Federated Learning for Cross-Institution Brain Network Analysis

Han Xie*, Yi Yang*, Hejie Cui, and Carl Yang†

Emory University, 400 Dowman Drive, Atlanta, USA

ABSTRACT

Recent advancements in neuroimaging techniques have sparked a growing interest in understanding the complex interactions between anatomical regions of interest (ROIs), forming into brain networks that play a crucial role in various clinical tasks, such as neural pattern discovery and disorder diagnosis. In recent years, graph neural networks (GNNs) have emerged as powerful tools for analyzing network data. However, due to the complexity of data acquisition and regulatory restrictions, brain network studies remain limited in scale and are often confined to local institutions. These limitations greatly challenge GNN models to capture useful neural circuitry patterns and deliver robust downstream performance. As a distributed machine learning paradigm, federated learning (FL) provides a promising solution in addressing resource limitation and privacy concerns, by enabling collaborative learning across local institutions (i.e., clients) without data sharing. While the data heterogeneity issues have been extensively studied in recent FL literature, cross-institutional brain network analysis presents unique data heterogeneity challenges, that is, the inconsistent ROI parcellation systems and varying predictive neural circuitry patterns across local neuroimaging studies. To this end, we propose FedBrain, a GNN-based personalized FL framework that takes into account the unique properties of brain network data. Specifically, we present a federated atlas mapping mechanism to overcome the feature and structure heterogeneity of brain networks arising from different ROI atlas systems, and a clustering approach guided by clinical prior knowledge to address varying predictive neural circuitry patterns regarding different patient groups, neuroimaging modalities and clinical outcomes. Comparing to existing FL strategies, our approach demonstrates superior and more consistent performance, showcasing its strong potential and generalizability in cross-institutional connectome-based brain imaging analysis.

Keywords: Brain Connectome Analysis, Computer-aided Diagnosis, Federated Learning

Extended Abstract

In recent years, neuroscience research has focused on unraveling the human brain’s intricacies and its links to disorders like bipolar disorder (BP), HIV, Autism, and Parkinson’s disease (PD). Neuroimaging techniques, such as functional magnetic resonance imaging (fMRI) and diffusion tensor imaging (DTI), create brain networks that are essentially weighted connected graphs with nodes representing anatomical regions of interest (ROIs) and edges denoting functional or structural connections. Analyzing these networks provides insights into neural system structures, aiding disorder detection and advancing neuroscience research. Graph Neural Networks (GNNs) gained popularity for analyzing network data. In neuroimaging study, GNNs can help predict diseases, identify biomarkers, and discover neural patterns. However, neuroimaging datasets are typically small in sample size which easily leads to model overfitting and limited generalization, posing challenges for GNN training.

Recently, federated learning (FL) has emerged as a promising solution to address the challenges of limited training data and computation resources in local studies. FL operates by collaboratively training a centralized model on the central server, based on data privately stored by multiple local clients. The approach offers two notable advantages. First, it ensures privacy preservation since clients solely communicate model parameters with the server, ensuring that individual data remains confidential and is never shared with other parties. Second, it facilitates knowledge discovery and generalization by aggregating information from multiple data owners into a centralized model. By doing so, the central model can adapt and mitigate the overfitting issues typically associated with learning on a single small dataset.

*Equal contributions; †Corresponding author (j.carlyang@emory.edu)
One significant challenge in FL is data heterogeneity, wherein the data distributions significantly differ across local data owners. Several FL algorithms have been proposed to tackle the data heterogeneity challenge, such as incorporating a proximal term or control variate to mitigate the diverging optimization trajectory between local and global models. However, these methods mostly focus on label distributions and fail to address the unique data heterogeneity scenarios in cross-institutional brain network analysis which can manifest in two key aspects. First, since network parcellation is traditionally an ad hoc process carried out by domain experts, it is difficult to assume or require all different institutions to conform to the same ROI atlas mapping systems when preprocessing their neuroimaging data. As a result, this leads to misalignment in network structures and ROI features across clients. Second, different institutions collect brain network data for different patient groups, with different neuroimaging techniques and towards different clinical purposes, which results in varying underlying predictive neural circuitry patterns.

In this work, we propose FedBrain, a personalized FL framework designed for GNN-based brain network analysis. Our framework comprises three key components: a GNN-based FL backbone, a federated atlas mapping mechanism, and a guided client clustering mechanism. To build our FL platform, we use the well-established FedAvg as a foundation, and our default GNN structure is an optimized GCN model adopted from BrainGB. To address the feature- and structure-wise heterogeneity issue due to potentially different atlas mapping systems used across local institutions, we introduce an autoencoder-based atlas mapping mechanism, which aims to project diverse ROI profiles onto a uniform sharable embedding space. To handle heterogeneous predictive neural circuitry patterns due to various patient groups neuroimaging modalities and clinical outcomes, we design a knowledge-guided client clustering mechanism by incorporating prior clinical knowledge (e.g., data modalities and outcome relations) into the dynamic clustering process of clients with similar data during FL.

APPENDIX A. THE FedBrain FRAMEWORK

A.1 The FL Backbone

The backbone FL structure of FedBrain is based on federated averaging (FedAvg) proposed by McMahan et al. The core principle of FedAvg is to aggregate the updated model parameters from local clients through a process of weighted averaging. These averaged parameters are then disseminated back to each client in the subsequent communication round. Specifically, when aggregating parameters, the server assigns a weight to each client in proportion to their respective sample size. The process is outlined in Algorithm 1.

```
Algorithm 1 Federated Averaging (FedAvg)

Input: Number of communication rounds $T$, set of total available clients $\mathbb{C} \leftarrow \{C_i\}_{i=1}^N$, set of total available client data $\mathbb{D} \leftarrow \{D_i\}_{i=1}^N$, learning rate $\eta$

Output: The final server model $w_G^T$

1: Randomly initialize the server model $w_G^0$
2: for $t \leftarrow 1$ to $T$ do
3: Sample a subset $\{C_m, D_m\}$ of $m$ participating clients from $\{\mathbb{C}, \mathbb{D}\}$
4: for each participant $\{C_j, D_j\} \in \{C_m, D_m\}$ in parallel do
5: Download model parameter from server $w_G^{t-1} \leftarrow w_G^{t-1}$
6: Update local parameter $w_C^{t} \leftarrow w_C^{t} - \eta \nabla w_C^{t} \mathcal{L}(w_C^{t}; D_j)$
7: Update the server model $w_G^{t} \leftarrow \sum_{\{C_j, D_j\} \in \{C_m, D_m\}} |D_j| \frac{|D_m|}{|D_m|} w_C^{t}$
```

We utilize an optimized GCN, as proposed by Cui et al., to parameterize both the server and client models. The ROI (i.e., node) features are initialized with the connection profiles (i.e., adjacency). That is, the feature matrix $X$ is equivalent to the adjacency $A$ ($X \equiv A$), where $A$ is parameterized by the node set $V = \{v_n\}_{n=1}^N$ and the weighted edge set $E = V \times V$. 


A.2 Federated Atlas Mapping

Motivation. For brain network data, the ROI (i.e., node) parcellation is determined by the brain atlas. Once a template is chosen, all brain networks within a dataset share the same ROI identities. However, in our cross-institutional setting, different institutions may utilize different parcellation systems. This leads to heterogeneity in both sizes and structures of the parcellated networks, as well as divergent meanings of ROI features (i.e., connectivity profiles). While it is possible to manually convert between atlases, this process is laborious and requires extensive domain expertise. Therefore, we propose a data-driven transformation that aims to align brain network features and structures across institutions, ensuring consistency in network dimensions and physical interpretations of features.

Autoencoder framework. To achieve uniform feature dimensions and network sizes, we employ a one-layer linear autoencoder (AE) to learn a dataset-specific projection. Given a target dimension $M$ that is consistent across all datasets and an input feature $X \in \mathbb{R}^{N \times N} \ (N > M)$, the objective is to learn a linear projection $W \in \mathbb{R}^{N \times M}$, such that the projected representations preserve as much information as possible from the original features. The AE is optimized using the mean-squared-error (MSE) reconstruction objective, denoted as $L_{\text{rec}} = (1/N) \| X - WXW^T \|^2$. Intuitively, the projection $W$ transforms initial features by applying a weighted linear combination on the original dimensions. Consequently, the columns of $W$ learn to assign original dimensions into $M$ groups. We exploit this concept to condense the network structure. To reduce the computational complexity, we formulate an assignment matrix $Z \in \mathbb{R}^{N \times M}$ such that $Z_{i,j} = \mathbb{1}[W_{i,j} \in \arg\max_k (\text{col}_j(W))]$. The matrix $Z$ records the top-$k$ greatest entries per each column in $W$ and zeros out the rest. Ultimately, given a graph adjacency matrix $A (\cong X)$, we construct a compressed network $A'$ by evaluating $A' = Z^T AZ$.

Federated training. Apart from dataset-specific projections, aligning the physical interpretations of projected features across datasets is equally vital to mitigate structure- and feature-level heterogeneity. To achieve this, we leverage the FL approach to train the autoencoders with the intention of obtaining a global atlas projection. However, the architectural sizes of autoencoders across clients can vary due to the differing original data dimensions across local datasets, which makes it challenging to communicate model parameters between local clients and the global server.

To address this issue, we propose a unified mapping method that aims to reduce the size of the global model, which is assumed to be the largest, to accommodate the varying dimensionality of each local dataset. In particular, this approach is intended to be implemented before the client downloads the parameters. Given a global projection $W_G \in \mathbb{R}^{N_G \times M}$ based on the most detailed parcellation template with $N_G$ defined ROIs, and a coarser template with $N_L$ defined ROIs ($N_L < N_G$) employed for local data, our goal is to derive an assignment matrix $P_L \in \mathbb{R}^{N_L \times N_G}$, which ensures the local projection $W_L \in \mathbb{R}^{N_L \times M}$ is distributed through the mapping $W_L = P_L W_G$. To achieve this, we leverage the 3D coordinates of the ROIs, denoted as $D_G \in \mathbb{R}^{N_G \times 3}$ for the global parcellation template and $D_L \in \mathbb{R}^{N_L \times 3}$ for the local template. We first calculate a distance matrix $S \in \mathbb{R}^{N_L \times N_G}$, where $S_{i,j} = d(\text{row}_i(D_L), \text{row}_j(D_G))$ represents the pairwise Euclidean distance between ROIs from the two templates. We then designate $P_{L_{i,j}} = \mathbb{1}[S_{i,j} = \arg\min_j (\text{col}_j(S))]$. This implied that we only consider the minimum entry per each column of $S$. Essentially, we enable $P_L$ to learn a mapping that groups ROIs in the global template with those in the local template, based on their spatial proximity. During each communication round, clients start by downloading the server’s parameter by applying the mapping $W_L = P_L W_G$. Subsequently, each client sends their updated parameters back to the server, employing the inverse mapping $W_L^* = P_L^T W_G^*$.

A.3 Guided Clustering

Motivation. Beyond the discrepancies in network parcellation systems, another significant source of heterogeneity originates from the variability in predictive neural circuitry patterns, encompassing data modalities and clinical outcomes. These variances can result in a suboptimal adaptation of the generalized global model to specific local objectives. Therefore, our aim is to strike a balance between global generalization and local personalization. Moreover, as shown in Table 1, we notice that similar neural patterns are shared among certain client institution subgroups. This motivates us to integrate client clustering into the FL process.
Clustered FL. When data distributions are similar among local clients, the average global model can optimize all their objectives concurrently, resulting in client gradients approaching zero as they reach their local optima. However, in instances of heterogeneity, where the global model fails to adapt to local optimizations, local models stop improving, and their gradients become stationary. Consequently, there is a need for a criterion to recognize this occurrence. Given a set of clients \( C = \{C_i\}_{i=1}^N \), their data distributions \( D = \{D_i\}_{i=1}^N \), gradients \( \Theta = \{\Delta \theta_i\}_{i=1}^N \), and a hyperparameter \( \epsilon_1 \), we define the criterion as follows:

\[
0 \leq \left\| \sum_{i=1}^N \frac{|D_i|}{|D|} \Delta \theta_i \right\| < \epsilon_1
\]  

(1)

Simultaneously, if the gradient norms of some clients deviate significantly from the stationary point, suggestive of high heterogeneity, an additional criterion is necessary. For this, we introduce a second hyperparameter \( \epsilon_2 \), and define the additional criterion as follows:

\[
\max (\|\Delta \theta_i\|) > \epsilon_2 > 0
\]

(2)

Clustering commences once both criteria are satisfied. Specifically, we employ a bottom-up hierarchical approach to merge clients into sub-clusters and sub-clusters into larger clusters. The distances between clients are calculated using pairwise cosine similarities of layer-wise gradient norms, while cluster distances are determined upon average linkage. Each cluster then initiates a dedicated FL subroutine with cluster-specific model aggregation. Clusters may further subdivide according to the criteria evaluated at each communication round.

Constrained clustering. While gradient-based client clustering effectively addresses the stationary point issue and improves performance over the basic FedAvg, the method is entirely data-driven, lacking consideration of shared clinical prior knowledge related to the neural circuitry patterns of each client. Consequently, heterogeneity may still exist within the formed clusters, necessitating further division of clusters. This often leads to the creation of singleton clusters, undermining the essence of collaborative learning. This phenomenon is demonstrated in Figure 2 (Section B.4). Based on these observations, we propose a refined variant of the clustering method that incorporates shared prior knowledge to guide the clustering process. For instance, in terms of data modalities, it is intuitive to group clients with similar ROI connectivities and MRI data. Likewise, with regard to clinical outcomes, FL on a cluster level could benefit from learning similar objectives. To this end, we create must-links between pairs of clients that exhibit highly similar neural patterns and define cannot-links for those that don’t. We introduce a weighted reward \( \lambda_{\text{must}} \) and penalty \( \lambda_{\text{cannot}} \) term, which are multiplied by the pairwise client similarity measure when must-links and cannot-links are identified, respectively. We detail this entire process in Algorithm 2.

APPENDIX B. EXPERIMENTS

Datasets. We evaluate our framework using six real-world brain network datasets: BP,12 HIV,13 PPMI,14 PNC,15 ABIDE,16 and ABCD.17 We present key statistics for each dataset in Table 1. Among them, BP, HIV, and PPMI contain multiple data modalities. In light of this, we propose to employ every such modality to be learned on a separate FL client. Based on the available label information, we define two possible tasks – disease prediction (i.e., patients vs. health controls) and gender prediction – both in the form of binary classification.

To ensure the safety and privacy of the participants, all data used in this study strictly adhere to the Good Clinical Practice guidelines and U.S. 21 CFR Part 50 (Protection of Human Subjects) and are approved by the Institutional Review Board (IRB) with no personally identifiable information being used or disclosed.

Baselines. We begin by comparing our proposed framework with self-train, a non-FL baseline. This comparison aims to validate whether individual client performance can be enhanced through collaborative training. Additionally, we benchmark FedBrain against three commonly used FL baselines: FedAvg,4 FedProx,5 and SCAFFOLD.6 It is worth noting that the latter two baselines are specifically designed to handle generic data and system heterogeneity, and their effectiveness in adapting to brain network learning is yet to be explored.
Algorithm 2 Guided Clustering.

Input: Set of total available clients \( C \leftarrow \{C_i\}_{i=1}^M \) to consider, their respective layer-wise gradient norms \( \Theta \leftarrow \{\Delta \theta_i\}_{i=1}^M \), their respective shared neural circuitry patterns \( \Psi \leftarrow \{\psi_i\}_{i=1}^M \), reward \( \lambda_{\text{must}} \) and penalty \( \lambda_{\text{cannot}} \) weight, desired number of clusters to produce \( r \).

Output: Cluster assignments \( S \leftarrow \{s_1, s_2, \ldots, s_r\} \).

1: \hspace{1em} \textbf{procedure} Clustering(\( C, \Theta, \Psi, \lambda_{\text{must}}, \lambda_{\text{cannot}}, r \))
2: \hspace{2em} Make every client its own cluster \( S \leftarrow \{|C_1, \Delta \theta_1, \psi_1\}, \ldots, \{|C_M, \Delta \theta_M, \psi_M\}| \)
3: \hspace{2em} \textbf{while} \( |S| > r \) \hspace{2em} \triangleright \text{Check if number of clusters is greater than } r
4: \hspace{3em} \textbf{for} every pair of clusters \( (S_i, S_j) \) \textbf{in} \( S \) \hspace{2em} \triangleright \text{Such that } i, j \in |S|, i \neq j
5: \hspace{4em} Calculate the inter-cluster linkage distance \( d(S_i, S_j) \leftarrow \text{LINKAGE}(S_i, S_j) \)
6: \hspace{4em} Find the pair with min linkage distance \( \min(d(S_i, S_j)) : i, j \in |S|, i \neq j \) and merge
7: \hspace{2em} \textbf{return} \( S : |S| = r \)
8: \hspace{1em} \textbf{procedure} Linkage\( (S_1, S_2) \)
9: \hspace{2em} \textbf{for} every pair of clients \( \{|C_p, \Delta \theta_p, \psi_p\}, \{|C_q, \Delta \theta_q, \psi_q\}\} \) \textbf{in} \( (S_1, S_2) \) \hspace{2em} \triangleright p \in |S_1|, q \in |S_2|
10: \hspace{3em} Calculate the cosine distance \( \cos(p, q) \leftarrow 1 - ((\Delta \theta_p \cdot \Delta \theta_q)/(\|\Delta \theta_p\|\|\Delta \theta_q\|)) \)
11: \hspace{3em} Determine if the pair forms must- or cannot-link \( \text{link}(p, q) \leftarrow \text{VALIDATE}(\psi_p, \psi_q) \)
12: \hspace{3em} Shorten their cosine distance if must-link \( \cos(p, q) \leftarrow \lambda_{\text{must}} \cdot \cos(p, q) \)
13: \hspace{3em} Increase their cosine distance if cannot-link \( \cos(p, q) \leftarrow \lambda_{\text{cannot}} \cdot \cos(p, q) \)
14: \hspace{2em} \textbf{return} \text{Averaged distance } \text{avg}(\cos(p, q)) : \forall p \in |S_1|, \forall q \in |S_2|)
15: \hspace{1em} \textbf{procedure} Validate\( (\psi_1, \psi_2) \) \hspace{2em} \triangleright \text{An example to determine must- or cannot-links}
16: \hspace{2em} \textbf{return} \text{Must-link if overlapping attributes in } (\psi_1, \psi_2) \text{ exceeds 80%}
17: \hspace{2em} \textbf{return} \text{Cannot-link if overlapping attributes in } (\psi_1, \psi_2) \text{ subceeds 20%}

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Modality</th>
<th>Sample Size</th>
<th>Atlas</th>
<th>Network Size</th>
<th>Outcome</th>
<th>Class Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP</td>
<td>fMRI, DTI</td>
<td>97</td>
<td>Brodmann 82</td>
<td>82 × 82</td>
<td>Disease</td>
<td>2</td>
</tr>
<tr>
<td>HIV</td>
<td>fMRI, DTI</td>
<td>70</td>
<td>AAL 90</td>
<td>90 × 90</td>
<td>Disease</td>
<td>2</td>
</tr>
<tr>
<td>PPMI</td>
<td>PICo, Hough, FSL</td>
<td>754</td>
<td>Desikan-Killiany 84</td>
<td>84 × 84</td>
<td>Disease</td>
<td>2</td>
</tr>
<tr>
<td>PNC</td>
<td>fMRI</td>
<td>503</td>
<td>Power 264</td>
<td>264 × 264</td>
<td>Gender</td>
<td>2</td>
</tr>
<tr>
<td>ABIDE</td>
<td>fMRI</td>
<td>1009</td>
<td>Craddock 200</td>
<td>200 × 200</td>
<td>Disease</td>
<td>2</td>
</tr>
<tr>
<td>ABCD</td>
<td>fMRI</td>
<td>7901</td>
<td>HCP 360</td>
<td>360 × 360</td>
<td>Gender</td>
<td>2</td>
</tr>
</tbody>
</table>
Parameter setup. The GCN\textsuperscript{18} model contains a hidden size of 32, ReLU activations, and dropout layers with a probability of 80%. The graph-level representations are obtained through sum pooling. The downstream classifier consists of a single-layer MLP, and we use the negative log-likelihood loss as the optimization objective and accuracy as the evaluation metric.

Throughout our experiments, we employ a batch size of 32 and use the Adam\textsuperscript{19} optimizer with a learning rate of $1 \cdot 10^{-4}$ and an L2 regularization weight of $5 \cdot 10^{-4}$. In the case of all FL baselines, a complete training procedure encompasses 80 communication rounds, with each local epoch set to 1. For the self-train baseline, each local model is trained for 80 epochs. The $\mu$ value of FedProx is set to 0.01. Regarding FedBrain, we retain only the top 3 entries in each column of the atlas mapping projection matrix for network transformation, and use the most detailed HCP 360 template to define the global model for our federated training of AEs. The clustering criteria $\epsilon_1$ and $\epsilon_2$ of FedBrain are set to 1.50 and 0.05, respectively, and the weighted terms $\lambda_{\text{must}}$ and $\lambda_{\text{cannot}}$ are set to 0.5 and 2.0, respectively. Lastly, our algorithm is aimed to produce a minimum of 2 clusters.

Research questions. To comprehensively evaluate the effectiveness and contribution of our proposed framework, we formulate four research questions as follows that will guide our empirical investigations:

- **RQ1**: How does FedBrain compare to other widely adopted FL frameworks in cross-institutional brain network analysis?
- **RQ2**: How do the proposed federated atlas mapping and guided clustering mechanisms individually contribute to the overall performance?
- **RQ3**: How effective is federated atlas mapping in addressing structure- and feature-level heterogeneity arising from inconsistent ROI parcellation systems?
- **RQ4**: How does the incorporation of clinical prior knowledge guidance contribute to the formation of clusters and impact the overall performance?

The following sections B.1 - B.4 answer these research questions separately.

B.1 Overall performance comparison (RQ1)

Table 2. Overall performance comparison. We present performance for each client averaged from 10-fold cross-validation, and a combined performance averaged across all clients. We highlight the best result in bold and the runner-up is underlined.

<table>
<thead>
<tr>
<th>Clients</th>
<th>BP-fMRI</th>
<th>BP-DTI</th>
<th>HIV-fMRI</th>
<th>HIV-DTI</th>
<th>PPMI-PICo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.5463(±0.019)</td>
<td>0.6037(±0.073)</td>
<td>0.6084(±0.117)</td>
<td>0.5800(±0.021)</td>
<td>0.5037(±0.037)</td>
</tr>
<tr>
<td>self-train</td>
<td>0.5012(±0.082)</td>
<td>0.5158(±0.013)</td>
<td>0.5853(±0.085)</td>
<td>0.6400(±0.132)</td>
<td>0.6479(±0.097)</td>
</tr>
<tr>
<td>FedAvg</td>
<td>0.5286(±0.035)</td>
<td>0.5457(±0.153)</td>
<td>0.6200(±0.132)</td>
<td>0.6629(±0.057)</td>
<td>0.7925(±0.002)</td>
</tr>
<tr>
<td>FedProx</td>
<td>0.4571(±0.140)</td>
<td>0.5000(±0.078)</td>
<td>0.6029(±0.097)</td>
<td>0.6629(±0.057)</td>
<td>0.7925(±0.002)</td>
</tr>
<tr>
<td>SCAFFOLD</td>
<td>0.6394(±0.034)</td>
<td>0.7925(±0.002)</td>
<td>0.7925(±0.002)</td>
<td>0.7778(±0.0000)</td>
<td>0.5789(±0.066)</td>
</tr>
<tr>
<td>FedBrain</td>
<td>0.7389(±0.066)</td>
<td>0.7500(±0.077)</td>
<td>0.7857(±0.071)</td>
<td>0.8143(±0.070)</td>
<td>0.8102(±0.010)</td>
</tr>
</tbody>
</table>

We present a comprehensive performance comparison in Table 2. We include the client (i.e., dataset) name, along with its modality name if it contains multiple; average accuracy per each client; combined accuracy across all clients; and the minimum client-wise gain over the self-train baseline. To ensure fair comparisons, we apply the same GNN architecture and parameter setup to all methods. Our analysis reveals several key observations.

Firstly, FL baselines show significant improvement over self-train, with an average relative gain of 15.34% across all clients. Notably, clients with smaller sample sizes, like BP, HIV, and PNC, experience the most substantial performance enhancement, with an average relative gain of 19.31%. This highlights the valuable
Table 3. Performance comparison on atlas mapping and its variants.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>average</th>
<th>min gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Atlas Mapping</td>
<td>0.6845(±0.068)</td>
<td>–</td>
</tr>
<tr>
<td>Atlas Mapping</td>
<td>0.7246(±0.063)</td>
<td>0.0039</td>
</tr>
<tr>
<td>Federated Atlas Mapping</td>
<td>0.7605(±0.052)</td>
<td>0.0214</td>
</tr>
</tbody>
</table>

Table 4. Performance comparison on guided clustering and its variants.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>average</th>
<th>min gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Clustering</td>
<td>0.6921(±0.071)</td>
<td>–</td>
</tr>
<tr>
<td>Non-guided Clustering</td>
<td>0.7231(±0.065)</td>
<td>0.0000</td>
</tr>
<tr>
<td>Guided Clustering</td>
<td>0.7605(±0.052)</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

effect of collaborative learning and cross-institutional knowledge generalization in overcoming model overfitting on limited training resources. Moreover, FL training also results in slight performance improvements on larger datasets, such as PPMI, ABIDE, and ABCD, underscoring the positive impact of a global optimization scheme in enhancing local performance. However, it is worth noting that among the chosen FL baselines, there is a slightly increased performance variance across clients, mainly due to underlying heterogeneity arising from the unique characteristics of brain network data.

Secondly, among all the selected FL baselines, SCAFFOLD stands out as the top performer, exhibiting an impressive average gain of 5.89% over its competitors. This result highlights the robustness of SCAFFOLD in addressing client heterogeneity through controlled gradient correction. Additionally, along with FedProx, which is also capable of handling data and system heterogeneity, the performance variance is reduced compared to FedAvg. This further aligns with our motivation to develop a specialized solution for reducing brain network-specific heterogeneity, which is aimed to unleash the full potential of collaborative learning, reflected through enhanced performance across multiple datasets at greater consistency.

Lastly, FedBrain outperforms SCAFFOLD by a relative margin of 14.29%, while also significantly reducing performance variance across clients, indicating the value of tailoring FL approaches to consider the unique properties and characteristics of brain network data. Moreover, FedBrain demonstrates statistically significant improvements over the compared baselines, as validated by passing the paired t-test with \( p = 0.05 \) in comparison to all methods.

B.2 Ablation studies (RQ2)

We analyze the two key components of FedBrain: federated atlas mapping and guided clustering. To highlight the contribution of each, we keep the best configuration of one component fixed while evaluating the other. The results are presented in Table 3 and Table 4, where we present an averaged performance across all clients. Regarding the analysis for atlas mapping, we investigate its impact on overall performance both without the entire module and without federated training. When atlas mapping is not applied, we add a learnable linear projection head to the client’s GNN model that is excluded from the FL process. In general, we make two main observations: (1) Ensuring consistency in feature and network dimensions reflects in a relative gain of 6.12% compared to the uncompressed baseline. (2) Aligning the physical meanings of projected features further boosts performance by 4.95%, showcasing its effectiveness in countering incongruous ROI parcellation systems.

Regarding client clustering, we explore the impact on overall performance by comparing two scenarios: one without clustering and one without shared prior knowledge guidance. Our key observations are as follows: (1) Personalizing client optimization through similarity-based clustering leads to a significant enhancement in downstream performance, with a relative margin of 4.48%. (2) By integrating clinical prior knowledge and applying relevant constraints, we further enhance cluster-specific learning and knowledge generalization, resulting in a relative gain of 5.17% and a reduction in performance variance.

B.3 Heterogeneity analysis of federated atlas mapping (RQ3)

To validate the contribution of the proposed federated atlas mapping in reducing structure- and feature-level heterogeneity, we employ two distinct quantitative metrics\(^1\) to evaluate the averaged heterogeneity measure among brain networks across every pair of datasets. Firstly, regarding structure-level heterogeneity, we leverage the Anonymous Walk Embeddings (AWEs)\(^\text{20}\) technique to generate representations for each brain network graph. We then calculate the Jensen-Shannon distance between every pair of AWE representations. Secondly, regarding feature-level heterogeneity, we analyze the empirical distribution of feature similarity between all pairs of linked
Figure 1. Pairwise structure- (upper) and feature-level (lower) heterogeneity measures across all datasets compared on brain networks processed without atlas mapping (left), with atlas mapping but without federated training (mid), and full federated atlas mapping (right). The smaller the numeric measure, the less heterogeneity exists within the investigated pair.

nodes (ROIs) present in each graph. We then compute the Jensen-Shannon divergence between each pair of these distributions. We present our findings in Figure 1. Specifically, we compare the heterogeneity measures among brain networks and features processed under three scenarios: without federated atlas mapping, with atlas mapping but without federated training, and with full federated atlas mapping. Our observation suggests that atlas mapping along with federated training significantly reduces the level of heterogeneity across datasets in both network structures and ROI features.

In addition, we investigate the individual influence of the transformed network structure and ROI features on downstream performance. The summarized results can be found in Table 5. We observe that learning from either transformed network structures or ROI features leads to an average relative gain of 4.68% over the non-transformation baseline. The best performance is achieved when learning from both transformed structures and features, further validating the robustness of our design in reducing heterogeneity and enhancing task-wise performance simultaneously. Furthermore, we observe a significant reduction in time complexity when learned on transformed data. Given the original network and feature dimension $N$, a transformed dimension $M$ ($M < N$), and a hidden size of $F$ of the $l$-layer GNN model, the bounded complexity reduces from $O(l(N^2F + NF^2))$ to $O(l(M^2F + MF^2))$. Reflecting this to actual FL training with 80 communication rounds, the transformation reduces the time consumption from roughly 612 seconds to 266 seconds in completion time.

B.4 Clustering analysis of guided clustering (RQ4)

We investigate the impact of the guided clustering approach on cluster formation. We focus on evaluating the effectiveness of this mechanism in grouping institutions (i.e., clients) with similar neural circuitry patterns while also maintaining reasonable cluster sizes. We compare the outcomes with those obtained from the standard hierarchical clustering. We show a dendrogram visualization of the cluster results in Figure 2. Specifically,
Table 5. Performance comparison on transformed structure and feature.

<table>
<thead>
<tr>
<th>Transformation</th>
<th>average</th>
<th>min gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.6845(±0.068)</td>
<td>–</td>
</tr>
<tr>
<td>Structure</td>
<td>0.7042(±0.070)</td>
<td>-0.0126</td>
</tr>
<tr>
<td>Feature</td>
<td>0.7288(±0.060)</td>
<td>0.0357</td>
</tr>
<tr>
<td>Structure &amp; Feature</td>
<td>0.7605(±0.052)</td>
<td>0.0417</td>
</tr>
</tbody>
</table>

Table 6. Performance comparison on must- or cannot-link constrained clustering.

<table>
<thead>
<tr>
<th>Link</th>
<th>average</th>
<th>min gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.7231(±0.065)</td>
<td>–</td>
</tr>
<tr>
<td>Cannot</td>
<td>0.7337(±0.061)</td>
<td>0.0089</td>
</tr>
<tr>
<td>Must</td>
<td>0.7445(±0.057)</td>
<td>0.0148</td>
</tr>
<tr>
<td>Cannot &amp; Must</td>
<td>0.7605(±0.052)</td>
<td>0.0235</td>
</tr>
</tbody>
</table>

Figure 2. Dendrogram visualization of cluster results from standard hierarchical clustering (left) and prior knowledge guided clustering (right). We list the client names alongside its neural circuitry attributes, namely clinical outcomes (e.g., disease/gender) and data modalities (e.g., functional/structural connectivities).

the linked branches depict the hierarchical relationships, with blue-colored lines representing singleton clusters, and other colors highlighting cluster assignments. Our observations indicate that incorporating clinical prior knowledge guidance substantially enhances the capability to identify and group clients with similar or near identical neural circuitry patterns. Our approach also avoids the production of singleton clusters, which were prominent when using the standard method.

Moreover, we study the impact on downstream performance when using clustering guidance that exclusively relies on either must- or cannot-link information. The results are presented in Table 6. We observe that sole cannot-link constraints lead to a relative gain of 1.47% over standard clustering. When guided by must-links alone, we achieve a further improvement of 1.53%, bringing the performance to within a mere 2.10% difference from considering both constraints. The findings suggest that must-link information plays a slightly more influential role in identifying similar neural circuitry patterns. On the other hand, cannot-link information proves valuable in averting additional intra-cluster heterogeneity, thereby reducing the likelihood of further cluster division and the formation of singleton clusters.

APPENDIX C. RELATED WORK

GNNs for Brain Network Analysis. GNNs have gained significant attention for their effectiveness in analyzing graph-structured data, with several pioneering models applied to brain network analysis. Notable examples include BrainGNN, which uses ROI-aware graph convolutional and ROI-selection pooling layers to predict neurological biomarkers from fMRI data. Another approach, BrainNetCNN, adopts a CNN framework with various convolutional filters designed to leverage the topological locality of structural brain networks.
BrainNetTF\textsuperscript{25} introduces a transformer architecture with an orthonormal clustering readout that considers ROI similarity within functional modules. Existing studies\textsuperscript{26–29} have demonstrated GNNs can substantially improve performance in brain disorder predictions when sufficient data is available. However, the difficulty emerges when dealing with limited training samples in practical scenarios, especially for particular clinical studies.\textsuperscript{30} This limitation hinders the full potential of GNNs for modeling brain network data, motivating designs capable of overcoming data scarcity and heterogeneity and improving performance in real clinical tasks.

**FL on Graphs.** FL has gained significant attention for collaboratively training deep learning models while preserving data privacy in various domains, including images, text, and multi-modalities.\textsuperscript{31–34} Recently, it has also been proven to be effective in the context of graphs. Some of the pioneering works have explored modeling clients as nodes in graphs,\textsuperscript{35–37} and benchmark surveys\textsuperscript{38, 39} have contributed to the understanding of GNN-based FL across graphs in diverse data domains. FL on graphs can face a unique challenge, graph data heterogeneity. Some previous related works include FedCG\textsuperscript{37} which addresses the challenge of statistical heterogeneity in FL by leveraging GNN models to extract interactions across domains; GCFL\textsuperscript{11} which studies the specific graph-level heterogeneity across domains and proposes a dynamic clustered graph FL framework; and FedLit\textsuperscript{40} which proposes a way to dynamically cluster the latent link types of graphs in FL to address the link-level heterogeneity across graphs. Nonetheless, the distinct ways in which heterogeneity manifests in brain network studies, such as the variance in parcellation systems and neural circuitry patterns, make most FL frameworks that emphasize generic graph structure learning inapplicable. While research on GNN-based FL for neuroimaging data has shown promise,\textsuperscript{41, 42} existing techniques focus on privacy preservation or domain adaptation.\textsuperscript{43} These objectives inherently diverge from our approach, which aspires to bolster data alignment and augment client personalization.

**ACKNOWLEDGMENTS**

This research was partially supported by the National Science Foundation under Award Number 2312502 and 2319449, as well as the Collaborative Research Grant funding provided by the Halle Institute for Global Research at Emory University and the Office of International Cooperation and Exchanges at Nanjing University. A shorter version of this work was also presented at the 2024 Pacific Symposium on Biocomputing (PSB 2024).

**REFERENCES**


