Interpretable Graph Attention for Autism Brain Networks

Teo Benarous
teo.benarous@mail.mcgill.ca
McGill University
Montreal, Quebec, Canada

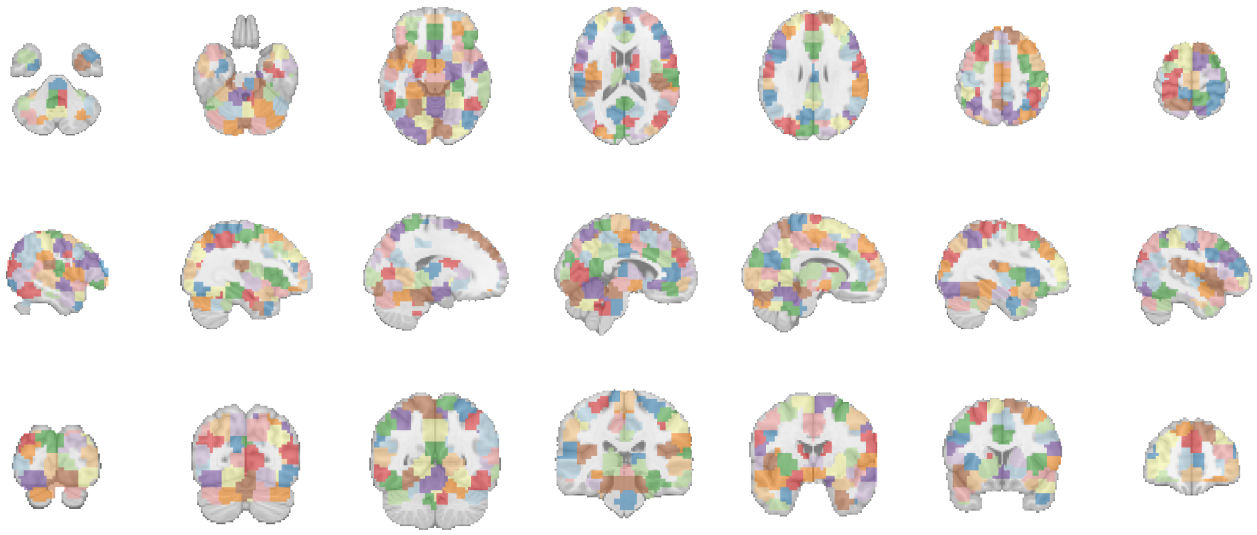


Figure 1: Slices of the Craddock 200 atlas.

Abstract

Autism Spectrum Disorder (ASD) presents heterogeneous manifestations, complicating imaging-based classification. We develop an interpretable Graph Neural Network to classify ASD using resting-state fMRI data from the Autism Brain Imaging Data Exchange. Each participant's brain is segmented into 200 regions of interest (ROIs), and correlations between the corresponding ROIs' time series are used to construct weighted adjacency matrices. The model combines Graph Attention Networks with hierarchical pooling to learn graph-level representations. Interpretability is achieved through GNNExplainer, which identifies the subgraphs and node features that drive model predictions.

Keywords

Graph Neural Networks, Autism Spectrum Disorder, Brain Connectomics, Interpretability

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
COMP 511 Final Projects, Montreal, QC, CA
© 2025 Copyright held by the owner/author(s).

1 Introduction and Motivation

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by difficulties in social communication, restricted interests, and repetitive behaviors. Given ASD's heterogeneous manifestations, identifying robust biomarkers remains a significant challenge in clinical neuroscience. Resting-state functional magnetic resonance imaging (rs-fMRI) [1] offers a window into intrinsic brain activity by measuring spontaneous fluctuations in the blood-oxygen-level-dependent (BOLD) signal without task-driven stimuli. These fluctuations reflect neural dynamics and form the basis for functional connectivity analysis.

Statistical dependencies between regional BOLD time series typically define functional connectivity. By parcellating the brain into discrete regions of interest (ROIs), one can compute mean time series within each region and use these to infer connectivity patterns. These patterns can be represented as graphs, where nodes correspond to ROIs and edges encode functional relationships. In this form, the brain's dynamics become legible to models that operate not just on data but on the structure itself.

We propose an interpretable graph neural network (GNN) framework for ASD classification using rs-fMRI data from the Autism Brain Imaging Data Exchange (ABIDE) [5]. Our approach integrates two complementary measures of functional connectivity: Pearson correlation and partial correlation. While Pearson correlation captures global co-activation patterns, partial correlation isolates direct statistical dependencies by conditioning out the influence of other

regions. We leverage this complementarity by using partial correlations to define the edge connectivity and Pearson correlations to define node features. This design is inspired by recent work in multi-view and multi-graph neuroimaging, which demonstrates that integrating distinct connectivity views enhances discriminative power and interpretability [14].

Incorporating attention mechanisms and hierarchical pooling within the GNN allows us to identify salient substructures in the brain graph that contribute to classification. Post hoc explanation tools, such as GNNExplainer [16], highlight influential nodes and edges, offering neuroscientific insight into model predictions. Through this pipeline, we aim to build an interpretable classifier for ASD based on resting-state brain connectivity.

2 Related Work

Graph Neural Networks for Brain Connectomics. Brain networks naturally lend themselves to graph-based representations, where ROIs map to nodes and functional or structural connections map to edges. Conventional graph-theoretical metrics (e.g., clustering coefficient, modularity) provided early insights into neurological conditions, yet their hand-engineered nature may miss subtle multi-regional interactions [2, 11]. Graph Convolutional Networks (GCNs) [8] introduced a learnable mechanism for node feature aggregation across the network, improving performance on various classification tasks, including neurological disorders [9].

Attention and Pooling in Graph Neural Networks. Uniform neighborhood aggregation in standard GCN layers can dilute critical edges. Graph Attention Networks (GATs) [15] address this by assigning learnable attention coefficients to each edge, refining the influence of highly relevant connections in the embedding process. Hierarchical pooling approaches like SAGPool [10] and DiffPool [17] further reduce graph complexity by aggregating nodes into representatives, intuitively aligning with the multi-scale organization of functional brain networks. While these pooling approaches have been validated on benchmark datasets, questions remain about their utility and optimal configuration for real-world neuroimaging data, such as multi-site cohorts like ABIDE.

Interpretable Graph Neural Networks in Neuroimaging. Deep learning approaches in healthcare demand explanation: clinicians not only need to see whether a model performs well but also why it outputs a particular decision [12]. GNNExplainer [16] uncovers which subgraph and features drive a GNN's prediction, yielding insights that can guide neuroscientific interpretations and help validate model outputs against established knowledge of ASD-related networks.

3 Problem Definition

Each participant $k \in \{1, \ldots, M\}$ is assigned a binary label $y^{(k)} \in \{0, 1\}$ indicating ASD diagnosis (1) or typical development (0). We represent the brain as a graph $\mathcal{G}^{(k)} = (\mathcal{V}, \mathcal{E}^{(k)})$ constructed from preprocessed rs-fMRI data. The brain is parcellated into N regions of interest (ROIs), giving rise to $|\mathcal{V}| = N$ nodes. Let $\mathbf{T}^{(k)} \in \mathbb{R}^{T \times N}$ denote the matrix of mean time series of each ROI for subject k, where T is the number of timepoints.

To define edge connectivity, we compute the partial correlation matrix $\mathbf{\Pi}^{(k)}$ by inverting the empirical covariance matrix $\Sigma^{(k)}$ of the of the ROIs' time series. The partial correlation between ROI i and j is given by:

$$\pi_{ij}^{(k)} = -\frac{\Theta_{ij}^{(k)}}{\sqrt{\Theta_{ii}^{(k)}\Theta_{jj}^{(k)}}} \tag{1}$$

where $\mathbf{\Theta}^{(k)}$ is the precision matrix. To enforce sparsity and reduce noise, we retain only the top 10% of connections based on the absolute partial correlation values to form the adjacency matrix $\mathbf{A}^{(k)} \in [0, 1]^{N \times N}$:

$$\mathbf{A}_{ij}^{(k)} = \begin{cases} \left|\pi_{ij}^{(k)}\right| & \text{if } \left|\pi_{ij}^{(k)}\right| > \tau^{(k)} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\tau^{(k)}$ is the threshold corresponding to the 90-th percentile of the absolute non-zero partial correlations.

Node features are defined by combining multiple types of ROI-level information. We first consider the Pearson correlation coefficients. To enrich each ROI's representation, we include scalar features derived from neural dynamics. These include: (1) the amplitude of low-frequency fluctuations (fALFF) [19], which quantifies the relative power of spontaneous BOLD signal fluctuations in the 0.01–0.1 Hz band and is associated with regional spontaneous activity; (2) regional homogeneity (ReHo) [18], which measures the similarity of a voxel's time series to those of its immediate neighbors and reflects local synchronization; and (3) the mean BOLD signal amplitude, representing the average signal intensity over time within a region, which may capture baseline regional activity levels. Let $a_i^{(k)}$, $r_i^{(k)}$, and $m_i^{(k)}$ denote the fALFF, ReHo, and mean amplitude for ROI i, respectively. The node feature vector is then given by:

$$\mathbf{x}_i^{(k)} = \left(\rho_{i1}^{(k)}, \rho_{i2}^{(k)}, \ldots, \rho_{iN}^{(k)}, a_i^{(k)}, r_i^{(k)}, m_i^{(k)}\right)^\top \in \mathbb{R}^{N+3} \tag{3}$$

This composite feature vector integrates global co-activation profiles with local region-specific properties, enabling the model to learn network topology and regional functional characteristics.

The resulting graph $\mathcal{G}^{(k)}$ with adjacency matrix $\mathbf{A}^{(k)}$ and node features $\mathbf{X}^{(k)} \in \mathbb{R}^{N \times (N+3)}$ is passed through a multi-layer GNN architecture composed of GAT layers and hierarchical pooling modules. A final readout layer produces a graph-level embedding used to predict the diagnostic label $y^{(k)}$. Interpretability is facilitated by GNNExplainer, which identifies relevant subgraphs and node features that drive the model's prediction.

4 Dataset Description

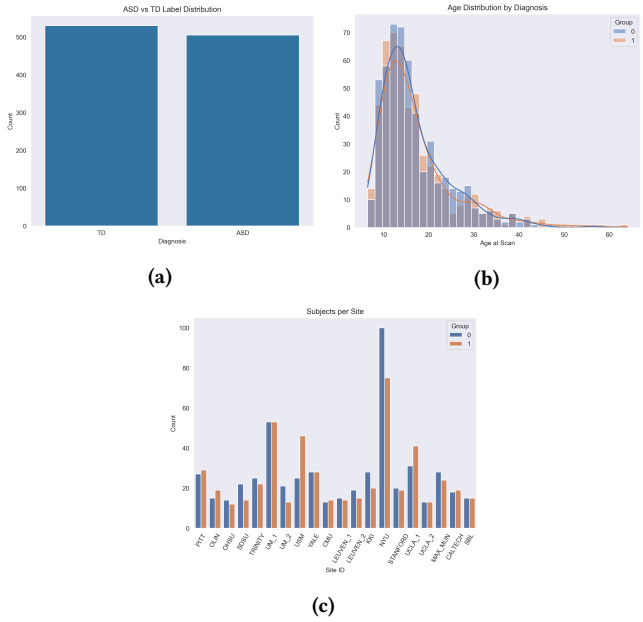
The Autism Brain Imaging Data Exchange (ABIDE) [5] is a large multi-site repository containing over one thousand resting-state fMRI scans from individuals with ASD and matched TD controls. We utilize the ABIDE Preprocessed repository [3] to ensure preprocessing consistency, specifically selecting the CPAC-preprocessed pipeline with bandpass filtering enabled and global signal regression disabled. All fMRI scans are normalized to MNI152 space, a standardized anatomical brain template, and temporally filtered to retain frequencies in the 0.01–0.1 Hz range.

The brain is parcellated using the Craddock 200 (CC200) functional atlas [4], which divides the brain into $N = 200$ spatially coherent ROIs. For each participant, we extract the mean BOLD time series

within each ROI, forming the basis for defining our framework's edge connectivity and node features.

5 Methodology

After restricting the subjects to those who passed quality assessment by all raters, the ABIDE dataset remains slightly unbalanced, with more TD than ASD participants (Figure 2.a). The age distribution (Figure 2.b) is centered around adolescence, with both groups spanning a broad developmental range. The number of subjects per site (Figure 2.c) highlights the multi-site nature of ABIDE and variability in per-site sample sizes.



(a) (b)

(c)

Figure 2: Distribution of patients. (a) Per diagnosis. (b) Per age. (c) Per site.

Figure 3 shows the Craddock 200 atlas overlaid on a brain template, illustrating the spatial distribution of ROIs used in all subsequent analyses. The model pipeline is illustrated in Figure 4.

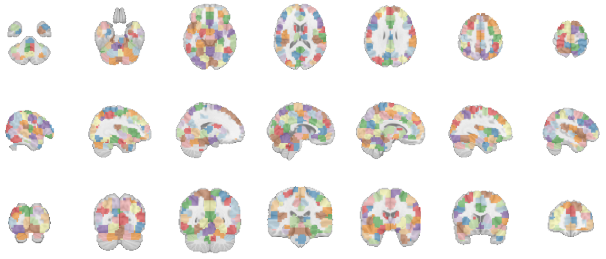


Figure 3: Slices of the Craddock 200 atlas. (Top) Along the x-axis. (Middle) Along the y-axis. (Bottom) Along the z-axis.

We present visualizations from a single subject (Subject 0) to further illustrate our graph construction framework. Node features are

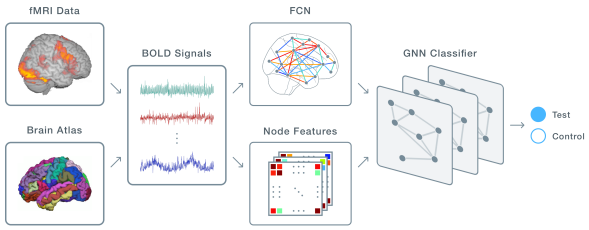
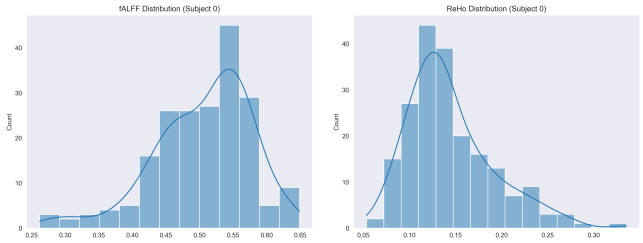


Figure 4: Model pipeline. Adapted from [13].

derived from Pearson correlation coefficients and three ROI-level metrics: fALFF, ReHo, and mean BOLD amplitude. The distributions of fALFF and ReHo across ROIs are shown in Figures 5.a and 5.b, respectively. Figure 6 displays five time series from randomly selected ROIs.



(a) (b)

Figure 5: Distribution of node features for Subject 0. (a) fALFF Distribution. (b) ReHo Distribution.

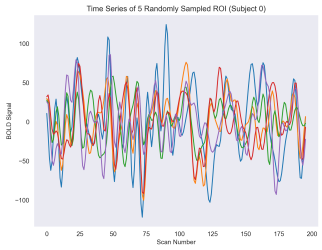


Figure 6: Time series of five randomly sampled ROI for Subject 0.

Functional connectivity is quantified using both Pearson and partial correlation. Figure 7 shows the corresponding matrices and the binarized adjacency matrix formed by thresholding the partial correlations at the 90th percentile. The final subject-specific graph is visualized in Figure 8, with nodes positioned anatomically and edges representing retained partial correlations. This example highlights the structure input graphs used for model training.

6 Experimental Setup

Baseline model. We implement a two-layer Graph Convolutional Network (GCN) using PyTorch Geometric [6]. Each layer has 64

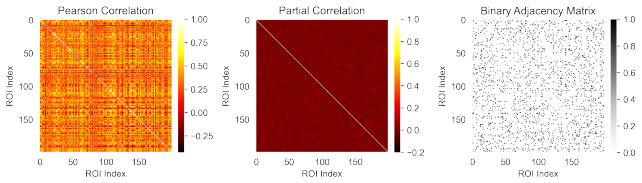
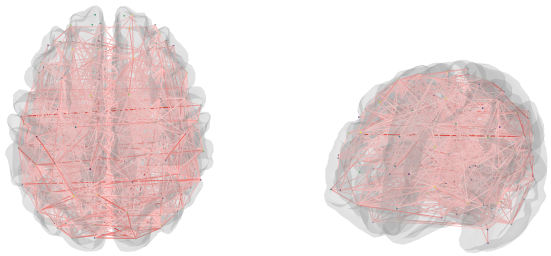


Figure 7: Functional Connectivity Matrices for Subject 0.



(a) (b)

Figure 8: Connectome of Subject 0. Red links indicate positive partial correlations, and blue links negative partial correlations. (a) Top. (b) Left.

hidden neurons with ReLU activations, followed by global mean pooling and a linear layer for binary classification. Dropout regularization, with probability $p = 0.3$, is applied after the first GCN layer. The model is trained using Adam optimizer [7], with a learning rate of 0.001, and binary cross-entropy loss with logits that account for class imbalance through a positive class weight.

Model training and evaluation. Model performance is evaluated using two approaches: (1) an 80/20 random train/test split and (2) leave-one-site-out (LOSO) cross-validation to assess generalization across sites. Performance metrics include Accuracy, Precision, Recall, F1 Score, and ROC-AUC.

7 Results

To assess structural consistency across the cohort, we compute an edge frequency matrix indicating how often each edge appears in the binarized graphs across all subjects. As shown in Figure 9, most edges occur infrequently, with a small subset of pairs of ROIs appearing consistently across individuals.

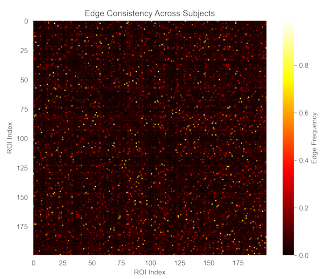
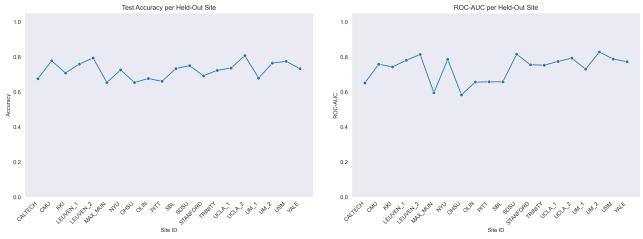


Figure 9: Edge connectivity across subjects.

As a baseline, we implement a two-layer GCN using global mean pooling and dropout regularization. When evaluated with an 80/20 random split repeated over five runs, the model achieves an average accuracy of about 60%. Classification metrics are reported in Table 1. Under a leave-one-site-out (LOSO) evaluation, performance slightly varies across sites (Figure 10).

Table 1: Baseline GCN performance, averaged over 5 runs.

Metric	Value
Accuracy	0.600 ± 0.026
Precision	0.555 ± 0.060
Recall	0.681 ± 0.170
F1 Score	0.596 ± 0.057
ROC-AUC	0.651 ± 0.045

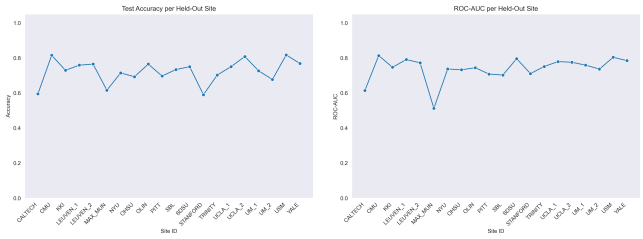


(a) (b)

Figure 10: Performance per Held-Out Site. (a) Accuracy. (b) ROC-AUC.

Table 2: GAT performance, averaged over 5 runs.

Metric	Value
Accuracy	0.678 ± 0.004
Precision	0.710 ± 0.025
Recall	0.676 ± 0.075
F1 Score	0.689 ± 0.024
ROC-AUC	0.754 ± 0.010



(a) (b)

Figure 11: Performance per Held-Out Site. (a) Accuracy. (b) ROC-AUC.

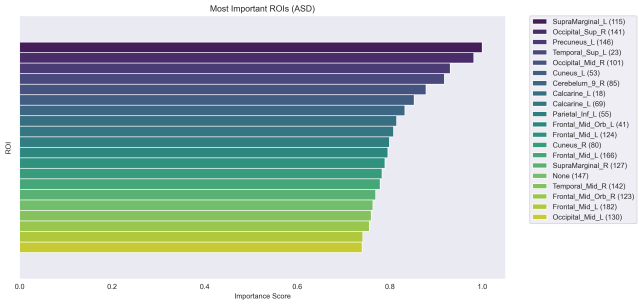


Figure 12: Top ROIs (ASD).

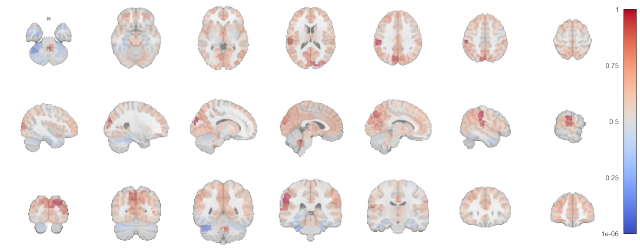


Figure 13: Top ROIs (ASD).

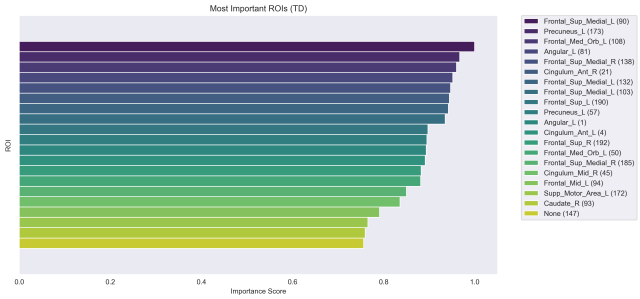


Figure 14: Top ROIs (TD).

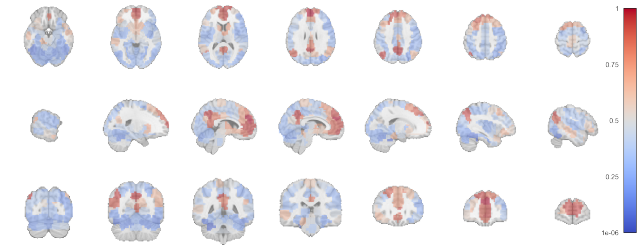


Figure 15: Top ROIs (ASD).

The ROI importance patterns identified in this analysis strongly resonate with established neuroscientific findings on autism spectrum disorder (ASD), lending both biological plausibility and interpretability to the model's outputs. The most predictive regions for ASD—such as the precuneus, supramarginal gyrus, and inferior parietal lobule—are all deeply embedded within networks governing social cognition, self-referential processing, and sensory integration, functions that are consistently disrupted in individuals with autism. In contrast, the top regions distinguishing typically developing individuals, including the medial orbitofrontal cortex, anterior cingulate, and superior medial frontal gyrus, are core hubs of emotional regulation and executive control—capacities often found to be diminished in ASD. This clear neuroanatomical divide between the groups not only mirrors decades of ASD research but also highlights the model's capacity to uncover meaningful, network-level biomarkers, potentially guiding more targeted interventions and enhancing the biological grounding of neuroimaging-based diagnostics.

8 Discussion

9 Conclusion

References

[1] Bharat Biswal, F Zerrin Yetkin, Victor M Haughton, and James S Hyde. 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic resonance in medicine* 34, 4 (1995), 537–541.

[2] Ed Bullmore and Olaf Sporns. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience* 10, 3 (2009), 186–198.

[3] Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li, Michael Milham, et al. 2013. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics* 7, 27 (2013), 5.

[4] R Cameron Craddock, G Andrew James, Paul E Holtzheimer III, Xiaoping P Hu, and Helen S Mayberg. 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human brain mapping* 33, 8 (2012), 1914–1928.

[5] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry* 19, 6 (2014), 659–667.

[6] Matthias Fey and Jan Eric Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. *arXiv preprint arXiv:1903.02428* (2019).

[7] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[8] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[9] Sofia Ira Ktena, Sarah Parisot, Enzo Ferrante, Martin Rajchl, Matthew Lee, Ben Glocker, and Daniel Rueckert. 2017. Distance metric learning using graph convolutional networks: Application to functional brain networks. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. Springer, 469–477.

[10] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. 2019. Self-attention graph pooling. In *International conference on machine learning*. pmlr, 3734–3743.

[11] Mikail Rubinov and Olaf Sporns. 2010. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 3 (2010), 1059–1069.

[12] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*. Vol. 11700. Springer Nature.

[13] Jonathan Shobrook and Paul C. Bogdan. 2020. DeepFCN. https://github.com/shobrook/DeepFCN. Created as part of research in the Dolcos Lab at the Beckman Institute for Advanced Science and Technology.

[14] Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. 2011. Network modelling methods for FMRI. *Neuroimage* 54, 2 (2011), 875–891.

[15] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[16] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems* 32 (2019).

[17] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. 2018. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems* 31 (2018).

[18] Yufeng Zang, Tianzi Jiang, Yingli Lu, Yong He, and Lixia Tian. 2004. Regional homogeneity approach to fMRI data analysis. *Neuroimage* 22, 1 (2004), 394–400.

[19] Qi-Hong Zou, Chao-Zhe Zhu, Yihong Yang, Xi-Nian Zuo, Xiang-Yu Long, Qing-Jiu Cao, Yu-Feng Wang, and Yu-Feng Zang. 2008. An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF. *Journal of neuroscience methods* 172, 1 (2008), 137–141.