dFCExpert: Learning Dynamic Functional Connectivity Patterns with Modularity and State Experts

Tingting Chen,[1,2*] Hongming Li,[1,2] Hao Zheng,[3] Jintai Chen,[4] Yong Fan[1,2]

[1]Center for Biomedical Image Computing and Analytics, Philadelphia, PA 19104, USA

[2]Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104, USA

[3]School of Computing and Informatics, University of Louisiana at Lafayette, Lafayette, LA 70503, USA

[4] AI Thrust, and Information Hub of HKUST(GZ), Guangzhou, Guangdong 511400, China

*Correspondence: ttchen0603@gmail.com

December 20, 2024

Keywords: fMRI, dynamic functional connectivity learning, brain modularity organization, dynamic FC states, mixture of experts.

Abstract: Characterizing brain dynamic functional connectivity (dFC) patterns from functional Magnetic Resonance Imaging (fMRI) data is of paramount importance in neuroscience and medicine. Recently, many graph neural network (GNN) models, combined with transformers or recurrent neural networks (RNNs), have shown great potential for modeling the dFC patterns. However, these methods face challenges in effectively characterizing the modularity organization of brain networks and capturing varying dFC state patterns. To address these limitations, we propose dFCExpert, a novel method designed to learn robust representations of dFC patterns in fMRI data with modularity experts and state experts. Specifically, the modularity experts optimize multiple experts to characterize the brain modularity organization during graph feature learning process by combining GNN and mixture of experts (MoE), with each expert focusing on brain nodes within the same functional network module. The state experts aggregate temporal dFC features into a set of distinctive connectivity states using a soft prototype clustering method, providing insight into how these states support different brain activities or are differentially affected by brain disorders. Experiments

on two large-scale fMRI datasets demonstrate the superiority of our method over existing alternatives. The learned dFC representations not only show improved interpretability but also hold promise for enhancing clinical diagnosis. The code can be accessed at MLDataAnalytics/dFCExpert on GitHub. [1]

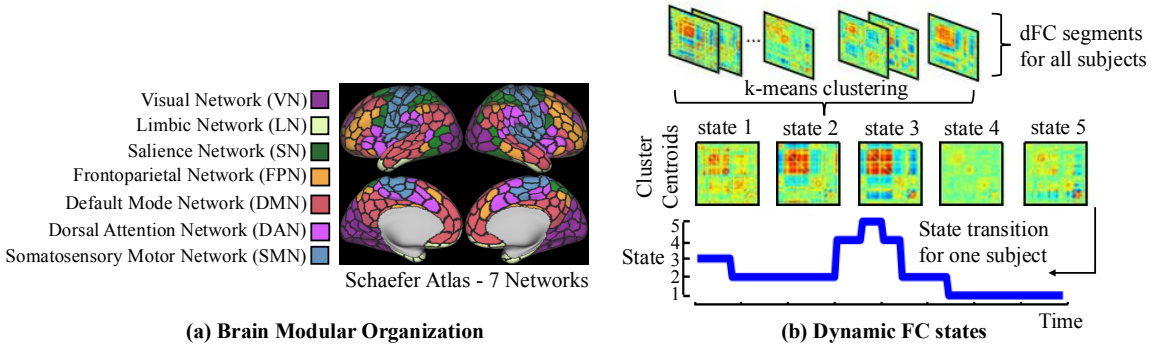


Figure 1: (a) The Schaefer atlas (Schaefer et al., 2018) with seven functional networks (modules). Each black-line-bordered region represents a brain node, while regions of the same color belong to a single functional module, collectively associated with specific brain functions. (b) dFC states represent a set of recurring patterns of dFC measures, capturing the evolving FC dynamics of the brain.

1 Introduction

The human brain is a dynamic network system that generates complex spatiotemporal dynamics of brain activity. Analyzing such dynamics holds promise to provide insights into the brain's functional organization and its relationship with human cognition (Greicius, 2008; K. Wang et al., 2007), behaviors (Smith et al., 2015), and brain disorders (Greicius, 2008; K. Wang et al., 2007). Among many brain imaging techniques, functional Magnetic Resonance Imaging (fMRI) is a particularly powerful tool that models the spatiotemporal patterns of brain activity by measuring fluctuations in blood-oxygen level-dependent (BOLD) signals (Matthews & Jezzard, 2004). Since brain activity often exhibits strong spatial correlations, the BOLD signals generally are parcellated into a collection of pre-defined brain regions (ROIs), and the pairwise correlations between those ROIs are referred as functional connectivity (FC), which has emerged as a key tool for understanding the brain function.

Based on the FC, the brain can be modeled as a functional network with graph theory approaches, where brain regions of interest serve as network nodes and the FC strengths between them act as edges. Leveraging this graph-structured nature of the brain, many fMRI data analytic methods have employed

[1] This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible

graph neural networks (GNNs) to learn brain network representations, enabling tasks such as decoding human traits or diagnosing diseases (Kan et al., 2022; Kawahara et al., 2017; X. Li et al., 2021). Generally, these methods fall into two categories: *static FC* and *dynamic FC methods*. The *static-FC methods* characterize the FC between nodes based on the entire fMRI scan, i.e., the full time series (B.-H. Kim & Ye, 2020; Ktena et al., 2018). However, these methods are not equipped to characterize the dynamic properties of FCs (fluctuate over time), which are crucial for capturing the brain's evolving states. Differently, the *dynamic-FC methods* split the whole fMRI time series into temporal segments and quantify the FC based on each segment so that time-varying FC measures can be derived. A typical pipeline for dynamic FC methods involves extracting brain network representations using GNNs for each temporal FC segment, followed by the applications of RNNs or transformers for temporal dynamics learning (Jing et al., 2023; B.-H. Kim et al., 2021; Lin et al., 2023; Q. Wang et al., 2023). Although promising, these methods still face challenges, primarily due to the unique characteristics of brain functional networks and dFC measures.

Firstly, **most GNN brain functional network analysis methods overlook the intrinsic brain modularity organization**, leading to suboptimal graph representation learning. Both theoretical and empirical studies highlight that the brain functions as a modular system, comprising specialized cognitive and topological modules. Each module consists of tightly connected brain network nodes (ROIs) that collaboratively perform specific functions (Bertolero et al., 2015; Sporns & Betzel, 2016), as illustrated in Figure 1(a). However, existing GNN-based methods typically treat all brain nodes uniformly, employing the same aggregation mechanism regardless of variations in node features. This limitation prevents these methods from effectively capturing modularity-specific features. *Secondly*, **most known approaches fail to capture distinctive dFC state patterns.** Evidence suggests that dFC measures often correspond to distinctive dynamic states (Figure 1(b)), which can be identified by clustering dFC measures from temporal segments of fMRI scans (Damaraju et al., 2014). Studies on brain disorders have revealed that disease-specific alterations are confined to certain dynamic states (Damaraju et al., 2014), underscoring that capturing dFC states can potentially improve the detection of functional brain changes associated with brain disorders.

To address these challenges, we propose a novel GNN-based dFC learning framework, called dFCExpert, aiming to enhance the representation learning of dynamic functional connectome of fMRI data. The framework consists of two key components: modularity experts and state experts, specifically designed to characterize the brain modularity and the dFC state patterns. The modularity experts are built upon GNN and Mixture of Experts (MoE) for learning brain graph features of each FC segment, which characterize the brain modularity organization in the graph learning process by optimizing multiple experts at each graph layer. Each of the experts focuses on specific subsets of nodes that exhibit similar behaviors

and interactions, enabling fine-grained modeling of the brain's modular structure. On top of the graph representations learned by the modularity experts, the state experts aggregate the temporal features of dFCs into a compact set of distinctive states based on a soft prototype clustering method, where each state has similar FC patterns and reflects different activities of the dFCs related to human behaviors or brain diseases. Finally, the clustering-derived state features are used for task prediction. With the modularity and state experts, our dFCExpert introduces two novel strategies for learning-based fMRI analysis: modeling the brain functional network modularity and dynamic FC states. This pioneering approach presents a first-of-its-kind method for learning informative and explainable features of dFC patterns from fMRI data, offering significant potential for advancing neuroscience research and clinical applications.

We evaluated the performance of dFCExpert on two large-scale fMRI datasets: the Human Connectome Project (HCP) (Van Essen et al., 2013) and the Adolescent Brain Cognitive Development (ABCD) (Casey et al., 2018). Across tasks such as sex classification and cognitive intelligence prediction, dFCExpert consistently outperformed known methods, achieving state-of-the-art performance. To validate the framework's design, we conducted comprehensive ablation studies to analyze the contribution of individual components of dFCExpert. Visualization results revealed that the modularity experts effectively target distinct brain modules and the state experts identify unique dynamic brain states, indicating that the method was capable of learning interpretable representations of dFC measures for prediction tasks. Moreover, we demonstrated that the brain modules and dynamic states identified by dFCExpert are biologically meaningful, bridging the gap between deep learning-based functional network modeling and traditional functional network analytic methods, highlighting the potential of dFCExpert to deliver both robust predictive performance and valuable insights into the brain's functional organization.

2 Related Works

2.1 Dynamic Functional Connectivity Learning

To capture the time-varying FC patterns of the brain, a spatio-temporal graph convolutional network (ST-GCN) is developed to incorporate both spatial and temporal convolutions to model the non-stationary nature of dFC measures (Gadgil et al., 2020). Many methods have adopted a GNN-RNN pipeline to learn graph-level features from each FC segment with the GNNs, followed by RNNs or transformers to capture temporal FC patterns (Campbell et al., 2022; B.-H. Kim et al., 2021; Said et al., 2023; Q. Wang et al., 2023). Particularly, STAGIN utilizes GNNs with spatial attention and transformers with

temporal attention to model the dynamics of brain networks (B.-H. Kim et al., 2021). DynDepNet introduces a dynamic graph structure learning method to capture the time-varying structures of fMRI data (Campbell et al., 2022). NeuroGraph systematically evaluates the effectiveness of different GNN designs for modeling dynamic functional networks (Said et al., 2023). While these methods have shown promise in modeling the dynamics of brain functional connectivity, they often overlook the brain's intrinsic modular organization and fail to capture distinctive dFC states. This limitation hinders their ability to provide deeper insights into the brain's functional organization and its evolving connectivity states.

2.2 Brain Modularity Organization

The human brain functions as a modular system composed of many cognitive/topological modules. Brain regions within the same modules are often tightly connected and tend to perform similar functions (Q. Wang et al., 2023). For instance, salience network (SN) and default mode network (DMN) are two crucial neurocognitive modules in the brain, where SN mainly detects internal or external stimuli and coordinates the brain's response to those stimuli, and DMN is responsible for self-related cognitive functions (Bertolero et al., 2015; Sporns & Betzel, 2016). A few existing methods have incorporated the brain modular organization into graph representation learning. BRAINNETTF introduces a novel graph readout function that leverages modular-level similarities between brain nodes to pool graph-level embeddings from clusters of functionally similar nodes (Kan et al., 2022). MSGNN (Q. Wang et al., 2023) develops a modularity-constrained GNN that enforces node embeddings to align with three functional network modules (i.e., central executive network, SN, DMN), by applying a modularity constraint loss after the graph layers to encourage similarity between node-level embeddings within the same module. However, since these methods incorporate the brain modularity only after the graph model, all brain network nodes are still processed uniformly within each graph layer. This may be suboptimal for effectively learning representations of nodes belonging to distinct brain modules. To address this, we propose modularity experts, a combination of GNN and MoE, to mimic the brain's modular organization, which employs multiple experts at each graph layer to guide brain network nodes with similar functions towards the same experts.

2.3 Dynamic FC States

The dynamic brain states are a compact set of unique patterns derived from dFC measures, which represent connectivity patterns that repetitively occur during the fMRI acquisition (Preti et al., 2017). Existing studies typically cluster the dFC measures into different states using clustering methods (e.g.,

k-means) and then conduct statistic analyses to investigate the relationships between these states and biological or behavior measures of interest (Damaraju et al., 2014). However, this two-step strategy has significant limitation that the features used for clustering are not explicitly optimized for the downstream statistical analyses, leading to a potential mismatch between the identified states and their informativeness for specific tasks. In contrast, our state experts overcome this limitation by adopting an end-to-end learning framework, simultaneously learning distinctive dFC states and optimizing representations of dFC graphs for specific tasks. By integrating state discovery and task optimization, our method ensures that the resulting states are both highly informative and tailored to the requirements of the analysis, enhancing their interpretability and utility in studying dFC patterns.

2.4 Expert Models

The concept of Mixture of Experts (MoE) is a well-established machine learning technique that employs multiple expert layers, each specializing in solving a specific subtask or learning a particular sub-structure (Jacobs et al., 1991). Over the years, MoE has demonstrated remarkable success and been widely applied across diverse domains, including multi-modal learning (Akbari et al., 2024), computer vision (Jain et al., 2024), and machine translation (Shen et al., 2019). In the context of graph classification, TopExpert integrates MoE on top of extracted graph features, leveraging topology-specific prediction experts for molecule property prediction (S. Kim et al., 2023). Moving beyond applying MoE post-feature extraction, a graph MoE (GMoE) model incorporates multiple experts within each graph layer to scale GNN model's capacity (H. Wang et al., 2024). Building upon the principles of MoE, we propose modularity experts, a novel approach designed to capture the brain's modular organization within the graph learning process. By integrating modularity experts directly into the learning of dynamic FC graphs, our method functions just like a brain, advancing the representation learning of dFCs.

3 Method

We begin by introducing the preliminaries of the proposed dFCExpert framework, including the concepts of GNNs and the construction of dFC graphs from fMRI data (Section 3.1), followed by a detailed description of the framework, including an overview, modularity experts, state experts, and the overall loss function (Section 3.2).

3.1 Preliminaries

3.1.1 Graph Neural Networks

To glean useful information from graph-structured data, GNNs iteratively compute node features by aggregating information of neighbor nodes and updating the node features with non-linear functions in a layer-wise manner. The propagation mechanism for a node i at the (l)-th GNN layer is formulated as:

$$h_i^{(l)} = F_u^{(l)}\left(h_i^{(l-1)}, F_a^{(l)}\left(\{(h_i^{(l-1)}, h_j^{(l-1)}, e_{ij}) : j \in N_i\}\right)\right), \tag{1}$$

where $h_i^{(l-1)}$ denotes the features of node i at the $(l-1)$ layer, N_i is the set of neighbors of node i, e_{ij} denotes the edge between nodes i and j, and F_a and F_u are differentiable functions for aggregating information and updating node features, respectively. Different choices of these two functions lead to various GNN architectures. For instance, Graph Isomorphism Network (GIN) (Xu, Hu, et al., 2018), a variant of GNN, uses summation as F_a and a multi-layer perceptron (MLP) as F_u:

$$h_i^{(l)} = M^{(l)}(h_i^{(l-1)}, e_{ij}, W^{(l)}) = \mathrm{MLP}^{(l)}\left((1+\epsilon^{(l)}) \cdot h_i^{(l-1)} + \sum_{j \in N_i} h_j^{(l-1)}\right). \tag{2}$$

We simplify the two functions described above into a single operation $M^{(l)}$, where $W^{(l)}$ represents the trainable weight of the multi-layer perceptron (MLP), and ϵ is a learnable parameter initialized to zero. Given its strong ability for graph representation learning, we adopt GIN as our dFC graph feature extractor, following the approach in (B.-H. Kim et al., 2021).

3.1.2 Dynamic Graph Definition

To construct dFC graphs from fMRI scans, we first use a brain atlas to transform the 4D fMRI data into a time-course matrix $\mathbf{P} \in \mathbb{R}^{N \times T_{\max}}$. This matrix is created by averaging the fMRI signals within each brain region defined by the atlas, yielding fMRI signals of N brain network nodes at each time point. Next, using a sliding-window approach, the time-course matrix is divided into $T = \lfloor T_{\max} - \Gamma/S \rfloor$ temporal segments, where Γ is the window length and S is the stride size. For each segment t, an FC matrix $X_t \in \mathbb{R}^{N \times N} (t = 1, \ldots, T)$ is computed by calculating Pearson correlation coefficients between the time series of all brain region pairs, yielding T FC matrices. Following (Said et al., 2023), the correlation matrices can be informative node features, where the features for the i-th node in segment t correspond to the elements of the i-th row in X_t. Additionally, a binary adjacent matrix $A_t \in \{0,1\}^{N \times N}$ is derived from X_t by retaining the top 30-percentile correlation values as connected and setting the remaining

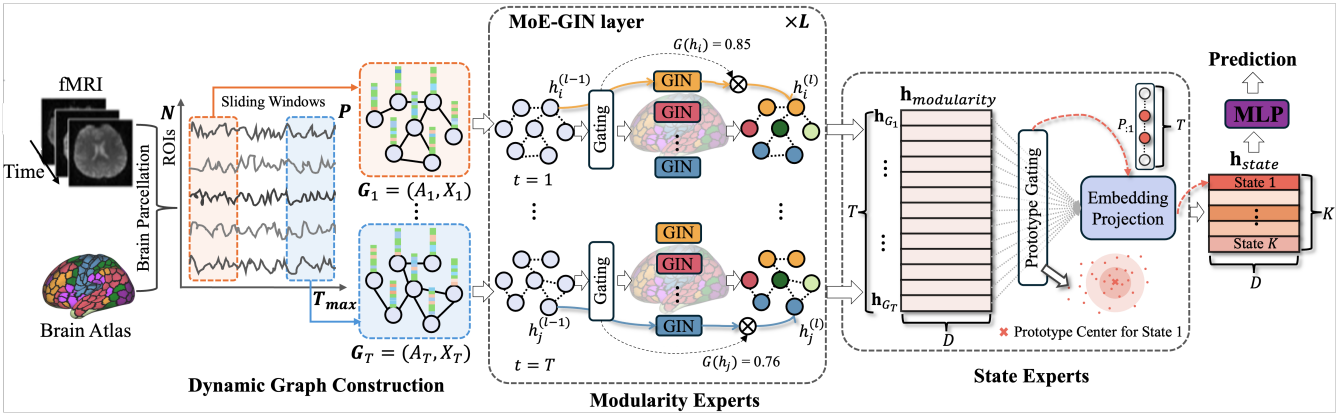


Figure 2: The dFCExpert framework consists of modularity experts and state experts. The modularity experts include L layers of MoE-GIN, which route node features h_i to a specific GIN expert with a gating score $G(h_i)$ for learning graph-level features $\mathbf{h}_{G_t}$ for each temporal segment t, yielding $\mathbf{h}_{modularity} = (\mathbf{h}_{G_1}, \ldots, \mathbf{h}_{G_T})$ across all T segments. The state experts aggregate the learned temporal graph features into dynamic states using prototype gating and embedding projection. Finally, the state features $\mathbf{h}_{state}$ are processed through an MLP layer for task prediction.

values as unconnected, as described in (B.-H. Kim & Ye, 2020). Thus, the input dFC graphs for each subject are represented as $\mathbf{G}_t = \{A_t, X_t\}(t = 1, \ldots, T)$ (Figure 2).

3.2 dFCExpert

3.2.1 Overview

As illustrated in Figure 2, dFCExpert consists of modularity experts and state experts. Taking the dFC graphs as input, the modularity experts leverage a combination of GIN and MoE to learn brain graph features for each FC segment, aiming to capture the brain modularity mechanism effectively. Building upon the outputs of the modularity experts, the state experts adaptively group the temporal graph features into distinctive states using a soft prototype clustering method. This approach allows the model to learn expressive state features by assigning soft clusters to the temporal graph features. Finally, the learned state features are passed through an MLP layer to predict a specific task, such as predicting sex in a classification setting or predicting an intelligence measure in a regression setting. Formally, the objective of dFCExpert is to train a neural network $f : (\mathbf{G}_1, \ldots, \mathbf{G}_T) \rightarrow \mathbf{h}_{state}$, where $\mathbf{G}_t = \{A_t, X_t\}(t = 1, \ldots, T)$ represents the sequence of constructed dFC graphs with T segments, and $\mathbf{h}_{state} \in \mathbb{R}^{K \times D}$ is the output features from the state experts. The framework combines the modularity experts m and state experts s to define $f = m \circ s$, where m learns dFC graph representations $\mathbf{h}_{modularity} = (\mathbf{h}_{G_1}, \ldots, \mathbf{h}_{G_T})$ for each of

T segments and s aggregates these learned temporal graph features into the final state features $\mathbf{h}_{state}$:

$$\begin{aligned} m &: (\mathbf{G}_1, \ldots, \mathbf{G}_T) \rightarrow (\mathbf{h}_{G_1}, \ldots, \mathbf{h}_{G_T}), \\ s &: (\mathbf{h}_{G_1}, \ldots, \mathbf{h}_{G_T}) \rightarrow \mathbf{h}_{state}. \end{aligned} \tag{3}$$

3.2.2 Modularity Experts

The modularity experts are designed to emulate the human brain's modular organization, enabling effective learning of brain graph representations. The modularity experts are implemented using a combination of GIN and MoE, referred to as MoE-GIN, consisting of L layers of MoE-GIN. Specifically, each MoE-GIN layer includes a gating function and multiple GIN experts, as illustrated in Figure 2. The gating function determines which experts are most suitable for processing a given node, while the multiple experts are independent GIN propagation functions (Eq. (2)), each with its own trainable parameters. This design of MoE-GIN enables nodes that are tightly connected or behaviorally similar to be routed to the same expert. Consequently, each GIN expert specializes in learning representations for a specific brain module to effectively capture distinct brain activities. By organizing the brain nodes into specialized groups, the modularity experts mimic the way the brain operates, where tightly connected regions interact to perform specific functions or activities. This design allows the model to reflect the brain's modularity, improving its ability to learn meaningful and interpretable representations of brain networks. Formally, the operation of a single MoE-GIN layer is described as:

$$h_i^{(l)} = \sum_{c=1}^{C} G_c^{(l)}(h_i^{(l-1)}) M_c^{(l)}(h_i^{(l-1)}, e_{ij}, W^{(l)}) \tag{4}$$

where C is the number of experts, and M denotes the node propagation function of GIN as defined in Eq. (2). G is a gating function to generate assignment scores based on the input of node features $h_i^{(l-1)}$, and $G_c(h_i^{(l-1)})$ denotes the score for the c-th expert. We implement a simple gating function by multiplying the input with a trainable weight matrix W_g, followed by the application of the $\mathrm{softmax}$ function (Jordan & Jacobs, 1994): $G(h_i) = \mathrm{softmax}(h_i W_g)$. Given that each node typically belongs to a single brain module, we adopt a simplified top-1 gating scheme as in (Fedus et al., 2022), where each node is routed to only one expert, corresponding to the expert with the highest gating score. This simplification reduces computational overhead of routing while maintaining model quality and interpretability.

The graph-level representation $h_{G_t}^{(l)}$ for segment t at layer l is computed by averaging the updated node features $\{h_i^{(l)}, i \in N\}$, and the final graph representation $\mathbf{h}_{G_t}$ is then obtained by concatenating the graph representations from all L layers (Xu, Li, et al., 2018) followed by an MLP layer for dimension

reduction: $\mathbf{h}_{G_t} = \mathrm{MLP}(\mathrm{concatenate}(h_{G_t}^{(l)} | l \in \{1, \cdots, L\})) \in \mathbb{R}^D$. Finally, we obtain $\mathbf{h}_{modularity} = (\mathbf{h}_{G_1}, \ldots, \mathbf{h}_{G_T}) \in \mathbb{R}^{T \times D}$.

Auxiliary Loss Functions for Training MoE-GIN. If the model is trained solely using the task-specific prediction loss, the gating network may converge to trivial solutions where only a few experts are consistently selected (Shazeer et al., 2017). This imbalance in expert selection becomes self-reinforcing, as the favored experts dominate the learning process, further increasing their frequency of selection. Another trivial situation is that the gating network may produce similar assignment scores across all experts for a given node, resulting in a lack of specialization in experts. In our scenario, we expect the gating network to be capable of effectively distinguishing nodes of different brain functional modules and enable sparse gating so that the selected expert receives a significantly higher score than the others. To achieve this, we introduce two auxiliary loss functions to promote balanced loading and enforce sparse gating, respectively. Given a batch B, each containing T graphs, the loading balance loss is computed as:

$$L_{balance} = \frac{1}{BTC} \sum^{B} \sum^{T} \sum_{c=1}^{C} p_c \log(p_c), \qquad p_c = \frac{1}{N} \sum_{i=1}^{N} G_c(h_i), \tag{5}$$

where N is the number of nodes, and p_c represents the fraction of nodes routed to expert c. By minimizing the negative entropy of the p_c distribution, the $L_{balance}$ encourages a uniform distribution of nodes across all experts. To promote sparse gating in $G(h_i)$, we introduce a sparse gating loss:

$$L_{sparse} = \frac{1}{BTN} \sum^{B} \sum^{T} \sum^{N} \frac{1}{C} \sum_{c=1, g \in G(h_i)}^{C} -g_c log(g_c), \tag{6}$$

where g_c represents the gating score for assigning a given node to expert c. Minimizing the entropy of the gating scores across all experts encourages sparsity, ensuring each node is strongly associated with a specific expert. We implement these two loss functions similar to (Huang et al., 2020), as they are fundamentally deep clustering-based losses and act in opposition during training.

3.2.3 State Experts

The state experts follow the traditional dFC analysis in neuroscience to adaptively group the temporal graph features $\mathbf{h}_{modularity}$ into a small number of states, each reflecting distinct dynamic states of dFC measures associated with human behaviors or brain diseases. From another perspective, representing dFC measures with a smaller set of states simplifies the model, reduces data complexity, and makes it easier to interpret the FC dynamics. Specifically, we design the state experts using a soft prototype clustering method, where temporal FC graph features are softly assigned to clusters in an unsupervised manner. As

illustrated in Figure 2, distinctive dFC states are characterized through prototype gating and embedding projection. The prototype gating introduces K trainable prototype centroids $U = [u_1, \ldots, u_K]$, each with D dimensions ($U \in \mathbb{R}^{K \times D}$), to adaptively learn the feature centers of each state. Given the learned graph features $\mathbf{h}_{modularity} = (\mathbf{h}_{G_1}, \ldots, \mathbf{h}_{G_T})$, the prototype gating calculates the assignment score P_{tk} for assigning segment t to state k using a Softmax projection:

$$P_{tk} = \frac{e^{\langle \mathbf{h}_{G_t}, u_k \rangle}}{\sum_{k'}^{K} e^{\langle \mathbf{h}_{G_t}, u_{k'} \rangle}}, \tag{7}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. The embedding projection aggregates the temporal graph features $\mathbf{h}_{modularity} \in \mathbb{R}^{T \times D}$ into state features $\mathbf{h}_{state}$ under the guidance of the obtained soft assignment score $P \in \mathbb{R}^{T \times K}$. Specifically, to compute features for a single state (e.g., State 1 as in Figure 2), the embedding projection integrates features of the temporal segments assigned to that state using soft gating, performed as $P_{:1}^{\top} \mathbf{h}_{modularity}$. This process ensures that the state features are derived by aggregating the temporal FC features with high probabilities assigned to a specific state, thus enabling the model to effectively capture the dynamic nature of FC patterns and associate them with interpretable states.

Since the clustering lacks explicit guidance labels, we optimize the temporal FC features and cluster centroids in a self-supervised manner by leveraging orthonormal initialization of U (Kan et al., 2022) and optimizing the cluster distribution for cohesive clustering. Specifically, we define a target distribution Q_{tk} for each $\mathbf{h}_{G_t}$ based on the current cluster assignment distribution P_{tk}, which therefore enhances high-confident assignments (Xie et al., 2016) through a squaring operation, defined as:

$$Q_{tk} = \frac{P_{tk}^2 / \sum_t P_{tk}}{\sum_{k'} (P_{tk'}^2 / \sum_t P_{tk'})}. \tag{8}$$

This target distribution strengthens the cluster cohesion, refining temporal FC representations to become more discriminative with respect to their state patterns. Finally, we minimize a KL divergence loss to align P and Q, boosting the cohesion of the clusters:

$$L_{state} = \mathbf{KL}(Q||P) = \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K} Q_{tk} \log \frac{Q_{tk}}{P_{tk}}. \tag{9}$$

3.2.4 Overall Loss Function

The final loss combines a task-specific loss (L_{task}), two auxiliary losses for the modularity experts ($L_{balance}$ and L_{sparse}), and a loss for state experts (L_{state}), yielding the overall optimization objective. Two scaling

Table 1: Performance comparison with alternatives and baselines. ("M" = Modularity, "S" = State)

Method	HCP				ABCD				FC Type
	Sex		Intelligence		Sex		Cognition		
	ACC(%)	AUC(%)	MSE(↓)	CORR(↑)	ACC(%)	AUC(%)	MSE(↓)	CORR(↑)	
BrainNetCNN (Kawahara et al., 2017)	84.56±1.87	91.97±1.33	1.130±0.039	0.200±0.071	85.38±0.07	92.67±0.33	0.474±0.031	0.442±0.011	Static
BRAINNETTF (Kan et al., 2022)	85.42±2.36	92.51±2.01	1.015±0.052	0.205±0.074	84.55±0.64	91.38±0.21	0.462±0.022	0.465±0.024	Static
NeuroGraph (Said et al., 2023)	84.38±2.60	92.03±2.39	1.321±0.043	0.198±0.065	85.71±0.60	93.75±0.61	0.512±0.013	0.423±0.015	Static
STAGIN (B.-H. Kim et al., 2021)	87.74±1.75	92.09±0.93	1.002±0.042	0.215±0.078	86.93±0.45	91.57±0.57	0.452±0.012	0.477±0.008	Dynamic
MSGNN (Q. Wang et al., 2023)	86.25±2.03	92.37±1.42	1.073±0.057	0.206±0.070	87.13±0.63	92.67±0.42	0.453±0.024	0.483±0.017	Dynamic
NeuroGraph (Said et al., 2023)	84.66±3.30	93.14±2.23	1.017±0.032	0.207±0.069	87.37±1.08	94.64±0.69	0.441±0.031	0.482±0.005	Dynamic
Baseline	88.11±4.26	96.68±1.71	0.978±0.079	0.238±0.086	87.25±0.70	94.86±0.70	0.442±0.021	0.484±0.013	Dynamic
Baseline + M	89.61±4.22	96.91±1.68	0.946±0.041	0.272±0.059	88.09±0.76	95.16±0.54	0.425±0.026	0.499±0.018	Dynamic
Baseline + S	89.98±2.83	97.23±1.32	0.967±0.074	0.248±0.083	88.56±1.25	95.81±0.53	0.419±0.014	0.496±0.013	Dynamic
dFCExpert	**91.03±1.71**	**97.40±1.30**	**0.932±0.036**	**0.296±0.049**	**89.28±0.83**	**95.99±0.52**	**0.412±0.012**	**0.513±0.009**	Dynamic

factors, α and β, are used to balance the contributions of these components:

$$L_{overall} = L_{task} + \alpha(L_{balance} + L_{sparse}) + \beta L_{state}. \tag{10}$$

After experimenting with various scaling ratios, we set $\alpha = 1$ and $\beta = 10$ for optimal performance. Note that the auxiliary losses for modularity experts are computed as the summation across all MoE-GIN layers.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets

We conducted experiments on two fMRI datasets. (a) *Human Connectome Project (HCP)*: We used data from the HCP S1200 release (Van Essen et al., 2013), which includes preprocessed and ICA denoised (Glasser et al., 2013) resting-state fMRI scans of 1071 individuals with 578 females and 493 males. Following (B.-H. Kim et al., 2021), we selected data from the first run and excluded scans with fewer than 1200 time points. All the resting-state fMRIs used in this study contain exactly 1200 time points. To build functional networks, we parcellated the brain into 400 regions using the Schaefer atlas (Schaefer et al., 2018), which is organized into seven intrinsic connectivity networks. (b) *Adolescent Brain Cognitive Development Study (ABCD)*: The study is a multi-site investigation of brain development and its relationship with behavioral outcomes in children aged 9-10 years (Casey et al., 2018). In our study, we used preprocessed baseline resting-state fMRI data from the ABCD BIDS Community Collection (ABCC) (Feczko et al., 2021). As in (Keller et al., 2023), participants with incomplete data, excessive head motion, or fewer than 600 remaining time points after motion censoring were excluded. After quality

control, our final dataset included resting-state fMRI scans of 6,165 children, with the time points ranging from 626 to 3516. The Gordon atlas with 352 regions ($N = 352$) was used for brain parcellation, as defined by the ABCC.

Both datasets are publicly available, with all identification information anonymized. We created 5 stratified splits for both datasets with a train-validation-test ratio of 7:1:2. Average results and standard deviations across the 5 splits were reported.

4.1.2 Targets and Metrics

We selected two evaluation tasks: sex classification and cognitive intelligence prediction ("fluid intelligence" for HCP and "general cognition" for ABCD). The sex classification task is a binary classification problem, and the cognitive intelligence prediction is a regression problem. These tasks were chosen to explore brain-biology and brain-cognition associations. For the sex classification task, we used accuracy (ACC) and AUC (Area Under ROC Curve) as evaluation metrics. For the regression task, the regression targets were z-normalized and the performance was evaluated using Mean Squared Error (MSE) and correlation coefficient (CORR) between the measured and predicted values.

4.1.3 Implementation Details

Our method was implemented using PyTorch (Paszke et al., 2019) and trained on an NVIDIA A100 GPU with 80 GB memory. We set the number of MoE-GIN layers to $L = 3$ and the embedding dimension to $D = 256$. For dFC graph construction, we used a window length of $\Gamma = 50$ and a window stride of $S = 3$, following common protocols for sliding-window-based dFC analyses (Preti et al., 2017; Zalesky & Breakspear, 2015). Consistent with (B.-H. Kim et al., 2021), we randomly selected 600 time points at each training step for the dFC graph computation. This procedure reduced computational and memory overhead while augmenting the training data. For testing, we utilized the full time-course matrix $\mathbf{P} \in \mathbb{R}^{N \times T_{\max}}$ to construct dFC graphs and evaluate the model. The sex classification and cognitive intelligence regression tasks were trained using cross-entropy and MSE losses, respectively. To ensure a fair comparison, we trained both dFCExpert and known methods under comparison in an end-to-end supervised fashion using a similar configuration:

- *Optimizer*: Adam.
- *Learning rate*: A learning rate of $5e^{-4}$ for training the classification task, and a learning rate of $1e^{-3}$ for training the regression task.

Table 2: Impact of the number of experts in the modularity experts on performance.

#Experts	HCP				ABCD			
	Sex		Intelligence		Sex		Cognition	
	ACC(%)	AUC(%)	MSE(↓)	CORR(↑)	ACC(%)	AUC(%)	MSE(↓)	CORR(↑)
1 (baseline)	88.11±4.25	96.68±1.71	0.978±0.079	0.238±0.086	87.25±0.70	94.86±0.70	0.442±0.021	0.484±0.013
3	88.68±2.57	96.46±1.58	0.961±0.067	0.266±0.061	87.57±0.94	95.16±0.70	0.435±0.019	0.481±0.005
5	89.25 ±3.46	96.41±1.27	0.953±0.054	**0.279±0.042**	**88.09±0.76**	**95.16±0.54**	**0.425±0.026**	**0.499±0.018**
7	**89.61±4.22**	**96.91±1.68**	**0.946±0.041**	0.272±0.059	87.80±1.06	94.85±0.65	0.429±0.023	0.481±0.016
9	88.73±3.48	95.26±2.10	0.963±0.036	0.224±0.051	87.65±0.87	94.98±0.69	**0.425±0.020**	0.479±0.010
17	87.83±3.83	95.87±1.47	0.976±0.032	0.209±0.069	87.19±1.30	94.84±0.53	0.431±0.024	0.477±0.020

- *Mini-batch size*: mini-batch size of 8 for the HCP dataset and mini-batch size of 16 for the ABCD dataset.
- *Epochs*: 50 epochs for the HCP dataset, and 30 epochs for the ABCD dataset.
- *Number of experts*: $C = 7$ for HCP dataset and $C = 5$ for ABCD dataset.
- *Number of states*: $K = 7$ for HCP dataset and $K = 5$ for ABCD dataset.

Based on our experiments, C can be set to 5 or 7, and K can range from 5 to 7. These values for C and K fall within a reasonable and effective range for our model's performance.

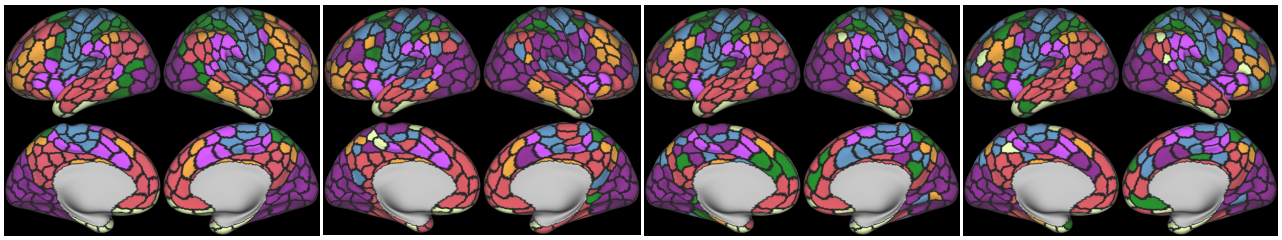
4.2 Sex Classification and Cognitive Intelligence Regression Results

4.2.1 Comparison with Existing Methods

We evaluated the performance of dFCExpert against alternative *static-* and *dynamic-FC methods* on both the sex classification and cognitive intelligence regression tasks. The results of these alternative methods are summarized in the top two blocks of Table 1. For the sex classification task, dFCExpert outperformed all the alternative methods under comparison on both the HCP and ABCD datasets. On the regression task, our method demonstrated a significant performance improvement, particularly on the HCP dataset, achieving an increase of nearly 7 points. These quantitative results highlight the importance and effectiveness of incorporating brain modularity and learning state patterns to model brain dynamics in fMRI data, underscoring the advantages of our approach.

4.2.2 Ablation Studies

We conducted ablation experiments to evaluate the contributions of individual components of dFCExpert, specifically the modularity experts and state experts. For fair comparison, the baseline model was a 3-



(a) Schaefer Atlas - 7 Networks (b) Averaged (c) Male (d) Female

Figure 3: Visualization of the learned brain modules (nodes assignment results) averaged across all subjects (b) and for two randomly selected individuals (c) and (d) from the test dataset, compared to the seven intrinsic functional networks from the atlas (a). Regions with matching colors in (a) and (b)(c)(d) indicate consistent assignments.

layer GIN followed by an MLP layer. The GIN was used to learn graph-level features, and the MLP made predictions for each segment t, which were then averaged to generate the final result. The baseline model was chosen for the following reasons: 1) Generating predictions for each temporal segment provides stronger supervision, potentially improving prediction compared to the alternatives that use RNNs or transformers on top of the GIN layer (e.g., STAGIN (B.-H. Kim et al., 2021) and NeuroGraph (Said et al., 2023), as shown in Table 1); 2) Since the state experts aggregate the temporal graph features into K state features, the use of RNNs or transformers for final predictions is unnecessary.

Based on this baseline, we introduced modularity experts ("Baseline + M") by replacing GIN with MoE-GIN and state experts ("Baseline + S") by aggregating temporal features before the MLP layer, respectively. The third block of Table 1 summarizes the ablation study results, from which several key observations can be made: 1) Compared with the baseline model, the modularity experts achieved significantly improved performance, particularly for the regression task on the HCP dataset, highlighting the effectiveness of characterizing brain modularity for learning more informative brain graph features; 2) By grouping the temporal FC measures into several states, the state experts achieved better performance across both tasks and both datasets, underscoring the benefit of using this new approach to explore the dynamics of FC measures; 3) Incorporating both the modularity and state experts, dFCExpert outperformed the baseline model, validating the overall effectiveness of our method. These results demonstrate that both components play a crucial role in improving the model's ability to capture dynamic brain connectivity and its association with cognitive and biological outcomes.

4.3 Modularity Experts Analysis

We evaluated the effectiveness of our modularity experts from the following perspectives: 1) the impact of the number of experts in the MoE-GIN layer; 2) the ability of modularity experts to learn specific

brain modules; 3) the performance of modularity experts when applied to *static-FC* measures; and 4) the effect of different scaling ratios for the auxiliary losses.

4.3.1 Different Number of Experts

To evaluate the effect of modularity experts on model performance, we investigated the impact of the number of experts (C) in the MoE-GIN layer. Specifically, we tested C values of 3, 5, 7, 9, 17, where $C = 7$ and $C = 17$ are widely used settings in large-scale functional network studies (Yeo et al., 2011). The results are summarized in Table 2 and reveal the following insights: 1) On the HCP dataset, the model's performance improved as the number of experts increased from 3 to 7 but began to decline when the number of experts increased further from 7 to 17, suggesting that a smaller number of experts is not only optimal for performance but also beneficial for reducing computational cost and model complexity. 2) On the ABCD dataset, the modularity experts outperformed the baseline across all settings. While the number of experts had a smaller effect on the classification task, the performance on the regression task improved significantly as C increased, peaking at $C = 9$. From these results, $C = 5$ or $C = 7$ emerged as reasonable choices, aligning with the typical number of functional modules in neuroscience. We ultimately selected $C = 7$ for the HCP dataset and $C = 5$ for the ABCD dataset. The differing optimal number of experts between the datasets may be attributed to two factors: the brain atlas used for brain parcellation and the intrinsic characteristics of the datasets, as HCP (young adults) and ABCD (young kids) represent different age groups and developmental stages.

4.3.2 Learned Brain Modules Visualization

To further evaluate the effectiveness of the modularity experts, we visualized the node assignment results produced by the modularity experts with $C = 7$ on the classification task (HCP dataset) and compared them with the Schaefer Atlas (Schaefer et al., 2018) that defines 7 functional networks (Figure 3 (a)). Specifically, the modularity experts generated 7 learned modules, each consisting of nodes assigned to a specific expert. Since the correspondence between these expert-learned modules and the 7 atlas networks was unknown, we matched them based on the maximum overlap of their respective nodes, and the matched results are visualized in Figure 3. The visualization demonstrates a strong overlap between the Schaefer atlas networks and the modules learned by the modularity experts, indicating that the modularity experts effectively grouped tightly connected nodes into the same module, allowing each expert to specialize in capturing nodes corresponding to a particular brain functional module. Moreover, because the modularity experts were optimized for specific tasks, they could potentially learn node-expert

Table 3: Performance of modularity experts (MoE-GIN) applied to **static-FC** measures.

Method	GNN Type	HCP				ABCD			
		Sex		Intelligence		Sex		Cognition	
		ACC(%)	AUC(%)	MSE(↓)	CORR(↑)	ACC(%)	AUC(%)	MSE(↓)	CORR(↑)
NeuroGraph (Said et al., 2023)	GIN	84.38±2.60	92.03±2.39	1.321±0.043	0.198±0.065	85.71±0.60	93.75±0.61	0.512±0.013	0.423±0.015
NeuroGraph (Said et al., 2023)	MoE-GIN	87.30±1.33	95.56±1.68	0.989±0.035	0.243±0.052	87.08±0.74	93.12±0.52	0.486±0.021	0.477±0.017

Table 4: Performance of modularity experts with different scaling ratios α (α_b and α_s represent the scaling ratios for $L_{balance}$ and L_{sparse}, respectively).

α	α_b, α_s	Sex		Intelligence	
		ACC(%)	AUC(%)	MSE(↓)	CORR(↑)
0	$\alpha_b = 0, \alpha_s = 0$	88.56±3.45	94.67±1.68	0.962±0.022	0.256±0.048
0.1	$\alpha_b = 0.1, \alpha_s = 0.1$	89.16±4.12	95.52±2.01	0.948±0.032	0.270±0.036
1	$\alpha_b = 1, \alpha_s = 1$	89.61±4.22	96.91±1.68	0.946±0.041	0.272±0.059
-	$\alpha_b = 1, \alpha_s = 0$	87.83±4.05	94.84±2.19	0.965±0.028	0.252±0.050
-	$\alpha_b = 0, \alpha_s = 1$	88.65±3.52	95.13±1.53	0.956±0.023	0.261±0.043

(or node-network) assignments that go beyond the prior knowledge encoded in the Schaefer Atlas. From Figure 3 (c)(d), we observed individual differences in the resulting learned modules across subjects, highlighting the potential of modularity experts to learn personalized functional networks (H. Li et al., 2017, 2023), adapting to the unique FC patterns of each individual.

4.3.3 Modularity Experts in Static-FC Method

To investigate the effectiveness of modularity experts in static FC analysis, we integrated the MoE-GIN layer into Neurograph framework(Said et al., 2023) that employs a GNN architecture with residual connections and combines hidden representations from message passing at each layer by concatenation. These combined representations are then processed by an MLP layer for predictions. In Table 3, incorporating the modularity experts into a *static-FC* method enhanced brain graph representations and improved performance, particularly on the HCP dataset. This highlights the versatility of modularity experts in various FC analyses.

4.3.4 Scaling Ratio α Analysis for Auxiliary Losses

We analyzed the impact of different scaling ratios α for the two auxiliary losses in the modularity experts on the HCP dataset, experimenting with $\alpha \in \{0.1, 1\}$. The results are summarized in Table 4. When $\alpha = 0.1$, performance improved on both tasks, highlighting the importance of the auxiliary losses for balance expert loading and sparse gating. With $\alpha = 1$, classification performance further improved, although the enhancement on the regression task was less significant. For consistency, we used $\alpha = 1$ in all subsequent experiments. To further validate the effectiveness of the proposed loading balance loss

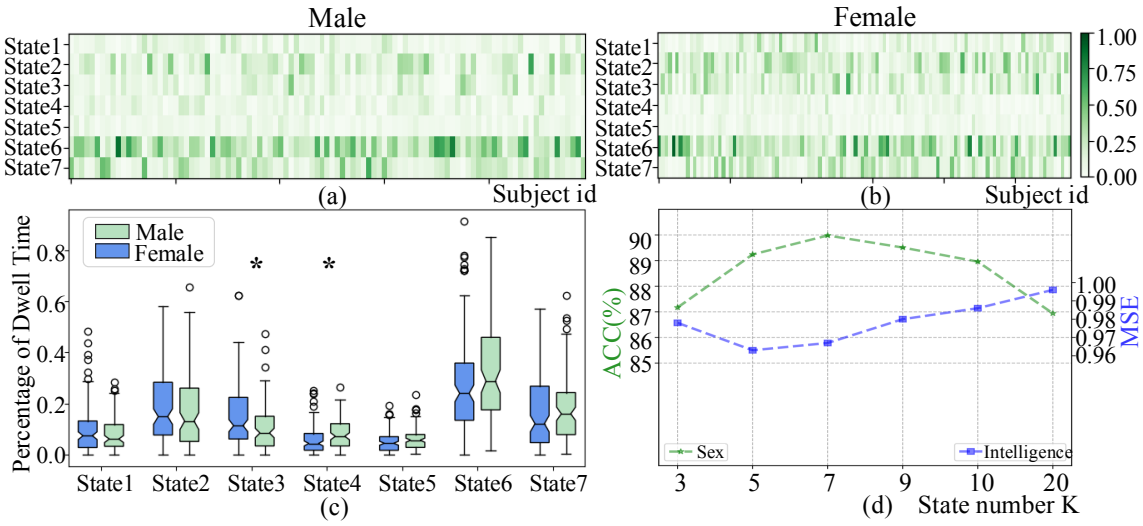


Figure 4: **a)** and **b)** Probabilities of temporal segments of individual males and females assigned to specific states, with the x-axis representing individuals in the testing dataset; **c)** Boxplots summarizing **a)** and **b)**, where each box corresponds to a row in **a)** or **b)**, and $\star$ indicates statistically significant difference; **d)** Model performance with different numbers of the dFC states.

$L_{balance}$ and sparse gating loss L_{sparse}, we conducted experiments by removing either $L_{balance}$ or L_{sparse}. The results showed that the model's performance decreased, underscoring their role in enabling balanced selection and specialization of experts.

4.4 State Experts Analysis

To assess the impact of prototype learning for identifying dFC states, we conducted experiments on HCP dataset to explore the performance of different numbers of states, visualize how these learned states provide insights for interpretability, and analyze the effect of the scaling factor β for L_{state}.

4.4.1 Different Numbers of States

We examined the effect of varying the number of states K within the range of 3, 5, 7, 9, 10, and 20. Figure 4(d) shows how the classification and prediction performance changed with the number of states. Consistent with observations from the modularity analysis, performance initially improved as K increased but began to degrade beyond a certain number of states. The best performance was achieved with $K = 7$ for the classification task and $K = 5$ for the regression task. This also aligns with the findings in existing dFC studies, which suggest that the typical number of dynamic states ranges from 5 to 7 (Damaraju et al., 2014).

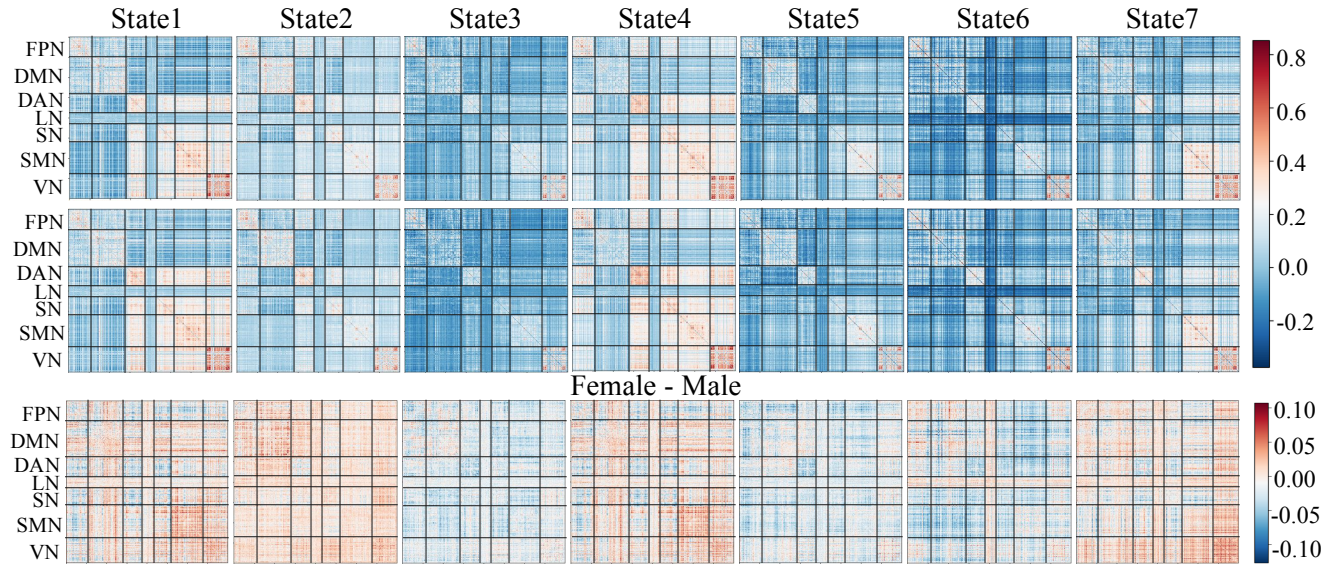


Figure 5: Visualization of the averaged FC measures of temporal segments for males and females assigned to a specific state. The first row represents the female group, the second row represents the male group, and the third row illustrates the FC differences (female-male) between the two groups for each state. The y-axis corresponds to the seven networks: FPN, DMN, DAN, LN, SN, SMN, and VN (also see Figure 1(a)).

4.4.2 Visualization of Learned dFC States

We visualized the state assignment results to assess whether the learned states provide interpretable evidence for distinguishing males from females. Specifically, we averaged the soft assignment results P across the T dimension, such that each resulting value represented the fraction of T temporal segments for a subject assigned to a specific state (e.g., the k-th state). We computed these values for all subjects in the HCP test dataset and visualized them separately for males and females. The results shown in Figure 4 (a) and (b), along with the boxplots in Figure 4 (c) reveal clear patterns. Male dFC graphs (temporal segments) were more frequently assigned to States of 4 and 6, while female dFC graphs were more frequently allocated to States of 2 and 3. Additionally, Wilcoxon rank-sum tests showed significant differences at State 3 and State 4 between males and females, with p-values less than 0.05. These findings validate that the dFC states reflect explainable evidence of sex differences, highlighting distinct brain FC patterns associated with each state. Overall, these results demonstrate the effectiveness of grouping dynamic FC measures into a smaller set of dFC states. They also illustrate that the state experts can identify distinctive dFC states comparable to findings from conventional dFC studies, despite utilizing a learning strategy that differs from traditional approaches.

We further visualized the averaged FCs of temporal segments assigned to each state for males and

Table 5: Results for state experts with different scaling ratios β.

β	Sex		Intelligence	
	ACC(%)	AUC(%)	MSE(↓)	CORR(↑)
0	86.75±2.43	93.52±1.12	1.023±0.065	0.210±0.072
1	88.23±2.12	96.28±1.85	0.974±0.044	0.268±0.063
10	89.98±2.83	97.23±1.32	0.967±0.074	0.248±0.083

females. As shown in Figure 5, the state experts effectively captured distinct dynamic patterns in the dFC measures. For instance, dFC graphs (networks) for females exhibited stronger FC in State 2 and also in States 1, 4, and 7, while males had more stronger FC in States 3, 5, and 6. These observed sex differences and FC strength differences in these averaged state FC measures underscore the interpretability of state experts, highlighting the state experts's ability to capture meaningful variations in dFC.

4.4.3 Scaling Factor β Analysis for L_{state}

We investigated the impact of the scaling factor β for the clustering loss L_{state} on the HCP dataset. As shown in Table 5, $\beta = 1$ resulted in improved performance compared to $\beta = 0$, demonstrating the critical role of the clustering loss L_{state} in learning meaningful state patterns. Since L_{state} is relatively small in magnitude, we also tested a larger scaling factor, $\beta = 10$, to amplify its influence. The resulting performance showed further improvement, indicating enhanced state assignment.

5 Conclusions and Discussions

We developed dFCExpert, a novel method for learning effective representations of dFC measures, which consists of modularity experts and state experts, designed to reflect the brain's modular organization and dFC states that have been extensively studied in fMRI research on functional neuroanatomy, brain development and aging, and brain disorders, but underexplored in the machine learning community for functional brain networks. As demonstrated by extensive experiments on two large-scale fMRI datasets, the modularity experts automatically routed network nodes with similar brain functions to the same expert, enabling specialization of experts and promoting effective representation learning of dFC measures. Meanwhile, the state experts grouped the learned temporal graph features into distinctive states, each characterized by similar dFC patterns. This facilitated the effective modeling of temporal dynamics in brain functional networks, revealing insights into different brain states and biological characteristics. These results highlight not only the superior performance of dFCExpert compared to state-of-the-art alternative methods but also its enhanced interpretability, offering a promising tool for understanding

and analyzing functional brain networks.

While dFCExpert demonstrates strong performance and interpretability, several limitations offer opportunities for future improvements. First, the model has been evaluated on two large-scale fMRI datasets (HCP and ABCD), but its generalizability to other datasets with different acquisition protocols, parcellation schemes, or population characteristics remains unverified. Future studies could evaluate the model on diverse datasets, including task-based fMRI, multimodal imaging (e.g., combining fMRI with EEG), or smaller datasets to assess robustness and adaptability. Second, while the modularity and state experts improve interpretability, connecting these findings to specific clinical or neurological outcomes remains a challenge. Integrating the model with clinical datasets and leveraging the learned features for diagnostic or prognostic tasks would enhance its clinical relevance. Third, the framework currently uses predefined atlases, which limit the resolution of brain parcellation. High-resolution parcellation could increase the number of nodes and exacerbate computational challenges. Future research could focus on adapting dFCExpert for finer-grained parcellations or voxel-wise analysis to enable fine-grained investigations. Fourth, while the model captures dFC states effectively, it does not explicitly model state transitions or the temporal relationships between states. Incorporating temporal sequence models, such as Markov chains or advanced sequence-learning architectures (e.g., transformers or RNNs), could provide deeper insights into the dynamics of state transitions. Lastly, the current implementation for sliding window adopts a window length of 50, and window stride of 3. Although it's a standard setting for dFC analysis used in many methods (B.-H. Kim et al., 2021; Preti et al., 2017), different settings of sliding windows need to be further explored, since these settings may yield different dynamics and fluctuations for the brain activities or FCs and have different performance. Addressing these limitations through future research will enhance the robustness, efficiency, and applicability of dFCExpert, paving the way for broader adoption in both scientific and clinical settings.

6 Data and Code Availability

The fMRI data used for the experiments of this paper should be downloaded from the Human Connectome Project and ABCD. The code can be accessed at MLDataAnalytics/dFCExpert on GitHub.

7 Funding

This research was supported in part by funding from the National Institutes of Health under grants EB022573 and AG066650.

References

Akbari, H., Kondratyuk, D., Cui, Y., Hornung, R., Wang, H., & Adam, H. (2024). Alternating gradient descent and mixture-of-experts for integrated multimodal perception. *Advances in Neural Information Processing Systems*, *36*.

Bertolero, M. A., Yeo, B. T., & D'Esposito, M. (2015). The modular and integrative functional architecture of the human brain. *Proceedings of the National Academy of Sciences*, *112*(49), E6798–E6807.

Campbell, A., Zippo, A. G., Passamonti, L., Toschi, N., & Lio, P. (2022). Dyndepnet: Learning time-varying dependency structures from fmri data via dynamic graph structure learning. *arXiv preprint arXiv:2209.13513*.

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., et al. (2018). The adolescent brain cognitive development (abcd) study: Imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, *32*, 43–54.

Damaraju, E., Allen, E. A., Belger, A., Ford, J. M., McEwen, S., Mathalon, D., Mueller, B., Pearlson, G., Potkin, S., Preda, A., et al. (2014). Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *NeuroImage: Clinical*, *5*, 298–308.

Feczko, E., Conan, G., Marek, S., Tervo-Clemmens, B., Cordova, M., Doyle, O., Earl, E., Perrone, A., Sturgeon, D., Klein, R., et al. (2021). Adolescent brain cognitive development (abcd) community mri collection and utilities. *BioRxiv*, 2021–07.

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, *23*(120), 1–39.

Gadgil, S., Zhao, Q., Pfefferbaum, A., Sullivan, E. V., Adeli, E., & Pohl, K. M. (2020). Spatio-temporal graph convolution for resting-state fmri analysis. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23*, 528–538.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, *80*, 105–124.

Greicius, M. (2008). Resting-state functional connectivity in neuropsychiatric disorders. *Current opinion in neurology*, *21*(4), 424–430.

Huang, J., Gong, S., & Zhu, X. (2020). Deep semantic clustering by partition confidence maximisation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8849–8858.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, *3*(1), 79–87.

Jain, Y., Behl, H., Kira, Z., & Vineet, V. (2024). Damex: Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets. *Advances in Neural Information Processing Systems*, *36*.

Jing, R., Lin, X., Ding, Z., Chang, S., Shi, L., Liu, L., Wang, Q., Si, J., Yu, M., Zhuo, C., et al. (2023). Heterogeneous brain dynamic functional connectivity patterns in first-episode drug-naive patients with major depressive disorder. *Human Brain Mapping*, *44*(8), 3112–3122.

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the em algorithm. *Neural computation*, *6*(2), 181–214.

Kan, X., Dai, W., Cui, H., Zhang, Z., Guo, Y., & Yang, C. (2022). Brain network transformer. *Advances in Neural Information Processing Systems*, *35*, 25586–25599.

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

Keller, A. S., Pines, A. R., Shanmugan, S., Sydnor, V. J., Cui, Z., Bertolero, M. A., Barzilay, R., Alexander-Bloch, A. F., Byington, N., Chen, A., et al. (2023). Personalized functional brain network topography is associated with individual differences in youth cognition. *Nature communications*, *14*(1), 8411.

Kim, B.-H., & Ye, J. C. (2020). Understanding graph isomorphism network for rs-fmri functional connectivity analysis. *Frontiers in neuroscience*, *14*, 545464.

Kim, B.-H., Ye, J. C., & Kim, J.-J. (2021). Learning dynamic graph representation of brain connectome with spatio-temporal attention. *Advances in Neural Information Processing Systems*, *34*, 4314–4327.

Kim, S., Lee, D., Kang, S., Lee, S., & Yu, H. (2023). Learning topology-specific experts for molecular property prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*, 8291–8299.

Ktena, S. I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., & Rueckert, D. (2018). Metric learning with spectral graph convolutions on brain connectivity networks. *NeuroImage*, *169*, 431–442.

Li, H., Satterthwaite, T. D., & Fan, Y. (2017). Large-scale sparse functional networks from resting state fmri. *NeuroImage*, *156*, 1–13.

Li, H., Srinivasan, D., Zhuo, C., Cui, Z., Gur, R. E., Gur, R. C., Oathes, D. J., Davatzikos, C., Satterthwaite, T. D., & Fan, Y. (2023). Computing personalized brain functional networks from fmri using self-supervised deep learning. *Medical Image Analysis*, *85*, 102756.

Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L. H., Ventola, P., & Duncan, J. S. (2021). Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, *74*, 102233.

Lin, X., Jing, R., Chang, S., Liu, L., Wang, Q., Zhuo, C., Shi, J., Fan, Y., Lu, L., & Li, P. (2023). Understanding the heterogeneity of dynamic functional connectivity patterns in first-episode drug naıve depression using normative models. *Journal of Affective Disorders*, *327*, 217–225.

Matthews, P. M., & Jezzard, P. (2004). Functional magnetic resonance imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, *75*(1), 6–12.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Preti, M. G., Bolton, T. A., & Van De Ville, D. (2017). The dynamic functional connectome: State-of-the-art and perspectives. *Neuroimage*, *160*, 41–54.

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

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, *28*(9), 3095–3114.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Shen, T., Ott, M., Auli, M., & Ranzato, M. (2019). Mixture models for diverse machine translation: Tricks of the trade. *International conference on machine learning*, 5719–5728.

Smith, S. M., Nichols, T. E., Vidaurre, D., Winkler, A. M., Behrens, T. E., Glasser, M. F., Ugurbil, K., Barch, D. M., Van Essen, D. C., & Miller, K. L. (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature neuroscience*, *18*(11), 1565–1567.

Sporns, O., & Betzel, R. F. (2016). Modular brain networks. *Annual review of psychology*, *67*, 613–640.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The wu-minn human connectome project: An overview. *Neuroimage*, *80*, 62–79.

Wang, H., Jiang, Z., You, Y., Han, Y., Liu, G., Srinivasa, J., Kompella, R., Wang, Z., et al. (2024). Graph mixture of experts: Learning on large-scale graphs with explicit diversity modeling. *Advances in Neural Information Processing Systems*, *36*.

Wang, K., Liang, M., Wang, L., Tian, L., Zhang, X., Li, K., & Jiang, T. (2007). Altered functional connectivity in early alzheimer's disease: A resting-state fmri study. *Human brain mapping*, *28*(10), 967–978.

Wang, Q., Wu, M., Fang, Y., Wang, W., Qiao, L., & Liu, M. (2023). Modularity-constrained dynamic representation learning for interpretable brain disorder analysis with functional mri. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 46–56.

Xie, J., Girshick, R., & Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. *International conference on machine learning*, 478–487.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.-i., & Jegelka, S. (2018). Representation learning on graphs with jumping knowledge networks. *International conference on machine learning*, 5453–5462.

Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*.

Zalesky, A., & Breakspear, M. (2015). Towards a statistical test for functional connectivity dynamics. *Neuroimage*, *114*, 466–470.