributes $O(N^2) = O(B^2n^2)$ in total, constituting the primary memory bottleneck and scaling quadratically with both batch size and the number of ROIs. The dense adjacency tensor in TLGCM contributes $O(Bn^2)$, which is smaller by a factor of B, representing the secondary bottleneck. All remaining tensors are at most $O(Nd_h)$, which is negligible in comparison.

Optimization note. The dominant memory term $O(N^2) = O(B^2n^2)$ arises from NIEM performing global self-attention over the entire batch as a single sequence. Partitioning the sequence per graph (as originally intended) would reduce the attention memory from $O(B^2n^2)$ to $O(Bn^2)$, matching the TLGCM bottleneck; adopting sparse or linear attention would further reduce it to $O(N)$, significantly improving scalability for brain graphs with large numbers of ROIs.

Table 7: Per-training-step time complexity of the full model.

#	Module	Operation	Time
1	NIEM	Linear proj $F \to H_L$	$O(NFH_L)$
		Self-attention (global, seq. len. N)	$\mathbf{O(n_L N^2 H_L)}$
		FFN (d_{ff}=2048)	$O(n_L N H_L d_{\text{ff}})$
		Gate head $H_L \to F$ + scaling	$O(NH_L F)$
2	GCN Layer 0 (bottom-level)	Linear $F \to d_h$ + message passing	$O(NFd_h + Ed_h)$
		BN + ReLU + Dropout	$O(Nd_h)$
3	GCN Layer 1 (bottom-level, residual)	Linear $d_h \to d_h$ + message passing	$O(Nd_h^2 + Ed_h)$
		BN + ReLU + residual add	$O(Nd_h)$
4	TLGCM	Cluster projection: Linear $d_h \to K$	$O(Nd_hK)$
		Dense adjacency materialization $\mathbf{A} \in \mathbb{R}^{B\times n\times n}$	$O(E + Bn^2)$
		$\mathbf{S}^\top \mathbf{X}_{\text{conv}}$ (node aggregation)	$O(BKnd_h)$
		$\mathbf{S}^\top \mathbf{A}\mathbf{S}$ (adj. coarsening)	$O(BKn^2)$
5	GCN Layer (cluster-level, residual)	Linear $d_h \to d_h$ + msg. passing (N'=BK)	$O(BKd_h^2 + BK^2 d_h)$
		BN + ReLU + residual add	$O(BKd_h)$
6	Readout + MLP	Global mean pooling	$O(BKd_h)$
		FC1: $d_h \to 32$ + BN	$O(32Bd_h)$
		FC2: $32 \to N_c$	$O(32BN_c)$
7	CCM (training only)	Laplacian construction per graph	$O(Bn^2)$
		Spectral decomposition $\mathbf{L} \in \mathbb{R}^{n\times n}$ per graph	$\mathbf{O(Bn^3)}$

Table 8: Per-module space complexity (activations).

#	Module	Tensor / Activation	Space
1	NIEM	Projected sequence $[1, N, H_L]$	$O(NH_L)$
		Attention weight matrix $[1, N, N]$	$\mathbf{O(N^2)}$
		Gate output $[N, F]$	$O(NF)$
2	GCN Layer 0 (bottom-level)	Node embeddings $[N, d_h]$	$O(Nd_h)$
3	GCN Layer 1 (bottom-level)	Node embeddings + residual	$O(Nd_h)$
4	TLGCM	Assignment matrix $\mathbf{S} \in \mathbb{R}^{N\times K}$	$O(NK)$
		Dense adjacency $[B, n, n]$	$\mathbf{O(Bn^2)}$
		Coarsened adj. $[B, K, K]$	$O(BK^2)$
5	GCN Layer (cluster-level)	Pooled embeddings $[BK, d_h]$	$O(BKd_h)$
6	Readout + MLP	Graph embeddings $[B, d_h]$	$O(Bd_h)$

E.4 Compute Resources

All experiments were conducted on a single NVIDIA RTX 6000 Ada Generation GPU (48 GB VRAM) with dual Intel Xeon Gold 6226R CPUs (2×16 cores @ 2.90 GHz, 64 threads total) and 251 GB system RAM. No distributed or cloud computing was used.

Per-run and total compute for main comparison experiments. Each experimental run consists of 50 training epochs with batch size $B = 32$, repeated for $R = 10$ independent runs per dataset. Table 9 reports the estimated floating-point operations (FLOPs), wall-clock time per run, and total GPU-hours for each dataset.

Table 9: Compute statistics for the main comparison experiments (full model, 10 runs $\times$ 50 epochs each).

Dataset	FLOPs/run (TFLOPs)	Time/run (s)	Total FLOPs (TFLOPs)	Total (GPU-h)
ADHD-200	17.97 ± 0.004	143.91 ± 3.33	179.74	0.400
OpenNeuro-ds002424	18.82 ± 0.566	104.17 ± 1.28	188.17	0.289
OpenNeuro-ds000030	7.363 ± 0.000	32.84 ± 0.23	73.63	0.091
Subtotal	—	—	441.54	0.780

Ablation studies. The ablation study evaluates 4 model configurations across 3 datasets with 10 independent runs each (120 runs in total). Using per-dataset run times from Table 9, the estimated total wall-clock time is approximately $4 \times (10 \times 143.91 + 10 \times 104.17 + 10 \times 32.84) \approx 11{,}243$ s ≈ 3.12 GPU-hours, corresponding to an estimated $4 \times 441.54 \approx 1{,}766$ TFLOPs.

Hyperparameter sensitivity analysis. The sensitivity analysis over $K \in \{6, \ldots, 10\}$ spans 5 configurations $\times$ 3 datasets $\times$ 10 runs (150 runs in total), yielding an estimated $5 \times (10 \times 143.91 + 10 \times 104.17 + 10 \times 32.84) \approx 14{,}046$ s ≈ 3.90 GPU-hours and $5 \times 441.54 \approx 2{,}208$ TFLOPs.

Total reported compute. Summing all three experiment types, the total reported compute is approximately $0.780 + 3.12 + 3.90 \approx 7.80$ GPU-hours, corresponding to $441.54 + 1{,}766 + 2{,}208 \approx 4{,}416$ TFLOPs (≈ 4.42 PFLOPs) on a single NVIDIA RTX 6000 Ada Generation GPU. In addition, preliminary hyperparameter exploration and architecture search conducted prior to the final reported runs consumed an estimated $\sim$8 GPU-hours of additional compute, bringing the total research compute to approximately $\sim$16 GPU-hours.

F LIMITATIONS

While the proposed bottom-up hierarchical brain graph network achieves state-of-the-art performance on multiple public benchmarks, we acknowledge several limitations that are common in the field of neuroimaging-based disease detection and do not detract from the core contributions of this work.

Dataset scale and diversity. The experiments are conducted on three publicly available datasets (ADHD-200, OpenNeuro-ds002424, and OpenNeuro-ds000030). Although these datasets represent standard benchmarks in the community, their sample sizes are modest by deep learning standards. This limitation is not unique to our method; it reflects the broader challenge of collecting large-scale, clinically labeled neuroimaging data. Our model achieves the highest mean Accuracy and F1-Score among the compared methods on all three datasets, but the reported one-sample tests do not establish statistically significant improvements over those methods or demonstrate data efficiency. We anticipate that access to larger multi-site cohorts would further improve performance and generalizability, and such validation is a natural extension of this work.

Generalization to other neuropsychiatric disorders. The current study focuses on ADHD and schizophrenia, which exhibit well-characterized functional connectivity alterations. Whether the proposed hierarchical brain graph architecture generalizes to other disorders—such as autism spectrum disorder (ASD), major depressive disorder, or Alzheimer's disease—remains an open empirical question. However, the core mechanisms (NIEM, TLGCM and CCM) are designed to capture generic properties of brain functional organization, and we see no principled barrier to their application across disorders. Future work will systematically evaluate transferability across the full spectrum of neuropsychiatric conditions.

Computational considerations. As analyzed in Appendix E, the NIEM introduces $O(n_L B^2 n^2 H_L)$ attention complexity in the forward pass, which can become a bottleneck for very large batch sizes or high-resolution brain parcellations. However, for the typical scale of brain network analysis ($n \leq 200$, $B = 32$), this is well within the capacity of modern GPUs (NVIDIA RTX 6000 Ada Generation used in our experiments). NIEM also operates during inference; only training-specific computations, including backpropagation and the CCM loss, can be omitted. For scenarios requiring extreme scalability, the architecture can be readily combined with sparse attention or linear Transformer variants without altering the core methodological contribution.

In summary, these limitations reflect the current state of the field rather than specific weaknesses of our approach. They do not invalidate the theoretical guarantees or empirical advantages of the proposed hierarchical brain graph framework, and each points to a clear, actionable direction for future research. We believe that open discussion of these issues strengthens, rather than undermines, the contribution of this work.

G Broader Impacts

Positive Societal Impacts. This work proposes a bottom-up hierarchical brain graph representation learning framework for brain functional connectivity analysis, with direct application to brain disease detection. Improved diagnostic accuracy and neurobiological interpretability of the learned brain graph representations may assist clinicians in identifying disorder-relevant brain patterns, potentially reducing diagnostic delays and supporting more targeted interventions for affected individuals. The biologically grounded clustering mechanism of the top-level brain graph (TLGCM + CCM) also offers a data-driven tool for neuroscientists to explore functional brain organization in neuropsychiatric disorders, contributing to a broader scientific understanding of conditions such as ADHD and schizophrenia.

Negative Societal Impacts and Limitations. As a foundational research contribution, this work is not directly deployed in clinical settings, and we do not foresee immediate negative societal impacts. However, several risks warrant consideration. First, if such models were deployed in diagnostic pipelines without sufficient clinical validation, incorrect predictions could lead to misdiagnosis, potentially harming patients through inappropriate treatment or stigmatization. Second, neuroimaging data is inherently sensitive; any downstream application must ensure strict data privacy protections and compliance with medical data regulations (e.g., HIPAA, GDPR). Third, as the model is trained on existing neuroimaging datasets, it may inherit demographic biases present in those datasets (e.g., imbalanced age, sex, or ethnicity distributions), which could reduce its generalizability across diverse populations and raise fairness concerns if applied unequally across groups.

Mitigation Strategies. We recommend that any clinical application of this or related methods undergo rigorous prospective validation on diverse, representative cohorts before deployment. Model predictions should be treated as decision-support tools rather than standalone diagnoses. Future work should incorporate fairness auditing across demographic subgroups and adopt privacy-preserving training paradigms to enable model development without centralizing sensitive patient data.