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, Cross-Instituion Learning, Federated Learning

Extended Abstract

In recent years, neuroscience research has been focused on unraveling the complexities of the human brain and its associations with intricate disorders such as bipolar disorder (BP) and Autism. Crucial tools in this pursuit are neuroimaging techniques like functional magnetic resonance imaging (fMRI) and diffusion tensor imaging (DTI), which play pivotal roles in measuring cerebral activities which could subsequently assist in the diagnosis of various diseases.[1] These techniques also facilitate the creation of brain networks, that are essentially weighted connected graphs where nodes represent anatomical regions of interest (ROIs) and edges denote their functional correlations or structural connections. Through the analysis of these networks, researchers gain valuable insights into the biological structures and functions of intricate neural systems, aiding in the early detection of neurological disorders and contributing to the advancement of fundamental neuroscience research.

Graph Neural Networks (GNNs) have garnered considerable attention for their effectiveness in analyzing graph-structured data, showcasing impressive performance across diverse domains such as social networks, recommender systems, and gene/protein interactions.[2,3] In the realm of neuroscience, GNNs find applications in brain network analysis, tackling tasks like disease prediction and neural pattern discovery.[4–13] However, deep learning models, including GNNs, heavily rely on extensive labeled datasets to achieve robust performance. Unfortunately, neuroimaging studies tend to be relatively small in sample quantities due to the inherent complexity of data acquisition, preprocessing, and annotation, resulting in notable model overfitting and limited generalization capabilities.[14,15] Notably, datasets for BP and HIV analysis, for example, consist of only a few dozen

*Equal contributions; †Corresponding author (*j.carlyang@emory.edu*) subjects,[16,17] posing a significant challenge for GNNs in effectively capturing crucial neural circuitry patterns from these noisy networks. Although there are relatively larger multi-site neuroimaging studies, they still pale in comparison to datasets in more typical machine learning domains.[18]

In recent times, federated learning (FL) has emerged as a highly promising solution for addressing the challenges associated with limited training data and computational resources in local studies.[19–21] FL functions through the collaborative training of a centralized server model using data privately stored by multiple local clients. During each communication round, the server transmits model parameters to each client to initialize local training. Subsequently, clients learn and update their local parameters based on their privately stored data. The server then receives and aggregates the updated local parameters in preparation for the next communication round. This approach boasts two significant advantages. Firstly, it ensures privacy preservation, as clients exclusively communicate model parameters, instead of data, with the server. Secondly, it facilitates knowledge generalization through the server aggregation, effectively mitigating overfitting issues commonly associated with learning from singular small datasets. These attributes have significantly contributed to the success of FL across diverse fields, including healthcare applications[22] and graph learning.[23]

A significant challenge in (FL) lies in data heterogeneity, where the data distributions remarkably vary among local data owners. Various FL algorithms[20,21] have been introduced to address this challenge. However, these methods primarily concentrate on label distributions and often overlook the distinctive data heterogeneity scenarios encountered in cross-institutional brain network analysis, which can manifest in two crucial aspects. Firstly, given that network parcellation is typically a specialized process carried out by domain experts, it becomes challenging to assume or impose a uniformity in the ROI atlas mapping systems adopted by different institutions during the preprocessing of their raw neuroimaging signals. In particular, the atlas templates can vary drastically in sizes, dimensions, and physical meanings of the defined ROIs. Consequently, this results in misalignment in network structures and ROI features across clients. Secondly, different institutions would collect brain network data for varying clinical purposes targeting at different diseases and patient groups. This leads to institutions utilizing different neuroimaging techniques and focusing on distinct clinical outcomes, leading to diverse underlying predictive neural circuitry patterns, such as data modalities, across studies.

To address the aforementioned challenges and effectively handle the distinctive data heterogeneities in cross-institutional brain network analysis, we present FedBrain, a personalized FL framework tailored for GNN-based brain network learning. Our framework is composed of three key elements: an FL backbone employing GNN-parameterized learning models, a federated atlas mapping mechanism, and a guided client clustering mechanism. In constructing our FL platform, we utilize the well-established `FedAvg` as the foundation, and our default GNN structure is an optimized graph convolutional network (GCN) model proposed by BrainGB.[4] To address the issue of feature- and structure-level heterogeneity resulting from potentially different atlas mapping systems employed by local institutions, we introduce an autoencoder-based atlas mapping mechanism as a data-driven pre-processing solution which aims to achieve data and network alignment across studies. Particularly, this module will project diverse ROI profiles onto a standardized, shared embedding space. To manage the heterogeneous predictive neural circuitry patterns arising from various neuroimaging modalities and clinical outcomes, we devise a knowledge-guided client clustering mechanism. This mechanism incorporates prior clinical knowledge into the dynamic client clustering during FL training. Specifically, the clustering decision will be heavily dependent, to a weighted extent, upon similarities in the sharable predictive neural circuitry information across local clients.

To illustrate the efficacy of FedBrain using real-world datasets sourced from diverse institutions, we perform comprehensive empirical evaluations, benchmarking our framework against state-of-the-art methods. The outcomes showcase FedBrain surpassing the baseline models uniformly across all client institutions, yielding a minimum relative gain of 21.36% in prediction accuracy. Additionally, we conduct ablation studies and specific case studies on the proposed federated atlas mapping and guided clustering mechanisms to thoroughly comprehend their contributions and robustness within the framework. The results highlight that the federated atlas mapping notably diminishes structure- and feature-level heterogeneity measures across all clients. In addition, the pre-processing also significantly helps reduce the computational runtime in the subsequent FL training. On the other hand, guided clustering exhibits a robust ability to dynamically identify similar groups of clients sharing overlapping degrees of predictive neural patterns.

ACKNOWLEDGMENTS

This research was partially supported by the National Science Foundation under Award Numbers 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 full version of this work was also presented at the 2024 Pacific Symposium on Biocomputing (PSB 2024).[24]

REFERENCES

[1] Yahata, N., Morimoto, J., Hashimoto, R., Lisi, G., Shibata, K., Kawakubo, Y., Kuwabara, H., Kuroda, M., Yamada, T., Megumi, F., et al., "A small number of abnormal brain connections predicts adult autism spectrum disorder," in [*Nat. Commun.*], (2016).

[2] Wu, S., Sun, F., Zhang, W., Xie, X., and Cui, B., "Graph neural networks in recommender systems: a survey," in [*ACM Comp. Surv.*], (2022).

[3] Fan, W., Ma, Y., Li, Q., He, Y., Zhao, E., Tang, J., and Yin, D., "Graph neural networks for social recommendation," in [*WWW*], (2019).

[4] Cui, H., Dai, W., Zhu, Y., Kan, X., Gu, A. A. C., Lukemire, J., Zhan, L., He, L., Guo, Y., and Yang, C., "Braingb: A benchmark for brain network analysis with graph neural networks," in [*IEEE TMI*], (2022).

[5] Kan, X., Cui, H., Lukemire, J., Guo, Y., and Yang, C., "Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation," in [*MIDL*], (2022).

[6] Luo, G., Li, C., Cui, H., Sun, L., He, L., and Yang, C., "Multi-view brain network analysis with cross-view missing network generation," in [*IEEE BIBM*], (2022).

[7] Kan, X., Dai, W., Cui, H., Zhang, Z., Guo, Y., and Yang, C., "Brain network transformer," in [*NeurIPS*], (2022).

[8] Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L. H., Ventola, P., and Duncan, J. S., "Braingnn: Interpretable brain graph neural network for fmri analysis," in [*Medical Image Analysis*], Elsevier (2021).

[9] Kawahara, J., Brown, C. J., Miller, S. P., Booth, B. G., Chau, V., Grunau, R. E., Zwicker, J. G., and Hamarneh, G., "Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment," in [*NeuroImage*], Elsevier (2017).

[10] Dai, W., Cui, H., Kan, X., Guo, Y., and Yang, C., "Transformer-based hierarchical clustering for brain network analysis," in [*IEEE International Symposium on Biomedical Imaging (ISBI)*], (2023).

[11] Kan, X., Gu, A. A. C., Cui, H., Guo, Y., and Yang, C., "Dynamic brain transformer with multi-level attention for functional brain network analysis," in [*International Conference on Biomedical and Health Informatics (IEEE-BHI)*], (2023).

[12] Kan, X., Li, Z., Cui, H., Yu, Y., Xu, R., Yu, S., Zhang, Z., Guo, Y., and Yang, C., "R-mixup: Riemannian mixup for biological networks," in [*ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*], (2023).

[13] Cui, H., Kan, X., Li, X., Guo, Y., He, L., Zhan, L., and Yang, C., "Brain network analysis with graph neural network," in [*IEEE International Symposium on Biomedical Imaging (ISBI)*], (2024).

[14] Yang, Y., Cui, H., and Yang, C., "Ptgb: Pre-train graph neural networks for brain network analysis," in [*CHIL*], (2023).

[15] Yang, Y., Zhu, Y., Cui, H., Kan, X., He, L., Guo, Y., and Yang, C., "Data-efficient brain connectome analysis via multi-task meta-learning," in [*ACM SIGKDD*], (2022).

[16] Cao, B., Zhan, L., Kong, X., Yu, P. S., Vizueta, N., Altshuler, L. L., and Leow, A. D., "Identification of discriminative subgraph patterns in fmri brain networks in bipolar affective disorder," in [*Brain Informatics and Health*], (2015).

[17] Ragin, A. B., Du, H., Ochs, R., Wu, Y., Sammet, C. L., Shoukry, A., and Epstein, L. G., "Structural brain alterations can be detected early in hiv infection," in [*Neurology*], AAN Enterprises (2012).

[18] Rossi, R. A. and Ahmed, N. K., "An interactive data repository with visual analytics," in [*ACM SIGKDD Explorations Newsletter*], (2016).

[19] McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A., "Communication-efficient learning of deep networks from decentralized data," in [*PMLR*], (2017).

[20] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V., "Federated optimization in heterogeneous networks," in [*MLSys*], (2020).

[21] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T., "Scaffold: Stochastic controlled averaging for federated learning," in [*ICML*], (2020).

[22] Wu, Q., Chen, X., Zhou, Z., and Zhang, J., "Fedhome: Cloud-edge based personalized federated learning for in-home health monitoring," in [*IEEE TMC*], (2020).

[23] Chen, M., Zhang, W., Yuan, Z., Jia, Y., and Chen, H., "Fede: Embedding knowledge graphs in federated setting," in [*IJCKG*], (2021).

[24] Yang, Y., Xie, H., Cui, H., and Yang, C., "Fedbrain: Federated training of graph neural networks for connectome-based brain imaging analysis," in [*PSB*], (2024).

[25] Brodmann, K., "Vergleichende lokalisationslehre der grosshirnrinde in ihren prinzipien dargestellt auf grund des zellenbaues," in [*Barth*], (1909).

[26] Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M., "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," in [*Neuroimage*], Elsevier (2002).

[27] Sattler, F., Müller, K.-R., and Samek, W., "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," in [*IEEE TNNLS*], (2020).

[28] Xie, H., Ma, J., Xiong, L., and Yang, C., "Federated graph classification over non-iid graphs," in [*NeurIPS*], (2021).

[29] Aleksovski, D., Miljkovic, D., Bravi, D., and Antonini, A., "Disease progression in parkinson subtypes: the ppmi dataset," in [*Neurol. Sci.*], (2018).

[30] Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M. E., Hopson, R., Jackson, C., Keefe, J., Riley, M., et al., "Neuroimaging of the philadelphia neurodevelopmental cohort," in [*Neuroimage*], Elsevier (2014).

[31] Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., et al., "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," in [*Molecular psychiatry*], Nature Publishing Group (2014).

[32] Casey, B., 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., "The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites," in [*Dev. Cogn. Neurosci.*], (2018).

[33] Kipf, T. N. and Welling, M., "Semi-supervised classification with graph convolutional networks," in [*ICLR*], (2017).

[34] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," in [*ICLR*], (2015).

[35] Ivanov, S. and Burnaev, E., "Anonymous walk embeddings," in [*ICML*], (2018).

[36] Liu, F., Wu, X., Ge, S., Fan, W., and Zou, Y., "Federated learning for vision-and-language grounding problems," in [*AAAI*], (2020).

[37] Shome, D. and Kar, T., "Fedaffect: Few-shot federated learning for facial expression recognition," in [*ICCVW*], (2021).

[38] Gao, L., Fu, H., Li, L., Chen, Y., Xu, M., and Xu, C.-Z., "Feddc: Federated learning with non-iid data via local drift decoupling and correction," in [*CVPR*], (2022).

[39] Lin, B. Y., He, C., Zeng, Z., Wang, H., Huang, Y., Dupuy, C., Gupta, R., Soltanolkotabi, M., Ren, X., and Avestimehr, S., "Fednlp: Benchmarking federated learning methods for natural language processing tasks," in [*NAACL Findings*], (2022).

[40] Lalitha, A., Kilinc, O. C., Javidi, T., and Koushanfar, F., "Peer-to-peer federated learning on graphs," in [*arXiv preprint*], (2019).

[41] Rizk, E. and Sayed, A. H., "A graph federated architecture with privacy preserving learning," in [*IEEE SPAWC*], (2021).

[42] Caldarola, D., Mancini, M., Galasso, F., Ciccone, M., Rodolà, E., and Caputo, B., "Cluster-driven graph federated learning over multiple domains," in [*CVPRW*], (2021).

[43] He, C., Balasubramanian, K., Ceyani, E., Yang, C., Xie, H., Sun, L., He, L., Yang, L., Philip, S. Y., Rong, Y., et al., "Fedgraphnn: A federated learning benchmark system for graph neural networks," in [*ICLR-DPML*], (2021).

[44] Liu, R., Xing, P., Deng, Z., Li, A., Guan, C., and Yu, H., "Federated graph neural networks: Overview, techniques and challenges," in [*arXiv preprint*], (2022).

[45] Xie, H., Xiong, L., and Yang, C., "Federated node classification over graphs with latent link-type heterogeneity," in [*WWW*], (2023).

[46] Stripelis, D., Gupta, U., Saleem, H., Dhinagar, N., Ghai, T., Sanchez, R., Anastasiou, C., Asghar, A., Steeg, G. V., Ravi, S., et al., "Secure federated learning for neuroimaging," in [*arXiv preprint*], (2022).

[47] Gürler, Z. and Rekik, I., "Federated brain graph evolution prediction using decentralized connectivity datasets with temporally-varying acquisitions," in [*IEEE TMI*], (2022).

[48] Darzidehkalani, E., Ghasemi-Rad, M., and van Ooijen, P., "Federated learning in medical imaging: part i: toward multicentral health care ecosystems," in [*Journal of the American College of Radiology*], Elsevier (2022).

[49] Li, X., Gu, Y., Dvornek, N., Staib, L. H., Ventola, P., and Duncan, J. S., "Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results," in [*Medical Image Analysis*], Elsevier (2020).

[50] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y., "Graph attention networks," in [*ICLR*], (2018).

[51] Xu, K., Hu, W., Leskovec, J., and Jegelka, S., "How powerful are graph neural networks?," in [*ICLR*], (2019).

[52] Hamilton, W., Ying, Z., and Leskovec, J., "Inductive representation learning on large graphs," in [*NeurIPS*], (2017).

[53] Cui, H., Dai, W., Zhu, Y., Li, X., He, L., and Yang, C., "Interpretable graph neural networks for connectome-based brain disorder analysis," in [*MICCAI*], (2022).

[54] Zhu, Y., Cui, H., He, L., Sun, L., and Yang, C., "Joint embedding of structural and functional brain networks with graph neural networks for mental illness diagnosis," in [*IEEE EMBC*], (2022).

[55] Yu, Y., Kan, X., Cui, H., Xu, R., Zheng, Y., Song, X., Zhu, Y., Zhang, K., Nabi, R., Guo, Y., et al., "Learning task-aware effective brain connectivity for fmri analysis with graph neural networks," in [*ISBI*], (2023).

[56] Li, A., Yang, Y., Cui, H., and Yang, C., "Brainsteam: A practical pipeline for connectome-based fmri analysis towards subject classification," in [*PSB*], (2024).

[57] Xu, R., Yu, Y., Ho, J. C., and Yang, C., "Weakly-supervised scientific document classification via retrieval-augmented multi-stage training," in [*ACM SIGIR*], (2023).

APPENDIX A. THE FEDBRAIN FRAMEWORK

A.1 The FL Backbone

The FL backbone of FedBrain is structured upon the federated averaging (`FedAvg`) framework proposed by McMahan et al.[19] The detailed procedure is presented in Algorithm 1. In particular, in each communication round, the server model distributes its model parameters to all clients as an initialization. The clients then update their respective models based on local data and training objectives. The clients transmit the updated parameters back to the server for a weighted aggregation. In the context of `FedAvg`, the weight for each client is determined by the fraction of the sample size of its local data w.r.t. the universal sample size across all clients.

As a powerful graph machine learning approach, GNNs have gained wide popularity in various applications due to its effectiveness in representing node- and graph-level information as well as connectivity structures at

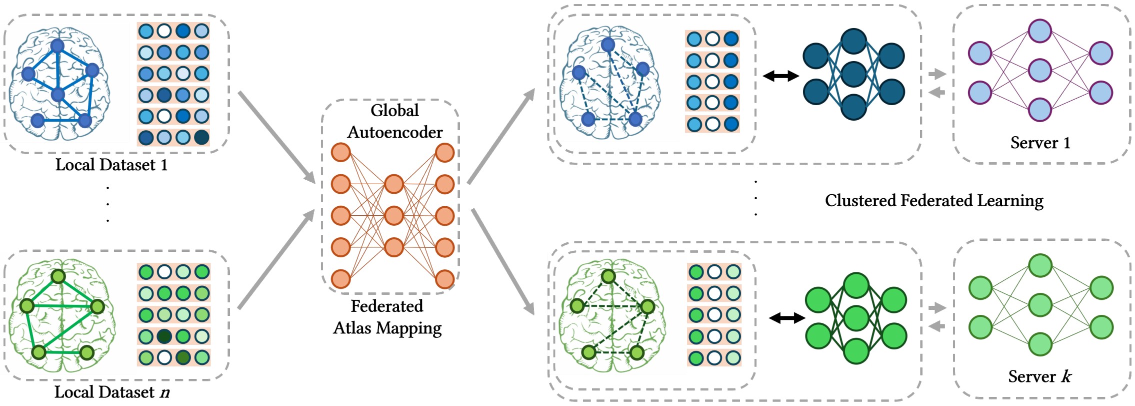

Figure 1. The comprehensive workflow of FedBrain. The first step involves the initial transformation and pre-processing of each set of local data through an atlas mapping autoencoder trained within a federated framework. Subsequently, during the GNN training phase, local clients undergo dynamic clustering into sub-groups characterized by dedicated server models and FL subroutines, determined by their similarities in neural patterns.

Algorithm 1 Federated Averaging (`FedAvg`)

```
Input: Number of communication rounds $T$, set of total available clients $\mathbb{C} \leftarrow \{\mathcal{C}_i\}_{i=1}^{N}$, set of total available client data $\mathbb{D} \leftarrow \{\mathcal{D}_i\}_{i=1}^{N}$, learning rate $\eta$
Output: The final server model $w_G^T$
Randomly initialize the server model $w_G^0$
for $t \leftarrow 1$ to $T$ do
    Sample a subset $\{\mathbb{C}_m, \mathbb{D}_m\}$ of $m$ participating clients from $\{\mathbb{C}, \mathbb{D}\}$
    for each participant $\{\mathcal{C}_j, \mathcal{D}_j\} \in \{\mathbb{C}_m, \mathbb{D}_m\}$ in parallel do
        Download model parameter from server: $w_{\mathcal{C}_j}^t \leftarrow w_G^{t-1}$
        Update local parameter: $w_{\mathcal{C}_j}^{t*} \leftarrow w_{\mathcal{C}_j}^t - \eta \nabla_{w_{\mathcal{C}_j}^t} \mathcal{L}\left(w_{\mathcal{C}_j}^t; \mathcal{D}_j\right)$
    Update the server model: $w_G^t \leftarrow \sum_{\{\mathcal{C}_j, \mathcal{D}_j\} \in \{\mathbb{C}_m, \mathbb{D}_m\}} \frac{|\mathcal{D}_j|}{|\mathbb{D}_m|} w_{\mathcal{C}_j}^{t*}$
```

varying scales. Specifically, given a graph $\mathcal{G}$, GNN learns node- (Eq. 1) and graph-level (Eq. 2) representations under the following general formulation:

$$\boldsymbol{h}_p^{(l+1)} = \textsc{update}^{(l)}\left(\boldsymbol{h}_p^{(l)},\ \textsc{aggregate}^{(l)}\left(\left\{\boldsymbol{h}_q^{(l)},\ \forall\, q \in \mathcal{N}(p)\right\}\right)\right), \tag{1}$$

$$\boldsymbol{h}_G = \textsc{readout}\left(\{\boldsymbol{h}_v,\ \forall\, v \in \mathcal{V}\}\right), \tag{2}$$

where $\boldsymbol{h}_p^{(l)}$ denotes node p's representation at layer l, $\mathcal{N}(p)$ refers to the neighborhood of node p, and update and aggregate can be learnable functions that differ among GNN variants. Graph representation $\boldsymbol{h}_G$ can be obtained by pooling from all node representations where the readout can be a permutation invariant function such as summation or mean. With the representations, one can perform downstream tasks such as classification using, for example, a Multi-Layer Perceptron (MLP).

FedBrain adopt an optimized GCN model, proposed by Cui et al.,[4] as our default GNN architecture for both the server and client models. Furthermore, it is worth recognizing that brain networks are distinct from other real-world graphs due to, most prominently, that brain networks are often non-attributed, meaning that they lack useful initial node (ROI) features. To this end, we initialize the node-level features with the connection profiles of the brain network. 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 $\mathcal{E} = \mathcal{V} \times \mathcal{V}$.

A.2 Federated Atlas Mapping

The ROI (*i.e.*, node) system for a brain network is determined by the atlas template[25, 26] chosen during the parcellation process. Once a template is selected for a particular study, all brain networks within the dataset share the same ROI identities. However, in our cross-institutional learning setting, different institutions may resort to different parcellation systems for varying clinical purposes. This leads to heterogeneity in both sizes and structures of the parcellated networks, as well as divergent physical meanings of ROI features (*i.e.,* connectivity profiles). Although manual conversion between atlas systems is feasible, it is a labor-intensive process requiring extensive domain expertise. Therefore, we propose a data-driven transformation, backboned by linear autoencoders, that aims to project the varying dimensions in brain network features and structures across institutions onto a uniform dimension. Furthermore, we aim to align the physical interpretations of the projected features across studies. To this end, we leverage the FL approach to train the autoencoers with the intention of obtaining a global atlas projection. We illustrate the two components in more details in the following subsections.

Autoencoder framework. We employ a one-layer linear autoencoder (AE) to learn a dataset-specific simultaneous projection of ROI features and network structures. Given a target dimension M that is shared across all datasets and an input feature $\boldsymbol{X} \in \mathbb{R}^{N \times N}$ ensuring that $N > M$, the objective is to learn a linear transformation $\boldsymbol{W} \in \mathbb{R}^{N \times M}$, such that the mean-squared-error (MSE) reconstruction objective, denoted as $\mathcal{L}_{rec} = (1/N)\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^\top\|^2$, is minimized. In other words, the projected representations $\boldsymbol{X'} = \boldsymbol{X}\boldsymbol{W}$ preserves as much information from $\boldsymbol{X}$. Mathematically, $\boldsymbol{W}$ can be considered as a weighted linear combination of the column space of $\boldsymbol{X}$. Consequently, $\boldsymbol{W}$ learns an assignment of $\mathrm{col}(\boldsymbol{X})$ 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 $\boldsymbol{Z}_{i,j} = \mathbb{1}[\boldsymbol{W}_{i,j} \in \arg\mathrm{top}\, k\,(\mathrm{col}_j(\boldsymbol{W}))]$. Namely, 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. We use the FL framework, specifically `FedAvg` for ease of implementation, to jointly train dataset-specific autoencoder models to help generalize the projection schemes into a global projection. This approach allows for improved alignment of data semantics from diverse institutions through the application of the global projection. However, the architectural sizes of autoencoders may differ across clients due to variations in the original data dimensions. This discrepancy poses a challenge in communicating model parameters effectively between local clients and the global server.

To tackle this challenge, we introduce a unified mapping method designed to resize the global model, assumed to be the largest, to accommodate 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 ($N_L < N_G$) utilized for local data, our objective 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 accomplish this, we leverage the physical 3D coordinates of the ROIs, denoted as $D_G \in \mathbb{R}^{N_G \times 3}$ for the global template and $D_L \in \mathbb{R}^{N_L \times 3}$ for the local template. Initially, we calculate a distance matrix $\boldsymbol{S} \in \mathbb{R}^{N_L \times N_G}$, where $\boldsymbol{S}_{i,j} = d(\mathrm{row}_i(D_L), \mathrm{row}_j(D_G))$ represents the pairwise Euclidean distance between ROIs from the two templates. We then define $\boldsymbol{P}_{L_{i,j}} := \mathbb{1}[\boldsymbol{S}_{i,j} = \arg\min\,(\mathrm{col}_j(\boldsymbol{S}))]$. This implies 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 into virtual ROIs resembling the order and identities presented in the local template. During each communication round, clients start by downloading the server's parameter by first applying $\boldsymbol{W}_L = \boldsymbol{P}_L \boldsymbol{W}_G$. Subsequently, each client sends their updated parameters back to the server, employing the inverse mapping $\boldsymbol{W}_L^* = \boldsymbol{P}_L^\top \boldsymbol{W}_L^*$. Upon completion of the federated training, each local client will proceed to download the global parameters for further fine-tuning conducted independently over a few additional training epochs.

A.3 Guided Clustering

Another significant source of heterogeneity arises from the diversity in predictive neural circuitry patterns due to differences in the clinical purposes that motivates the studies. In particular, the inconsistencies are manifested in the various imaging modalities and patient outcomes characterized for each dataset. These variations can lead to a less-than-optimal local adaptation and knowledge generalization of the global model due to the arising phenomenon of client drifting. Hence, our goal is to find a balance between global generalization (*i.e.,* full federated training) and local personalization (*i.e.,* separate self training). Fortunately, as shown in Table 1, we observe that specific neural patterns are shared among distinct client institution subgroups. This observation encourages us to incorporate client clustering[27,28] into the FL process.

To be specific, 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.[27] Consequently, there is a need for a criterion to recognize this occurrence. Given a set of clients $\mathbb{C} = \{\mathcal{C}_i\}_{i=1}^N$, their data distributions $\mathbb{D} = \{\mathcal{D}_i\}_{i=1}^N$, gradients $\Theta = \{\Delta\theta_i\}_{i=1}^N$, and a hyperparameter ϵ_1, we define the criterion as follows:

$$0 \leqslant \left\| \sum_{i=1}^{N} \frac{|\mathcal{D}_i|}{|\mathbb{D}|} \Delta\theta_i \right\| < \epsilon_1. \tag{3}$$

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 ϵ_2, and define the additional criterion as follows:

$$\max\left(\|\Delta\theta_i\|\right) > \epsilon_2 > 0. \tag{4}$$

Clustering starts 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 server model and aggregation procedure. Clusters may further subdivide according to the criteria evaluated at each communication round.

Constrained clustering. A fundamental concern with the aforementioned base clustering mechanism is its disregard for shared clinical prior knowledge concerning the neural circuitry patterns of each client. Specifically, the heterogeneity issue may persist within the formed clusters due to significantly divergent clinical patterns, necessitating further division of clusters. This often results in the creation of singleton clusters, undermining the collaborative learning essence. This phenomenon is illustrated in Figure 3 (refer to Appendix B.4). Based on these observations, we propose an enhanced version of the clustering method that incorporates shared prior knowledge to guide the clustering process. For instance, concerning data modalities, it is intuitive to group clients with similar ROI connectivities and MRI data. Similarly, regarding clinical outcomes, federated learning (FL) on a cluster level could benefit from learning similar objectives. To achieve this, we establish must-links between pairs of clients that exhibit highly similar neural patterns and define cannot-links for those that do not. We introduce a weighted reward term λ_{must} and a penalty term λ_{cannot}, which are multiplied by the pairwise client similarity measure when must-links and cannot-links are identified, respectively. The complete process is detailed in Algorithm 2.

APPENDIX B. EXPERIMENTS

Datasets. We evaluate our framework using six real-world brain network datasets: BP,[16] HIV,[17] PPMI,[29] PNC,[30] ABIDE,[31] and ABCD;[32] of which, the BP and HIV datasets are strictly private, and the rest are publicly accessible to authorized users. We present the key statistics for each dataset in Table 1. Among them, BP, HIV, and PPMI contain multiple imaging modalities which we consider them to be trained on separate FL clients. Additionally, evident inconsistencies exist in their chosen parcellation systems, leading to varied network structures and dimensions. The sample sizes also exhibit significant variation among institutions, with the ABCD dataset, a multisite collaborative study, containing substantially more trainable data than other studies. Based on the available clinical outcomes, we define two potential 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.

Algorithm 2 Guided Clustering.

```
Input: Set of total available clients $\mathbb{C} \leftarrow \{\mathcal{C}_i\}_{i=1}^{M}$ 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 $\mathbb{S} \leftarrow \{s_1, s_2, \cdots, s_r\}$.
procedure CLUSTERING($\mathbb{C}$, $\Theta$, $\Psi$, $\lambda_{\text{must}}$, $\lambda_{\text{cannot}}$, $r$)
    Make every client its own cluster $\mathbb{S} \leftarrow \{\{\mathcal{C}_1, \Delta\theta_1, \psi_1\}, \cdots, \{\mathcal{C}_M, \Delta\theta_M, \psi_M\}\}$
    while $|\mathbb{S}| > r$ do                                    ▷ Check if number of clusters is greater than $r$
        for every pair of clusters $(\mathcal{S}_i, \mathcal{S}_j)$ in $\mathbb{S}$ do              ▷ Such that $i, j \in |\mathbb{S}|, i \neq j$
            Calculate the inter-cluster linkage distance $d(\mathcal{S}_i, \mathcal{S}_j) \leftarrow$ LINKAGE$(\mathcal{S}_i, \mathcal{S}_j)$
        Find the pair with min linkage distance $min(d(\mathcal{S}_i, \mathcal{S}_j)) : i, j \in |\mathbb{S}|, i \neq j$ and merge
    return $\mathbb{S} : |\mathbb{S}| = r$
procedure LINKAGE($\mathcal{S}_1$, $\mathcal{S}_2$)
    for every pair of clients $(\{\mathcal{C}_p, \Delta\theta_p, \psi_p\}, \{\mathcal{C}_q, \Delta\theta_q, \psi_q\})$ in $(\mathcal{S}_1, \mathcal{S}_2)$ do     ▷ $p \in |\mathcal{S}_1|, q \in |\mathcal{S}_2|$
        Calculate the cosine distance $cos(p, q) \leftarrow 1 - ((\Delta\theta_p \cdot \Delta\theta_q)/(\|\Delta\theta_p\|\|\Delta\theta_q\|))$
        Determine if the pair forms must- or cannot-link $link(p, q) \leftarrow$ VALIDATE$(\psi_p, \psi_q)$
        Shorten their cosine distance if must-link $cos(p, q) \leftarrow \lambda_{\text{must}} \cdot cos(p, q)$
        Increase their cosine distance if cannot-link $cos(p, q) \leftarrow \lambda_{\text{cannot}} \cdot cos(p, q)$
    return Averaged distance $avg(cos(p, q)) : \forall p \in |\mathcal{S}_1|, \forall q \in |\mathcal{S}_2|$
procedure VALIDATE($\psi_1$, $\psi_2$)                        ▷ An example to determine must- or cannot-links
    return Must-link if overlapping attributes in $(\psi_1, \psi_2)$ exceeds 80%
    return Cannot-link if overlapping attributes in $(\psi_1, \psi_2)$ subceeds 20%
```

Table 1. Dataset summarization.

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

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`,[19] `FedProx`,[20] and `SCAFFOLD`.[21] 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.

Default parameters. The optimized GCN[33] model contains a default hidden size of 32, with 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 classification accuracy as the evaluation metric.

Throughout our experiments, we employ a batch size of 32 and use the Adam[34] optimizer with a learning rate of $1 \cdot 10^{-4}$ and an ℓ_2 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 μ value of `FedProx` is set to 0.01. Option II. of `SCAFFOLD`, which reuses previously computed gradients, is used to update the local control variates. 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 autoencoders. In other words, the autoencoder model for ABCD would simply require an identity mapping from the server model for each federated communication. The clustering criteria ϵ_1 and ϵ_2 are set to 1.50 and 0.05, respectively, and the weighted terms λ_{must} and λ_{cannot} are set to 0.5 and 2.0, respectively. Lastly, the proposed guided clustering algorithm is aimed to ensure the production of 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)

We present a comprehensive performance comparison in Table 2 and remark on two key observations:

1. All FL-based algorithms demonstrate a notable improvement in accuracy compared to `self-train`, with a reported average relative gain of 15.34% across all client data. Particularly, clients with smaller sample sizes, such as BP, HIV, and PNC, experience the most significant enhancement, with an average relative gain of 19.31%. This underscores the valuable advantage of collaborative learning in generalizing knowledge across institutions to address limited training resources and alleviate model overfitting. Additionally, FedBrain outperforms, even in comparison to the most robust baseline `SCAFFOLD`, by a relative margin of 14.29%, while also substantially reducing performance variance across clients. This underscores the importance of tailoring FL approaches to account for the unique heterogeneity properties of brain network data. Furthermore, the improvements achieved by FedBrain are statistically significant, validated by passing the paired t-test with a threshold p value of 0.05 in comparison to all selected methods.

2. It is noteworthy that, among the chosen FL baselines, except for FedBrain, there is a slightly increased performance variance across clients compared to `self-train`. This variance primarily stems from the

Table 2. Performance comparison. We present classification accuracy for each client averaged from 10-fold cross-validation along with standard deviations, and a combined accuracy averaged across all clients.

Clients	BP-fMRI	BP-DTI	HIV-fMRI	HIV-DTI	PPMI-PICo	PPMI-Hough	PPMI-FSL	PNC	ABIDE	ABCD	
Accuracy			average					average			combine
self-train	0.5463(±0.019)	0.5012(±0.082)	0.5286(±0.035)	0.4571(±0.140)	0.6394(±0.034)	0.6570(±0.054)	0.6852(±0.041)	0.5034(±0.052)	0.5025(±0.007)	0.5342(±0.002)	0.5555(±0.073)
FedAvg	0.6037(±0.073)	0.5158(±0.013)	0.5457(±0.153)	0.5000(±0.078)	0.7925(±0.002)	0.7633(±0.031)	0.7925(±0.002)	0.5434(±0.008)	0.5044(±0.012)	0.5167(±0.017)	0.6078(±0.118)
FedProx	0.6084(±0.117)	0.5853(±0.085)	0.6200(±0.132)	0.6029(±0.097)	0.7925(±0.002)	0.7536(±0.037)	0.7925(±0.002)	0.6057(±0.018)	0.5594(±0.003)	0.5700(±0.020)	0.6490(±0.088)
SCAFFOLD	0.5800(±0.120)	0.6400(±0.049)	0.6343(±0.070)	0.6629(±0.057)	0.7778(±0.000)	0.7944(±0.014)	0.7889(±0.014)	0.6015(±0.009)	0.5765(±0.090)	0.5980(±0.045)	0.6654(±0.084)
FedBrain	**0.7389(±0.066)**	**0.7500(±0.077)**	**0.7857(±0.071)**	**0.8143(±0.070)**	**0.8102(±0.010)**	**0.8102(±0.010)**	**0.8095(±0.010)**	**0.7275(±0.044)**	**0.6549(±0.034)**	**0.7033(±0.033)**	**0.7605(±0.052)**

Table 3. Performance of atlas mapping and its variants.

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. Performance of guided clustering and its variants.

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

unique heterogeneity characteristics arising from brain network data, that are left unaddressed by the more generically applicable state-of-the-arts. On the other hand, SCAFFOLD emerges as the top performing baseline, showcasing an impressive average gain of 5.89% over its competitors. This outcome underscores the robustness of SCAFFOLD in addressing client heterogeneity through controlled gradient correction. Additionally, alongside FedProx, which is also adept at handling data and system heterogeneity, the performance variance is reduced compared to FedAvg, which does not consider data heterogeneity issues at all.

B.2 Ablation studies (RQ2)

In this section, we study the contributions made by the two constituent components of FedBrain — guided clustering and federated atlas mapping — through separate investigations. The results are detailed in Table 3 and Table 4, in which we include the minimum client-wise gain over the raw baseline ("min gain"). To underscore the impact of each component, we maintain the best configuration of one while evaluating the other.

Regarding client clustering, we explore the impact on overall performance both without clustering and without shared prior knowledge guidance. We observe that personalizing client optimization through similarity-based clustering leads to a significant enhancement in downstream performance, with a relative margin of 4.48%. Moreover, by integrating clinical prior knowledge and applying relevant constraints, we further enhance cluster-specific learning and knowledge generalization, resulting in an additional relative gain of 5.17% and a reduction in performance variance across participating clients.

Regarding atlas mapping, we assess its influence on overall performance both without the entire module and without federated training. When atlas mapping is not applied, we attach a learnable linear projection head to the client's GNN model, which do not join the FL process. The results reflect that ensuring consistency in feature and network dimensions, through un-federated atlas mapping, results in a relative gain of 6.12% in performance accuracy. With federated trainig which facilitate the alignment of the physical meanings of projected features would further enhance performance by a margin of 4.95%, demonstrating its effectiveness in countering incongruous ROI parcellation systems.

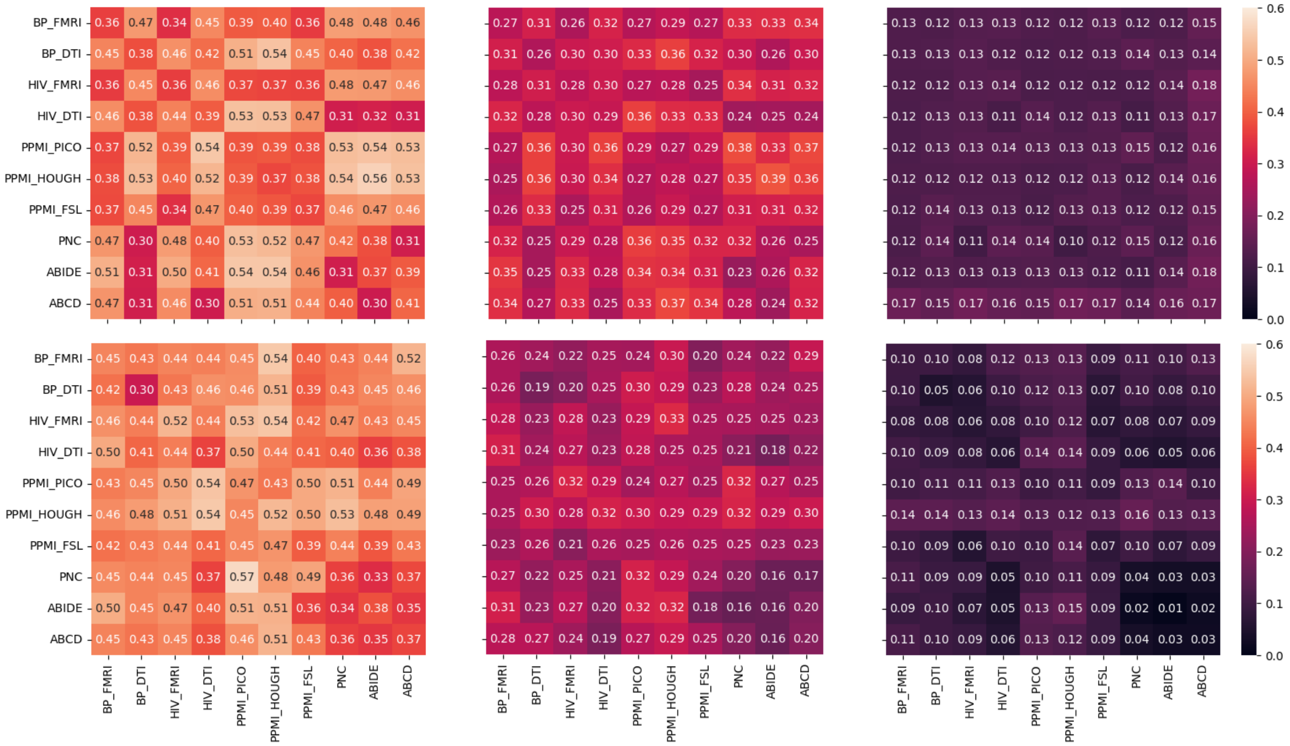

Figure 2. Pairwise structure- (upper) and feature-level (lower) heterogeneity measures across all datasets compared on brain networks processed without atlas mapping (left), with atlas mapping but without federated training (mid), and full federated atlas mapping (right). The smaller the value, the less heterogeneity exists within the investigated pair.

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

In this section, we substantiate the impact of the proposed federated atlas mapping in alleviating structure- and feature-level heterogeneity. Our findings, illustrated in Figure 2, compare heterogeneity measures among brain networks and features processed under three scenarios: without the entire module, without federated training, and with full federated atlas mapping. To quantify our evaluations, we utilize two distinct metrics[28] to measure the averaged pair-wise heterogeneity among datasets. For structure-level heterogeneity, we employ the Anonymous Walk Embeddings (AWEs)[35] technique to generate representations for each brain network graph. Subsequently, we calculate the Jensen-Shannon distance between each pair of AWE representations from two different datasets. For feature-level heterogeneity, we examine the empirical distribution of features between all pairs of connected nodes (ROIs) within each network. We then compute the Jensen-Shannon distance between every pair of these distributions extracted from different networks belonging to different datasets. Our observations indicate that the integration of atlas mapping and federated training significantly reduces the level of heterogeneity across datasets in both network structures and ROI features.

Moreover, by transforming the network structure and ROI features, we observe an improved downstream performance as reflected in Table 5. In particular, transforming the network structures alone would result in an average relative gain of 2.73% over the raw baseline, but with deteriorated performance on a few clients. When further integrating feature-level alignment, we observe an average relative gain of 8.67% with all clients receiving positive performance increase. Furthermore, our analysis indicates a significant reduction in time complexity when training on transformed data, bringing down the actual runtime from approximately 612 seconds to 266 seconds over 80 communication rounds, which is almost three times faster. Adding the overhead of completing the atlas mapping pre-processing, which finishes in roughly 74 seconds, this efficiency improvement is still significant.

Table 5. Performance of transformed structure and feature.

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. Performance of constrained clustering.

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

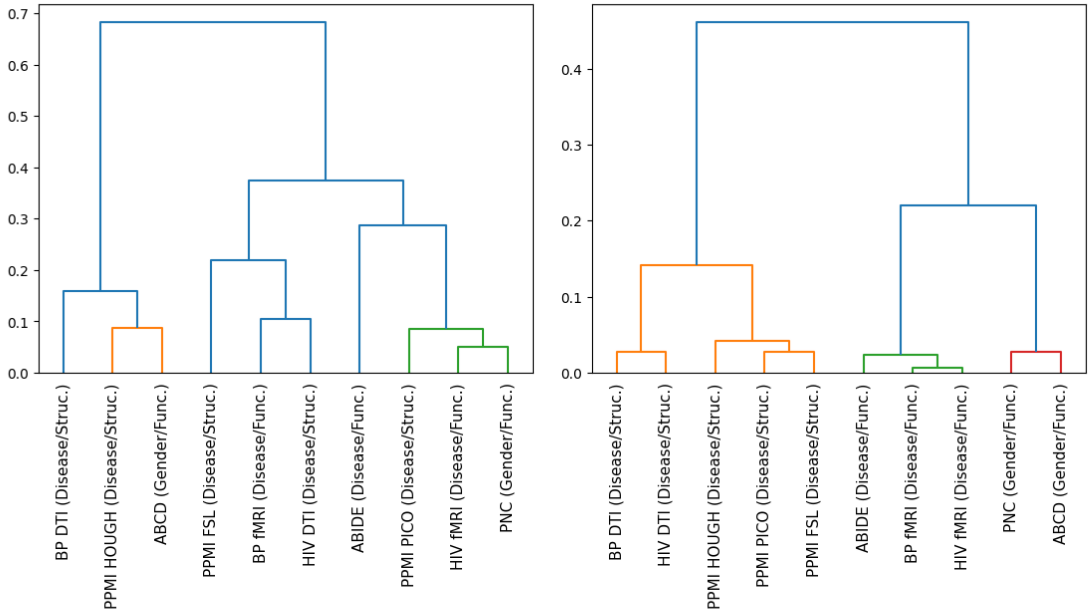

Figure 3. Dendrogram visualization of cluster results from standard hierarchical clustering (left) and prior knowledge guided clustering (right). We list the client names alongside its neural circuitry attributes, namely clinical outcomes (*e.g.,* disease/gender) and data modalities (*e.g.,* functional/structural connectivities).

B.4 Clustering analysis of guided clustering (RQ4)

In this section, we explore the influence of guided clustering on cluster formation. We compare the outcome with a standard hierarchical approach, which we illustrate our visualizations in Figure 3. Specifically, the linked branches depict the hierarchical relationships, with blue-colored lines representing singleton clusters, and other cluster assignments are highlighted under different color cues. We observe that with clinical prior knowledge guidance, our approach is significantly more effective in grouping studies (*i.e.,* clients) with similar imaging modalities and patient outcomes. Since the neural circuitry patterns are given higher priorities in the clustering process, the guided approach would separate datasets into different clusters belonging to the same study. Fortunately, it produces fewer number of clusters with each holding a reasonable amount of clients, avoiding the production of singletone clusters, which is prominent when utilizing the standard approach.

Moreover, the results presented in Table 6 depict our investigation into the downstream performance impact when utilizing prior knowledge guidance exclusively based on either must-link or cannot-link information. We observe that imposing cannot-link constraints alone leads to a relative gain of 1.47% over standard clustering. When solely guided by must-links, we achieve a further improvement of 1.53%, bringing the performance to within a mere 2.10% difference from implementing the fully guided approach. 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

FL on Graphs. FL has gained significant popularity in various domains including images, text, and multi-modality learning,[36–39] for its capabilities in collaboratively training deep learning models while preserving data privacy of its participants (*i.e.,* clients). Recently, graph-level FL is also remarked by significant advancements. Lalitha et al.,[40] Rizk et al.,[41] and Caldarola et al.[42] are some trailblazing literatures that proposed to model clients as nodes in graphs where the collaborative training is analogous to neighborhood aggregation in graph data learning. FedGraphNN and Liu et al.[43,44] are prominent benchmark surveys that have contributed to examine the applications and theoretical insights into GNN-based FL across graphs in diverse data domains. However, graph FL encounters unique challenges stemming from graph-specific heterogeneities, such as inconsistent node- and edge-level semantics. In response to this, GCFL[28] investigates graph-level heterogeneity across domains and proposes a clustered graph FL framework, which serves as a significant influence on the development of FedBrain. Another approach, FedLit,[45] suggests to cluster clients based on the latent link types of graphs to address link-level heterogeneity. Despite these efforts, the distinct manifestations of heterogeneity in brain network studies, including variances in parcellation systems and neural circuitry patterns, render most existing graph FL frameworks inapplicable. While research on GNN-based FL for neuroimaging data has shown promise,[46,47] current techniques tend to focus on privacy preservation[48] or domain adaptation,[49] often overlooking the importance of data-level alignment and client personalization in addressing data heterogeneity.

GNNs for Brain Network Analysis. Inspired by the recent successes of GNNs in learning graph-structured data,[33,50–52] numerous pioneering efforts have emerged in applying GNN models to brain network analysis. BrainGNN[8] utilizes ROI-aware graph convolutional and pooling layers to predict neurological biomarkers from fMRI data. BrainNetTF[7] introduces a transformer architecture with an orthonormal clustering capable of considering ROI similarity within functional modules. Existing litertatures[5,53–56] have showcased that when sufficient data is available, GNNs can substantially improve performance in disorders prediction. However, in most practical scenarios, training samples are often limited especially for clinical studies.[57] This limitation hinders the GNNs for effective modeling of brain network data, motivating research in overcoming data scarcity and heterogeneity to improve performance in real clinical tasks.