FedBrain: Federated Training of Graph Neural Networks for Connectome-based Brain Imaging Analysis

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

Department of Computer Science, Emory University, Atlanta, Georgia, USA

†*E-mail: j.carlyang@emory.edu*

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. Compared 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. The implementation is available here.

Keywords: Brain Connectome Analysis; Digital Health; Federated Learning

1. Introduction

In recent years, research in neuroscience has been driven to unravel the intricacies of the human brain and its connection to complex disorders such as bipolar disorder (BP) and Autism. Neuroimaging techniques, including fMRI and DTI, have emerged as crucial tools for facilitating the diagnosis of various diseases.[1] These techniques enable the construction of brain networks, which are essentially weighted connected graphs, where nodes represent anatomical regions of interest (ROIs) and edges represent their functional correlations or

structural connections. By analyzing these networks, researchers gain valuable insights into the biological structures and functions of complex neural systems, aiding in the early detection of neurological disorders and advancing fundamental neuroscience research.

Graph Neural Networks (GNNs) have gained significant popularity in analyzing graph-structured data, demonstrating impressive performance across various domains like social networks, recommender systems, and gene/protein interactions.[2,3] In neuroscience, GNNs have been applied to brain network analysis, addressing tasks such as disease prediction and neural pattern discovery[4–9] However, deep learning models, including GNNs, heavily rely on large labeled datasets to obtain strong performance. Unfortunately, neuroimaging datasets are often relatively small due to the high complexity of data acquisition, preprocessing, and annotation, leading to significant model overfitting and limited generalization power.[10,11] For instance, the popular datasets for BP and HIV analysis consist of only a few dozen subjects,[12,13] making it particularly challenging for GNNs to effectively capture important neural circuitry patterns from the noisy networks. While there exist several relatively large multi-site neuroimaging studies, the these are still small compared to datasets in typical ML domains.[14]

Recently, federated learning (FL) has emerged as a promising solution to address the challenges of limited training data and computation resources in local studies.[15–17] FL operates by collaboratively training a centralized server model 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. Second, it facilitates knowledge generalization by client aggregatiion which can mitigate the overfitting issues typically associated with learning on small datasets. These aspects have contributed to the success of FL in various fields including healthcare applications[18] and graph learning.[19]

One significant challenge in FL is data heterogeneity, wherein the data distributions significantly differ across local data owners. Several FL algorithms[16,17] have been proposed to tackle the data heterogeneity challenge. 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.[4] 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 heteroge-

neous predictive neural circuitry patterns due to various neuroimaging modalities and clinical outcomes, we design a knowledge-guided client clustering mechanism by incorporating prior clinical knowledge into the dynamic clustering process of clients with similar data during FL.

To showcase the effectiveness of FedBrain on real-world datasets from different institutions, we conduct extensive empirical evaluations, comparing our framework to state-of-the-art methods. The results demonstrate that FedBrain outperforms the baselines across all clients, with a minimum relative gain of 21.36% in accuracy. Moreover, we conduct ablation studies and specific analyses on the proposed federated atlas mapping and guided clustering mechanisms to fully understand their contribution and robustness within the framework. The results confirmed the necessity of these components in improving overall model performance.

2. Related Work

GNNs for Brain Network Analysis. GNNs have gained significant attention for their effectiveness in analyzing graph-structured data,[20–22] with several pioneering models applied to brain network analysis. Notable examples include BrainGNN,[8] which uses ROI-aware graph convolutional and ROI-selection pooling layers to predict neurological biomarkers from fMRI data. Another approach, BrainNetCNN,[9] adopts a CNN framework with various convolutional filters designed to leverage the topological locality of structural brain networks. BrainNetTF[7] introduces a transformer architecture with an orthonormal clustering readout that considers ROI similarity within functional modules. Existing studies[5,23–25] 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.[26] 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. Recently, it has been proven to be effective in the context of graphs. Some of the pioneering works have explored modeling clients as nodes in graphs,[27,28] and benchmark surveys[29] 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[28] which addresses the challenge of statistical heterogeneity in FL by leveraging GNN models to extract interactions across domains; GCFL[30] which studies the specific graph-level heterogeneity across domains and proposes a dynamic clustered graph FL framework; and FedLit[31] 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, existing techniques focus on privacy preservation[32] or domain adaptation.[33] These objectives inherently diverge from our approach, which aspires to bolster data alignment and augment client personalization.

3. The FedBrain Framework

3.1. *The FL Backbone*

The backbone FL structure of FedBrain is based on federated averaging (`FedAvg`).[15] The essence 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.

We utilize an optimized GCN[4] as backbone for both the server and client models. The ROI (*i.e.,* node) features are initialized with the connection profiles (*i.e.,* adjacency).[4] That is, the feature matrix $\boldsymbol{X}$ is equivalent to the adjacency $\boldsymbol{A}$ ($\boldsymbol{X} \equiv \boldsymbol{A}$), where $\boldsymbol{A}$ is parameterized by the node set $\mathcal{V} = \{v_n\}_{n=1}^{N}$ and the weighted edge set $\mathcal{E} = \mathcal{V} \times \mathcal{V}$.

3.2. *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, as a pre-processing mechanism, that aims to align 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 $\boldsymbol{X} \in \mathbb{R}^{N \times N}$ ($N > M$), the objective is to learn a linear projection $\boldsymbol{W} \in \mathbb{R}^{N \times M}$, 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 $\mathcal{L}_{rec} = (1/N)\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^\top\|^2$. Intuitively, the projection $\boldsymbol{W}$ transforms initial features by applying a weighted linear combination on the original dimensions. Consequently, the columns of $\boldsymbol{W}$ learns 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 $\boldsymbol{Z} \in \mathbb{R}^{N \times M}$ such that $\boldsymbol{Z}_{i,j} = \mathbb{1}[\boldsymbol{W}_{i,j} \in \arg\operatorname{top} k\,(\mathrm{col}_j(\boldsymbol{W}))]$. The matrix $\boldsymbol{Z}$ records the top-k greatest entries per each column in $\boldsymbol{W}$ and zeros out the rest. Ultimately, given a graph adjacency matrix $\boldsymbol{A}\,(\equiv \boldsymbol{X})$, we construct a compressed network $\boldsymbol{A}'$ by evaluating $\boldsymbol{A}' = \boldsymbol{Z}^\top \boldsymbol{A} \boldsymbol{Z}$.

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,

which makes it challenging to communicate model parameters.

To address this issue, we propose a unified mapping method that aims to adapt the size of the global model to the varying dimensionality of each local dataset. Given a global projection $\boldsymbol{W}_G \in \mathbb{R}^{N_G \times M}$ 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 $\boldsymbol{P}_L \in \mathbb{R}^{N_L \times N_G}$, which ensures the local projection $\boldsymbol{W}_L \in \mathbb{R}^{N_L \times M}$ is distributed through the mapping $\boldsymbol{W}_L = \boldsymbol{P}_L \boldsymbol{W}_G$. 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 $\boldsymbol{S} \in \mathbb{R}^{N_L \times N_G}$, where $\boldsymbol{S}_{i,j} = d(\text{row}_i(D_L), \text{row}_j(D_G))$ represents the pairwise Euclidean distance between ROIs from the two templates. We then designate $\boldsymbol{P}_{L_{i,j}} = \mathbb{1}[\boldsymbol{S}_{i,j} = \arg\min(\text{col}_j(\boldsymbol{S}))]$. This implied that we only consider the minimum entry per each column of $\boldsymbol{S}$. Essentially, we enable $\boldsymbol{P}_L$ 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 $\boldsymbol{W}_L = \boldsymbol{P}_L \boldsymbol{W}_G$. Subsequently, each client sends their updated parameters back to the server, employing the inverse mapping $\boldsymbol{W}_L^* = \boldsymbol{P}_L^\top \boldsymbol{W}_L^*$.

3.3. *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[30,34] into the FL process.

Clustered FL. When data distributions are similar among local clients, the average global model can achieve convergence for all local objectives. However, in instances of heterogeneity, the global model fails to adapt to local optimizations, resulting in stationary point convergence.[34] To mitigate stationary convergence, clients can be assigned to clusters with homogeneous data distributions, thereby initiating cluster-specific FL subroutines.

Constrained clustering. While gradient-based 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 4.4). 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 λ_{must} and penalty λ_{cannot} term, which are multiplied to the pairwise client similarity measure when creating must- and cannot-links.

4. Experiments

Datasets. We evaluate our framework using six real-world brain network datasets: BP,[12] HIV,[13] PPMI,[35] PNC,[36] ABIDE,[37] and ABCD.[38] 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.

Table 1. Dataset statistics.

Dataset	Modality	Sample Size	Atlas	Network Size	Outcome	Class Number
BP	fMRI, DTI	97	Brodmann 82	82×82	Disease	2
HIV	fMRI, DTI	70	AAL 90	90×90	Disease	2
PPMI	PICo, Hough, FSL	754	Desikan-Killiany 84	84×84	Disease	2
PNC	fMRI	503	Power 264	264×264	Gender	2
ABIDE	fMRI	1009	Craddock 200	200×200	Disease	2
ABCD	fMRI	7901	HCP 360	360×360	Gender	2

Parameter setup. The downstream classifier consists of a single-layer MLP, and we use the negative log-likelihood measure as the optimization objective and accuracy as the evaluation metric. In the case of all FL baselines, a complete training procedure encompasses 80 communication rounds. For the `self-train` (*i.e.,* non-FL) baseline, each local model is trained for 80 epochs. Regarding FedBrain, we retain 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.

Empirical analyses. The following sections are structured to assess (1) the performance of FedBrain in comparison to widely adopted FL frameworks, and (2) the contribution of the key components to the overall performance, supplemented by case studies.

4.1. *Overall performance comparison (RQ1)*

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 effect of collaborative learning and cross-institutional knowledge generalization in overcoming model overfitting on limited training

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

Clients	BP-fMRI	BP-DTI	HIV-fMRI	HIV-DTI	PPMI-PICo
Accuracy			average		
`self-train`	0.5463(±0.019)	0.5012(±0.082)	0.5286(±0.035)	0.4571(±0.140)	0.6394(±0.034)
`FedAvg`	0.6037(±0.073)	0.5158(±0.013)	0.5457(±0.153)	0.5000(±0.078)	0.7925(±0.002)
`FedProx`	0.6084(±0.117)	0.5853(±0.085)	0.6200(±0.132)	0.6029(±0.097)	0.7925(±0.002)
`SCAFFOLD`	0.5800(±0.120)	0.6400(±0.049)	0.6343(±0.070)	0.6629(±0.057)	0.7778(±0.000)
FedBrain	**0.7389(±0.066)**	**0.7500(±0.077)**	**0.7857(±0.071)**	**0.8143(±0.070)**	**0.8102(±0.010)**

PPMI-Hough	PPMI-FSL	PNC	ABIDE	ABCD		
		average			combine	min gain
0.6570(±0.054)	0.6852(±0.041)	0.5034(±0.052)	0.5025(±0.007)	0.5342(±0.002)	0.5555(±0.073)	–
0.7633(±0.031)	0.7925(±0.002)	0.5434(±0.008)	0.5044(±0.012)	0.5167(±0.017)	0.6078(±0.118)	-0.032
0.7536(±0.037)	0.7925(±0.002)	0.6057(±0.018)	0.5594(±0.003)	0.5700(±0.020)	0.6490(±0.088)	0.067
0.7944(±0.014)	0.7889(±0.014)	0.6015(±0.009)	0.5765(±0.090)	0.5980(±0.045)	0.6654(±0.084)	0.120
0.8102(±0.010)	**0.8095(±0.010)**	**0.7275(±0.044)**	**0.6549(±0.034)**	**0.7033(±0.033)**	**0.7605(±0.052)**	**0.214**

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.

Table 3. Atlas mapping comparisons.

Accuracy	average	min gain
No Atlas Mapping	0.6845(±0.068)	–
Atlas Mapping	0.7246(±0.063)	0.0039
Federated Atlas Mapping	0.7605(±0.052)	0.0214

Table 4. Guided clustering comparisons.

Accuracy	average	min gain
No Clustering	0.6921(±0.071)	–
Non-guided Clustering	0.7231(±0.065)	0.0000
Guided Clustering	0.7605(±0.052)	0.0000

4.2. *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 compare two scenarios: without clustering and 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 constraints, we further enhance cluster-specific learning and knowledge generalization, resulting in a relative gain of 5.17% and a reduction in performance variance.

4.3. *Heterogeneity analysis of federated atlas mapping (RQ3)*

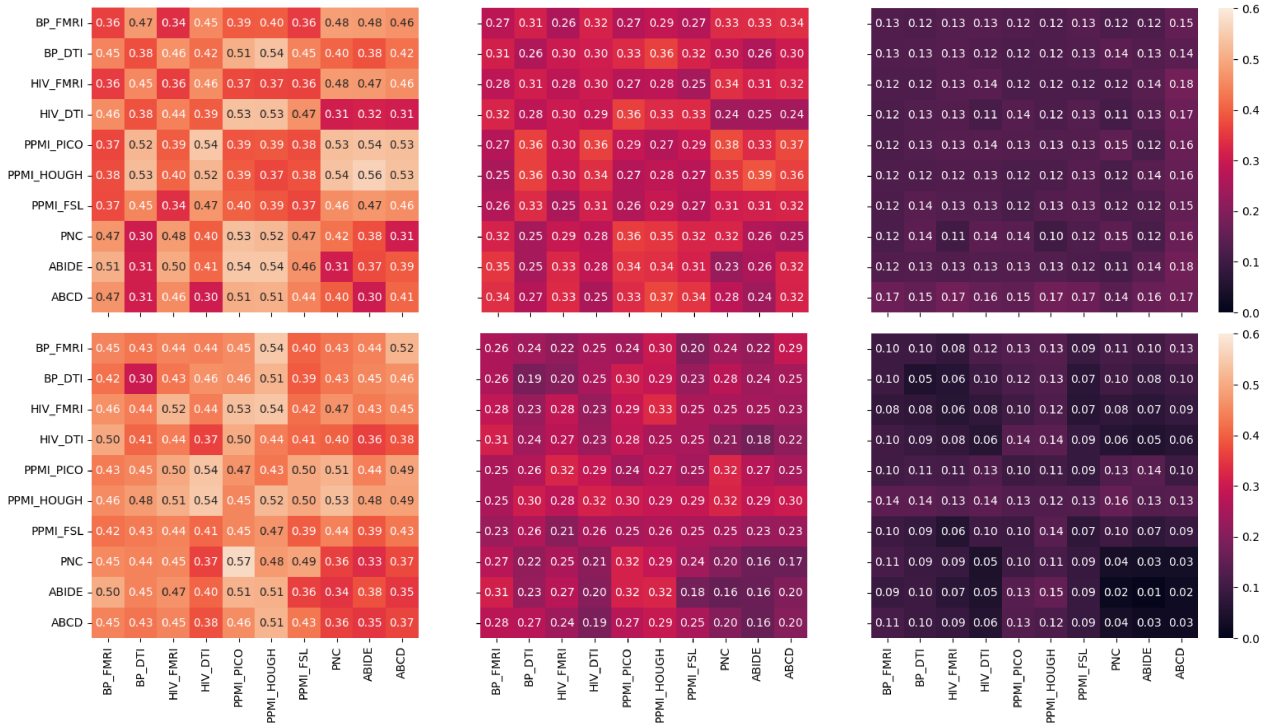

Fig. 1. Pairwise structure- (upper) and feature-level (lower) heterogeneity measures across all datasets compared on brain networks processed without atlas mapping (left), with atlas mapping but without federated training (mid), and full federated atlas mapping (right). The smaller the numeric measure, the less heterogeneity exists within the investigated pair.

To validate the contribution of the proposed federated atlas mapping in reducing structure- and feature-level heterogeneity, we employ two distinct quantitative metrics[30] 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)[39] 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 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.

Table 5. Network transformation comparisons.

Transformation	average	min gain
None	0.6845(±0.068)	–
Structure	0.7042(±0.070)	-0.0126
Feature	0.7288(±0.060)	0.0357
Structure & Feature	0.7605(±0.052)	0.0417

Table 6. Cluster constraints comparisons.

Link	average	min gain
None	0.7231(±0.065)	–
Cannot	0.7337(±0.061)	0.0089
Must	0.7445(±0.057)	0.0148
Cannot & Must	0.7605(±0.052)	0.0235

4.4. *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, 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 produc-

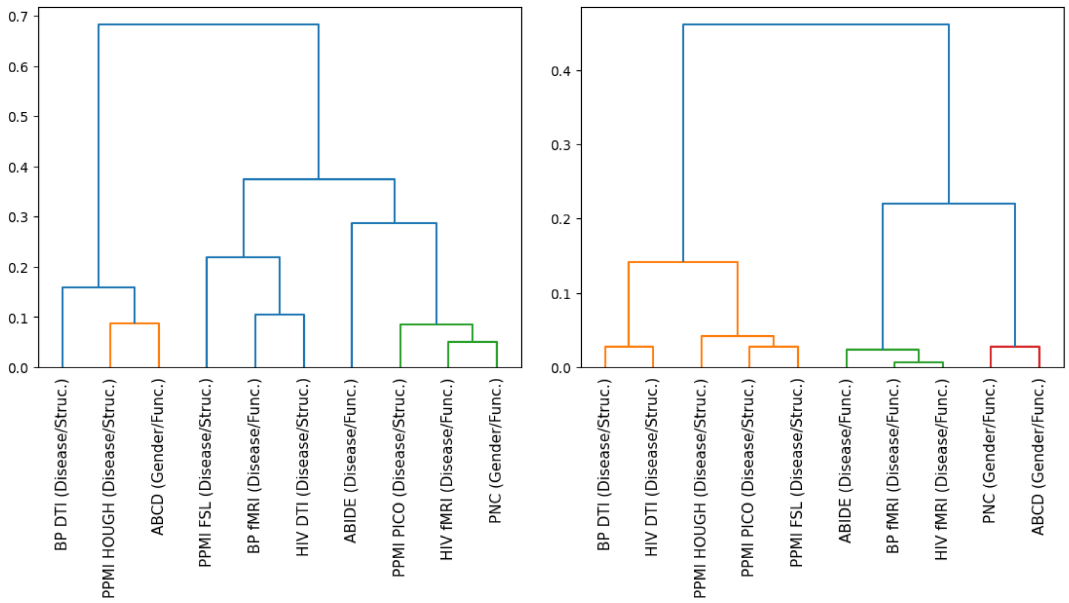

Fig. 2. Dendrogram visualization of cluster results from standard hierarchical clustering (left) and prior knowledge guided clustering (right). We list the client names alongside its clinical outcomes (*e.g.,* disease/gender) and data modalities (*e.g.,* functional/structural connectivities).

tion 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.

5. Conclusion

Cross-institutional brain network analysis has been a challenging task for conventional FL frameworks and GNN models. The presence of unique data heterogeneity, particularly in terms of inconsistent ROI parcellation systems and predictive neural circuitry patterns, poses a significant obstacle to effective collaborative training and knowledge generalization. To tackle these challenges, we propose FedBrain, a personalized GNN-based FL framework. Specifically, we leverage a data-driven atlas mapping mechanism to address the issue of incompatible ROI parcellation systems. Moreover, we incorporate clustered FL to enhance client personalization and integrate clinical prior knowledge to guide the clustering process. We conducted extensive experiments on multiple real-world brain network studies, demonstrating the superior performance of FedBrain compared to various state-of-the-art FL baselines.

We direct our future efforts to enhance FedBrain by addressing current limitations.

Firstly, we'll expand data considerations to include a wider array of atlas templates, clinical tasks, and clients with multi-modal data. Secondly, we'll optimize computational efficiency as the framework becomes more sophisticated. Thirdly, we'll delve into theoretical investigations to ensure strong privacy guarantees. Lastly, we plan to broaden empirical investigations by incorporating a broader set of data to validate the framework's robustness.

References

1. N. Yahata, J. Morimoto, R. Hashimoto, G. Lisi, K. Shibata, Y. Kawakubo, H. Kuwabara, M. Kuroda, T. Yamada, F. Megumi *et al.*, A small number of abnormal brain connections predicts adult autism spectrum disorder, in *Nat. Commun.*, (2016).
2. S. Wu, F. Sun, W. Zhang, X. Xie and B. Cui, Graph neural networks in recommender systems: a survey, in *ACM Comp. Surv.*, (2022).
3. W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang and D. Yin, Graph neural networks for social recommendation, in *WWW*, (2019).
4. H. Cui, W. Dai, Y. Zhu, X. Kan, A. A. C. Gu, J. Lukemire, L. Zhan, L. He, Y. Guo and C. Yang, Braingb: A benchmark for brain network analysis with graph neural networks, in *IEEE TMI*, (2022).
5. X. Kan, H. Cui, J. Lukemire, Y. Guo and C. Yang, Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation, in *MIDL*, (2022).
6. G. Luo, C. Li, H. Cui, L. Sun, L. He and C. Yang, Multi-view brain network analysis with cross-view missing network generation, in *IEEE BIBM*, (2022).
7. X. Kan, W. Dai, H. Cui, Z. Zhang, Y. Guo and C. Yang, Brain network transformer, in *NeurIPS*, (2022).
8. X. Li, Y. Zhou, N. Dvornek, M. Zhang, S. Gao, J. Zhuang, D. Scheinost, L. H. Staib, P. Ventola and J. S. Duncan, Braingnn: Interpretable brain graph neural network for fmri analysis, in *Medical Image Analysis*, (2021).
9. J. Kawahara, C. J. Brown, S. P. Miller, B. G. Booth, V. Chau, R. E. Grunau, J. G. Zwicker and G. Hamarneh, Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment, in *NeuroImage*, (2017).
10. Y. Yang, H. Cui and C. Yang, Ptgb: Pre-train graph neural networks for brain network analysis, in *CHIL*, (2023).
11. Y. Yang, Y. Zhu, H. Cui, X. Kan, L. He, Y. Guo and C. Yang, Data-efficient brain connectome analysis via multi-task meta-learning, in *ACM SIGKDD*, (2022).
12. B. Cao, L. Zhan, X. Kong, P. S. Yu, N. Vizueta, L. L. Altshuler and A. D. Leow, Identification of discriminative subgraph patterns in fmri brain networks in bipolar affective disorder, in *Brain Informatics and Health*, (2015).
13. A. B. Ragin, H. Du, R. Ochs, Y. Wu, C. L. Sammet, A. Shoukry and L. G. Epstein, Structural brain alterations can be detected early in hiv infection, in *Neurology*, (2012).
14. R. A. Rossi and N. K. Ahmed, An interactive data repository with visual analytics, in *ACM SIGKDD Explorations Newsletter*, (2016).
15. B. McMahan, E. Moore, D. Ramage, S. Hampson and B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in *PMLR*, (2017).
16. T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar and V. Smith, Federated optimization in heterogeneous networks, in *MLSys*, (2020).
17. S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich and A. T. Suresh, Scaffold: Stochastic controlled averaging for federated learning, in *ICML*, (2020).
18. Q. Wu, X. Chen, Z. Zhou and J. Zhang, Fedhome: Cloud-edge based personalized federated learning for in-home health monitoring, in *IEEE TMC*, (2020).

19. M. Chen, W. Zhang, Z. Yuan, Y. Jia and H. Chen, Fede: Embedding knowledge graphs in federated setting, in *IJCKG*, (2021).
20. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio and Y. Bengio, Graph attention networks, in *ICLR*, (2018).
21. K. Xu, W. Hu, J. Leskovec and S. Jegelka, How powerful are graph neural networks?, in *ICLR*, (2019).
22. T. N. Kipf and M. Welling, Semi-supervised classification with graph convolutional networks, in *ICLR*, (2017).
23. H. Cui, W. Dai, Y. Zhu, X. Li, L. He and C. Yang, Interpretable graph neural networks for connectome-based brain disorder analysis, in *MICCAI*, (2022).
24. Y. Zhu, H. Cui, L. He, L. Sun and C. Yang, Joint embedding of structural and functional brain networks with graph neural networks for mental illness diagnosis, in *IEEE EMBC*, (2022).
25. Y. Yu, X. Kan, H. Cui, R. Xu, Y. Zheng, X. Song, Y. Zhu, K. Zhang, R. Nabi, Y. Guo *et al.*, Learning task-aware effective brain connectivity for fmri analysis with graph neural networks, in *ISBI*, (2023).
26. R. Xu, Y. Yu, J. C. Ho and C. Yang, Weakly-supervised scientific document classification via retrieval-augmented multi-stage training, in *ACM SIGIR*, (2023).
27. A. Lalitha, O. C. Kilinc, T. Javidi and F. Koushanfar, Peer-to-peer federated learning on graphs, in *arXiv preprint*, (2019).
28. D. Caldarola, M. Mancini, F. Galasso, M. Ciccone, E. Rodolà and B. Caputo, Cluster-driven graph federated learning over multiple domains, in *CVPRW*, (2021).
29. C. He, K. Balasubramanian, E. Ceyani, C. Yang, H. Xie, L. Sun, L. He, L. Yang, S. Y. Philip, Y. Rong *et al.*, Fedgraphnn: A federated learning benchmark system for graph neural networks, in *ICLR-DPML*, (2021).
30. H. Xie, J. Ma, L. Xiong and C. Yang, Federated graph classification over non-iid graphs, in *NeurIPS*, (2021).
31. H. Xie, L. Xiong and C. Yang, Federated node classification over graphs with latent link-type heterogeneity, in *WWW*, (2023).
32. E. Darzidehkalani, M. Ghasemi-Rad and P. van Ooijen, Federated learning in medical imaging: part i: toward multicentral health care ecosystems, in *Journal of the American College of Radiology*, (2022).
33. X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola and J. S. Duncan, Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results, in *Medical Image Analysis*, (2020).
34. F. Sattler, K.-R. Müller and W. Samek, Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints, in *IEEE TNNLS*, (2020).
35. D. Aleksovski, D. Miljkovic, D. Bravi and A. Antonini, Disease progression in parkinson subtypes: the ppmi dataset, in *Neurol. Sci.*, (2018).
36. T. D. Satterthwaite, M. A. Elliott, K. Ruparel, J. Loughead, K. Prabhakaran, M. E. Calkins, R. Hopson, C. Jackson, J. Keefe, M. Riley *et al.*, Neuroimaging of the philadelphia neurodevelopmental cohort, in *Neuroimage*, (2014).
37. A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto *et al.*, The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism, in *Molecular psychiatry*, (2014).
38. B. Casey, T. Cannonier, M. I. Conley, A. O. Cohen, D. M. Barch, M. M. Heitzeg, M. E. Soules, T. Teslovich, D. V. Dellarco, H. Garavan *et al.*, The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites, in *Dev. Cogn. Neurosci.*, (2018).
39. S. Ivanov and E. Burnaev, Anonymous walk embeddings, in *ICML*, (2018).