Evaluating the explainability of BrainGNN on prediction of ASD diagnosis

Matheo Angelo Pereira Dantas, André Carlos Ponce de Leon Ferreira de Carvalho

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 – 13560.970 – São Carlos – SP – Brazil

matheoangelo@usp.br, andre@icmc.usp.br

Abstract. *Due to the absence of a biological marker for Autism Spectrum Disorder (ASD), most of recent research attempts to uncover the neurological patterns of ASD using Deep Learning, and these patterns, hidden in the latent feature space of neural networks, have to be interpreted with the use of Explainable AI. However, although many of the models proposed for the problem report results of explainability, they are not evaluated with any metric, so their reliability is unknown. The objective of this paper is to propose an evaluation framework to fill this gap, and here, we focus on a detailed analysis of a well-known model from the literature, BrainGNN. We trained BrainGNN in varying hyperparameter settings that influence explainability and analyzed our findings for each case.*

1. Introduction

ASD (Autism Spectrum Disorder) is characterized by communication impairments and repetitive behavior patterns, in different contexts, since childhood [American Psychiatric Association 2013]. Individuals with ASD are more likely to experience depression and other conditions of psychic suffering [American Psychiatric Association 2013], and ASD is legally recognized as a disability [Brazil 2012]. Therefore, it is crucial to ensure quality of life for autistic individuals through adequate therapeutic support and disability rights.

These benefits can only be accessed with formal medical diagnosis. However, ASD does not have any biomarker known to science, and has to be diagnosed through a clinical evaluation with psychological tests and interviews with the patient and their close relatives [American Psychiatric Association 2013]. This method is flawed, being susceptible to delayed diagnosis and misdiagnosis [Huang et al. 2020].

The problem has recently prompted the search for biomarkers of ASD in different forms of biological data, and one of the most investigated forms of data is fMRI (Functional Magnetic Resonance Imaging [Zhang and Chiang-shan 2012]). An fMRI exam is a three-dimensional filming of the BOLD (Blood-Oxygen-Level-Dependend) signal, mapping brain activity through the flow of blood in the brain.

Since this data format is highly complex and difficult to manually investigate, researchers often rely on the use of deep neural networks trained by deep learning, usually GNNs (Graph Neural Networks [Zhang et al. 2023a]. These algorithms are used to induce predictive models able to automatically extract patterns associated with the functional connectivity of the brain. These models can be later deployed to classify a new fMRI image as either presenting or not presenting a pattern associated with ASD, providing support for the diagnosis by a specialist. The use of GNNs has been part of many

breakthroughs in recent life science research [Wang et al. 2023]. These previous contributions show that they can be a promising tool to support the investigation of neurological aspects associated with ASD, as well as other mental disorders.

Despite of their potential, a major challenge that need to be overcome for their validation and acceptance is the lack of interpretability in these models. This deficiency undermines the transparency of the model's decision process, compromising the trustworthiness needed for their deployment in the public healthcare system.

To deal with this limitation, the scientific community has been investing in the creation of explainability methods [Yuan et al. 2022], as well as self-explainable neural networks that have built-in interpretable mechanisms, out of which it is expected to derive explanations that are more faithful [Dai et al. 2024]. Scientists seek to not only create machine learning models that are highly precise, but also to create tools that are informative to the diagnostic process and can be leveraged to gain new insights into the etiology of the health conditions that undergo their investigation.

Within the scope of this problem, although extensive literature has been developed in explainability, a key aspect is generally neglected: the evaluation of explanations. It is not a standard practice to provide metrics of explainability along with the lists of important brain regions, so biomarker suggestions are not guaranteed to be reliable. Such is the case of BrainGNN, the most popular GNN model in our application domain.

In this paper, we investigate the use of BrainGNN and evaluate its explainability, using metrics that could be extended to other models in a future comparative analysis. The code was directly sourced from the BrainGNN repository, with changes made to include functions for explainability along with the results of our experiments, and it is available on GitHub[1].

1.1. Related Work

1.1.1. Explainability in GNNs

Explainability approaches in GNNs are divided into two types: self-explainable GNNs and post-hoc explainability [Dai et al. 2024].

Self-explainable GNNs are designed for explainability and make predictions using interpretable internal mechanisms. Generally, this is done by creating a separate model to filter only the relevant parts of the graph before making predictions [Wu et al. 2020] or by generating prototypes for each class and making predictions through direct comparisons with these prototypes [Zheng et al. 2024a].

On the other hand, post-hoc explainability involves applying ready-made methods after the GNN has been trained [Yuan et al. 2022]. Initially, these methods are subdivided into instance-level methods, which explain a specific prediction, and global-level methods, which explain the GNN's decision-making as a whole. Global-level methods may involve generating prototypes for each class based on the GNN's behavior or aggregating instance-level results. Meanwhile, instance-level methods can rely on computing the gradient of the input with respect to the output (Gradients/Features), decomposing the output through the GNN's weights (Decomposition), searching for minimal changes in the graph

[1] https://github.com/matheo-angelo/BrainGNN_Pytorch

that affect the model's output (Perturbation), or creating simple interpretable models that approximate the local decision boundary for a specific instance (Surrogate).

1.1.2. GNN Explainability for ASD diagnosis from fMRI data

Systematic literature reviews on the topic [Luo et al. 2024, Zhang et al. 2023b] show that there are two distinct data pipelines for predicting ASD (Autism Spectrum Disorder) diagnosis using GNNs and fMRI scans. The first, which is the focus of this research [Li et al. 2021], represents each individual's brain as a graph, where each node corresponds to a Region of Interest (ROI) according to a brain atlas, and each edge represents a pair of ROIs with strong positive correlation in their BOLD (Blood-Oxygen-Level-Dependent) time series from the fMRI scan [Zhang and Chiang-shan 2012]; some variations might combine fMRI data with other modalities, such as Diffusion Tensor Imaging (DTI) [Yang et al. 2023], or model the fMRI exam as a dynamic graph and apply Spatio-Temporal Graph Neural Networks [Yan et al. 2022]. The second approach [Lin et al. 2022] combines each patient's brain connectivity representation into a population graph, where each node represents an individual and each edge encodes feature similarity (e.g., brain connectivity, genetic data, and demographic information).

Regarding explainability (at least restricting our scope to static subject graphs constructed from fMRI data only), current studies focus on implementing novel self-explainable architectures tailored to the specific application or using *ad-hoc* methods specific to the model architecture, rather than applying post-hoc methods or even using off-the-shelf self-explainable models. This explainability can operate at the node level, to infer critical brain regions [Li et al. 2021], or at the edge level, to highlight significant brain connections [Zheng et al. 2024b]. In both cases, model-level explainability is achieved by aggregating instance-level results, aiming to uncover neurological features common to all autistic individuals that could later be cataloged as biological markers. Biomarker suggestions are generally trusted based on the model accuracy and how well they match prior knowledge on the neurology of autism, but they lack the use of specific metrics to assess the reliability of these suggestions, such as Fidelity [Yuan et al. 2022].

2. Methodology

This section describes the main steps followed for the experiments carried out in this study.

2.1. Dataset

For this study, we use the ABIDE I dataset, which has fMRI exams from 539 autistic individuals and 573 typical controls, sampled from several sites in the United States and Europe. It was obtained from the ABIDE I Preprocessed Connectomes Project (PCP)[2], which has several download options with the exams pre-processed using different brain atlases, as will be explained in the next section.

2.2. Pipeline overview

To make a prediction for one instance, the fMRI exam is first parceled into several regions, called Regions Of Interest (ROI), using a brain atlas (here, we chose the Harvard-Oxford

[2]http://preprocessed-connectomes-project.org/abide/

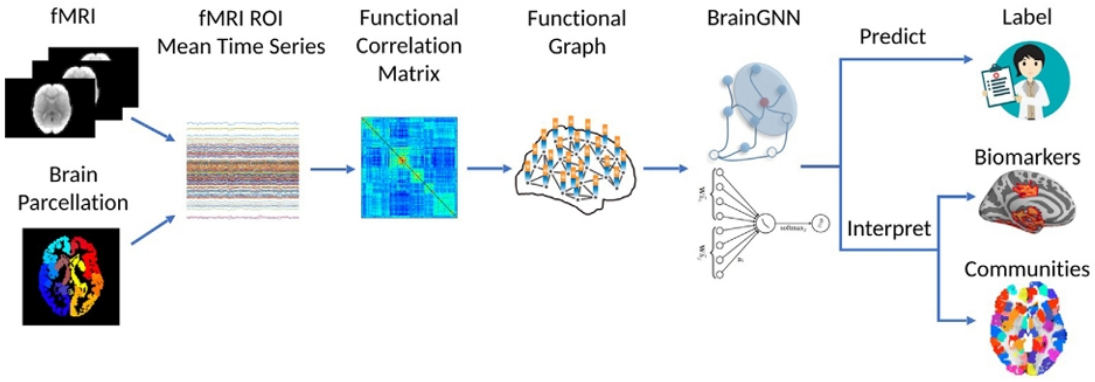

Figure 1. Illustration of the BrainGNN model pipeline [Li et al. 2021].

atlas [Smith et al. 2004], with 110 labeled ROIs). We then average the intensity of the BOLD signal in each ROI, obtaining a time series for each ROI. We then calculate the correlation matrix of all ROI fMRI series X and use it as the feature matrix of the graph, and apply a threshold to X to obtain the adjacency matrix A, with edge weights corresponding to the pairwise correlation value. These matrices are then fed into a Graph Neural Network $f(X, A)$ which outputs a label for either Autistic or Typical Control and provides explainability by pointing to ROIs that were relevant for the decision of the model. The full pipeline was executed on Google Colab for this paper, and it is illustrated in Figure 1.

2.3. BrainGNN

The model is composed of two stacked message-passing layers, each followed by a node pooling layer, and after those, we have a global pooling layer that obtains a vector representation of the whole graph and feeds it into a MLP (Multi-Layer Perceptron) classifier. The code used[3] was directly resourced from the original BrainGNN repository, with modifications to implement the specific evaluative experiments of this paper.

2.3.1. Ra-GConv Layer

The ROI-Aware Graph Convolutional Layer (Ra-GConv) is a heterogeneous graph message-passing layer with 8 basis transformations [Schlichtkrull et al. 2018], where different ROIs are assigned different weight matrices, to avoid the node permutation invariance that applies to traditional graph neural networks [Bronstein et al. 2017]. It follows the scheme in the equation:

$$\tilde{h}_i^{(l+1)} = ReLU\left(W_i^{(l)} h_i^{(l)} + \sum_{j \in N_i^{(l)}} e_{ij} W_j^{(l)} h_j^{(l)}\right)$$

Where $W_i^{(l)}$ is the weight matrix associated with node i in the l-th message passing layer, $h_i^{(l)}$ is the hidden embedding of node i in the l-th layer, $N_i^{(l)}$ is the neighborhood of

[3] https://github.com/matheo-angelo/BrainGNN_Pytorch

node i in the l-th layer, and e_{ij} is the edge weight between nodes i and j, normalized such that $\sum_{j \in N_i^{(l)}} e_{ij} = 1$ for every node i.

2.3.2. TopK Pooling Layer

TopK Pooling [Gao and Ji 2019] [Cangea et al. 2018] uses a learnable projection vector $p^{(l)}$ to drop all nodes in the graph except for the ones that are the most similar to $p^{(l)}$.

$$\begin{aligned} s^{(l)} &= \tilde{H}^{(l+1)} p^{(l)} \,/\, \| \, p^{(l)} \, \| \\ i &= top_k(s^{(l)}) \\ H^{(l+1)} &= (\tilde{H}^{(l+1)} \odot sigmoid(s^{(l)}) \,)_i \\ A^{(l+1)} &= A^{(l)}_{i,i} \end{aligned}$$

Where top_k returns the indices of the k highest values in the vector, $\odot$ is the Hadamard broadcasting product, X_i is the X matrix using only the rows whose indices belong to i, and $A_{i,i}$ is the adjacency matrix A using only the rows and columns in i.

2.3.3. Global Pooling Layer

After all message-passing layers, we use the node embeddings of all layers to obtain a graph-level representation:

$$\begin{aligned} z^{(l)} &= mean \, H^{(l)} \, \| \, max \, H^{(l)} \\ z &= z^{(1)} \, \| \, z^{(2)} \, \| \, ... \, \| \, z^{(L)} \end{aligned}$$

Where $\|$ denotes concatenation and L is the number of layers in the GNN.

2.3.4. Loss function

The loss function of BrainGNN is composed of several functions, that may be targeted at optimizing classification or explainability:

Classification loss - We optimize model classification on the binary cross entropy loss:

$$L_{classification} = -\frac{1}{M} \sum_{m=1}^{M} \sum_{c=1}^{C} y_{m,c} \, log(\hat{y}_{m,c})$$

Where M is the number of instances, C is the number of classes (here, $C = 2$), $y_{m,c}$ is a binary value that represents whether instance m belongs to class c, and $\hat{y}_{m,c}$ is the probability output of the model for estimating $y_{m,c}$.

Group-level consistency loss - In order to enhance model-level interpretability, we may want to force the model to always select similar node pooling sets for graphs of the same class, in order to achieve coherent biomarker suggestions. The loss function is calculated as following:

$$L_{consistency} = \sum_{c=1}^{C} \sum_{m,n \in F_c} \|s_m^{(1)} - s_n^{(1)}\|^2$$

Where F_c is the set of graphs belonging to class c in the current batch and $s_m^{(1)}$ is the vector of pooling scores in the first TopK Pooling layer for instance m.

TopK Pooling loss - Optimizes the model to give high pooling scores to nodes that are selected in the pooling layers and, likewise, low scores to nodes that are dropped, using cross entropy. We optimize it to approximate a binary value that indicates whether a node will be pooled.

$$L_{pooling}^{(l)} = -\frac{1}{M} \sum_{m=1}^{M} \frac{1}{N^{(l)}} \left(\sum_{i=1}^{N^{(l)}} \rho_{m,i}^{(l)} \, log(s_{m,i}^{(l)}) \right)$$

Where $N^{(l)}$ is the number of nodes in layer l, $\rho_{m,i}^{(l)}$ indicates whether node i in graph m was pooled in layer l, and $s_{m,i}^{(l)}$ is the pooling score of that node.

The total loss function, therefore, is:

$$L_{total} = L_{classification} + \lambda_1 \sum_{l=1}^{L} L_{pooling}^{(l)} + \lambda_2 \, L_{consistency}$$

Where λ_1 and λ_2 are hyperparameters.

2.3.5. Explainability

Individual-level explainability comes in the form of a node mask, provided by the last set of pooled nodes remaining before the Global Pooling layer of the GNN, and model-level explainability is constructed from counting the occurrences of each ROI in individual-level explanations and selecting the most frequent ROIs. We create biomarker suggestions for ASD by looking at the frequency of each ROI in the explanation masks of the model for each sample of the test set assigned as autistic by the model, and the 10 most frequent ones are suggested as biomarkers.

Although it is not the target explanation modality of this paper, BrainGNN also has an implicit community detection mechanism. As a heterogeneous GNN [Schlichtkrull et al. 2018], it composes the weight matrix of each node type (hereby, the weight matrix $W_i^{(l)}$ associated with each ROI i) from a linear combination of basis transformations $(B_1^{(l)}, B_2^{(l)}, ..., B_K^{(l)})$. The weight $\alpha_{iu}^{(l)}$ of basis u for node i in layer l can be

understood as the intensity of the relation between the node i and the ROI community u. Here, we use $K = 8$ and therefore we have 8 implicit communities in every graph.

2.3.6. Hyperparameters

We tested different combinations for the hyperparameter values, to study their effects in explainability, adjusting them with the following values:

- **Pooling ratio:** The fraction of nodes that remain in the graph after each TopK Pooling layer. Default: 0.5.
- **TopK Pooling regularization:** Value of λ_1 in the TopK Pooling loss. Default: 0.1.
- **Group consistency regularization:** Value of λ_2 in the group consistency loss. Default: 0.1.

The other hyperparameters of the model were set to their default value, which can be found in the source code.

2.4. Explainability metrics

We applied different statistics to evaluate the quality of explanations, including established metrics from the literature and *ad-hoc* metrics.

Sparsity [Yuan et al. 2022]: explanations must be efficient in summarizing the decision of the model using only a small portion of the whole graph. Here, N is the number of explanations to be evaluated, m_i is the node mask explanation for instance i and M_i is the set of nodes in the graph indexed by i. Sparsity is the average portion of the graph left out of the explanations.

$$Sparsity = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{|m_i|}{|M_i|}\right)$$

Fidelity$_+$ and Fidelity$_-$ [Yuan et al. 2022]: explanations must be faithful to the model. Fidelity$_+$ measures if the model includes relevant information in the explanation, by perturbating the nodes in the explanation mask and measuring the change in the model output. Fidelity$_-$, on the other hand, measures if no information is lost by ignoring nodes out of the explanation mask.

In the notation presented next, $f(G_i)$ is the probability output of the model for graph G_i, $G_i^{m_i}$ is the graph G_i with the node features assigned as 0 for all nodes, except for those in the mask m_i, and $1 - m_i$ is the complementary set to m_i with respect to the full node set of G_i.

$$Fidelity_+ = \frac{1}{N}\sum_{i=1}^{N}|f(G_i) - f(G_i^{1-m_i})|$$

$$Fidelity_- = 1 - \frac{1}{N}\sum_{i=1}^{N}|f(G_i) - f(G_i^{m_i})|$$

Biomarker consistency: to have reliable model-level explanations, besides having faithful instance-level explanations, they must consistently point to the ROIs that will be assigned as a biomarker suggestion. In the Consistency equation, M is the amount of ROIs in our brain atlas, B is the number of ROIs to be chosen as biomarkers and n_i is the amount of occurrences of the i-th ROI in explanation masks.

$$Consistency = \frac{\sum_{i=1}^{B} n_i}{\sum_{i=1}^{M} n_i}$$

3. Experimental results

The experimental results obtained by the model trained for each hyperparameter combination are shown in Table 1.

It must be observed thar the training accuracy in the table is the highest training accuracy of all epochs, and it might not correspond to the model applied to the test set, which was selected based on validation accuracy.

The accuracy of the models in the test set, although demonstrating capacity to capture patterns in autistic neurology, was relatively low, while the accuracy in the training set was much higher, suggesting that more data is necessary to avoid overfitting and enable the successful application of Deep Learning in the problem. It is also lower than the 79.8% accuracy reported in the original BrainGNN paper [Li et al. 2021], which used a task-fMRI dataset, and possibly had an experimental setting where external stimuli triggered more evident neurological manifestations of ASD.

The low accuracy impacted the fidelity measures: $Fidelity_+$ was very low while $Fidelity_-$ was high, meaning that the model generally fails to leverage positional information and therefore ROI explanations are not informative. $Consistency$ was also low, as the node mask explanations were scattered across brain ROIs. $Sparsity$, however, was satisfyingly high across all trained models; for this specific architecture, it a function of the *Pooling Ratio* hyperparameter, where slightly decreasing its value did not seem to decrease the model's accuracy while also increasing sparsity.

As for the influence of hyperparameter choice in explainability, group-level regularization did not influence biomarker consistency, while increasing explanation sparsity slightly improved it. TopK-Pooling regularization also did not have any influence.

4. Conclusions and Future research directions

This work investigated the use of GNNs, in particular the BrainGNN neural network architecture based on graphs, as classification models to support ASD diagnosis. For such, we used a public domain dataset, ABIDE, for the experiments. Given the importance of explaining the decisions made by the model for its acceptance in public healthcare, we also investigated how these explanations can be evaluated. The experimental results obtained show that, although the models seem reasonably capable of extracting patterns, more research should focus on making model results more reproductible and accurate, and with that, be able to extract reliable biomarkers from accurate and faithful explanations.

Table 1. Results from BrainGNN experiments. Each column represents a hyperparameter combination, and each row corresponds to a performance metric.

	DEFAULT	*Ratio* = 0.3	*Ratio*=0.3, *TopK*=0.5	*Group* = 0.5
Accuracy (training)	0.82	0.87	0.82	0.83
Accuracy (testing)	0.51	0.57	0.59	0.56
Fidelity$_+$	0.050	0.025	0.032	0.078
Fidelity$_-$	0.79	0.85	0.88	0.85
Sparsity	0.75	0.90	0.90	0.75
Consistency	0.23	0.34	0.26	0.19
Biomarker	• Left Parahippocampal Gyrus; posterior division • Left Heschl's Gyrus (includes H1 and H2) • Left Frontal Orbital Cortex • Right Temporal Fusiform Cortex; anterior division • Right Lingual Gyrus • Right Planum Temporale • Right Supracalcarine Cortex • Left Intracalcarine Cortex • Left Temporal Fusiform Cortex; posterior division • Left Putamen	• Left Supramarginal Gyrus; posterior division • Left Frontal Medial Cortex • Left Superior Parietal Lobule • Right Thalamus • Left Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex) • Left Angular Gyrus • Right Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex) • Left Planum Temporale • Left Inferior Frontal Gyrus; pars triangularis • Right Amygdala	• Left Supramarginal Gyrus; posterior division • Right Angular Gyrus • Left Frontal Medial Cortex • Right Subcallosal Cortex • Right Superior Temporal Gyrus; anterior division • Right Amygdala • Right Paracingulate Gyrus • Left Planum Temporale • Left Superior Parietal Lobule • Left Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex)	• Left Supramarginal Gyrus; posterior division • Right Angular Gyrus • Left Planum Temporale • Left Superior Parietal Lobule • Right Occipital Pole • Right Subcallosal Cortex • Left Frontal Medial Cortex • Right Thalamus • Right Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex) • Right Amygdala

As future work to expand this study and improve the results of related studies, we understand that further research could be carried out in the following directions:

- Investigate different varieties of the BrainGNN self-explainability (e. g. using the set of pooled nodes from the first *TopK Pooling* layers rather than the final remaining node set);
- Test other hyperparameter combinations;
- Evaluate the community detection mechanism in BrainGNN;

- Apply the same explainability metrics to other self-explainable models;
- Compare the results of self-explainable mechanisms with state-of-the-art *post-hoc* explainers such as SubgraphX [Yuan et al. 2021].
- Test hyperparameter variations with more robust statistical evaluations (such as hypothesis testing and cross-validation).

5. Acknowledgements

This paper is part of a project supported by a scholarship from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), through the grant of number 24/09181-2. The project, called "Explainability in Graph Neural Networks for Autism Assessment Using fMRI Analysis", is advised by professor André Carlos Ponce de Leon Ferreira de Carvalho (ICMC-USP).

References

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition*. American Psychiatric Publishing, Washington, D.C., 5 edition.

Brazil (2012). Law no. 12,764, of december 27, 2012. establishes the national policy for the protection of the rights of persons with autism spectrum disorder. Known as "Berenice Piana Law" or "Brazilian Autism Law". It recognizes autism spectrum disorder as a disability for all legal purposes in Brazil.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42.

Cangea, C., Veličković, P., Jovanović, N., Kipf, T., and Liò, P. (2018). Towards sparse hierarchical graph classifiers. *arXiv preprint arXiv:1811.01287*.

Dai, E., Zhao, T., Zhu, H., Xu, J., Guo, Z., Liu, H., Tang, J., and Wang, S. (2024). A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *Machine Intelligence Research*, 21(6):1011–1061.

Gao, H. and Ji, S. (2019). Graph u-nets. In *international conference on machine learning*, pages 2083–2092. PMLR.

Huang, Z.-A., Zhu, Z., Yau, C. H., and Tan, K. C. (2020). Identifying autism spectrum disorder from resting-state fmri using deep belief network. *IEEE Transactions on neural networks and learning systems*, 32(7):2847–2861.

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

Lin, Y., Yang, J., and Hu, W. (2022). Denoising fmri message on population graph for multi-site disease prediction. In *International Conference on Neural Information Processing*, pages 660–671. Springer.

Luo, X., Wu, J., Yang, J., Xue, S., Beheshti, A., Sheng, Q. Z., McAlpine, D., Sowman, P., Giral, A., and Yu, P. S. (2024). Graph neural networks for brain graph learning: a survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8170–8178.

Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., and Welling, M. (2018). Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, pages 593–607. Springer.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., and Matthews, P. M. (2004). Advances in functional and structural mr image analysis and implementation as fsl. *NeuroImage*, 23:S208–S219.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Wu, T., Ren, H., Li, P., and Leskovec, J. (2020). Graph information bottleneck. *Advances in Neural Information Processing Systems*, 33:20437–20448.

Yan, J., Chen, Y., Xiao, Z., Zhang, S., Jiang, M., Wang, T., Zhang, T., Lv, J., Becker, B., Zhang, R., et al. (2022). Modeling spatio-temporal patterns of holistic functional brain networks via multi-head guided attention graph neural networks (multi-head gagnns). *Medical Image Analysis*, 80:102518.

Yang, Y., Cui, H., and Yang, C. (2023). Ptgb: Pre-train graph neural networks for brain network analysis. *arXiv preprint arXiv:2305.14376*.

Yuan, H., Yu, H., Gui, S., and Ji, S. (2022). Explainability in graph neural networks: A taxonomic survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5782–5799.

Yuan, H., Yu, H., Wang, J., Li, K., and Ji, S. (2021). On explainability of graph neural networks via subgraph explorations. In *International conference on machine learning*, pages 12241–12252. PMLR.

Zhang, S. and Chiang-shan, R. L. (2012). Functional connectivity mapping of the human precuneus by resting state fmri. *Neuroimage*, 59(4):3548–3562.

Zhang, S., Yang, J., Zhang, Y., Zhong, J., Hu, W., Li, C., and Jiang, J. (2023a). The combination of a graph neural network technique and brain imaging to diagnose neurological disorders: A review and outlook. *Brain Sciences*, 13(10):1462.

Zhang, S., Yang, J., Zhang, Y., Zhong, J., Hu, W., Li, C., and Jiang, J. (2023b). The combination of a graph neural network technique and brain imaging to diagnose neurological disorders: A review and outlook. *Brain Sciences*, 13(10):1462.

Zheng, K., Yu, S., Chen, L., Dang, L., and Chen, B. (2024a). Bpi-gnn: Interpretable brain network-based psychiatric diagnosis and subtyping. *NeuroImage*, 292:120594.

Zheng, K., Yu, S., Li, B., Jenssen, R., and Chen, B. (2024b). Brainib: Interpretable brain network-based psychiatric diagnosis with graph information bottleneck. *IEEE Transactions on Neural Networks and Learning Systems*.