ARE FOUNDATION MODELS ENOUGH FOR DATA-SCARCE FUNCTIONAL CONNECTIVITY DIAGNOSIS? REVISITING TASK-SPECIFIC LEARNING WITH BRAINLRR

Anonymous authors
Paper under double-blind review

ABSTRACT

Functional connectivity (FC) derived from Resting-state fMRI captures disruptions in brain-network organization, making FC-based classification a promising non-invasive approach to disease diagnosis. However, FC classification is constrained by small, heterogeneous, and noisy datasets. Whether foundation models (FMs) can transfer effectively under such label scarcity remains unclear. We benchmark general-purpose LLMs, vision backbones, and neuroimaging FMs under zero-shot, few-shot, and target-dataset fine-tuning protocols. Learned alignment improves LLM-based classification from a mean AUROC of 53.3 to 70.8, but encoder-only models using the same learned FC representation achieve comparable performance, suggesting pretrained language representations do not provide a consistent advantage. We therefore introduce **Brain L**ow-**R**ank **R**egularization (BrainLRR), a compact task-specific model combining region-of-interest (ROI)-profile masking with low-rank regularization of subject representations. Across four classification tasks spanning Autism spectrum disorder (ASD), Parkinson's disease (PD), and Alzheimer's disease (AD), BrainLRR demonstrates favorable sample efficiency and outperforms the evaluated FM baselines across the examined data-scarcity settings. The design components of BrainLRR also generalize to other architectures, improving BrainNetCNN across all four tasks. These results show that, in data-scarce FC classification, FM utility depends on modality alignment and pretraining domain, while carefully designed task-specific models can provide an effective alternative. **Code:** `anon-research-code/BrainLRR`

1 INTRODUCTION

Resting-state functional magnetic resonance imaging (rs-fMRI) provides a non-invasive way to study spontaneous neural activity and large-scale brain network organization. Functional connectivity (FC) estimated from rs-fMRI has become a widely used representation for studying brain function and dysfunction (Biswal et al., 1995; Bullmore & Sporns, 2009; Sporns, 2013). Since many neurological and psychiatric disorders involve disrupted interactions among distributed brain systems rather than isolated regional abnormalities, FC-based classification has attracted increasing attention for disease characterization and biomarker discovery (Khosla et al., 2019; Tessitore et al., 2019; Greicius et al., 2004).

Reliable FC-based diagnosis remains challenging because rs-fMRI datasets are often small, heterogeneous, and noisy, while FC representations are high-dimensional. This imbalance increases the risk of overfitting and site- or scanner-related confounding, particularly under label scarcity. Effective FC classification therefore requires structure-preserving representations and regularization tailored to small-sample learning.

Foundation models (FMs) offer a promising but underexplored approach to data-scarce FC classification. Although large-scale pretraining supports transfer in language, vision, and neuroimaging (Bommasani et al., 2021; He et al., 2016; Dosovitskiy et al., 2020; Ortega Caro et al., 2024; Wang et al., 2026; Cox et al., 2024), existing pretraining inputs are poorly aligned with ROI-level

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.

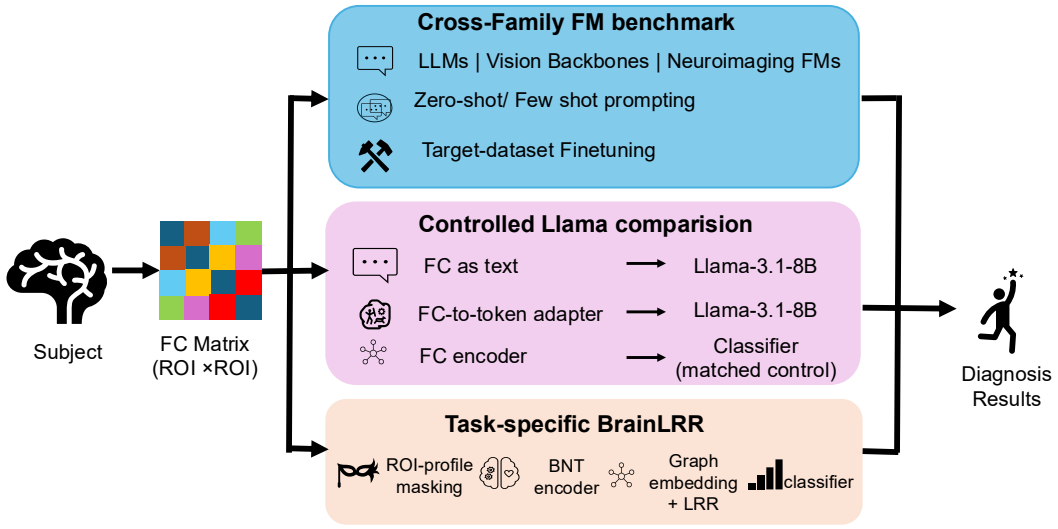


Figure 1: FM adaptation, controlled comparisons, and BrainLRR for data-scarce FC classification.

FC matrices. Thus, it remains unclear whether transfer gains arise from pretrained representations or effective FC alignment. This motivates our central question: *Are foundation models enough for data-scarce functional connectivity diagnosis?*

To answer this question, we evaluate three complementary routes from a common FC representation, as illustrated in 1: (i) a cross-family FM benchmark under zero-shot, few-shot, and target-dataset fine-tuning protocols; (ii) a controlled comparison among FC-as-text, a learned FC-to-token adapter for Llama-3.1-8B, and a matched encoder-only classifier; and (iii) the task-specific BrainLRR we further evaluate selected models using 25/50/75/100% of the ABIDE (Di Martino et al., 2014) and the Alzheimer's Disease Neuroimaging Initiative (ADNI) normal control versus mild cognitive impairment (NC–MCI) task.

Results reveal more than a simple domain gap: they expose important limitations of the evaluated FMs for data-scarce FC diagnosis. General-purpose FMs often produce unstable or single-class predictions under zero- and few-shot prompting, while fine-tuned neuroimaging FMs remain dataset-dependent and exhibit sensitivity–specificity imbalances. Learned FC alignment raises Llama's four-task mean AUROC from 53.3 to 70.8, but the matched encoder-only control also achieves 70.8, showing no consistent benefit from pretrained language representations beyond the learned FC encoder.

Motivated by these observations, we propose **Brain L**ow-**R**ank **R**egularization (BrainLRR), a compact task-specific architecture for data-scarce FC classification. Built upon the brain-network transformer backbone (Kan et al., 2022b), BrainLRR combines two parameter-free mechanisms: ROI-profile masking, which zeros selected FC rows during training to discourage reliance on individual connectivity profiles, and low-rank regularization, which promotes intra-class compactness and inter-class separation among graph-level embeddings.

We evaluate BrainLRR across four classification tasks spanning Autism spectrum disorder (ASD), Parkinson's disease (PD), and Alzheimer's disease (AD). BrainLRR shows strong overall and sample-efficient performance against task-specific models and fine-tuned general-purpose, vision, and neuroimaging FMs. Ablation and sensitivity analyses assess its components, while ROI attribution identifies regions that most influence its predictions.

In summary, our contributions are:

- We benchmark language, vision, and neuroimaging FMs under zero-shot, few-shot, and target-dataset fine-tuning, including a controlled Llama comparison that separates FC alignment from pretrained language representations.

- We introduce BrainLRR, a compact FC classifier combining ROI-profile masking and low-rank regularization for limited-data learning.
- We evaluate BrainLRR across four tasks on three brain-disorder datasets and analyze its components, sample efficiency, and ROI-level attributions.

2 Related Work

Functional connectivity classification. FC-based disorder classification typically represents each subject by an ROI-by-ROI connectivity matrix and predicts diagnostic status from inter-regional network patterns. Early approaches often vectorized FC matrices and applied conventional machine learning classifiers, while later deep learning methods introduced architectures tailored to brain connectivity (Khosla et al., 2019). BrainNetCNN models connectivity matrices using specialized edge-to-edge, edge-to-node, and node-to-graph operations (Kawahara et al., 2017). More recently, the Brain Network Transformer (BNT) uses attention to model relationships among brain regions and has become a strong task-specific baseline for FC classification (Kan et al., 2022b). These methods show that FC data benefit from architectures that respect brain-network structure. However, their performance can still be limited by small sample sizes, high-dimensional connectivity features, and overfitting to dataset-specific ROI patterns.

Foundation models for medical and neuroimaging AI. Foundation models have achieved strong transfer performance in language and vision through large-scale pretraining (Bommasani et al., 2021; He et al., 2016; Dosovitskiy et al., 2020), motivating their use in medical AI. In neuroimaging FMs, BrainLM pretrains on ROI-level fMRI time series using self-supervised masked prediction to model brain activity dynamics (Ortega Caro et al., 2024). BrainSegFounder studies 3D foundation modeling for neuroimage segmentation (Cox et al., 2024). These models motivate brain-domain pretraining, but their transferability to ROI-level FC classification remains unknown. In particular, their input representations and pretraining objectives are not necessarily aligned with ROI-level FC matrices. General-purpose LLMs and vision backbones may face a larger modality gap because FC matrices differ substantially from the text and natural images used during pretraining. Our work evaluates these FMs under zero-shot, few-shot, and fine-tuning protocols and uses a matched Llama comparison to disentangle FC representation alignment from pretrained language representations.

Regularization in data-scarce neuroimaging. Regularization is critical for FC classification because neuroimaging datasets are often small relative to the dimensionality of connectivity features. Dropout-style perturbations and graph masking strategies can improve robustness by preventing models from relying on a small subset of features or nodes (You et al., 2020; Hou et al., 2022). Recent work has used topology-guided masked autoencoding to refine population graphs and learn individual representations from limited neuroimaging data (Li et al., 2025). Unlike population-graph masked reconstruction, BrainLRR applies ROI-profile masking and low-rank regularization to subject-level FC representations.

3 Methodology

3.1 Problem formulation

Each subject is represented by an ROI-level functional connectivity (FC) matrix $X \in \mathbb{R}^{N \times N}$, where X_{ij} is the Pearson correlation between the BOLD time series of ROIs i and j. For binary classification, we learn $f_\theta(X)$ to predict $y \in \{0, 1\}$.

3.2 BrainLRR overview

BrainLRR comprises ROI-profile masking, a transformer encoder, and cluster pooling (Figure 2). Each row $x_i \in \mathbb{R}^N$ represents the connectivity profile of ROI i and is randomly masked during training to reduce reliance on individual ROIs. The transformer models global dependencies via multi-head self-attention (Vaswani et al., 2017), and cluster pooling aggregates ROI representations into a graph-level embedding for classification. Low-rank regularization is applied to this embedding during training (Liu et al., 2012).

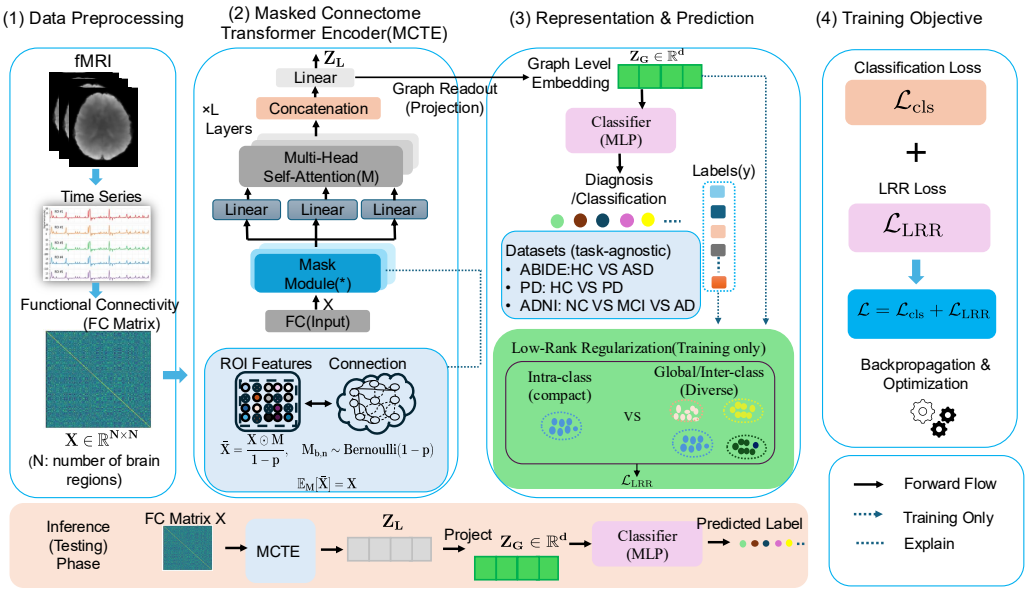


Figure 2: Overview of BrainLRR. Given an ROI-level FC matrix, BrainLRR applies ROI-profile masking to reduce overfitting, followed by a transformer encoder to learn attention-enhanced ROI representations. Cluster pooling then produces a compact graph-level embedding for classification. During training, low-rank regularization encourages compact within-class and discriminative cross-class representations.

3.3 Stochastic ROI-profile masking

Standard dropout perturbs individual features independently (Srivastava et al., 2014), but ignores the anatomical meaning of ROI nodes. In contrast, BrainLRR applies structured ROI-profile masking at the ROI level. This design is related to structured perturbation strategies in graph learning, where removing edges or masking graph components can improve robustness and generalization (Hou et al., 2022). For each subject, we sample a binary mask $m \in \{0,1\}^N$, where each ROI is retained with probability $1-p$:

$$m_i \sim \text{Bernoulli}(1-p), \tag{1}$$

where p is the node masking probability. Since each row of X is treated as the connectivity profile of an ROI, the masked input is defined as

$$\tilde{X} = \frac{m \odot X}{1-p}, \tag{2}$$

where the mask is broadcast along the feature dimension. The rescaling preserves the expected activation magnitude during training. By zeroing complete ROI profiles, it discourages over-reliance on individual connectivity profiles.

3.4 Transformer encoder

After ROI-profile masking, the connectivity matrix is processed by stacked multi-head self-attention layers to model global inter-regional dependencies. Given the masked input $\tilde{X} \in \mathbb{R}^{N\times N}$, we compute attention as

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{3}$$

where $Q = \tilde{X}W_Q$, $K = \tilde{X}W_K$, and $V = \tilde{X}W_V$ are learned projections of Queries, Keys, and Values, and d is the feature dimension. We adopt an L-layer multi-head self-attention (MHSA) module to learn expressive node representations. For each layer l,

$$Z^l = Concat\left(h^{l,1}, ..., h^{l,M}\right) W_O^l, \tag{4}$$

where $h^{l,m} = Attention\left(W_Q^{l,m}, W_K^{l,m}, W_V^{l,m}\right)$, M denotes the number of heads and $W_Q^{l,m}, W_K^{l,m}, W_V^{l,m}, W_O^l$ are learnable parameters. Stacking L layers yields the final node embeddings $Z_L \in \mathbb{R}^{N\times N}$.

3.5 CLUSTER POOLING

Directly flattening node embeddings may ignore the modular organization of brain networks and lead to overfitting under limited data. To obtain a compact and biologically meaningful graph-level representation, we employ a structured readout mechanism. Instead of using simple global pooling, we adopt the *Orthonormal Clustering Readout*(OCREAD) (Kan et al., 2022b).

Given the attention-enhanced node embeddings $Z_L \in \mathbb{R}^{N\times N}$, OCREAD performs differentiable clustering by softly assigning nodes to a set of learnable cluster centers $E \in \mathbb{R}^{C\times N}$, where $C < N$. Formally,

$$S = \mathrm{Softmax}(Z_L E^\top), \quad Z_G = S^\top Z_L, \tag{5}$$

where $S \in \mathbb{R}^{N\times C}$ denotes the soft assignment matrix and $Z_G \in \mathbb{R}^{C\times N}$ is the clustered graph representation. The clustering mechanism is inspired by deep embedded clustering (Xie et al., 2016) and enforces orthonormal constraints to improve cluster separability and stability. This design aligns with the modular organization of brain networks (Sporns & Betzel, 2016), where functionally related ROIs form subnetworks.

3.6 LOW-RANK REGULARIZATION

To encourage compact and discriminative representations under limited data, BrainLRR applies low-rank regularization to the graph-level embeddings. For a mini-batch of embeddings $H = [h_1, \ldots, h_B]^\top$ with labels y_i, let H_c denote the subset of embeddings belonging to class c. We define the low-rank regularization loss as

$$\mathcal{L}_{\mathrm{LRR}} = \frac{1}{B}\left(\sum_{c\in\mathcal{C}_B} \|H_c\|_* - \|H\|_*\right), \tag{6}$$

where $\mathcal{C}_B$ contains the classes present in the mini-batch. The class-specific terms are not separately normalized by class size; when a batch contains only one class, the two terms cancel. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{LRR}}\mathcal{L}_{\mathrm{LRR}}, \tag{7}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the cross-entropy loss and λ_{LRR} controls the contribution of the low-rank regularizer. For imbalanced datasets, cross-entropy uses class weights computed from the training labels, and mini-batches are otherwise sampled randomly rather than class-balanced.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Evaluation Metrics. We evaluate BrainLRR on three brain disorder datasets across four binary classification tasks: ABIDE for ASD vs. healthy control (HC), a pooled PD cohort for PD vs. HC, and ADNI for normal control (NC) vs. mild cognitive impairment (MCI) and NC vs. Alzheimer's Disease(AD). BrainLRR and static-FC baselines use CC200 (Craddock et al., 2012) atlas. Cohort statistics and preprocessing are provided in supplementary Section A.

Experimental Setting. All methods use the same 70%/10%/20% train/validation/test split (Kan et al., 2022b; Bannadabhavi et al., 2023; Peng et al., 2024; Chen et al., 2025), stratified by site for ABIDE, site×label for PD, and label for ADNI. For models trained in our study, we report mean $\pm$ standard deviation across five optimization seeds on each fixed subject partition. BrainLRR uses fixed hyperparameters across datasets. BrainLRR training details and FM hyperparameter-selection procedures are provided in Supplementary Section B.

Baselines and Protocols. We compare BrainLRR with foundation-model and task-specific FC baselines. For the ABIDE FM benchmark, we evaluate general-purpose LLMs, vision backbones, and

Table 1: Target-dataset comparison across four FC classification tasks. TS denotes ROI-level fMRI time-series inputs.

Model	Setting	Trainable/Total Params	AUROC	Acc	Sens	Spec
ABIDE: ASD versus HC						
Llama-3.1-8B	FC-as-text	41.9M / 8.1B	53.0 ± 4.5	52.7 ± 3.2	46.0 ± 13.6	59.8 ± 14.4
Llama-3.1-8B	FC adapter	44.4M / 8.1B	63.7 ± 1.2	58.9 ± 3.0	54.6 ± 17.5	63.5 ± 16.6
FC encoder	classifier only	2.5M / 2.5M	60.2 ± 1.7	56.7 ± 3.0	70.6 ± 11.1	42.0 ± 14.0
BrainLM	fine-tuned (FC)	86.2M / 86.2M	55.3 ± 4.9	54.9 ± 3.7	58.5 ± 14.4	51.0 ± 16.7
BrainLM	fine-tuned (TS)	86.2M / 86.2M	63.4 ± 3.6	59.5 ± 3.7	$\mathbf{72.9 \pm 21.7}$	45.3 ± 17.2
BrainSegFounder	fine-tuned	19.1M / 19.1M	66.7 ± 1.9	62.0 ± 1.8	68.3 ± 9.2	55.3 ± 11.1
BrainLRR (Ours)	task-specific	28.0M / 28.0M	$\mathbf{79.1 \pm 1.8}$	$\mathbf{72.1 \pm 2.4}$	71.7 ± 3.3	$\mathbf{72.5 \pm 2.6}$
PPMI: PD versus HC						
Llama-3.1-8B	FC-as-text	41.9M / 8.1B	58.3 ± 11.6	56.2 ± 5.9	61.8 ± 16.3	48.0 ± 25.6
Llama-3.1-8B	FC adapter	44.4M / 8.1B	62.9 ± 7.7	58.9 ± 2.3	63.6 ± 13.3	52.0 ± 20.2
FC encoder	classifier only	2.5M / 2.5M	$\mathbf{71.0 \pm 3.8}$	64.3 ± 7.0	63.6 ± 17.3	$\mathbf{65.3 \pm 12.8}$
BrainLM	fine-tuned (FC)	86.2M / 86.2M	66.4 ± 1.6	61.1 ± 1.5	81.8 ± 5.6	30.7 ± 8.9
BrainLM	fine-tuned (TS–CC200)	86.2M / 86.2M	57.5 ± 2.0	$\mathbf{66.5 \pm 4.5}$	$\mathbf{81.8 \pm 4.6}$	44.0 ± 10.1
BrainLM	fine-tuned (TS–A424)	86.2M / 86.2M	56.2 ± 3.0	57.8 ± 4.9	69.1 ± 17.7	41.3 ± 15.2
BrainSegFounder	fine-tuned	19.1M / 19.1M	67.3 ± 1.2	63.2 ± 2.4	72.7 ± 5.6	49.3 ± 13.8
BrainLRR (Ours)	task-specific	28.0M / 28.0M	68.0 ± 2.3	63.8 ± 3.1	71.8 ± 2.0	52.0 ± 8.7
ADNI: NC versus MCI						
Llama-3.1-8B	FC-as-text	41.9M / 8.1B	54.6 ± 1.5	58.3 ± 7.7	64.5 ± 18.5	47.7 ± 11.1
Llama-3.1-8B	FC adapter	44.4M / 8.1B	69.7 ± 6.5	63.0 ± 7.4	63.3 ± 9.3	62.6 ± 8.2
FC encoder	classifier only	2.5M / 2.5M	68.5 ± 3.5	67.0 ± 3.1	76.4 ± 8.9	50.8 ± 11.1
BrainLM	fine-tuned (FC)	86.2M / 86.2M	64.4 ± 4.8	57.2 ± 5.7	53.1 ± 10.4	64.1 ± 8.7
BrainSegFounder	fine-tuned	19.1M / 19.1M	72.4 ± 3.9	64.3 ± 5.1	60.3 ± 16.1	71.3 ± 15.2
BrainLRR (Ours)	task-specific	28.0M / 28.0M	$\mathbf{85.5 \pm 1.3}$	$\mathbf{78.7 \pm 1.4}$	$\mathbf{77.9 \pm 2.6}$	$\mathbf{80.0 \pm 2.7}$
ADNI: NC versus AD						
Llama-3.1-8B	FC-as-text	41.9M / 8.1B	47.4 ± 7.3	50.5 ± 7.8	51.1 ± 41.7	50.0 ± 42.8
Llama-3.1-8B	FC adapter	44.4M / 8.1B	86.7 ± 2.5	78.2 ± 4.0	71.9 ± 16.1	82.6 ± 12.7
FC encoder	classifier only	2.5M / 2.5M	83.6 ± 2.6	78.5 ± 6.3	73.3 ± 10.6	82.1 ± 12.8
BrainLM	fine-tuned (FC)	86.2M / 86.2M	74.4 ± 4.9	68.0 ± 6.5	60.7 ± 11.9	73.2 ± 17.1
BrainSegFounder	fine-tuned	19.1M / 19.1M	75.9 ± 2.6	68.9 ± 2.0	51.9 ± 18.3	81.1 ± 10.1
BrainLRR (Ours)	task-specific	28.0M / 28.0M	91.6 ± 1.9	$\mathbf{87.1 \pm 1.8}$	$\mathbf{77.8 \pm 2.6}$	$\mathbf{93.7 \pm 3.0}$

Note: The FC-adapter and encoder-only conditions share the same FC encoder, data splits, five seeds, and checkpoint-selection criterion; their comparison isolates the contribution of the pretrained LLM after modality alignment.

neuroimaging FMs under zero-shot, few-shot, linear-probing, and/or fine-tuning protocols, as applicable, full table in the Supplement Table 6. The LLM baseline uses Llama-3.1-8B, which is further adapted using LoRA (Hu et al., 2022). Vision baselines include ResNet and ViT (He et al., 2016; Dosovitskiy et al., 2020), while Neuroimaging FMs include BrainLM (Ortega Caro et al., 2024), and BrainSegFounder (Cox et al., 2024). We additionally conduct a controlled comparison among FC-as-text, a learned FC-to-token adapter for Llama-3.1-8B, and a matched FC encoder-only classifier in Table 1. We further compare with representative CNN-, GNN-, and transformer-based FC classifiers in Table 2. To isolate the contributions of stochastic ROI-profile masking and low-rank regularization, we use a BNT-style backbone (Kan et al., 2022b) implemented within the same BrainLRR framework, with both proposed components removed. The full FM benchmark is conducted on ABIDE and ADNI NC-MCI, and representative neuroimaging FMs are evaluated across datasets to assess cross-disorder generalization.

4.2 BRAINLRR VS. FMS

We compare BrainLRR with the FM baselines under a common static-FC setting and additionally evaluate BrainLM using ROI-level time-series inputs on ABIDE and PD, where such data are available. Table 1 reports the results. NeuroSTORM (Wang et al., 2026) is not included because it requires whole-brain voxelwise 4D fMRI ($1 \times 96 \times 96 \times 96 \times 20$), whereas our study evaluates models using 200×200 static-FC representations. Detailed adaptation pipelines are provided in Supplementary Figures 7 and 8, and Table 19.

Llama-3.1-8B with FC-as-text reaches a mean AUROC of 53.3 across the four tasks, with substantial variation in per-seed sensitivity and specificity, consistent with near-degenerate single-class predictions and limited evidence of meaningful discrimination. Replacing FC-as-text with a learned FC-

Table 2: Performance comparison of task-specific models across datasets.

Dataset	Model	AUROC	Acc.	Sens.	Spec.
ABIDE: ASD versus HC (493/516)					
CNN-based					
	BrainNetCNN (Kawahara et al., 2017)	74.9 ± 2.4	67.8 ± 2.7	63.8 ± 9.7	71.0 ± 10.2
GNN-based					
	BrainNetGNN (Mahmood et al., 2021)	55.3 ± 1.9	51.2 ± 5.4	67.7 ± 37.5	33.9 ± 34.2
	BrainGNN (Li et al., 2021)	62.4 ± 3.5	59.4 ± 2.3	36.7 ± 24.0	70.7 ± 19.3
	BrainGB (Cui et al., 2022)	69.7 ± 3.3	63.6 ± 1.9	63.7 ± 8.3	60.4 ± 10.1
	FBNETGNN (Kan et al., 2022a)	75.6 ± 1.2	68.0 ± 1.4	64.7 ± 8.7	62.4 ± 9.2
	DGM (Kazi et al., 2022)	52.7 ± 3.8	60.7 ± 12.6	53.8 ± 41.2	51.1 ± 40.9
Transformer-based					
	BolT (Bedel et al., 2023)	69.8 ± 2.0	62.6 ± 3.5	55.3 ± 5.5	70.0 ± 3.3
	BrainWaveNet (Jeong et al., 2024)	73.8 ± 2.0	67.5 ± 2.0	66.4 ± 9.1	68.3 ± 9.8
	BNT-style (Kan et al., 2022b)	75.7 ± 2.9	67.6 ± 2.6	67.8 ± 3.3	67.4 ± 2.3
	GBT (Peng et al., 2024)	78.3 ± 4.1	71.5 ± 5.8	$\mathbf{75.5 \pm 14.7}$	68.2 ± 12.8
	Com-BrainTF (Bannadabhavi et al., 2023)	77.3 ± 4.1	71.6 ± 4.5	75.1 ± 11.9	67.4 ± 9.3
	BrainLRR (Ours)	$\mathbf{79.1 \pm 1.8}$	$\mathbf{72.1 \pm 2.4}$	71.7 ± 3.3	$\mathbf{72.5 \pm 2.6}$
PD versus HC (73/110)					
	BNT-style	63.8 ± 1.1	62.2 ± 2.7	70.9 ± 2.5	49.3 ± 6.0
	BrainLRR (Ours)	$\mathbf{68.0 \pm 2.3}$	$\mathbf{63.8 \pm 3.1}$	$\mathbf{71.8 \pm 2.0}$	$\mathbf{52.0 \pm 8.7}$
ADNI: NC versus MCI (191/336)					
	BNT-style	83.2 ± 3.7	76.8 ± 4.8	$\mathbf{79.1 \pm 4.6}$	72.8 ± 7.4
	BrainLRR (Ours)	$\mathbf{85.5 \pm 1.3}$	$\mathbf{78.7 \pm 1.4}$	77.9 ± 2.6	$\mathbf{80.0 \pm 2.7}$
ADNI: NC versus AD (191/132)					
	BNT-style	$\mathbf{93.5 \pm 1.6}$	84.9 ± 1.3	73.3 ± 4.1	93.2 ± 1.4
	BrainLRR (Ours)	91.6 ± 1.9	$\mathbf{87.1 \pm 1.8}$	$\mathbf{77.8 \pm 2.6}$	$\mathbf{93.7 \pm 3.0}$

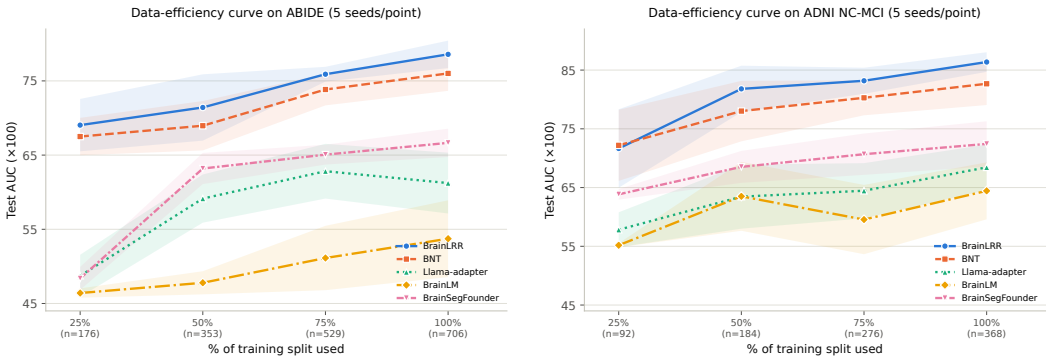
to-token adapter raises Llama's four-task mean AUROC from 53.3 to 70.8. However, the matched FC encoder-only classifier achieves 70.8 as well, despite using 2.5M parameters against Llama's 8.1B. The Llama adapter leads on three tasks but trails on PD. Thus, the learned FC representation substantially improves upon FC-as-text, whereas the evaluated pretrained LLM provides no consistent incremental benefit beyond the same FC encoder in this controlled comparison. Additional pretrained-versus-random-initialization controls are reported in Supplementary Table 9.

BrainLM and BrainSegFounder outperform the FC-as-text baseline on all four tasks, with four-task mean AUROCs of 65.1 and 70.6, respectively. Their sensitivity and specificity, however, remain poorly balanced after fine-tuning. For example, BrainLM reaches 81.8 sensitivity but only 30.7 specificity on PD, while BrainSegFounder shows the opposite pattern on NC–AD, with 51.9 sensitivity and 81.1 specificity. BrainLM also behaves differently across datasets when the input modality is changed. On ABIDE, using ROI-level time series increases AUROC from 55.3 with static FC to 63.4. On PD, the opposite trend is observed: static FC reaches 66.4 AUROC, compared with 57.5 for CC200 and 56.2 for A424 time-series inputs. Thus, using the modality or atlas seen during pretraining does not necessarily improve transfer to a new dataset.

BrainLRR achieves the highest AUROC of 79.1 on ABIDE, 85.5 on ADNI NC–MCI, and 91.6 on ADNI NC–AD, and the highest four-task mean AUROC of 81.1, compared with 70.8 for the best FM. Sensitivity and specificity are also well balanced on three of the four tasks. On PD, the FC encoder-only control achieves a slightly higher AUROC than BrainLRR, 71.0 versus 68.0, although the difference is within one standard deviation. PD is also the smallest cohort in our study, with 73 of 110 subjects, making the lower-capacity encoder a competitive baseline in this setting. Across these tasks, none of the evaluated pretrained FM configurations matches the performance of BrainLRR.

4.3 BrainLRR vs. task-specific models

We compare BrainLRR with a matched BNT-style backbone without ROI-profile masking and LRR. The two models use the same architecture, data splits, seeds, and training protocol, with the two



(a) ABIDE. (b) ADNI NC–MCI.

Figure 3: Data-efficiency comparison using 25/50/75/100% of the labeled training split.

proposed components as the only differences.Table 2 summarizes the results. BrainLRR improves AUROC on three of the four tasks, with gains of 3.4 points on ABIDE, 4.2 on PD, and 2.3 on ADNI NC–MCI. It also improves accuracy and specificity on all four tasks. On ADNI NC–AD, the backbone has a higher AUROC than BrainLRR, 93.5 versus 91.6, but BrainLRR has higher accuracy, sensitivity, and specificity, improving them from 84.9 to 87.1, 73.3 to 77.8, and 93.2 to 93.7, respectively. The AUROC difference occurs on the task with the highest baseline performance, where both models already achieve strong class separation. BrainLRR also has a lower seed-to-seed standard deviation overall. Thus, adding ROI-profile masking and LRR to the same backbone improves performance on most tasks.

4.4 Ablation and Sensitivity Analysis

Data-scarcity evaluation. To examine how the models perform with less training data, we retrain BrainLRR, the matched BNT-style backbone, the Llama FC-adapter, BrainLM, and BrainSeg-Founder using 25%, 50%, 75%, and 100% of the training split, with five seeds at each fraction. The validation and test sets remain fixed across all fractions, so only the amount of training data changes. As shown in Figure 3, the two task-specific models remain ahead of the three FM-based conditions across the training fractions. On ABIDE, BrainLRR trained with 25% of the training split already exceeds all FM conditions trained with the full split. The same pattern is observed on ADNI NC–MCI. Increasing the amount of labeled training data therefore does not eliminate the performance gap between the task-specific models and the evaluated FM conditions. The comparison with the matched BNT-style backbone is closer. BrainLRR achieves higher AUROC at every training fraction on ABIDE and from 50% onward on ADNI NC–MCI.

Component ablation. Table 3 evaluates the contributions of ROI profile masking and low-rank regularization on ABIDE. Removing ROI-profile masking causes the largest drop, reducing AUROC from 79.1 to 75.9 and accuracy from 72.1 to 69.5, indicating that ROI-level masking is the primary driver of robustness and reduces over-reliance on specific brain regions. Removing LRR results in a smaller performance drop, with AUROC decreasing from 79.1 to 77.9 and accuracy decreasing from 72.1 to 70.8, suggesting that LRR provides complementary refinement beyond masking.

Table 3: Ablation study of BrainLRR on ABIDE.

Metric	w/o ROI-profile masking	w/o LRR	BrainLRR
AUROC	75.9 ± 2.9	77.9 ± 2.0	$\mathbf{79.1 \pm 1.8}$
Acc.	69.5 ± 3.4	70.8 ± 3.1	$\mathbf{72.1 \pm 2.4}$
Sens.	68.0 ± 3.3	69.1 ± 3.0	$\mathbf{71.7 \pm 3.3}$
Spec.	71.1 ± 4.2	72.5 ± 5.7	$\mathbf{72.5 \pm 2.6}$

Masking ratio p. We further analyze the masking ratio p. As shown in Figure 4, moderate ROI-profile masking improves performance, while excessive masking degrades it, with the best result at $p = 0.4$. These results confirm that stochastic ROI-profile masking provides effective regularization under limited data.

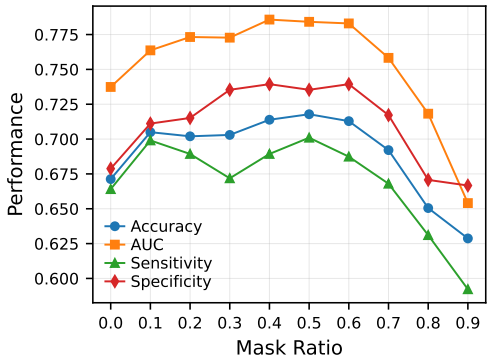


Figure 4: Sensitivity of ABIDE ASD classification performance to the row-only ROI-profile masking probability p.

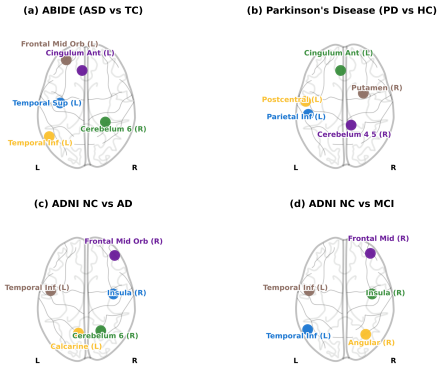


Figure 5: Top-five ROI importance maps identified by BrainLRR for ASD and PD classification tasks.

LRR coefficient λ_{LRR}. We sweep λ_{LRR} over $[0, 5 \times 10^{-4}]$ on ABIDE and ADNI NC–MCI using validation AUROC (Supplementary Fig 6). Low-to-moderate nonzero values generally improve over $\lambda_{\mathrm{LRR}} = 0$, although the best value differs between datasets. We fix $\lambda_{\mathrm{LRR}} = 1.25 \times 10^{-4}$ for all datasets rather than tuning it separately for each dataset.

Portability to other backbones. Finally, incorporating ROI-profile masking and LRR into BrainNetCNN without modifying its convolutional backbone increases mean AUROC across all four tasks by 0.43–4.66 points, supporting portability beyond transformers. The results are reported in Supplementary Table 12.

4.5 Interpretability Analysis

Figure 5(a) shows that the top ASD regions include the anterior cingulate, orbitofrontal pole, superior temporal/insular cortex, and cerebellum, aligning with salience, social-cognition, and cerebellar circuits implicated in autism (Uddin & Menon, 2009; D'Mello & Stoodley, 2015). For PD (Figure 5(b)), highlighted regions, putamen, postcentral gyrus, cerebellum, and anterior cingulate are consistent with disrupted basal ganglia–thalamocortical and cerebello-cortical motor circuits (Postuma & Dagher, 2006; Wu & Hallett, 2013). Figure 5(c)(d) shows MCI and AD share important temporal and insula regions, suggesting common involvement along Alzheimer's disease progression (Braak & Braak, 1991). However, their patterns also diverge: the precuneus/calcarine cortex is more salient in AD, while the angular gyrus, a temporo-parietal default-mode region, is more salient in MCI (Buckner et al., 2008; Sorg et al., 2007). These findings suggest that BrainLRR captures biologically meaningful and stage-sensitive FC patterns, providing interpretable evidence about disease-relevant brain regions.

5 Conclusion

In this work, we revisited a fundamental question for neuroimaging AI: whether foundation models are sufficient on their own for data-scarce FC classification. Across general-purpose and neuroimaging foundation models, large-scale pretraining did not consistently yield effective transfer to FC-based diagnosis. Motivated by this finding, we introduced BrainLRR, a task-specific framework combining stochastic ROI-profile masking with low-rank regularization. Across three brain-disorder datasets, BrainLRR achieves competitive performance while providing interpretable identification of disease-relevant brain regions. Our results suggest that, in data-scarce neuroimaging settings, carefully designed task-specific inductive biases can remain more effective than substantially larger pretrained models, highlighting the continued importance of domain-aware model design alongside the growing paradigm of foundation models.

AI USE STATEMENT

AI assistance was used solely for grammar checking and language refinement, and was not used to generate experimental results or analyses. The authors verified all scientific claims, experiments, analyses, and the final text.

ETHICS STATEMENT

This study uses previously collected, de-identified neuroimaging data from public research datasets. Data collection, informed consent, and institutional oversight were handled by the original dataset providers, and we use the data in accordance with their respective access and usage requirements. No new human participants were recruited for this study. Because neuroimaging datasets may contain demographic, site, scanner, and cohort biases, the reported results may not generalize to underrepresented populations or unseen clinical settings. BrainLRR is intended for research purposes and should not be interpreted as a clinical diagnostic system. Furthermore, the reported ROI-importance patterns represent model-derived associations and should not be interpreted as causal or definitive clinical biomarkers.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the datasets, preprocessing procedures, data partitions, model architecture, training settings, baseline adaptation protocols, hyperparameter searches, and evaluation procedures in the main paper and appendix. Results for models trained in this study are reported across five optimization seeds on fixed subject partitions. The appendix additionally reports component ablations, hyperparameter-sensitivity analyses, bootstrap procedures, and dataset-specific implementation details. Anonymous source code and configuration files are provided in the paper to support reproduction of the experiments.

REFERENCES

Olfactory dysfunction and functional connectivity changes in parkinson's disease, 2017. OpenNeuro dataset ds000245.

Liviu Badea, Mihaela Onu, Tao Wu, Adina Roceanu, and Ovidiu Bajenaru. Exploring the reproducibility of functional connectivity alterations in parkinson's disease. *PLoS One*, 12(11): e0188196, 2017.

Anushree Bannadabhavi, Soojin Lee, Wenlong Deng, Rex Ying, and Xiaoxiao Li. Community-aware transformer for autism prediction in fmri connectome. In *International conference on medical image computing and computer-assisted intervention*, pp. 287–297. Springer, 2023.

Hasan A Bedel, Irmak Sivgin, Onat Dalmaz, Salman UH Dar, and Tolga Çukur. Bolt: Fused window transformers for fmri time series analysis. *Medical image analysis*, 88:102841, 2023.

Bharat Biswal, F. Zerrin Yetkin, Victor M. Haughton, and James S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic Resonance in Medicine*, 34(4):537–541, 1995. doi: 10.1002/mrm.1910340409.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Heiko Braak and Eva Braak. Neuropathological stageing of alzheimer-related changes. *Acta neuropathologica*, 82(4):239–259, 1991.

Randy L Buckner, Jessica R Andrews-Hanna, and Daniel L Schacter. The brain's default network: anatomy, function, and relevance to disease. *Annals of the new York Academy of Sciences*, 1124 (1):1–38, 2008.

Ed Bullmore and Olaf Sporns. Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009. doi: 10.1038/nrn2575.

Minheng Chen, Xiaowei Yu, Jing Zhang, Tong Chen, Chao Cao, Yan Zhuang, Yanjun Lyu, Lu Zhang, Tianming Liu, and Dajiang Zhu. Core-periphery principle guided state space model for functional connectome classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 236–246. Springer, 2025.

Joseph Cox, Peng Liu, Skylar E Stolte, Yunchao Yang, Kang Liu, Kyle B See, Huiwen Ju, and Ruogu Fang. Brainsegfounder: Towards 3d foundation models for neuroimage segmentation. *Medical Image Analysis*, 97:103301, 2024.

R Cameron Craddock, G Andrew James, Paul E Holtzheimer III, Xiaoping P Hu, and Helen S Mayberg. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human brain mapping*, 33(8):1914–1928, 2012.

Hejie Cui, Wei Dai, Yanqiao Zhu, Xuan Kan, Antonio Aodong Chen Gu, Joshua Lukemire, Liang Zhan, Lifang He, Ying Guo, and Carl Yang. Braingb: a benchmark for brain network analysis with graph neural networks. *IEEE transactions on medical imaging*, 42(2):493–506, 2022.

Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

Anila M D'Mello and Catherine J Stoodley. Cerebro-cerebellar circuits in autism spectrum disorder. *Frontiers in neuroscience*, 9:408, 2015.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Michael D. Greicius, Gaurav Srivastava, Allan L. Reiss, and Vinod Menon. Default-mode network activity distinguishes alzheimer's disease from healthy aging: Evidence from functional mri. *Proceedings of the National Academy of Sciences*, 101(13):4637–4642, 2004. doi: 10.1073/pnas.0308627101.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 594–604, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Ah-Yeong Jeong, Da-Woon Heo, Eunsong Kang, and Heung-Il Suk. Brainwavenet: wavelet-based transformer for autism spectrum disorder diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 56–66. Springer, 2024.

Xuan Kan, Hejie Cui, Joshua Lukemire, Ying Guo, and Carl Yang. Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation. In *International conference on medical imaging with deep learning*, pp. 618–637. PMLR, 2022a.

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

Jeremy Kawahara, Colin J. Brown, Steven P. Miller, Brian G. Booth, Vann Chau, Ruth E. Grunau, Jill G. Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017. doi: 10.1016/j.neuroimage.2016.09.046.

Anees Kazi, Luca Cosmo, Seyed-Ahmad Ahmadi, Nassir Navab, and Michael M Bronstein. Differentiable graph module (dgm) for graph convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1606–1617, 2022.

Aaron S Kemp, Journey Eubank, Yahya Younus, James E Galvin, Fred W Prior, and Linda J Larson-Prior. Resting state mri data from healthy control (hc), parkinson's disease with normal cognition (pd-nc), and parkinson's disease with mild cognitive impairment (pd-mci) cohorts. 2025.

Meenakshi Khosla, Keith Jamison, Gia H. Ngo, Amy Kuceyeski, and Mert R. Sabuncu. Machine learning in resting-state fmri analysis. *Magnetic Resonance Imaging*, 64:101–121, 2019. doi: 10.1016/j.mri.2019.05.031.

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

Yueying Li, Xiaotong Zhang, Shihan Guan, Guolin Ma, and Youyong Kong. Topology-guided graph masked autoencoder learning for population-based neurodevelopmental disorder diagnosis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184, 2012.

Usman Mahmood, Zening Fu, Vince D Calhoun, and Sergey Plis. A deep learning model for data-driven discovery of functional connectivity. *Algorithms*, 14(3):75, 2021.

Josue Ortega Caro, Antonio Henrique de Oliveira Fonseca, Syed Rizvi, Matteo Rosati, Christopher Averill, James Cross, Prateek Mittal, Emanuele Zappala, Rahul Dhodapkar, Chadi Abdallah, et al. Brainlm: A foundation model for brain activity recordings. In *International Conference on Learning Representations*, volume 2024, pp. 565–576, 2024.

Zhihao Peng, Zhibin He, Yu Jiang, Pengyu Wang, and Yixuan Yuan. Gbt: Geometric-oriented brain transformer for autism diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 142–152. Springer, 2024.

Ronald B Postuma and Alain Dagher. Basal ganglia functional connectivity based on a meta-analysis of 126 positron emission tomography and functional magnetic resonance imaging publications. *Cerebral cortex*, 16(10):1508–1521, 2006.

Christian Sorg, Valentin Riedl, Mark Muhlau, Vince D Calhoun, Tom Eichele, Leonhard Läer, Alexander Drzezga, Hans Förstl, Alexander Kurz, Claus Zimmer, et al. Selective changes of resting-state networks in individuals at risk for alzheimer's disease. *Proceedings of the National Academy of Sciences*, 104(47):18760–18765, 2007.

Olaf Sporns. Structure and function of complex brain networks. *Dialogues in Clinical Neuroscience*, 15(3):247–262, 2013.

Olaf Sporns and Richard F Betzel. Modular brain networks. *Annual review of psychology*, 67(1): 613–640, 2016.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Alessandro Tessitore, Marina Cirillo, and Rosa De Micco. Functional connectivity signatures of parkinson's disease. *Journal of Parkinson's Disease*, 9(4):637–652, 2019. doi: 10.3233/JPD-191592.

Lucina Q Uddin and Vinod Menon. The anterior insula in autism: under-connected and under-examined. *Neuroscience & Biobehavioral Reviews*, 33(8):1198–1203, 2009.

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

Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Changbae Bang, Lin Zhao, Jinglei Lv, Jorge Sepulcre, Carl Yang, Lifang He, Tianming Liu, Daniel Barron, Quanzheng Li, Randy Hirschtick, Byung-Hoon Kim, Xiang Li, and Yixuan Yuan. Towards a general-purpose foundation model for fmri analysis. *Nature Biomedical Engineering*, 2026.

Tao Wu and Mark Hallett. The cerebellum in parkinson's disease. *Brain*, 136(3):696–709, 2013.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pp. 478–487. PMLR, 2016.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5812–5823, 2020.

W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. doi: 10.1002/1097-0142(1950)3:1⟨32::AID-CNCR2820030106⟩3.0.CO;2-3.

Table 4: Demographic information for the PD and ADNI datasets. Age is reported as mean ± standard deviation.

Variable	PD		ADNI: NC vs. AD		ADNI: NC vs. MCI	
	PD	HC	AD	NC	MCI	NC
Subjects	110	73	132	191	336	191
Sex (F/M)	47/63	43/30	61/71	111/80	168/168	111/80
Age	67.46 ± 7.75	66.88 ± 8.23	75.54 ± 8.01	80.02 ± 7.05	76.68 ± 6.32	80.02 ± 7.05

A Supplementary Dataset Information

Datasets and preprocessing. We evaluate BrainLRR on three rs-fMRI datasets covering four binary FC classification tasks. ABIDE is used for ASD vs. HC classification with 493 HC and 516 ASD subjects (Di Martino et al., 2014). For Parkinson's disease classification, we construct a pooled PD cohort with 110 PD and 73 HC subjects from the Tao Wu, NEUROCON, and OpenNeuro ds005892/ds000245 datasets (Badea et al., 2017; Kemp et al., 2025; ope, 2017). For ADNI, we define two binary tasks: NC vs. AD, with 191 NC and 132 AD subjects, and NC vs. MCI, with 191 NC and 336 MCI subjects. Demographic information is summarized in 4.

Following standard rs-fMRI preprocessing, brain regions are defined using the Craddock 200 atlas (CC200) (Craddock et al., 2012). For each subject, ROI-level time series are used to compute pairwise functional connectivity, resulting in a 200×200 FC matrix. The resulting FC matrices are used as the common input representation for BrainLRR and the task-specific FC baselines.

Evaluation metrics and threshold calibration. We report AUROC, accuracy (ACC), sensitivity (SEN), and specificity (SPE). AUROC is threshold-independent and serves as the primary ranking metric. For PD and ADNI, where class imbalance is more pronounced, models are trained using balanced focal loss with inverse-frequency class weighting to reduce majority-class bias (Lin et al., 2017).

For threshold-dependent metrics, BrainLRR is evaluated using the default 0.5 decision threshold, equivalent to the argmax rule. For fine-tuned foundation-model baselines, which can collapse to a single class under the default threshold, we additionally calibrate the decision threshold for each seed on the validation set using Youden's J statistic before test-set evaluation (Youden, 1950).

B Supplementary implementation details

B.1 BrainLRR Training Details

BrainLRR was trained on an NVIDIA RTX PRO 6000 GPU. We used a fixed training configuration across all datasets: 200 epochs, learning rate 7.74×10^{-5}, weight decay 0.0235, and LRR loss weight 1.25×10^{-4}.

B.2 Hyperparameter Selection of Foundation-Model Baselines

To prevent test-set leakage, all hyperparameters were selected solely by validation AUROC. For each model–dataset pair, the search was conducted using the first experimental seed (seed 42). The configuration achieving the highest validation AUROC was then fixed and evaluated across all five reporting seeds. Table 5 summarizes the observed ranges and selected settings.

BrainLM. We conducted 10 Optuna trials independently for each dataset. The learning rate was sampled log-uniformly from 10^{-5} to 10^{-3}, and weight decay was sampled from 0 to 0.1. Each trial was trained for five epochs using AdamW, a batch size of 16, and a constant learning rate, with checkpoint selection based on validation AUROC.

Llama-3.1-8B adapter. We fixed the LoRA configuration to rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05. We then sampled the learning rate log-uniformly from 5×10^{-5} to 5×10^{-4} and weight decay from 0 to 0.1. Constant and cosine-decay schedules were compared, both with a 10%

Table 5: Validation-only hyperparameter selection for the foundation-model baselines. Spread denotes the difference between the maximum and minimum validation AUROC across the searched configurations.

Model	Dataset	Val. AUROC	Spread	Selected LR	Selected WD	Schedule
BrainLM: 10-trial Optuna search						
BrainLM	ABIDE	58.1–62.1	3.9	8.27×10^{-4}	0.0085	Constant
BrainLM	PD	76.6–84.4	7.8	6.27×10^{-4}	0.0336	Constant
BrainLM	ADNI NC–AD	46.6–56.7	10.1	3.02×10^{-5}	0.0895	Constant
BrainLM	ADNI NC–MCI	63.6–68.0	4.3	2.07×10^{-4}	0.0868	Constant
Llama-3.1-8B adapter: Optuna search						
Llama adapter	ABIDE	63.8–68.8	5.0	1.48×10^{-4}	0.0529	Constant
Llama adapter	PD	66.2–80.5	14.3	1.50×10^{-4}	0.0478	Constant
Llama adapter	ADNI NC–AD	70.0–81.0	10.9	7.96×10^{-5}	0.0228	Constant
Llama adapter	ADNI NC–MCI	69.7–78.8	9.1	1.74×10^{-4}	0.0775	Cosine
BrainSegFounder: 12-configuration grid search						
BrainSegFounder	ABIDE	56.4–62.0	5.6	3.00×10^{-5}	0	–
BrainSegFounder	PD	68.8–77.9	9.1	1.00×10^{-4}	0	–
BrainSegFounder	ADNI NC–AD	45.8–75.7	30.0	1.00×10^{-4}	0	–
BrainSegFounder	ADNI NC–MCI	63.8–86.1	22.3	3.00×10^{-5}	0	–

warmup period. Each trial used an effective batch size of 16 and was trained for up to 10 epochs, with checkpoint selection based on validation AUROC.

BrainSegFounder. We evaluated the Cartesian product of learning rates $\{10^{-5}, 3\times10^{-5}, 10^{-4}, 3\times10^{-4}\}$ and weight-decay values $\{0, 0.01, 0.05\}$, yielding 12 configurations per dataset. all other optimization settings were held fixed across configurations.

C Additional Foundation-Model Results

C.1 Additional foundation-model results on ABIDE

Table 6 provides the complete ABIDE evaluation of general-purpose LLMs, vision backbones, neuroimaging foundation models, and the task-specific BrainLRR model. These results complement the main-paper analysis by reporting zero-shot, few-shot, linear-probe, and fine-tuning settings in one table. The comparison shows that general-purpose FMs suffer from a substantial domain gap, neuroimaging FMs reduce this gap under fine-tuning, and BrainLRR achieves the strongest overall performance with substantially fewer trainable parameters.

C.2 Comparison with task-specific models across diseases

In addition to foundation-model baselines, we compare BrainLRR with existing task-specific models on the ABIDE benchmark in Table 2. These models include CNN-based, GNN-based, and transformer-based approaches for brain network classification. BrainLRR achieves the best AUROC of 79.1%, the highest accuracy of 72.1%, and the best specificity of 72.5%, while obtaining competitive sensitivity compared with prior transformer-based models. Although some baselines achieve higher sensitivity, their specificity is often lower, indicating less balanced operating-point behavior.

C.3 Data-Scarcity Sweep

We retain $\{25, 50, 75, 100\}\%$ of training subjects while holding validation and test splits fixed at full size, so differences reflect training-set size alone; subsampling is stratified by site and label, the subsets are nested, and all other settings match the full-data runs (main paper, Sec. 4.1) with 5 seed(s) per configuration. Table 8 shows the ordering between model families is stable across the sweep: on ABIDE, BrainLRR trained on a quarter of the subjects (69.0 AUROC) exceeds every foundation-

Table 6: Additional ABIDE results for foundation-model and backbone evaluation under zero-shot, few-shot, linear-probe, and fine-tuning settings. For ResNet and ViT, linear-probe results are obtained by freezing the pretrained backbone and fitting a linear classifier on ABIDE training features, with the best epoch selected by validation AUROC. TS denotes ROI-level fMRI time-series input. Metrics are reported in percentage.

Model	Setting	Total Params	AUROC	Acc	Sens	Spec
General-purpose FMs: large domain gap						
GPT-OSS-20B	zero-shot	20914.8M	55.3	52.4	0.0	100.0
GPT-OSS-20B	few-shot	20914.8M	50.3	52.4	0.0	100.0
Qwen3-32B	zero-shot	32762.1M	51.0	47.5	100.0	0.0
Qwen3-32B	few-shot	32762.1M	48.9	47.5	100.0	0.0
Llama-3.1-8B	zero-shot	8030.3M	46.3	47.5	100.0	0.0
Llama-3.1-8B	few-shot	8030.3M	51.0	47.5	100.0	0.0
Llama-3.1-8B	fine-tuned	8030.3M	53.0	52.7	46.0	59.8
Neuroimaging FMs: reduced domain gap						
BrainLM	zero-shot	86.2M	47.7	40.7	8.3	85.7
BrainLM	few-shot	86.2M	49.4	48.5	36.0	63.7
BrainLM	fine-tuned (TS)	86.2M	63.4	59.5	72.9	45.3
BrainSegFounder	zero-shot	19.1M	54.4	52.4	29.7	77.7
BrainSegFounder	few-shot	19.1M	51.4	52.9	43.9	65.0
BrainSegFounder	fine-tuned	19.1M	66.0	62.0	60.3	71.3
General vision backbones: reference baselines						
ResNet101	linear probe	42.5M	51.0	50.5	57.4	37.9
ResNet101	fine-tuned	42.5M	53.7	51.4	64.8	42.8
ResNet50	linear probe	23.5M	54.3	51.4	70.8	22.2
ResNet50	fine-tuned	23.5M	51.0	50.9	50.0	53.7
ViT-B/16	linear probe	85.8M	54.1	53.4	52.0	55.5
ViT-B/16	fine-tuned	85.8M	64.6	59.9	38.5	78.0
ViT-L/16	linear probe	303.3M	54.9	54.4	64.5	45.2
ViT-L/16	fine-tuned	303.3M	64.7	59.9	70.2	57.5
Task-specific models						
BrainLRR (Ours)	task-specific	28.0M	**79.1**	**72.1**	**71.7**	**72.5**

model baseline trained on the full set (61.2, 66.7, 53.7), and on ADNI NC vs. MCI, BrainLRR at 50% (81.8) surpasses all foundation-model baselines at 100%. we report trends unsmoothed.

C.4 Pretrained versus Randomly Initialized Models

To isolate pretrained initialization under a matched training configuration, we fine-tuned randomly initialized ("scratch") versions of BrainLM and BrainSegFoundercand compared them with their pretrained counterparts. Architecture, input representation, adaptation protocol, and trainable parameter count were held fixed. Both conditions used the learning rate and weight decay selected as optimal for the pretrained model through the dataset-specific hyperparameter search. Table 9 shows pretraining improves mean AUROC for BrainLM with ROI inputs on both ABIDE and ADNI NC–MCI, but not with time-series input on ABIDE. However pretraining doesnt improve much of the result for BrainSedgeFounder. These results suggest that under one fixed recipe, pretrained initialization can be beneficial, but its effect depends on the input representation and target dataset.

C.5 Paired Bootstrap Analysis of Model and Component Effects

To quantify held-out-subject sampling uncertainty, we computed paired BCa 95% confidence intervals on split 101 using 2,000 test-subject resamples, applying identical resampled indices to each matched model pair and training seed. As shown in Table 10, BrainLRR shows a positive AUROC difference from the matched BNT-style backbone on ABIDE (+2.35, [+0.51, +4.30]), PD (+3.32,

Table 7: Performance comparison of task-specific models on the ABIDE benchmark.

Model	AUROC	Acc.	Sens.	Spec.
CNN-based models				
BrainNetCNN (Kawahara et al., 2017)	74.9 ± 2.4	67.8 ± 2.7	63.8 ± 9.7	71.0 ± 10.2
GNN-based models				
BrainNetGNN (Mahmood et al., 2021)	55.3 ± 1.9	51.2 ± 5.4	67.7 ± 37.5	33.9 ± 34.2
BrainGNN (Li et al., 2021)	62.4 ± 3.5	59.4 ± 2.3	36.7 ± 24.0	70.7 ± 19.3
BrainGB (Cui et al., 2022)	69.7 ± 3.3	63.6 ± 1.9	63.7 ± 8.3	60.4 ± 10.1
FBNETGNN (Kan et al., 2022a)	75.6 ± 1.2	68.0 ± 1.4	64.7 ± 8.7	62.4 ± 9.2
DGM (Kazi et al., 2022)	52.7 ± 3.8	60.7 ± 12.6	53.8 ± 41.2	51.1 ± 40.9
Transformer-based models				
BolT (Bedel et al., 2023)	69.8 ± 2.0	62.6 ± 3.5	55.3 ± 5.5	70.0 ± 3.3
BrainWaveNet (Jeong et al., 2024)	73.8 ± 2.0	67.5 ± 2.0	66.4 ± 9.1	68.3 ± 9.8
BNT (Kan et al., 2022b)	75.7 ± 2.9	67.6 ± 2.6	67.8 ± 3.3	67.4 ± 2.3
GBT (Peng et al., 2024)	78.3 ± 4.1	71.5 ± 5.8	$\mathbf{75.5 \pm 14.7}$	68.2 ± 12.8
Com-BrainTF (Bannadabhavi et al., 2023)	77.3 ± 4.1	71.6 ± 4.5	75.1 ± 11.9	67.4 ± 9.3
BrainLRR (Ours)	$\mathbf{79.1 \pm 1.8}$	$\mathbf{72.1 \pm 2.4}$	71.7 ± 3.3	$\mathbf{72.5 \pm 2.6}$

Table 8: Data-scarcity sweep. Test AUROC (%) as a function of the fraction of training subjects retained; validation and test splits are held fixed at full size across all fractions. Subsampling is site- and label-stratified, and the subsets are nested. Results are averaged over five seeds. Best per column in **bold**.

	ABIDE: ASD vs. HC				ADNI: NC vs. MCI			
Model	25%	50%	75%	100%	25%	50%	75%	100%
BrainLRR (Ours)	**69.0**	**71.4**	**75.9**	**78.6**	71.7	**81.8**	**83.2**	**86.4**
BNT	67.5	69.0	73.8	76.0	**72.2**	78.0	80.3	82.7
Llama-adapter	48.7	59.1	62.8	61.2	57.8	63.5	64.5	68.4
BrainLM	46.4	47.8	51.1	53.7	55.2	63.5	59.6	64.4
BrainSegFounder	48.4	63.2	65.1	66.7	63.9	68.5	70.7	72.4

[+0.87, +6.91]), and ADNI NC-MCI (+2.27, [+0.26, +6.17]), with all 95% confidence intervals excluding zero. On ADNI NC-AD, the difference is -1.55, with a wide 95% confidence interval spanning zero ([-5.87, +3.72]). At the component level, ROI-profile masking shows a positive effect on ABIDE (+2.83, [+0.97, +4.61]), while LRR shows a positive effect on ADNI NC-AD (+2.15, [+0.50, +4.35]). The remaining intervals span zero.

C.6 Sensitivity and Formulation of ROI-Profile Masking

The relative effect of the two formulations is dataset-dependent. On the full ABIDE training set, symmetric masking reduces AUROC by 2.64 points, with the comparison lying at the conventional significance boundary ($p = 0.050$). Reducing the ABIDE training set from approximately 706 to 368 subjects narrows the difference to 1.72 points but does not reverse its direction ($p = 0.154$). The matched-size control therefore does not support training-set size alone as an explanation for the opposing ABIDE and ADNI NC–MCI results.

Across the other tasks, symmetric masking produces a small, nonsignificant improvement on PD classification ($+1.39$, $p = 0.604$) and virtually no change on ADNI NC–AD (-0.02, $p = 0.991$). ADNI NC–MCI is the only task on which symmetric masking yields an improvement at the uncorrected $p < 0.05$ level ($+3.17$, $p = 0.021$); And no result reaches significance. Neither formulation is uniformly superior.

Symmetric masking is not merely a symmetry-preserving implementation of row-only masking, but a stronger input perturbation. For off-diagonal FC entries, row-only and symmetric masking retain expected fractions of $1 - p$ and $(1 - p)^2$, respectively; at $p = 0.4$ these correspond to approximately 60% and 36% of the matrix entries. Row-only masking also preserves redundant connectiv-

Table 9: Comparison of pretrained and randomly initialized BrainLM and BrainSegFounder models. Results are reported as mean AUROC (%) ± standard deviation over five runs. Δ denotes pretrained minus scratch performance in AUROC points. Because the shared configuration was selected by searching on the pretrained model, the scratch arm may be under-tuned; Δ should therefore be read as an upper bound on the benefit of pretraining.

Model	Dataset	Pretrained	Scratch	Δ
BrainLM (ROI)	ABIDE	$\mathbf{53.64 \pm 5.44}$	47.30 ± 1.86	$+6.34$
BrainLM (time series)	ABIDE	63.40 ± 3.61	$\mathbf{64.38 \pm 0.81}$	-0.98
BrainLM	ADNI NC–MCI	$\mathbf{64.46 \pm 4.84}$	53.32 ± 2.15	$+11.14$
BrainSegFounder	ABIDE	$\mathbf{66.65 \pm 1.91}$	65.95 ± 2.61	$+0.70$
BrainSegFounder	ADNI NC–MCI	$\mathbf{72.43 \pm 3.85}$	70.62 ± 2.14	$+1.81$

Table 10: Paired subject-level BCa 95% confidence intervals for the mean AUROC difference between BrainLRR and each comparator on split 101. Identical indices were used to resample the test subjects for each matched model pair and training seed over 2,000 bootstrap resamples.

Dataset (test n)	vs. BNT	vs. w/o LRR	vs. w/o masking
ABIDE ($n = 202$)	$+2.35$ $[+0.51, +4.30]^*$	$+0.05$ $[-1.46, +1.44]$	$+2.83$ $[+0.97, +4.61]^*$
PD ($n = 37$)	-1.55 $[-5.87, +3.72]$	$+1.39$ $[-3.39, +6.52]$	-3.79 $[-11.50, +2.47]$
ADNI NC–AD ($n = 65$)	$+3.32$ $[+0.87, +6.91]^*$	$+2.15$ $[+0.50, +4.35]^*$	$+1.64$ $[-1.20, +4.94]$
ADNI NC–MCI ($n = 106$)	$+2.27$ $[+0.26, +6.17]^*$	$+0.57$ $[-1.19, +2.30]$	$+0.88$ $[-1.10, +3.14]$

Note: Values are mean AUROC differences in percentage points, computed as BrainLRR minus the comparator; positive values favor BrainLRR. An asterisk indicates that the 95% confidence interval excludes zero.

ity information through the profiles of unmasked ROIs, whereas symmetric masking removes this information from both views. This difference may contribute to the heterogeneous effects, although the present experiments do not establish the underlying mechanism. We consequently characterize ROI-profile masking as training-time regularization rather than as a simulation of missing ROIs or direct evidence of test-time robustness to ROI corruption.

C.7 LRR-Coefficient Sensitivity

We evaluate seven LRR coefficients,$\lambda_{\mathrm{LRR}} \in [0, 5 \times 10^{-4}]$, on ABIDE and ADNI NC–MCI using fixed data splits, hyperparameters, ROI-profile masking, and a 100-epoch training budget. As shown in FIgure 6, low-to-moderate nonzero coefficients generally preserve or improve validation AUROC, although the optimal value is dataset-dependent.

C.8 Portability to BrainNetCNN

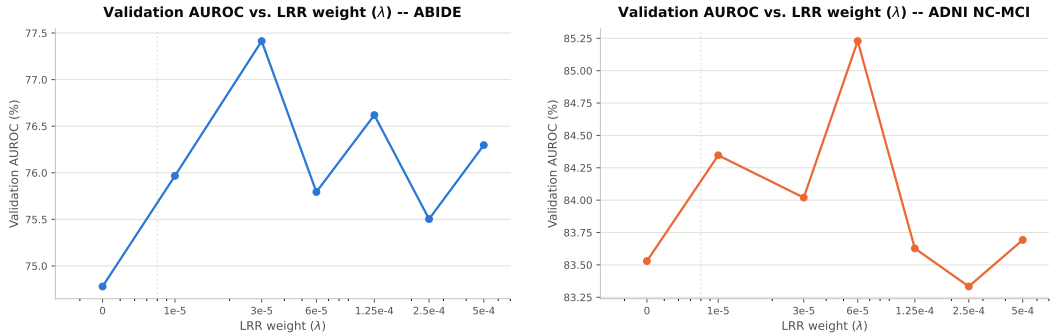
To evaluate portability beyond transformers, we incorporate ROI-profile masking and LRR into BrainNetCNN without modifying its convolutional backbone. As shown in Table 12, the augmented model achieves higher mean AUROC on all four tasks, with gains ranging from 0.43 to 4.66 points. Sensitivity decreases on PD and ADNI NC–MCI, indicating that the effects vary across metrics and datasets. Overall, these results provide evidence that the proposed components can transfer to a convolutional FC backbone.

C.9 Batch Sampling and LRR Normalization

We evaluate four conditions: A uses random sampling and the original LRR; B uses random sampling with class-size-normalized LRR; C uses class-balanced sampling with the original LRR; and D combines balanced sampling with class-size normalization. All conditions use identical architectures, hyperparameters, data splits, training budgets, and five matched seeds, differing only in sampling and LRR normalization.

Table 11: Comparison of row-only and symmetric ROI-profile masking. Results are mean AUROC (%) ± standard deviation over five paired seeds. Δ denotes symmetric minus row-only AUROC in percentage points. The p-values are from two-sided paired t-tests ; * denotes $p < 0.05$.

Dataset	**Row-only**	**Symmetric**	Δ	p
ABIDE (full, $n_{\text{train}} \approx 706$)	78.58 ± 1.85	75.94 ± 0.99	-2.64	0.050
ABIDE (matched, $n_{\text{train}} \approx 368$)	74.23 ± 2.46	72.51 ± 0.66	-1.72	0.154
PD	68.00 ± 2.26	69.39 ± 5.46	$+1.39$	0.604
ADNI NC–AD	90.84 ± 2.79	90.82 ± 1.50	-0.02	0.991
ADNI NC–MCI	86.37 ± 1.67	89.53 ± 1.48	$+3.17$	0.021^*



(a) ABIDE.

(b) ADNI NC–MCI.

Figure 6: Sensitivity to the LRR coefficient λ_{LRR} on the ABIDE and ADNI NC–MCI validation sets. Each point reports mean validation AUROC across five seeds under fixed data splits and a 100-epoch training budget.

Because balanced batches contain equal class counts, class-size normalization reduces to the original LRR formulation in our implementation; hence, Conditions C and D are algebraically equivalent and produce identical results.

None of the ABIDE accuracy or AUROC differences in Table 13 reaches $p < 0.05$. We further compare the two individual modifications across all four tasks in Table 14. Balanced sampling increases mean AUROC only on PD, while class-size normalization does not increase mean AUROC on any task. Thus, neither modification provides a consistent benefit across the evaluated datasets, supporting our use of random sampling with the original LRR formulation.

D Supplementary ROI Analysis

ROI-level importance analysis. To provide model-derived interpretability, we compute ROI-level importance scores from the trained BrainLRR model. Since BrainLRR processes FC matrices as ROI-level connectivity profiles and uses attention to model inter-regional dependencies, the learned attention patterns can be aggregated to estimate the contribution of each ROI to the final prediction. Specifically, we aggregate attention weights across heads and layers, averaging over the validation set to obtain a single ROI-level importance score per dataset. Scores are then averaged within each dataset and ranked to identify the top disease-relevant regions. These ROI-level analyses help identify disease-relevant brain regions. supporting BrainLRR's interpretability and potential value for biomarker discovery.

Supplementary ROI ranking. Tables 15–18 report the top-10 ranked ROIs for each classification task. Importance scores are computed from post-softmax attention values and averaged over attention heads, transformer layers, validation subjects, and training seeds. We also report ×Unif. relative to the uniform-attention baseline of $1/200 = 0.005$. AAL, Harvard–Oxford, and MNI centroid information is provided for anatomical reference.

Table 12: Portability of ROI-profile masking and LRR to BrainNetCNN. Results are mean ± standard deviation across five training seeds. Bold indicates the higher mean within each dataset and metric; tied rounded values are both bolded.

Dataset	Configuration	Acc.	AUROC	Sens.	Spec.
ABIDE	Baseline	67.23 ± 1.87	75.23 ± 2.03	65.24 ± 7.11	69.29 ± 6.76
	+ Masking + LRR	$\mathbf{68.61 \pm 1.77}$	$\mathbf{76.44 \pm 1.08}$	$\mathbf{67.57 \pm 2.88}$	$\mathbf{69.70 \pm 2.47}$
PD	Baseline	68.65 ± 5.92	67.39 ± 3.12	$\mathbf{88.18 \pm 5.18}$	40.00 ± 13.33
	+ Masking + LRR	$\mathbf{69.73 \pm 1.21}$	$\mathbf{67.82 \pm 1.54}$	81.82 ± 3.21	$\mathbf{52.00 \pm 2.98}$
ADNI NC–AD	Baseline	81.54 ± 1.88	88.36 ± 1.00	67.41 ± 8.84	91.58 ± 4.71
	+ Masking + LRR	$\mathbf{85.23 \pm 2.57}$	$\mathbf{89.10 \pm 0.93}$	$\mathbf{76.30 \pm 4.97}$	$\mathbf{91.58 \pm 2.20}$
ADNI NC–MCI	Baseline	74.15 ± 0.84	81.63 ± 3.12	$\mathbf{86.77 \pm 3.86}$	54.15 ± 7.79
	+ Masking + LRR	$\mathbf{79.25 \pm 2.91}$	$\mathbf{86.29 \pm 1.09}$	82.46 ± 4.69	$\mathbf{74.15 \pm 4.08}$

Table 13: Paired comparisons with Condition A on ABIDE. Differences are computed as the modified condition minus A; parentheses report p-values from paired t-tests across five matched seeds.

Comparison	Δ Acc. (p)	Δ AUROC (p)
B − A (class norm.)	-0.10 (0.940)	-1.70 (0.069)
C − A (balanced)	$+0.10$ (0.952)	-0.65 (0.487)
D − A (both)	$+0.10$ (0.952)	-0.65 (0.487)

E Prompt Templates and Foundation-Model Pipelines

We provide the exact LLM prompt templates and medical foundation-model input pipelines used in our experiments. For LLMs, FC is summarized as text and fed into prompts, whereas for neuroimaging FMs, FC matrices or ROI-level time-series signals are transformed into tensor inputs compatible with each pretrained encoder. These implementation is to ensure that the comparisons are reproducible and that each model receives inputs in a format consistent with its original design. This also helps clarify the distinction between language-based prompting and tensor-based neuroimaging adaptation in our evaluation.

E.1 LLM Prompt Templates

LLM prompting setup. For the LLM baselines, each subject-level FC matrix is first converted into a compact textual summary containing global FC statistics, edge-polarity information, top hub ROIs, and strongest connectivity edges. The same summary format is used across zero-shot, few-shot, and LoRA fine-tuning settings to ensure that differences in performance reflect the adaptation protocol rather than changes in input representation. 7 shows the exact prompt templates used in our experiments.

E.2 Dataset-Specific Prompt Substitutions

Task-specific label mapping. The prompt structure is shared across datasets, while the disease context and output label strings are changed according to the target classification task. Table 19 summarizes the positive and negative label substitutions used for ABIDE, PD, and the two ADNI tasks.

E.3 Medical Foundation-Model Input Pipelines

Tensor-based FM adaptation. Unlike LLMs, the medical foundation-model baselines do not use natural-language prompts. Instead, FC matrices or ROI-level time-series signals are transformed into tensor inputs compatible with each pretrained encoder. Figure 8 summarizes the preprocessing and

Table 14: Controlled ablations of class-balanced sampling and class-size-aware LRR normalization across four classification tasks. Results are mean ± standard deviation across five seeds. Condition D is omitted because it is algebraically equivalent to Condition C under balanced sampling and produces identical results.

Dataset	Condition	Acc.	AUROC	Sens.	Spec.
ABIDE	A: Random/orig.	71.88 ± 1.76	$\mathbf{78.58 \pm 1.85}$	$\mathbf{70.49 \pm 2.24}$	73.33 ± 2.21
	B: Random/class norm.	71.78 ± 1.89	76.88 ± 1.30	69.71 ± 2.51	73.94 ± 3.38
	C: Balanced/orig.	$\mathbf{71.98 \pm 2.73}$	77.93 ± 1.00	70.10 ± 3.85	$\mathbf{73.94 \pm 1.66}$
PD	A: Random/orig.	63.78 ± 3.08	68.00 ± 2.26	$\mathbf{71.82 \pm 2.03}$	52.00 ± 8.69
	B: Random/class norm.	$\mathbf{64.32 \pm 3.52}$	67.58 ± 3.95	68.18 ± 6.43	$\mathbf{58.67 \pm 5.58}$
	C: Balanced/orig.	62.70 ± 2.96	$\mathbf{70.30 \pm 1.09}$	69.09 ± 4.98	53.33 ± 0.00
ADNI NC–AD	A: Random/orig.	$\mathbf{85.54 \pm 1.75}$	$\mathbf{90.84 \pm 2.79}$	$\mathbf{76.30 \pm 4.22}$	$\mathbf{92.11 \pm 3.22}$
	B: Random/class norm.	82.77 ± 2.01	90.29 ± 1.45	74.81 ± 1.66	88.42 ± 3.00
	C: Balanced/orig.	83.69 ± 3.71	89.69 ± 2.85	$\mathbf{76.30 \pm 3.31}$	88.95 ± 4.71
ADNI NC–MCI	A: Random/orig.	$\mathbf{77.92 \pm 2.88}$	$\mathbf{86.37 \pm 1.67}$	$\mathbf{76.00 \pm 5.17}$	80.98 ± 5.29
	B: Random/class norm.	77.36 ± 1.33	85.92 ± 1.98	73.85 ± 3.61	$\mathbf{82.93 \pm 6.22}$
	C: Balanced/orig.	75.85 ± 1.08	85.07 ± 1.83	73.54 ± 3.34	79.51 ± 5.62

Table 15: Top-10 ROIs identified by BrainLRR attention-based importance analysis for ABIDE ASD vs. HC classification. Importance denotes mean post-softmax attention; ×Unif. reports the score relative to the uniform baseline $1/200$.

Rank	AAL Region	HO Region	Imp.	×Unif.	MNI (x, y, z)
1	Cingulum_Ant_L	Paracingulate Gyrus	0.005066	1.013×	$(-5.2, 36.2, 21.0)$
2	Frontal_Mid_Orb_L	Frontal Pole	0.005064	1.013×	$(-28.5, 51.7, -12.3)$
3	Temporal_Inf_L	Lateral Occipital Cortex, inferior division	0.005051	1.010×	$(-53.5, -60.7, -2.8)$
4	Temporal_Sup_L	Insular Cortex	0.005051	1.010×	$(-37.5, -11.5, -4.8)$
5	Cerebelum_6_R	Temporal Fusiform Cortex, posterior division	0.005049	1.010×	$(29.0, -39.4, -31.1)$
6	Temporal_Sup_R	Superior Temporal Gyrus, anterior division	0.005048	1.010×	$(57.7, -6.4, -9.1)$
7	Occipital_Inf_L	Lateral Occipital Cortex, inferior division	0.005047	1.009×	$(-42.2, -72.7, -12.0)$
8	Frontal_Mid_R	Frontal Pole	0.005046	1.009×	$(29.9, 48.7, 26.0)$
9	Cingulum_Mid_R	Paracingulate Gyrus	0.005045	1.009×	$(3.1, 12.7, 47.2)$
10	Temporal_Mid_R	Middle Temporal Gyrus, temporooccipital part	0.005045	1.009×	$(62.6, -39.9, -8.0)$

fine-tuning pipelines for BrainLM and BrainSegFounder, including normalization, resizing, channel or depth conversion, pretrained encoding, pooling, and linear classification.

F Limitations and social impact

Our primary evaluation uses a fixed 70%/10%/20% subject partition with five training seeds; therefore, the reported mean ± standard deviation primarily reflects optimization randomness. Paired subject-level BCa intervals on split 101 quantify finite-test-cohort uncertainty conditional on that partition, but neither analysis captures variability across alternative splits, sites, or scanners. The data-scarcity experiments cover only ABIDE and ADNI NC–MCI and a selected set of models. Moreover, the heterogeneous cross-family FM benchmark cannot causally isolate the effect of pretraining, although the controlled Llama and pretrained-versus-scratch comparisons partially address this limitation. ROI-profile masking is also a training regularizer rather than true graph-node removal or a simulation of missing ROIs.

BrainLRR computes LRR using randomly shuffled mini-batches without class-balanced sampling or class-size normalization. Although single-class batches are uncommon, they provide no inter-class LRR signal. In a full 2×2 ABIDE experiment, neither balanced sampling nor class-size normalization produced an accuracy or AUROC difference with $p < 0.05$ relative to the original

Table 16: Top-10 ROIs identified by BrainLRR attention-based importance analysis for PD vs. HC classification.

Rank	AAL Region	HO Region	Imp.	×Unif.	MNI (x, y, z)
1	Cerebelum_4_5_R	Brain-Stem	0.005148	$1.030\times$	$(10.2, -43.9, -19.6)$
2	Putamen_R	Right Putamen	0.005133	$1.027\times$	$(28.0, 2.9, 1.6)$
3	Postcentral_L	Postcentral Gyrus	0.005129	$1.026\times$	$(-56.4, -9.1, 31.0)$
4	Parietal_Inf_L	Postcentral Gyrus	0.005127	$1.025\times$	$(-52.5, -28.0, 41.3)$
5	Cingulum_Ant_L	Paracingulate Gyrus	0.005124	$1.025\times$	$(-5.2, 36.2, 21.0)$
6	Insula_L	Central Opercular Cortex	0.005120	$1.024\times$	$(-47.4, 6.0, 0.8)$
7	Caudate_L	Left Thalamus	0.005119	$1.024\times$	$(-1.5, -3.5, -8.8)$
8	Frontal_Inf_Oper_L	Precentral Gyrus	0.005117	$1.023\times$	$(-54.1, 7.6, 18.4)$
9	Precuneus_L	Precuneous Cortex	0.005116	$1.023\times$	$(-7.7, -69.8, 48.9)$
10	Supp_Motor_Area_R	Precentral Gyrus	0.005113	$1.023\times$	$(1.8, -21.2, 67.8)$

Table 17: Top-10 ROIs identified by BrainLRR attention-based importance analysis for ADNI NC vs. AD classification.

Rank	AAL Region	HO Region	Imp.	×Unif.	MNI (x, y, z)
1	Frontal_Mid_Orb_R	Frontal Pole	0.005792	$1.158\times$	$(42.6, 48.0, -8.0)$
2	Temporal_Inf_L	Middle Temporal Gyrus, anterior division	0.005594	$1.119\times$	$(-51.3, -3.5, -30.3)$
3	Calcarine_L	Precuneous Cortex	0.005402	$1.080\times$	$(-10.6, -65.9, 23.4)$
4	Insula_R	Central Opercular Cortex	0.005344	$1.069\times$	$(40.9, -8.6, 11.7)$
5	Cerebelum_6_R	Unlabeled / Mixed	0.005334	$1.067\times$	$(22.2, -62.4, -28.2)$
6	Cingulum_Ant_L	Paracingulate Gyrus	0.005329	$1.066\times$	$(-5.4, 45.3, 3.9)$
7	Frontal_Mid_R	Middle Frontal Gyrus	0.005302	$1.060\times$	$(41.3, 24.8, 38.7)$
8	Frontal_Inf_Tri_R	Inferior Frontal Gyrus, pars triangularis	0.005300	$1.060\times$	$(50.6, 32.0, 4.3)$
9	Precuneus_R	Precuneous Cortex	0.005286	$1.057\times$	$(2.3, -56.6, 37.5)$
10	Frontal_Inf_Oper_R	Precentral Gyrus	0.005286	$1.057\times$	$(43.8, 9.9, 31.9)$

implementation. Component effects are also dataset-dependent: paired BCa intervals excluded zero for ROI-profile masking only on ABIDE and for LRR only on ADNI NC–AD. Additional failure case include lower AUROC than BNT on ADNI NC–AD.

Finally, BrainLRR is intended for research rather than standalone clinical diagnosis. Dataset, demographic, site, scanner, and preprocessing biases may limit generalizability, while erroneous predictions could cause harm if used without clinical oversight. External and site-held-out validation, subgroup auditing, uncertainty assessment, privacy protection, and prospective clinical evaluation are required before deployment.

Table 18: Top-10 ROIs identified by BrainLRR attention-based importance analysis for ADNI NC vs. MCI classification.

Rank	AAL Region	HO Region	Imp.	×Unif.	MNI (x, y, z)
1	Frontal_Mid_R	Frontal Pole	0.005715	$1.143\times$	$(38.4, 51.1, 9.8)$
2	Temporal_Inf_L	Middle Temporal Gyrus, anterior division	0.005554	$1.111\times$	$(-51.3, -3.5, -30.3)$
3	Angular_R	Lateral Occipital Cortex, superior division	0.005503	$1.101\times$	$(31.6, -67.6, 44.8)$
4	Temporal_Inf_L	Lateral Occipital Cortex, inferior division	0.005485	$1.097\times$	$(-53.5, -60.7, -2.8)$
5	Insula_R	Central Opercular Cortex	0.005458	$1.092\times$	$(40.9, -8.6, 11.7)$
6	Temporal_Pole_Sup_R	Frontal Orbital Cortex	0.005424	$1.085\times$	$(29.3, 11.0, -17.2)$
7	SupraMarginal_R	Supramarginal Gyrus, posterior division	0.005421	$1.084\times$	$(60.4, -33.0, 29.6)$
8	Cerebelum_6_L	Temporal Fusiform Cortex, posterior division	0.005413	$1.083\times$	$(-29.8, -41.3, -31.5)$
9	Frontal_Inf_Tri_R	Inferior Frontal Gyrus, pars triangularis	0.005404	$1.081\times$	$(50.6, 32.0, 4.3)$
10	Frontal_Mid_Orb_R	Frontal Pole	0.005395	$1.079\times$	$(42.6, 48.0, -8.0)$

(a) Zero-shot prompt

[system] You are a careful medical classification assistant. Given a summary of a subject's brain functional connectivity, classify whether the subject has Autism Spectrum Disorder (ASD) or is a neurotypical Control. Respond with exactly one word: ASD or Control. Reasoning: low.
[user] Subject summary:
<FC summary: statistics + edge polarity + top hub ROIs + strongest FC edges>
Label:

(b) Few-shot prompt with k in-context examples

[system] *Identical to the zero-shot system prompt.*
[user] Example 1
Subject summary:
<FC summary>
Label: ASD
$\cdots$ *(k examples total, class-balanced)*
Now classify the next subject.
Subject summary:
<FC summary>
Label:

(c) Fine-tuning prompt for Llama-3.1-8B with LoRA

`<|begin_of_text|><|start_header_id|>system<|end_header_id|>`
You are a careful medical classification assistant. *<disease_ctx>* Reasoning: low. `<|eot_id|>`
`<|start_header_id|>user<|end_header_id|>`
Subject summary:
<FC summary: statistics + edge polarity + top hub ROIs + strongest FC edges>
Label: `<|eot_id|>`
`<|start_header_id|>assistant<|end_header_id|>`
Target completion: *<label_pos or label_neg>* `<|eot_id|>`

Figure 7: Prompt templates used for LLM-based evaluation. (a) Zero-shot classification directly predicts the label from a text summary of the subject's FC matrix. (b) Few-shot classification prepends k class-balanced demonstration examples before the query subject. (c) Fine-tuning uses the native Llama-3.1-8B chat template with LoRA adaptation. The `disease_ctx` field and label strings are substituted per dataset as summarized in Table 19. The FC summary encoding is identical across all three settings.

Table 19: Dataset-specific label substitutions used in the Llama-3.1-8B fine-tuning prompt template

Dataset	Classification target	Positive label	Negative label
ABIDE	ASD vs. neurotypical control	ASD	Control
PD	Parkinson's disease vs. healthy control	Parkinson's Disease	Healthy Control
ADNI NC vs. AD	Alzheimer's disease vs. normal control	Alzheimer's Disease	Normal Control
ADNI NC vs. MCI	Mild cognitive impairment vs. normal control	MCI	Normal Control

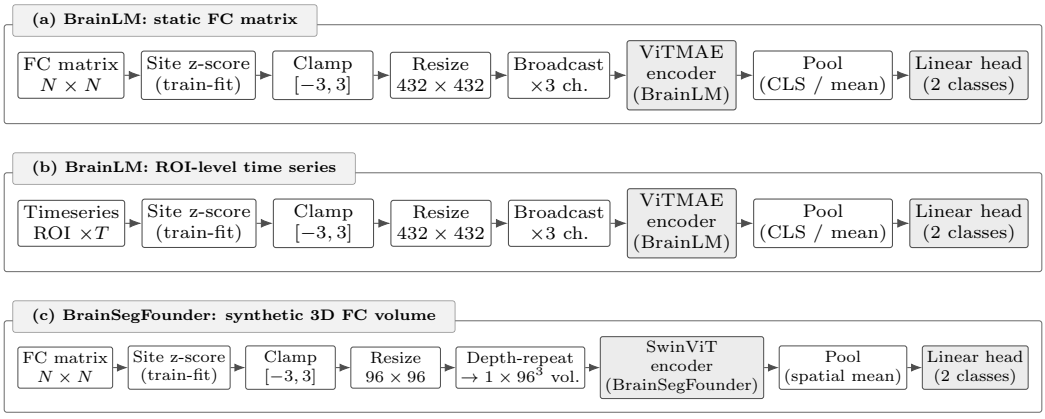


Figure 8: Input pipelines for the medical foundation-model fine-tuning baselines. Shaded boxes denote pretrained components initialized from the original checkpoints. BrainLM is evaluated with either static FC matrices or ROI-level time-series inputs, while BrainSegFounder receives a synthetic 3D volume formed by depth-repeating the 2D FC representation.