AREION: A FOUNDATION MODEL FOR ATLAS-AGNOSTIC REPRESENTATION LEARNING OF BRAIN CONNECTOME

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain foundation models have shown promising advances in learning generalizable representations from functional magnetic resonance imaging (fMRI) data. However, existing models are typically tied to specific brain atlases and cannot be applied to fMRI data represented using different atlases without full retraining, thereby severely limiting their broader utility across neuroimaging studies. To address this bottleneck, we introduce *AREION*, **a**tlas-agnostic **re**presentat**i**on learning of brain c**on**nectomes. AREION overcomes atlas dependency through two integrated innovations. First, a *Brain Region Contextualizing* technique maps brain regions and connectivity entries from diverse atlases into a unified representation space based on their spatial, functional, and morphological characteristics. Second, a *Hierarchical Brain Transformer Encoder* processes region-specific connectivity profiles as variable-length token sequences and jointly captures local functional structure and global brain functional organization, conditioned on brain regional context, without assuming a fixed number of regions. Together, these two techniques enable the learning of generalizable representations across brain atlases with different parcellation resolutions and connectivity dimensions. AREION is pretrained on a large-scale multi-atlas cohort of fMRI data via self-supervised modeling and evaluated across diverse downstream tasks. Extensive experiments show that it outperforms strong supervised baselines and fMRI foundation models while using substantially fewer parameters, demonstrating strong transferability to atlases unseen during pretraining, generalization to external datasets, and excellent interpretability at the brain regional level.

1 INTRODUCTION

Understanding the human brain is of fundamental importance to both neuroscience and clinical applications (Fornito et al., 2015; Drysdale et al., 2017). As a widely used tool in neuroscience, functional magnetic resonance imaging (fMRI) measures brain activity by capturing blood-oxygen-level-dependent (BOLD) signals. Driven by recent advances in artificial intelligence, advanced machine learning approaches have been developed for fMRI analysis (Kawahara et al., 2017; Li et al., 2021; Kan et al., 2022; Chen et al., 2026), leading to improved disease diagnosis and deeper insights into brain function organization. Despite this promise, machine learning models for fMRI analysis are often constrained by limited availability of labeled data and task-specific training paradigms, which restrict their transferability to unseen tasks and datasets. Foundation models offer a promising alternative by pretraining on large-scale unlabeled data to learn generalizable representations that can be adapted to diverse tasks, motivating a growing effort to build foundation models for neuroimaging data.

Recent studies have developed fMRI foundation models for both time-series (Ortega Caro et al., 2024; Dong et al., 2024) and functional connectivity (FC) (Yang et al., 2024; Hu et al., 2024) data. While time-series-based models learn representations directly from the temporal dynamics of fMRI signals, FC-based approaches leverage precomputed connectivity matrices to provide a compact characterization of functional interactions among brain regions, facilitating large-scale pretraining. However, the representations learned by existing FC-based models are tied to the brain atlas used. Such atlas dependence is a fundamental limitation because studies routinely differ in the atlases they

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.

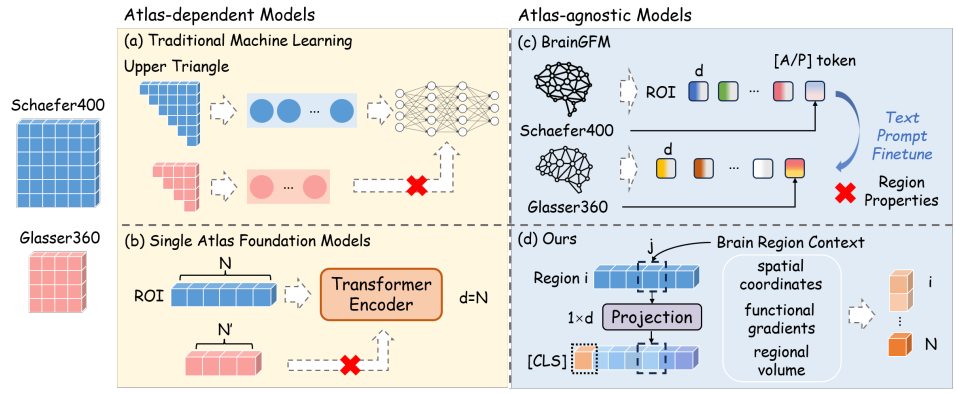


Figure 1: **Motivation.** Given inputs from distinct brain atlases, **(a)** traditional machine learning methods represent FC matrices as flattened upper triangular vectors, whose dimensions vary across atlases; **(b)** single-atlas foundation models set the model dimension to the number of regions, requiring an atlas-specific configuration; **(c)** BrainGFM enables cross-atlas adaptation by introducing language prompt tokens of atlas names but does not explicitly model the intrinsic characteristics of individual atlases; **(d)** our model maps regional connectivity profiles into a unified representation space by incorporating the spatial, functional, and morphological properties of individual brain regions, thereby enabling cross-atlas transfer.

adopt, and training a separate model for each parcellation is neither scalable nor efficient. A recent graph-based approach (Wei et al., 2026) takes a step toward addressing this limitation by pretraining across multiple brain atlases and incorporating atlas names as language prompt tokens to facilitate cross-atlas adaptation. However, its ability to generalize to fMRI data represented using unseen atlases from external datasets remains unexplored.

The key challenge in achieving atlas-agnostic representation learning is to construct representations that are transferable across distinct atlases, including those unseen during pretraining. This challenge arises from two related limitations of existing approaches. First, previous work uses textual descriptions of atlas names to derive language embeddings (Wei et al., 2026). However, these embeddings do not explicitly characterize the intrinsic spatial, functional, and morphological properties of individual brain regions of different atlases, particularly those unseen during pretraining. Building an atlas-agnostic fMRI foundation model therefore requires a more effective mechanism to encode regional properties into the learned representations. Second, existing methods typically encode the entire connectivity profile of each region as a single token (Yang et al., 2024; Hu et al., 2024). Because the model dimension depends on the number of regions defined by the adopted atlas, such a representation is tied to a specific parcellation and hinders transfer across atlases with different numbers of regions. Moreover, projecting an entire connectivity profile into a single token overlooks the rich structural dependencies among individual functional connections, limiting model interpretability. Figure 1 summarizes these limitations and the motivation for our framework.

To address this challenge, we propose AREION, an atlas-agnostic representation learning approach, and use it to build an fMRI foundation model pretrained on 26,000 resting-state fMRI samples from 11,189 unique subjects across five brain atlases. AREION incorporates two key innovations. First, we design Brain Region Contextualizing, a mechanism that encodes continuous information about each brain region by mapping its spatial, functional, and morphological metadata into a shared latent space, allowing the model to adapt to unseen parcellations at inference time without retraining. Second, we design a hierarchical encoder architecture that applies self-attention over variable-length region sequences of functional connectivity profiles rather than operating on a fixed number of regions. This design enables model training on data across multiple atlases while capturing both region-specific connectivity patterns and global brain organization. We conduct extensive experiments comparing AREION with traditional machine learning methods, supervised deep learning models, and pretrained foundation models across diverse downstream tasks, including disease and sex classification and the prediction of age and behavioral measures. Our contributions can be summarized as follows:

- We introduce Brain Region Contextualizing (BRC), a mechanism that represents brain regions using their spatial coordinates, functional gradients, and morphological properties. Unlike previous approaches that rely on atlas names to encode brain regions, BRC directly captures the intrinsic properties of individual regions and provides a unified representation space that can be directly used to characterize brain regions from previously unseen atlases.
- We develop a hierarchical Transformer architecture that separately models fine-grained, region-specific connectivity profiles and whole brain inter-regional dependencies. By treating regional connections as variable-length sequences, the architecture accommodates atlases with different numbers of regions.
- AREION outperforms strong traditional machine learning methods, supervised deep learning models, and existing SOTA fMRI foundation models across diverse downstream tasks and datasets. More importantly, it demonstrates strong transferability to brain atlases unseen during pretraining and provides interpretability at the level of individual brain regions.

2 Related Work

2.1 Brain Functional Network Analysis

Functional connectivity derived from fMRI data, which is typically estimated as the pairwise Pearson correlation between BOLD time series of regions of interest (ROIs), has been widely used as a biomarker for behavioral traits and neurological disorders. In recent years, deep learning methods have been increasingly applied to brain functional network analysis. Early approaches primarily relied on convolutional architectures, such as BrainNetCNN (Kawahara et al., 2017), which introduces specialized filters to capture the inherent structure of brain networks. Beyond CNNs, a growing number of graph neural networks (GNNs) have been developed (Li et al., 2019; Kim & Ye, 2020; Zhang et al., 2022; Chen et al., 2026), which explicitly model brain regions as nodes and their functional interactions as edges. For instance, BrainGNN (Li et al., 2021) proposes an ROI-aware graph convolutional layer to extract features from brain fMRI graphs, together with an ROI-pooling layer that selects the most informative nodes for downstream tasks. More recently, Brain Network Transformer (Kan et al., 2022) employs Transformer encoders to learn node features in brain networks, and proposes an orthonormal clustering readout layer to exploit modular-level similarities between ROIs. Despite these advances, these approaches are primarily designed for task-specific prediction, limiting the transferability of the learned representations across datasets.

2.2 Self-Supervised Learning and Masked Modeling

Self-supervised learning (SSL) has emerged as a powerful paradigm for learning robust representations from large amounts of unlabeled data through a variety of pretext tasks (Gidaris et al., 2018; He et al., 2020; Chen et al., 2020; 2021). By learning from the inherent structure of the data rather than from predefined downstream labels, SSL enables models to capture general patterns that can be transferred across tasks. Among these paradigms, masked autoencoder is a form of generative pretraining in which masked content is reconstructed from the remaining visible data. It was popularized in natural language processing by BERT (Devlin et al., 2019), and was later extended to computer vision by Masked Autoencoders (MAE) (He et al., 2022), which demonstrated that reconstructing masked image patches yields highly transferable visual representations. The success of masked modeling in the language and vision domains has motivated its adaptation to modeling of fMRI data, where large-scale unlabeled fMRI recordings provide an ideal setting for self-supervised pretraining.

2.3 fMRI Foundation Models

The success of large-scale pretraining has motivated a growing body of work on fMRI foundation models. Among them, FC-based models operate directly on precomputed FC matrices, providing a compact representation of brain functional organization. BrainMass (Yang et al., 2024) learns FC representations with a Transformer encoder using masked reconstruction and latent-alignment objectives, while BrainNPT (Hu et al., 2024) pretrains a Transformer encoder with replaced ROI prediction before the encoder is finetuned for downstream classification. Although both models

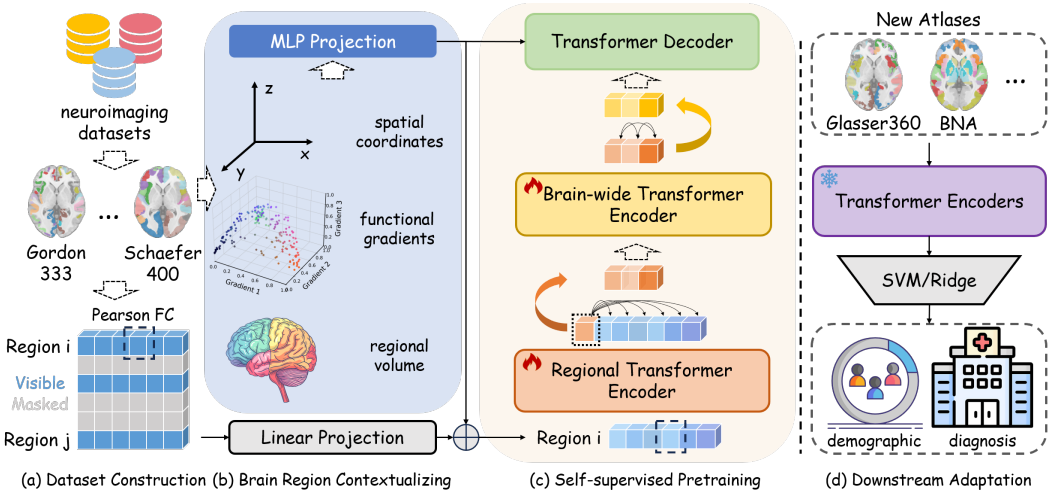


Figure 2: Overview: (a) A large-scale multi-atlas dataset is curated for pretraining; (b) Brain Region Contextualizing encodes the spatial, functional, and morphological properties of individual brain regions; (c) A hierarchical brain Transformer encoder is pretrained via masked autoencoding; (d) The pretrained Transformer encoders are frozen, and the resulting region representations are flattened into a feature vector for downstream prediction.

achieve competitive downstream performance, they are tied to fixed and predefined parcellations, limiting their direct transfer across atlases that differ in the number and definition of regions. More recently, BrainGFM (Wei et al., 2026) introduces a graph-based approach that enables pretraining across multiple brain atlases by incorporating atlas names as language prompt tokens for cross-atlas adaptation. However, the prompt tokens representing atlas names do not explicitly encode the intrinsic spatial, functional, and morphological properties of individual brain regions, motivating the need for a mechanism that maps brain regions from different atlases into a common representation space.

3 Method

To enable atlas-agnostic representation learning across different brain atlases, we propose AREION, consisting of four key components: (1) Brain Region Contextualizing to encode the spatial, functional, and morphological properties of individual brain regions; (2) Hierarchical Brain Transformer Encoder to capture both local connectivity patterns and whole brain functional organization; (3) Masked Autoencoding for self-supervised pretraining; and (4) Downstream Adaptation to adapt and evaluate the learned representations on diverse downstream tasks based on fMRI data represented using different atlases, including those unseen during pretraining. An overview of the proposed framework is illustrated in Figure 2.

3.1 Definition

Let $\mathcal{A} = \{1, \ldots, N\}$ denote a brain atlas consisting of N ROIs. For a given fMRI scan, we compute the functional connectivity (FC) matrix $\mathbf{F} \in \mathbb{R}^{N \times N}$, where each entry F_{ij} represents the Pearson correlation between the BOLD time series of regions i and j. Since different atlases yield matrices of different sizes (e.g., $N = 100, 200, 300, 400, \ldots$), we aim to learn an atlas-agnostic encoder $E_\theta : (\mathbf{F}, \{\mathbf{a}_i\}_{i=1}^N) \mapsto \{\mathbf{z}_i\}_{i=1}^N$ that maps an FC matrix and the associated atlas metadata to a set of regional representations $\mathbf{z}_i \in \mathbb{R}^d$, one for each region of a given atlas. The encoder is pretrained on unlabeled fMRI data from multiple atlases and subsequently adapted to downstream tasks.

3.2 Brain Region Contextualizing

A key challenge in atlas-agnostic representation learning of FC measures is to construct informative representations of brain regional FC profiles adaptive to different atlases. Existing approaches such

as BrainGFM incorporate atlas information through language embeddings of atlas or parcellation names. In contrast, we explicitly encode the intrinsic structural and functional properties of each region into a continuous latent space. Specifically, we describe each region i with a metadata vector $\mathbf{a}_i = [\mathbf{c}_i;\, \mathbf{g}_i;\, v_i] \in \mathbb{R}^{13}$, comprising its MNI coordinates $\mathbf{c}_i \in \mathbb{R}^6$, functional gradient projections $\mathbf{g}_i \in \mathbb{R}^6$, and regional volume $v_i \in \mathbb{R}$. Through BRC, the spatial, functional, and morphological metadata of an individual region i are mapped to a d-dimensional latent representation:

$$\mathbf{p}_i = \mathrm{MLP}([\mathbf{c}_i;\, \mathbf{g}_i;\, v_i]) \in \mathbb{R}^d, \tag{1}$$

where MLP consists of a linear projection, layer normalization, a GELU activation (Hendrycks & Gimpel, 2016), and a second linear projection with dropout. Here, $\mathbf{p}_i$ is computed from the continuous intrinsic properties of the brain region rather than from a fixed region index or an atlas-specific vocabulary, allowing the encoder to handle arbitrary atlases at inference time, and the three inputs of each region capture its intrinsic characteristics independent of atlases: 1) $\mathbf{c}_i \in \mathbb{R}^6$: the mean MNI coordinates (x, y, z) of the voxels and their corresponding standard deviations $(\sigma_x, \sigma_y, \sigma_z)$, capturing the *spatial* properties of the region; 2) $\mathbf{g}_i \in \mathbb{R}^6$: the first three functional gradient projections (g_1, g_2, g_3) and their corresponding standard deviations $(\sigma_{g_1}, \sigma_{g_2}, \sigma_{g_3})$, capturing the position of the region along the principal axes of *functional* organization in the brain; and 3) $v_i \in \mathbb{R}$: the regional volume, capturing the *morphological* properties of the region. Particularly, the functional gradients are informative for characterizing functional changes associated with age, cognition, and brain disease (Nguyen et al., 2024; Bethlehem et al., 2020). Since each functional connectivity value corresponds to the interaction between two brain regions, the BRC embeddings of regions i and j are combined additively as $\mathbf{p}_i + \mathbf{p}_j$ to encode their connectivity so that the properties of both regions are incorporated into their connectivity token.

3.3 Hierarchical Brain Transformer Encoders

Motivated by the multi-level organization of human brain networks (Meunier et al., 2009; Sporns & Betzel, 2016), we develop a hierarchical Transformer architecture to separately model regional FC profiles and whole brain FC organization, thereby preserving fine-grained regional level FC patterns while capturing global organization across atlases with varying numbers of brain regions.

3.3.1 Regional Transformer Encoder

The Regional Transformer Encoder (RTE) models the FC profile of each region by operating independently on each row of the FC matrix. For region i, its FC profile $\mathbf{F}_i = [F_{i1}, \ldots, F_{iN}] \in \mathbb{R}^N$ describes its connectivity to all other regions. We project each connectivity value F_{ij} into a d-dimensional token through a learned linear projection combined with the pairwise positional encoding:

$$\mathbf{t}_j^{(i)} = \mathbf{W}_s F_{ij} + \mathbf{p}_i + \mathbf{p}_j \in \mathbb{R}^d, \tag{2}$$

where $\mathbf{W}_s \in \mathbb{R}^{d \times 1}$ is a shared learnable projection. The token $\mathbf{t}_j^{(i)}$ thus encodes both the strength of the connection (F_{ij}) and the properties of its two connected regions. A learnable CLS token $\mathbf{e}_{\mathrm{cls}}$ is then prepended to the sequence to aggregate information across all connectivity tokens through self-attention. The resulting sequence is processed by a Transformer encoder:

$$[\mathbf{h}_i,\, \mathbf{T}'^{(i)}] = \mathrm{Transformer}\left([\mathbf{e}_{\mathrm{cls}},\, \mathbf{T}^{(i)}]\right), \tag{3}$$

where $\mathbf{T}^{(i)} = [\mathbf{t}_1^{(i)}, \ldots, \mathbf{t}_N^{(i)}] \in \mathbb{R}^{N \times d}$. The CLS output $\mathbf{h}_i \in \mathbb{R}^d$ serves as the region-level representation that summarizes the connectivity profile of region i.

3.3.2 Brain-wide Transformer Encoder

The RTE models the FC profile of each region independently but does not explicitly capture whole brain functional organization. To address this issue, we develop a Brain-wide Transformer Encoder (BTE) to jointly model all regional tokens for characterizing the global organization of the functional brain network.

The BTE takes the region-level representations $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]$ as input and adds the corresponding BRC embeddings $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_N]$ to form enriched input tokens. A deeper Transformer

encoder then processes this sequence through global self-attention across all brain regions:

$$\mathbf{Z} = \text{Transformer}\left(\mathbf{H} + \mathbf{P}\right), \tag{4}$$

where $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N] \in \mathbb{R}^{N \times d}$ denotes the resulting context-enriched region tokens. These representations capture both region-specific connectivity profiles and whole-brain functional organization and are used as the learned brain representations for downstream tasks.

3.4 Masked Autoencoding

We pretrain the Transformer encoders using a masked autoencoding paradigm. Given N regions, we randomly sample a visible subset $\mathcal{V} \subset \{1, \ldots, N\}$, where $|\mathcal{V}| = (1 - r)N$ and r denotes the masking ratio. The remaining regions, $\mathcal{M} = \{1, \ldots, N\} \setminus \mathcal{V}$, are masked. The visible regions are then processed by the hierarchical Transformer encoders to obtain regional representations $\{\mathbf{z}_i\}_{i \in \mathcal{V}}$, while the masked regions are excluded from the encoder input. This asymmetric design substantially reduces the computational cost of pretraining because the encoders process only a fraction, $1 - r$, of the N regions.

For reconstruction, each masked region is represented by a learnable mask embedding $\mathbf{e}_{\text{mask}} \in \mathbb{R}^d$. The corresponding BRC embedding $\mathbf{p}_i$ is added to the learnable embedding so that the decoder retains the spatial, functional, and morphological identity of the region even when the connectivity information is unavailable. The decoder input for region i is defined as:

$$\mathbf{x}_i = \begin{cases} \mathbf{z}_i + \mathbf{p}_i, & i \in \mathcal{V} \\ \mathbf{e}_{\text{mask}} + \mathbf{p}_i, & i \in \mathcal{M} \end{cases} \tag{5}$$

A Transformer decoder processes the complete sequence $\{\mathbf{x}_i\}_{i=1}^N$ through global self-attention. This allows masked regions to aggregate contextual information from visible region tokens, producing decoded representations $\tilde{\mathbf{z}}_i$ for all N regions. Because the FC matrix $\mathbf{F}$ is computed using Pearson correlations, which correspond to normalized inner products between regional BOLD time series, we reconstruct each element of $\mathbf{F}$ using a matching inner-product readout:

$$\hat{F}_{ij} = \frac{\left\langle \phi(\tilde{\mathbf{z}}_i),\, \phi(\tilde{\mathbf{z}}_j) \right\rangle}{d - 1}, \qquad \phi(\tilde{\mathbf{z}}) = \frac{\mathbf{W}_r \tilde{\mathbf{z}} - \mu(\mathbf{W}_r \tilde{\mathbf{z}})}{\sigma(\mathbf{W}_r \tilde{\mathbf{z}})}, \tag{6}$$

where $\mathbf{W}_r \in \mathbb{R}^{d \times d}$ is a learnable projection matrix, and $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and sample standard deviation computed across the d channels, respectively. The reconstruction loss is computed over the masked rows of FC matrix:

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{M}|N} \sum_{i \in \mathcal{M}} \sum_{j=1}^{N} (\hat{F}_{ij} - F_{ij})^2. \tag{7}$$

3.5 Downstream Adaptation and Applications

For downstream evaluation, we freeze the pretrained Transformer encoders and extract regional representations from FC matrices defined using any atlases. For a given input FC matrix with N regions, the encoders produce N region tokens of dimension d. These tokens are concatenated and flattened into an $N \times d$ feature vector, which is fed into a task-specific predictor: a support vector machine (SVM) for classification tasks or a ridge regressor for regression tasks. Although regional representations from atlases with varying numbers of regions can be integrated using Transformers, the present study focuses on evaluating each atlas separately using traditional machine learning models that require fixed-dimensional feature vectors.

4 Experiments

4.1 Datasets and Evaluation

Datasets. We pretrain our model on three fMRI datasets: the Human Connectome Project (HCP) (Van Essen et al., 2013), Human Connectome Project-Development (HCP-D) (Somerville et al.,

Table 1: Performance comparison across downstream tasks on the Schaefer400 atlas: sex classification (Sex), schizophrenia classification (SCHZ), age regression (Age), and executive function prediction (Exec_Func). The best and second-best results are highlighted in **bold** and underlined, respectively. ML, DL, and PT denote traditional machine learning methods, supervised deep learning models, and pretrained foundation models, respectively. **Note:** BrainGFM (Single) and AREION (Single) are pretrained using only the Schaefer400 atlas, whereas BrainGFM (Multi) and AREION (Multi) are pretrained using multiple atlases.

Type	Method	HCP (Sex)			PNC (Sex)			SCZ (SCHZ)		
		ACC ↑	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑
ML	MLP	$85.28_{\pm 2.53}$	$92.99_{\pm 1.55}$	$85.17_{\pm 2.61}$	$78.50_{\pm 3.36}$	$87.87_{\pm 2.67}$	$78.21_{\pm 3.38}$	$66.08_{\pm 8.46}$	$73.40_{\pm 6.79}$	$65.47_{\pm 8.27}$
	SVM	$88.50_{\pm 1.87}$	$95.89_{\pm 0.44}$	$88.41_{\pm 1.89}$	$83.69_{\pm 1.94}$	$92.09_{\pm 1.00}$	$83.05_{\pm 2.17}$	$71.55_{\pm 3.73}$	$79.30_{\pm 3.08}$	$70.65_{\pm 3.42}$
	RF	$76.79_{\pm 3.12}$	$86.69_{\pm 2.47}$	$76.43_{\pm 3.20}$	$73.59_{\pm 2.43}$	$83.83_{\pm 2.55}$	$70.56_{\pm 3.06}$	$69.26_{\pm 4.07}$	$74.81_{\pm 2.92}$	$68.29_{\pm 4.70}$
DL	BrainNetCNN	$91.09_{\pm 1.65}$	$96.94_{\pm 0.81}$	$91.03_{\pm 1.66}$	$82.98_{\pm 2.32}$	$93.09_{\pm 0.42}$	$82.89_{\pm 2.25}$	$67.39_{\pm 5.38}$	$79.03_{\pm 3.03}$	$64.39_{\pm 8.75}$
	BrainGNN	$79.19_{\pm 6.33}$	$92.52_{\pm 1.80}$	$78.24_{\pm 6.92}$	$80.66_{\pm 2.10}$	$89.79_{\pm 1.39}$	$79.85_{\pm 1.95}$	$49.51_{\pm 3.87}$	$74.62_{\pm 2.96}$	$35.33_{\pm 4.61}$
	BrainNetTF	$92.75_{\pm 0.73}$	$\underline{98.05_{\pm 0.61}}$	$92.72_{\pm 0.73}$	$85.29_{\pm 3.87}$	$\mathbf{95.85_{\pm 0.98}}$	$\underline{84.72_{\pm 4.63}}$	$65.15_{\pm 11.67}$	$\mathbf{83.75_{\pm 4.65}}$	$60.73_{\pm 15.51}$
PT	BrainMass	$85.39_{\pm 1.52}$	$93.35_{\pm 0.99}$	$85.29_{\pm 1.57}$	$79.22_{\pm 3.10}$	$87.86_{\pm 1.86}$	$78.38_{\pm 3.57}$	$68.74_{\pm 7.04}$	$76.68_{\pm 4.28}$	$67.73_{\pm 0.77}$
	BrainNPT	$86.32_{\pm 1.75}$	$93.90_{\pm 1.63}$	$86.23_{\pm 1.76}$	$82.25_{\pm 3.97}$	$91.04_{\pm 1.63}$	$81.40_{\pm 5.02}$	$71.33_{\pm 6.57}$	$79.62_{\pm 3.97}$	$69.98_{\pm 8.11}$
	BrainGFM (Single)	$91.92_{\pm 0.90}$	$96.79_{\pm 1.11}$	$91.84_{\pm 0.96}$	$82.61_{\pm 2.12}$	$90.30_{\pm 1.77}$	$81.86_{\pm 2.25}$	$68.18_{\pm 9.09}$	$77.89_{\pm 7.40}$	$67.09_{\pm 9.01}$
	BrainGFM (Multi)	$92.12_{\pm 0.51}$	$97.03_{\pm 0.65}$	$92.06_{\pm 0.47}$	$\mathbf{85.51_{\pm 0.28}}$	$90.24_{\pm 0.68}$	$\mathbf{85.13_{\pm 3.09}}$	$70.45_{\pm 3.69}$	$72.93_{\pm 6.18}$	$70.44_{\pm 3.28}$
	AREION (Single)	$\underline{93.00_{\pm 1.33}}$	$97.93_{\pm 0.46}$	$\underline{92.96_{\pm 1.33}}$	$84.70_{\pm 1.98}$	$94.96_{\pm 0.94}$	$83.91_{\pm 2.29}$	$\underline{72.47_{\pm 3.24}}$	$81.89_{\pm 3.18}$	$\underline{71.25_{\pm 3.66}}$
	AREION (Multi)	$\mathbf{93.16_{\pm 1.52}}$	$\mathbf{98.15_{\pm 0.39}}$	$\mathbf{93.11_{\pm 1.54}}$	$\underline{85.42_{\pm 1.81}}$	$\underline{95.04_{\pm 0.65}}$	$84.60_{\pm 2.20}$	$\mathbf{74.30_{\pm 3.43}}$	$\underline{82.45_{\pm 2.74}}$	$\mathbf{73.31_{\pm 4.24}}$

Type	Method	HCP (Age)			PNC (Age)			PNC (Exec_Func)		
		R^2 ↑	MAE ↓	ρ ↑	R^2 ↑	MAE ↓	ρ ↑	R^2 ↑	MAE ↓	ρ ↑
ML	MLP	$0.080_{\pm 0.068}$	$2.937_{\pm 0.088}$	$0.327_{\pm 0.064}$	$0.443_{\pm 0.044}$	$1.964_{\pm 0.122}$	$0.680_{\pm 0.030}$	$0.289_{\pm 0.075}$	$0.588_{\pm 0.053}$	$0.566_{\pm 0.057}$
	Ridge	$0.231_{\pm 0.025}$	$2.662_{\pm 0.060}$	$0.491_{\pm 0.019}$	$0.630_{\pm 0.033}$	$1.583_{\pm 0.067}$	$0.799_{\pm 0.019}$	$0.372_{\pm 0.022}$	$0.559_{\pm 0.032}$	$0.613_{\pm 0.019}$
	RF	$0.076_{\pm 0.010}$	$3.000_{\pm 0.081}$	$0.391_{\pm 0.049}$	$0.277_{\pm 0.040}$	$2.259_{\pm 0.104}$	$0.637_{\pm 0.065}$	$0.179_{\pm 0.023}$	$0.646_{\pm 0.041}$	$0.543_{\pm 0.043}$
DL	BrainNetCNN	$0.258_{\pm 0.027}$	$2.602_{\pm 0.039}$	$0.513_{\pm 0.023}$	$0.607_{\pm 0.037}$	$1.608_{\pm 0.074}$	$0.781_{\pm 0.025}$	$0.367_{\pm 0.029}$	$0.560_{\pm 0.039}$	$0.608_{\pm 0.023}$
	BrainGNN	$0.085_{\pm 0.029}$	$2.904_{\pm 0.124}$	$0.330_{\pm 0.028}$	$0.453_{\pm 0.017}$	$1.931_{\pm 0.060}$	$0.681_{\pm 0.016}$	$0.215_{\pm 0.040}$	$0.622_{\pm 0.027}$	$0.502_{\pm 0.046}$
	BrainNetTF	$0.224_{\pm 0.031}$	$2.692_{\pm 0.050}$	$0.480_{\pm 0.029}$	$0.628_{\pm 0.027}$	$1.577_{\pm 0.075}$	$0.794_{\pm 0.017}$	$0.381_{\pm 0.015}$	$\mathbf{0.549_{\pm 0.027}}$	$0.620_{\pm 0.010}$
PT	BrainMass	$0.100_{\pm 0.026}$	$2.898_{\pm 0.060}$	$0.334_{\pm 0.040}$	$0.521_{\pm 0.044}$	$1.819_{\pm 0.109}$	$0.729_{\pm 0.032}$	$0.324_{\pm 0.048}$	$0.579_{\pm 0.040}$	$0.577_{\pm 0.038}$
	BrainNPT	$0.131_{\pm 0.018}$	$2.811_{\pm 0.120}$	$0.408_{\pm 0.033}$	$0.612_{\pm 0.016}$	$1.597_{\pm 0.083}$	$0.783_{\pm 0.010}$	$0.329_{\pm 0.055}$	$0.580_{\pm 0.040}$	$0.593_{\pm 0.033}$
	BrainGFM (Single)	$0.160_{\pm 0.059}$	$2.754_{\pm 0.100}$	$0.431_{\pm 0.056}$	$0.466_{\pm 0.081}$	$1.902_{\pm 0.104}$	$0.684_{\pm 0.059}$	$0.225_{\pm 0.048}$	$0.612_{\pm 0.027}$	$0.496_{\pm 0.033}$
	BrainGFM (Multi)	$0.183_{\pm 0.044}$	$2.713_{\pm 0.081}$	$0.476_{\pm 0.030}$	$0.482_{\pm 0.055}$	$1.842_{\pm 0.033}$	$0.699_{\pm 0.036}$	$0.279_{\pm 0.052}$	$0.598_{\pm 0.051}$	$0.546_{\pm 0.043}$
	AREION (Single)	$\underline{0.287_{\pm 0.020}}$	$\underline{2.565_{\pm 0.064}}$	$\underline{0.542_{\pm 0.018}}$	$\mathbf{0.647_{\pm 0.023}}$	$\mathbf{1.550_{\pm 0.053}}$	$\underline{0.816_{\pm 0.012}}$	$\underline{0.383_{\pm 0.037}}$	$0.554_{\pm 0.042}$	$\underline{0.623_{\pm 0.030}}$
	AREION (Multi)	$\mathbf{0.298_{\pm 0.020}}$	$\mathbf{2.547_{\pm 0.068}}$	$\mathbf{0.553_{\pm 0.018}}$	$\underline{0.640_{\pm 0.025}}$	$\underline{1.570_{\pm 0.065}}$	$\mathbf{0.817_{\pm 0.014}}$	$\mathbf{0.392_{\pm 0.036}}$	$\underline{0.549_{\pm 0.044}}$	$\mathbf{0.630_{\pm 0.030}}$

2018), and Adolescent Brain Cognitive Development (ABCD) (Casey et al., 2018), with their fMRI data processed using multiple atlases. For external evaluation, we use the Philadelphia Neurodevelopmental Cohort (PNC) dataset (Satterthwaite et al., 2014), with its fMRI data processed using two atlases unseen during pretraining: Glasser360 atlas (Glasser et al., 2016) and the Brainnetome Atlas (BNA; 210 regions) (Fan et al., 2016). Additional dataset details are provided in Appendix A.

Evaluation Metrics and Interpretability Analysis. We evaluate classification performance using accuracy (ACC), area under curve (AUC), and macro-F1 score, and regression performance using the coefficient of determination (R^2), mean absolute error (MAE), and Pearson correlation coefficient (ρ). Interpretability analyses of the learned FC representations are provided in Appendix F.

4.2 Main Results

4.2.1 Comparison with Other Methods Using FC from the Same Atlases

We compare AREION with traditional machine learning (ML) methods, supervised deep learning (DL) models, and pretrained fMRI foundation models. For a fair comparison, all foundation models are retrained on our collected data, and all methods are evaluated using the same five-fold cross-validation. Because BrainGFM and AREION support multi-atlas pretraining, we evaluate both methods under single-atlas and multi-atlas settings. The single-atlas setting is consistent with that used for the other foundation model baselines. Additional details about baselines are provided in Appendix E.

As shown in Table 1, traditional ML methods are highly competitive. SVM and ridge regression outperform several supervised DL and pretrained foundation models across multiple tasks. For example, SVM outperforms BrainGNN and BrainMass on both sex and schizophrenia classification tasks, highlighting the importance of including strong traditional ML baselines. Under single-

Table 2: Generalization to unseen brain atlases. We evaluate the generalizability of AREION, SVM, ridge regression, and BrainGFM on the external PNC dataset using atlases unseen during pretraining.

Sex Classification						
Method	Glasser360			BNA		
	ACC ↑	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑
SVM	$83.55_{\pm 1.33}$	$91.47_{\pm 0.64}$	$82.86_{\pm 1.66}$	$82.25_{\pm 2.21}$	$92.00_{\pm 0.89}$	$81.48_{\pm 2.66}$
BrainGFM	$85.61_{\pm 0.99}$	$91.48_{\pm 0.84}$	$84.96_{\pm 1.05}$	$84.17_{\pm 3.38}$	$93.57_{\pm 1.99}$	$83.55_{\pm 3.34}$
AREION	$\mathbf{86.43_{\pm 2.32}}$	$\mathbf{95.11_{\pm 0.86}}$	$\mathbf{85.66_{\pm 2.69}}$	$\mathbf{85.27_{\pm 3.35}}$	$\mathbf{94.66_{\pm 0.67}}$	$\mathbf{84.39_{\pm 3.99}}$
Age Regression						
Method	Glasser360			BNA		
	R^2 ↑	MAE ↓	ρ ↑	R^2 ↑	MAE ↓	ρ ↑
Ridge	$0.636_{\pm 0.034}$	$1.571_{\pm 0.087}$	$0.802_{\pm 0.023}$	$0.620_{\pm 0.022}$	$1.596_{\pm 0.082}$	$0.794_{\pm 0.011}$
BrainGFM	$0.503_{\pm 0.062}$	$1.833_{\pm 0.111}$	$0.712_{\pm 0.043}$	$0.517_{\pm 0.045}$	$1.782_{\pm 0.149}$	$0.720_{\pm 0.032}$
AREION	$\mathbf{0.654_{\pm 0.024}}$	$\mathbf{1.533_{\pm 0.072}}$	$\mathbf{0.819_{\pm 0.015}}$	$\mathbf{0.653_{\pm 0.020}}$	$\mathbf{1.521_{\pm 0.062}}$	$\mathbf{0.817_{\pm 0.012}}$

atlas pretraining, AREION outperforms all foundation models evaluated under the same setting. BrainGFM and AREION also generally benefit from multi-atlas pretraining, with AREION achieving the strongest overall performance across the evaluated downstream tasks. These results support the potential of multi-atlas pretraining to learn more generalizable representations. Additional comparisons across atlas resolutions are provided in Appendix B.

4.2.2 Generalization to Unseen Atlases

To evaluate AREION's ability to generalize to unseen atlases, we conduct experiments on the external PNC dataset, with its fMRI data processed using the Glasser360 and Brainnetome atlases, neither of which is used during pretraining. We compare AREION with BrainGFM, which also supports multi-atlas pretraining. Both models are pretrained on our collected cohort. As shown in Table 2, AREION consistently outperforms both the conventional ML baselines and BrainGFM on both unseen atlases, demonstrating strong cross-atlas transferability. In contrast to BrainGFM, which encodes textual descriptions of atlas names, AREION uses Brain Region Contextualizing to directly encode the intrinsic spatial, functional, and morphological properties of individual brain regions. The results support the effectiveness of this regional contextualization in learning representations that transfer reliably to previously unseen atlases.

4.3 Ablation Studies

4.3.1 Ablation Study of Positional Encoding Methods

To investigate the contribution of BRC to cross-atlas generalization, we compare it with two commonly used positional encoding (PE) methods: sinusoidal positional encoding (Vaswani et al., 2017) and relative positional encoding (Shaw et al., 2018). All variants use the same hierarchical encoder architecture and are independently pretrained on the HCP dataset using the Schaefer100/200/300/400 atlases. We evaluate these variants on FC measures derived from the Schaefer400 atlas, which is seen during pretraining, and the Gordon333 atlas, which is unseen during pretraining.

As shown in Table 3, all positional encoding methods substantially outperform the model without positional encoding, demonstrating the importance of positional information. On the seen Schaefer400 atlas, BRC achieves the best performance across all metrics. Although sinusoidal PE achieves comparable results in this setting, its performance declines notably when transferred to the Gordon333 atlas, whereas BRC maintains more stable performance. These results indicate that, unlike conventional PE methods that rely on absolute or relative sequence positions, BRC encodes continuous spatial, functional, and morphological properties to produce more transferable representations of brain regions, particularly for generalization to unseen atlases.

Table 3: Ablation study of different positional encoding methods. All variants are independently pretrained on the HCP dataset and evaluated for sex classification using a seen atlas, Schaefer400, and an unseen atlas, Gordon333, during pretraining.

Positional Encoding	Schaefer400			Gordon333		
	ACC ↑	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑
w/o PE	$75.03_{\pm 1.20}$	$84.02_{\pm 2.29}$	$74.67_{\pm 1.23}$	$72.85_{\pm 2.63}$	$82.04_{\pm 2.45}$	$72.37_{\pm 2.63}$
Sinusoidal PE	$91.92_{\pm 2.16}$	$97.48_{\pm 0.59}$	$91.86_{\pm 2.18}$	$88.81_{\pm 1.72}$	$95.17_{\pm 0.37}$	$88.75_{\pm 1.73}$
Relative PE	$90.78_{\pm 0.39}$	$96.50_{\pm 0.61}$	$90.72_{\pm 0.41}$	$86.22_{\pm 1.49}$	$94.38_{\pm 0.47}$	$86.13_{\pm 1.49}$
BRC (Ours)	$\mathbf{92.96_{\pm 1.24}}$	$\mathbf{97.91_{\pm 0.45}}$	$\mathbf{92.94_{\pm 1.25}}$	$\mathbf{91.61_{\pm 1.20}}$	$\mathbf{97.17_{\pm 0.60}}$	$\mathbf{91.56_{\pm 1.19}}$

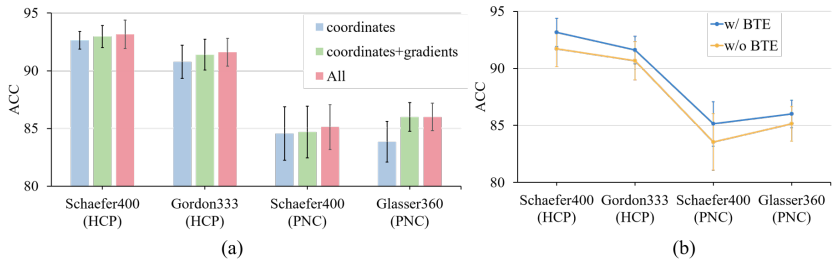


Figure 3: Ablation study of framework components. **(a)** Effects of the spatial coordinates, functional gradients, and regional volume of Brain Region Contextualizing. **(b)** Effect of Brain-wide Transformer Encoder. All models are evaluated on the sex classification task using HCP with the seen Schaefer400 atlas and PNC with the unseen Gordon333 and Glasser360 atlases during pretraining.

4.3.2 Ablation Study on Framework Components

We conduct ablation studies to investigate the contributions of the proposed framework components. All model variants are pretrained on the HCP dataset and evaluated under both in-distribution and cross-atlas settings. We first examine the contributions of the three components of BRC: spatial coordinates, functional gradients, and regional volume. As illustrated in Figure 3(a), incorporating functional gradients substantially improves performance, while the subsequent addition of regional volume consistently yields the highest overall accuracy. These results suggest that the three components provide complementary information about regional identity, enabling the model to generalize more effectively across distinct atlases.

Next, we compare AREION with its ablated version in which the Brain-wide Transformer Encoder is removed. As shown in Figure 3(b), the model with the BTE consistently outperforms its counterpart without BTE, demonstrating the effectiveness of the proposed hierarchical representation learning strategy. Without the BTE, the learned representations primarily capture region-specific connectivity patterns. Incorporating the BTE further models inter-regional dependencies and enriches the regional representations with global brain context.

5 Conclusion

In this paper, we introduce AREION, an fMRI foundation modeling approach for atlas-agnostic representation learning of brain connectomes. To overcome the dependence of existing approaches on fixed brain atlases, we propose two complementary techniques: Brain Region Contextualizing, which represents brain regions and their connectivity entries using their intrinsic spatial, functional, and morphological characteristics; and a Hierarchical Brain Transformer Encoder, which captures region-specific connectivity patterns and global inter-regional topology from variable-length connectivity sequences. Extensive experiments demonstrate that AREION achieves superior performance across diverse downstream tasks, generalizes effectively to atlases unseen during pretraining, and produces representations with biologically meaningful interpretability at the regional level.

AI USE STATEMENT

In this work, we used generative AI tools for aiding and polishing writing, as well as grammar checking. We have not used generative AI tools for research ideation or execution, drafting sections of the paper, generating synthetic datasets, or proving mathematical claims, and the remaining generative tasks are not applicable to this work. We have reviewed all AI-assisted work. Specifically, we manually reviewed all polished text to make sure that it is consistent with the content of our work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The implementation details of AREION are provided in Appendix C. The code and pretrained weights of AREION will be made publicly available upon acceptance.

REFERENCES

Richard AI Bethlehem, Casey Paquola, Jakob Seidlitz, Lisa Ronan, Boris Bernhardt, Kamen A Tsvetanov, Cam-CAN Consortium, et al. Dispersion of functional gradients across the adult lifespan. *Neuroimage*, 222:117299, 2020.

Betty Jo Casey, Tariq Cannonier, May I Conley, Alexandra O Cohen, Deanna M Barch, Mary M Heitzeg, Mary E Soules, Theresa Teslovich, Danielle V Dellarco, Hugh Garavan, et al. The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, 32:43–54, 2018.

Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2004–2013, 2021.

Jianzhong Chen, Angela Tam, Valeria Kebets, Csaba Orban, Leon Qi Rong Ooi, Christopher L Asplund, Scott Marek, Nico UF Dosenbach, Simon B Eickhoff, Danilo Bzdok, et al. Shared and unique brain network features predict cognitive, personality, and mental health scores in the abcd study. *Nature communications*, 13(1):2217, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607, 2020.

Tingting Chen, Hongming Li, Hao Zheng, and Yong Fan. dfcexpert: Learning dynamic functional connectivity patterns with modularity and state experts. *IEEE Transactions on Medical Imaging*, 45(3):1088–1098, 2026.

Zaixu Cui, Hongming Li, Cedric H Xia, Bart Larsen, Azeez Adebimpe, Graham L Baum, Matt Cieslak, Raquel E Gur, Ruben C Gur, Tyler M Moore, et al. Individual variation in functional topography of association networks in youth. *Neuron*, 106(2):340–353, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan T Nguyen, Joanna S Chong, Fang Ji, Nathanael R Tong, Christopher L Chen, and Juan H Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024.

Nico UF Dosenbach, Binyam Nardos, Alexander L Cohen, Damien A Fair, Jonathan D Power, Jessica A Church, Steven M Nelson, Gagan S Wig, Alecia C Vogel, Christina N Lessov-Schlaggar, et al. Prediction of individual brain maturity using fmri. *Science*, 329(5997):1358–1361, 2010.

Andrew T Drysdale, Logan Grosenick, Jonathan Downar, Katharine Dunlop, Farrokh Mansouri, Yue Meng, Robert N Fetcho, Benjamin Zebley, Desmond J Oathes, Amit Etkin, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature medicine*, 23 (1):28–38, 2017.

Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

Eric Feczko, Greg Conan, Scott Marek, Brenden Tervo-Clemmens, Michaela Cordova, Olivia Doyle, Eric Earl, Anders Perrone, Darrick Sturgeon, Rachel Klein, et al. Adolescent brain cognitive development (abcd) community mri collection and utilities. *BioRxiv*, pp. 2021–07, 2021.

Alex Fornito, Andrew Zalesky, and Michael Breakspear. The connectomics of brain disorders. *Nature reviews neuroscience*, 16(3):159–172, 2015.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.

Evan M Gordon, Timothy O Laumann, Babatunde Adeyemo, Jeremy F Huckins, William M Kelley, and Steven E Petersen. Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral cortex*, 26(1):288–303, 2016.

Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Jinlong Hu, Yangmin Huang, Nan Wang, and Shoubin Dong. Brainnpt: Pre-training transformer networks for brain network classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:2727–2736, 2024.

Rixing Jing, Peng Li, Zengbo Ding, Xiao Lin, Rongjiang Zhao, Le Shi, Hao Yan, Jinmin Liao, Chuanjun Zhuo, Lin Lu, et al. Machine learning identifies unaffected first-degree relatives with functional network patterns and cognitive impairment similar to those of schizophrenia patients. *Human brain mapping*, 40(13):3930–3939, 2019.

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

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.

Byung-Hoon Kim and Jong Chul Ye. Understanding graph isomorphism network for rs-fmri functional connectivity analysis. *Frontiers in neuroscience*, 14:630, 2020.

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

Hongming Li, Matthew Cieslak, Taylor Salo, Russell T Shinohara, Desmond J Oathes, Christos Davatzikos, Theodore D Satterthwaite, and Yong Fan. Uncovering functional connectivity patterns predictive of cognition in youth using interpretable predictive modeling. *Proceedings of the National Academy of Sciences*, 122(42):e2505600122, 2025.

Hongming Li, Zaixu Cui, Matthew Cieslak, Taylor Salo, Tyler M Moore, Raquel E Gur, Ruben C Gur, Russell T Shinohara, Desmond J Oathes, Christos Davatzikos, et al. Spatial heterogeneity and subtypes of functional connectivity development in youth. *Nature Communications*, 17(1): 1956, 2026.

Xiaoxiao Li, Nicha C Dvornek, Yuan Zhou, Juntang Zhuang, Pamela Ventola, and James S Duncan. Graph neural network for interpreting task-fmri biomarkers. In *International conference on medical image computing and computer-assisted intervention*, pp. 485–493, 2019.

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.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Audrey C Luo, Valerie J Sydnor, Adam Pines, Bart Larsen, Aaron F Alexander-Bloch, Matthew Cieslak, Sydney Covitz, Andrew A Chen, Nathalia Bianchini Esper, Eric Feczko, et al. Functional connectivity development along the sensorimotor-association axis enhances the cortical hierarchy. *Nature communications*, 15(1):3511, 2024.

David Meunier, Renaud Lambiotte, Alex Fornito, Karen Ersche, and Edward T Bullmore. Hierarchical modularity in human brain functional networks. *Frontiers in neuroinformatics*, 3:571, 2009.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

Thuan Tinh Nguyen, Xing Qian, Eric Kwun Kei Ng, Marcus Qin Wen Ong, Zhen Ming Ngoh, Shayne SP Yeo, Jia Ming Lau, Ai Peng Tan, Birit FP Broekman, Evelyn C Law, et al. Variations in cortical functional gradients relate to dimensions of psychopathology in preschool children. *Journal of the American Academy of Child & Adolescent Psychiatry*, 63(1):80–89, 2024.

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.

Raimon HR Pruim, Maarten Mennes, Daan Van Rooij, Alberto Llera, Jan K Buitelaar, and Christian F Beckmann. Ica-aroma: A robust ica-based strategy for removing motion artifacts from fmri data. *Neuroimage*, 112:267–277, 2015.

Theodore D Satterthwaite, Mark A Elliott, Kosha Ruparel, James Loughead, Karthik Prabhakaran, Monica E Calkins, Ryan Hopson, Chad Jackson, Jack Keefe, Marisa Riley, et al. Neuroimaging of the philadelphia neurodevelopmental cohort. *Neuroimage*, 86:544–553, 2014.

Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, 2018.

Leah H Somerville, Susan Y Bookheimer, Randy L Buckner, Gregory C Burgess, Sandra W Curtiss, Mirella Dapretto, Jennifer Stine Elam, Michael S Gaffrey, Michael P Harms, Cynthia Hodge, et al. The lifespan human connectome project in development: A large-scale study of brain connectivity development in 5–21 year olds. *Neuroimage*, 183:456–468, 2018.

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

Chandra Sripada, Saige Rutherford, Mike Angstadt, Wesley K Thompson, Monica Luciana, Alexander Weigard, Luke H Hyde, and Mary Heitzeg. Prediction of neurocognition in youth from resting state fmri. *Molecular psychiatry*, 25(12):3413–3421, 2020.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

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.

Xinxu Wei, kanhao zhao, Yong Jiao, Lifang He, and Yu Zhang. A brain graph foundation model: Pre-training and prompt-tuning across broad atlases and disorders. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yanwu Yang, Chenfei Ye, Guinan Su, Ziyao Zhang, Zhikai Chang, Hairui Chen, Piu Chan, Yue Yu, and Ting Ma. Brainmass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. *IEEE transactions on medical imaging*, 43(11):4004–4016, 2024.

Hao Zhang, Ran Song, Liping Wang, Lin Zhang, Dawei Wang, Cong Wang, and Wei Zhang. Classification of brain disorders in rs-fmri via local-to-global graph neural networks. *IEEE transactions on medical imaging*, 42(2):444–455, 2022.

Chuanjun Zhuo, Jiajia Zhu, Chunli Wang, Hongru Qu, Xiaolei Ma, Hongjun Tian, Mei Liu, and Wen Qin. Brain structural and functional dissociated patterns in schizophrenia. *BMC psychiatry*, 17(1):45, 2017.

A Dataset Details and Preprocessing

Human Connectome Project (HCP). The HCP dataset contains fMRI scans of healthy subjects aged 22–37 years. Details of the acquisition protocols have been published previously (Glasser et al., 2013). The minimally preprocessed outputs from the HCP pipeline were postprocessed using XCP-D. The processed fMRI scans of 965 subjects were included in this study. For each subject, two resting-state fMRI sessions were included, and two runs with left-to-right and right-to-left phase encoding were concatenated to compute the functional connectivity for each session.

Human Connectome Project-Development (HCP-D). The HCP-D dataset includes subjects recruited across four sites in the United States, to study healthy brain development in children and adolescents. The image acquisition and preprocessing procedures have been described previously (Luo et al., 2024). fMRI scans from 632 subjects were included in this study.

Adolescent Brain Cognitive Development (ABCD). The ABCD dataset contains fMRI scans from participants aged 9–10 years at baseline, collected across 21 sites in the United States. Resting-state fMRI scans of $9,592$ participants preprocessed by the ABCD BIDS Community Collection (ABCC) (Feczko et al., 2021) were used in this study.

Philadelphia Neurodevelopmental Cohort (PNC). The PNC dataset comprises a community-based sample of children and adolescents from the greater Philadelphia area. We included demographic measures, cognitive assessments, and fMRI scans from 693 participants (301 males, 392 females) aged 8–23 years. Subject inclusion criteria and image preprocessing followed previously established protocols (Cui et al., 2020).

Schizophrenia Dataset (SCZ). The SCZ dataset consists of 218 individuals, including 117 healthy controls and 101 schizophrenia patients, collected from previous studies (Jing et al., 2019; Zhuo et al., 2017). All subjects underwent structural MRI and resting-state fMRI scanning on a 3.0-Tesla MR system. The resting-state fMRI data were preprocessed using fMRIPrep (Esteban et al., 2019), and motion-related artifacts were reduced using ICA-AROMA (Pruim et al., 2015).

Table 4 provides an overview of the datasets used in our study. We categorize the datasets into four groups according to their roles: pretraining, internal testing, semi-external testing, and external testing. The pretraining group consists of three widely used fMRI datasets, totaling 11,189 unique subjects and 26,000 samples derived using five distinct brain atlases. The internal testing group includes settings in which both the dataset and the atlas are seen during pretraining. It is constructed from subsets of the pretraining datasets and is used to evaluate the in-distribution performance of our approach. The semi-external testing group includes settings in which either the dataset or the atlas is unseen during pretraining, while the other is seen, enabling the evaluation of cross-dataset or cross-atlas generalization. The external testing group includes settings in which both the dataset and atlas are unseen during pretraining, enabling the evaluation of generalization across both unseen datasets and unseen atlases. Collectively, the test groups cover diverse downstream tasks, including demographic prediction, such as age regression and sex classification, cognitive function prediction, such as executive function prediction, and disease diagnosis, such as schizophrenia classification. These tasks enable a comprehensive evaluation of the generalizability of the learned representations.

Table 4: Overview of neuroimaging datasets used in this study. We group datasets according to their roles in our pipeline: pretraining, internal testing, semi-external testing, and external testing. (1) **Pretraining**: Data used for self-supervised pretraining. (2) **Internal Testing**: Settings in which both the dataset and atlas are seen during pretraining. (3) **Semi-External Testing**: Settings in which either the dataset or the atlas is unseen during pretraining. (4) **External Testing**: Settings in which both the dataset and atlas are unseen during pretraining. The table lists datasets, atlases, number of unique subjects, total samples, and associated downstream tasks.

Datasets	**Atlases**	**Subjects**	**Samples**	**Task**
Pretraining				
HCP	Schaefer100/200/300/400	965	13,880	-
HCP-D	Schaefer100/200/300/400	632	2,528	-
ABCD	Gordon333	9,592	9,592	-
Internal Test				
HCP	Schaefer100/200/300/400	965	3,860	Multiple
Semi-External Test				
HCP	Gordon333	965	965	Multiple
PNC	Schaefer400	693	693	Multiple
SCZ	Schaefer100/200/300/400	218	872	Schizophrenia Classification
External Test				
PNC	Glasser360/BNA	693	1,386	Multiple
SCZ	Glasser360	218	218	Schizophrenia Classification

B Additional Results

B.1 Additional Comparison with Other Baselines

To further examine the effect of parcellation resolution on baseline performance, we compare the methods across multiple brain atlases. Figures 4 and 5 show their performance on the Schaefer100, Schaefer200, Schaefer300 and Schaefer400 atlases for HCP sex classification and age regression, respectively. For BrainGFM and AREION, we report results from models pretrained on multiple

atlases, whereas the other methods under comparison are trained or retrained separately for each corresponding atlas resolution. As shown in the figures, AREION maintains consistently strong performance across different parcellation resolutions. Moreover, most baseline methods exhibit improved performance as the parcellation resolution increases, suggesting that higher-resolution parcellations may provide finer-grained regional FC information that benefits downstream prediction.

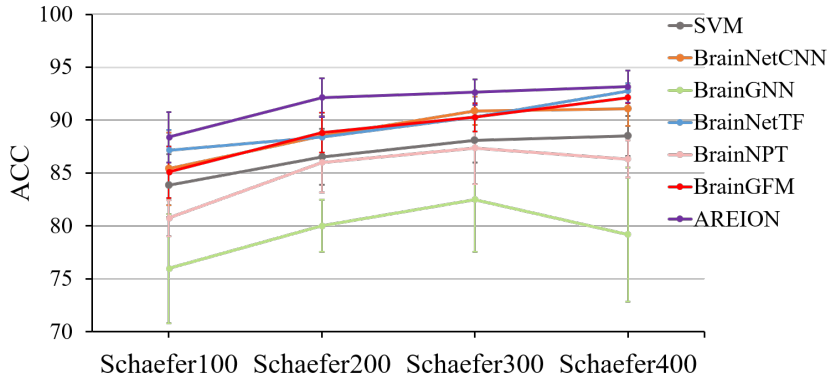


Figure 4: Performance comparison of different methods for HCP sex classification on Schaefer100/200/300/400.

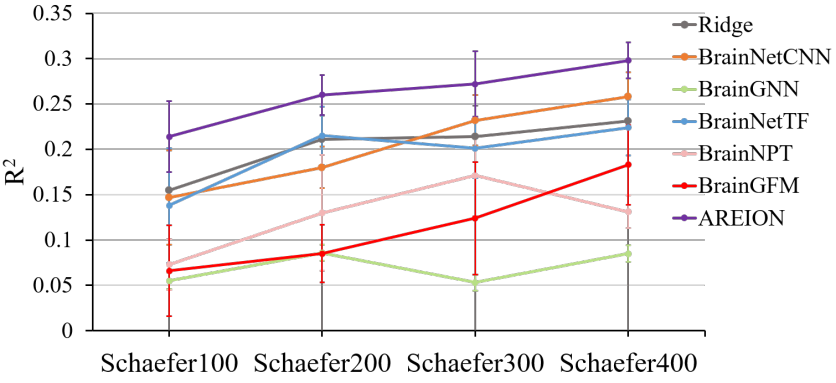


Figure 5: Performance comparison of different methods for HCP age regression on Schaefer100/200/300/400.

B.2 Additional Ablation Results

To complement the ablation results reported in the main paper, we further evaluate different positional encoding methods on the HCP sex classification task using the Schaefer100, Schaefer200, and Schaefer300 atlases. As shown in Table 5, BRC consistently achieves the best performance across all three parcellation resolutions and all evaluation metrics, further supporting its effectiveness.

C Implementation Details

For pretraining, we use the AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of 1.5×10^{-4} and a weight decay of 0.05. The learning rate undergoes a linear warmup over 40 epochs followed by a cosine decay schedule(Loshchilov & Hutter, 2016). The model is pretrained

Table 5: Ablation study of positional encoding methods. All variants are independently pretrained on the HCP dataset and evaluated on the HCP sex classification task using the Schaefer100/200/300 atlases.

Positional Encoding	Schaefer100			Schaefer200			Schaefer300		
	ACC ↑	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑
w/o PE	$73.47_{\pm 1.20}$	$80.31_{\pm 3.09}$	$73.15_{\pm 0.94}$	$73.68_{\pm 1.65}$	$81.35_{\pm 2.72}$	$73.32_{\pm 1.68}$	$74.72_{\pm 1.48}$	$82.80_{\pm 2.67}$	$74.33_{\pm 1.53}$
Sinusoidal PE	$87.98_{\pm 2.13}$	$95.61_{\pm 0.61}$	$87.90_{\pm 2.16}$	$91.61_{\pm 2.33}$	$96.71_{\pm 0.72}$	$91.56_{\pm 2.35}$	$92.54_{\pm 0.70}$	$97.46_{\pm 0.42}$	$92.51_{\pm 0.71}$
Relative PE	$83.52_{\pm 3.17}$	$92.57_{\pm 1.25}$	$83.39_{\pm 3.23}$	$86.74_{\pm 1.96}$	$94.69_{\pm 0.67}$	$86.64_{\pm 2.02}$	$88.70_{\pm 1.24}$	$95.99_{\pm 0.45}$	$88.64_{\pm 1.28}$
BRC (Ours)	$\mathbf{88.50_{\pm 2.35}}$	$\mathbf{95.77_{\pm 0.58}}$	$\mathbf{88.43_{\pm 2.37}}$	$\mathbf{91.81_{\pm 2.16}}$	$\mathbf{97.20_{\pm 0.66}}$	$\mathbf{91.76_{\pm 2.16}}$	$\mathbf{92.75_{\pm 1.47}}$	$\mathbf{97.73_{\pm 0.44}}$	$\mathbf{92.70_{\pm 1.48}}$

for 400 epochs with a mask ratio of 75% and a batch size of 8. All pretraining experiments are conducted using two NVIDIA A100 GPUs. Additional pretraining and architectural hyperparameters are provided in Table 6.

Table 6: Training and architectural hyperparameters used in the proposed framework.

Hyperparameter	**Value**	**Description**
`batch_size`	8	Number of samples per batch during training.
`learning_rate`	1.5×10^{-4}	Learning rate used by the optimizer.
`epochs`	400	Number of epochs.
`warmup_epochs`	40	Number of warmup epochs.
`weight_decay`	0.05	Optimizer weight decay coefficient.
`betas`	[0.9, 0.95]	Optimizer momentum.
`d`	256	Hidden dimension shared across the Brain Region Contextualizing, hierarchical encoders and MAE decoder.
`region_nhead`	4	Number of attention heads in Regional Transformer Encoder.
`region_layers`	4	Number of Transformer layers used in the Regional Transformer Encoder.
`brain_nhead`	8	Number of attention heads in Brain-wide Transformer Encoder.
`brain_layers`	6	Number of Transformer layers used in the Brain-wide Transformer Encoder.
`mask_ratio`	0.75	Fraction of regions masked.
`dropout`	0.1	Dropout rate used for regularization.

For downstream adaptation, we freeze the pretrained Transformer encoders and evaluate the output embeddings using SVM for classification tasks and ridge regression for regression tasks. The pretrained encoder is not finetuned for the downstream evaluation. More settings about the architectures of downstream models are provided in Table 7.

D Comparison of Used Atlases

Table 8 summarizes the seven brain atlases used in this study. Among them, five atlases are used during pretraining, while the remaining two atlases are reserved for evaluating generalization to unseen atlases. The Schaefer atlas series (Schaefer100/200/300/400) (Schaefer et al., 2018), one of the most widely used functional parcellation frameworks, was derived from resting-state fMRI data. Its four resolutions, ranging from 100 to 400 regions, provide progressively increasing levels of parcellation granularity while preserving a consistent functional organization, allowing us to evaluate the robustness of the learned representations to variations in the number and granularity of brain regions. The Gordon333 atlas (Gordon et al., 2016) is an independently developed functional atlas with 333 regions, introducing variation in regional definitions beyond the resolution differences

Table 7: Downstream evaluation hyperparameters used in the proposed framework.

Hyperparameter	Value	Description
SVM		
`kernel`	RBF	Kernel function used for classification.
`probability`	True	Enables probability estimates for AUC computation.
`C`	1.0	SVM regularization strength.
`gamma`	scale	RBF kernel coefficient.
Ridge		
`alpha`	$\{10^{-1}, \ldots, 10^{6}\}$	30 log-spaced candidate L2 regularizations. The best value is selected using leave-one-out cross-validation via `RidgeCV`.
`feature_scaling`	StandardScaler	Embeddings are standardized before fitting downstream models.

within the Schaefer series. For unseen-atlas evaluation, we use Glasser360, a multimodal cortical parcellation with 360 regions, to assess generalization to a distinct regional organization. We also include the Brainnetome Atlas (BNA), a connectivity-based brain atlas constructed from multimodal neuroimaging data that provides a fine-grained parcellation together with detailed anatomical and functional connectivity information. Together, these atlases span diverse parcellation resolutions and regional definitions, enabling a comprehensive evaluation of the atlas-agnostic generalizability of AREION.

Table 8: Brain atlases and parcellations used in our experiments.

Atlas	#ROI	Description
Schaefer100	100	Functional parcellation derived from resting-state fMRI and aligned with Yeo 7/17 networks.
Schaefer200	200	Higher-resolution functional parcellation with consistent network organization.
Schaefer300	300	Fine-grained functional parcellation with increased resolution.
Schaefer400	400	High-resolution functional parcellation in the Schaefer atlas series.
Gordon333	333	Functional parcellation derived from resting-state functional connectivity and network organization.
Glasser360	360	Multimodal cortical parcellation with 360 regions.
BNA	210	Connectivity-based brain atlas constructed from multimodal neuroimaging data; 210 cortical regions used in our experiments.

E BASELINES

To rigorously evaluate the effectiveness of AREION, we compare it with a comprehensive set of baselines, categorized into three groups: traditional machine learning methods, supervised deep learning models, and pretrained fMRI foundation models.

Traditional Machine Learning Methods. We include a multilayer perceptron (MLP) and random forest (RF) for both classification and regression tasks, along with a support vector machine (SVM) for classification and ridge regression for regression tasks. For SVM and ridge regression, we use the same settings as those used for the downstream adaptation of AREION (see Table 7). For the MLP, we use a three-hidden-layer network with ReLU activations (Nair & Hinton, 2010). The MLP

is trained using Adam (Kingma & Ba, 2014) with a learning rate of 10^{-3}. Additional training and architectural details are provided in Table 9.

Table 9: Training and architectural hyperparameters for the MLP and random forest baselines.

Hyperparameter	Value	Description
MLP		
`hidden_layer_sizes`	(256, 128, 64)	Number and width of hidden layers.
`activation`	ReLU	Hidden-layer activation function.
`solver`	Adam	Weight optimization algorithm.
`learning_rate_init`	10^{-3}	Initial learning rate.
`alpha`	$\{10^{-2}, \ldots, 10^{4}\}$	7 log-spaced candidate L2 regularizations. The best value is selected per fold using inner 3-fold cross-validation via `GridSearchCV`.
`max_iter`	500	Maximum number of training iterations.
Random Forest		
`n_estimators`	500	Number of trees in the forest.
`criterion`	Gini / squared error	Split criterion for classification / regression.
`max_depth`	None	Maximum depth of each tree.
`max_features`	sqrt	Uses $\sqrt{d}$ randomly selected features at each split.

Supervised Deep Learning Models. We include three supervised deep learning models, including BrainNetCNN (Kawahara et al., 2017), BrainGNN (Li et al., 2021), and BrainNetTF (Kan et al., 2022), covering CNN-, GNN-, and Transformer-based architectures, respectively. All models are trained end-to-end from scratch using supervised learning on the labeled downstream training data without pretraining, and their performance is evaluated on the atlas used for training.

Pretrained fMRI Foundation Models.

For a fair comparison, all foundation models are retrained on our collected data. BrainMass (Yang et al., 2024) uses the same task-specific predictors as AREION, whereas BrainNPT (Hu et al., 2024) and BrainGFM (Wei et al., 2026) are evaluated via end-to-end finetuning following their default protocols. Because BrainMass and BrainNPT support only single-atlas pretraining, we retrain each model separately on every atlas used in the corresponding experiments. Results for Schaefer400 are reported in Table 1, while results for other Schaefer resolutions are provided in Appendix B. BrainGFM, which supports multi-atlas pretraining, is retrained on the same multi-atlas cohort as AREION and evaluated on both Schaefer400 (Table 1) and atlases unseen during pretraining (Table 2). This protocol enables a direct comparison of cross-atlas generalization between BrainGFM and AREION. Figure 6 further compares the prediction performance and model size of the pretrained foundation models on the HCP Schaefer400 sex classification task. AREION achieves the best performance with only 6.4 million parameters, substantially fewer parameters than the other foundation models.

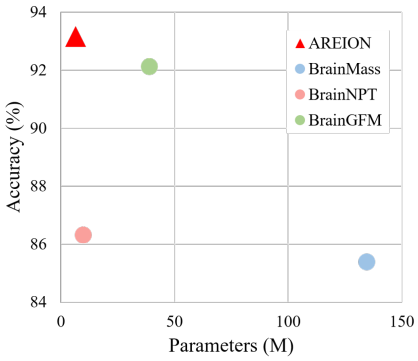


Figure 6: Comparison of performance and model size across different foundation models on the HCP Schaefer400 sex classification task.

F Interpretability of Learned FC Representations

To assess whether the FC representations learned by AREION capture functional brain organization and its associations with brain development and cognition, we examine regional importance

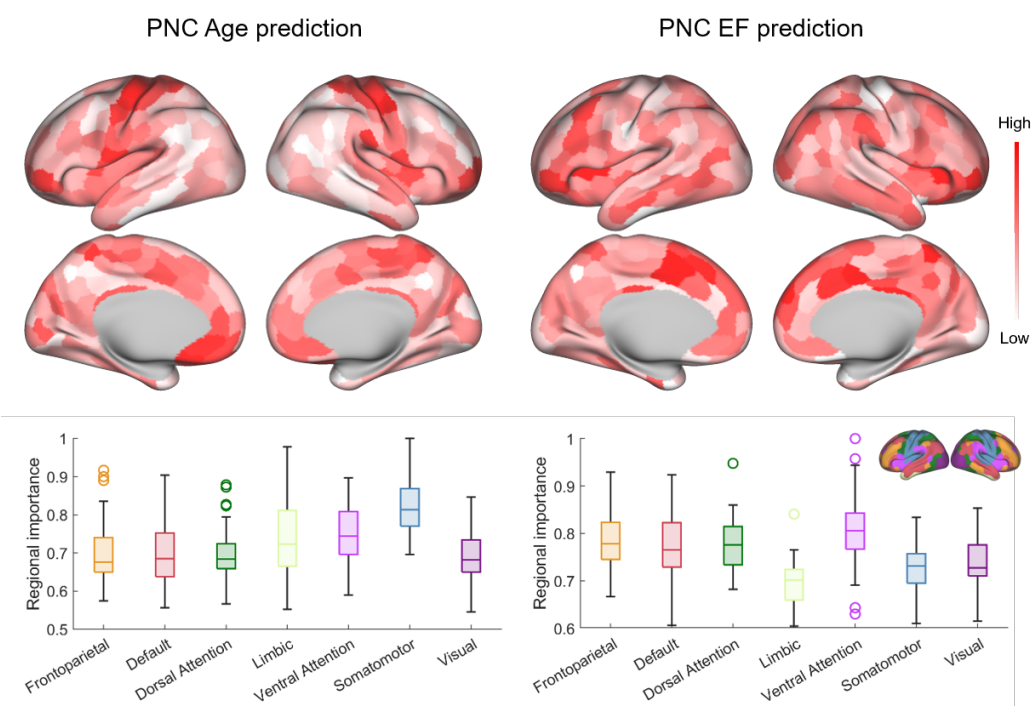


Figure 7: Importance of FC representations for age (left) and executive function (right) prediction across cortical regions (top) and large-scale functional networks (bottom) on the PNC cohort.

in models predicting age and executive function in the PNC cohort. Regional importance is computed as the region-wise sum of the absolute values of the Haufe-transformed feature weights (Haufe et al., 2014) from each probing model. As shown in Figure 7, cortical regions within the somatomotor network and the ventral attention network exhibit higher importance for age prediction, and regions within the ventral attention network, frontoparietal network, and dorsal attention network show higher importance for the cognition prediction, consistent with previous brain-wide association studies (Dosenbach et al., 2010; Cui et al., 2020; Li et al., 2026; Sripada et al., 2020; Chen et al., 2022; Li et al., 2025), suggesting that AREION captures biologically meaningful variations in its learned FC representations.