

Interpretable Graph Attention for Autism Brain Networks

Teo Benarous

teo.benarous@mail.mcgill.ca

McGill University

Montreal, Quebec, Canada

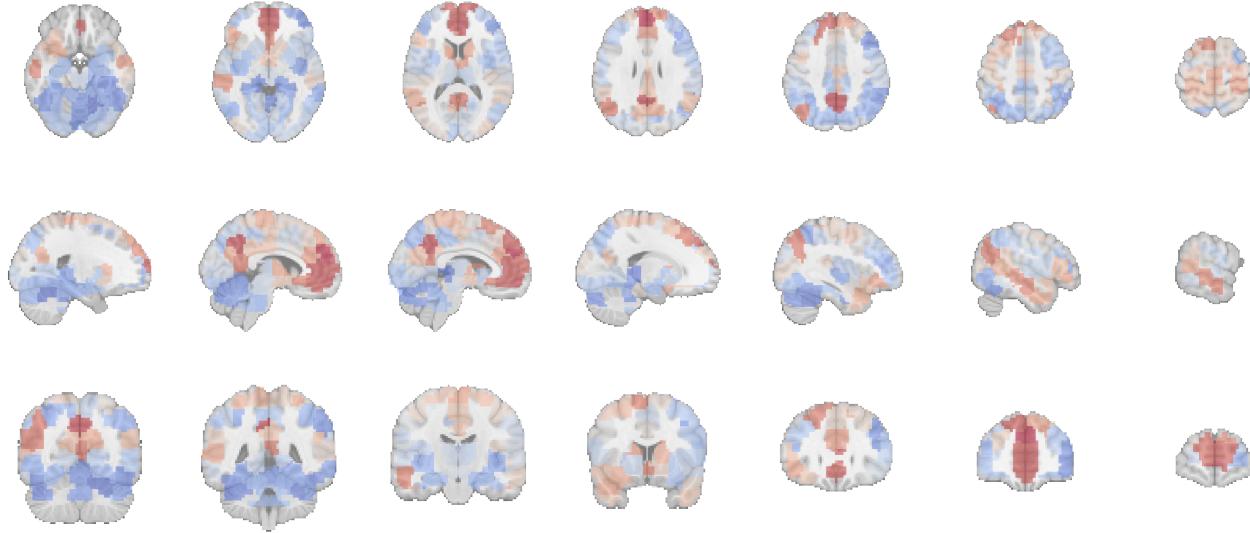


Figure 1: Influential ROIs in ASD Classification.

Abstract

Autism Spectrum Disorder (ASD) presents heterogeneous manifestations, complicating imaging-based classification. This study presents a case study on ASD classification using graph neural networks (GNNs) applied to rs-fMRI data from the Autism Brain Imaging Data Exchange. Brain networks were constructed using partial correlations for edge connectivity, with Pearson correlations and region-specific neural dynamics for node features. A graph attention model with hierarchical pooling was employed, reflecting the brain's multi-scale functional architecture. While classification accuracy aligns with prior GNN-based approaches, the primary focus of this work is interpretability. Using GNNExplainer, we identify influential regions of interest consistent with the ASD literature.

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.

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 the signal properties but on the topology of inter-regional interactions.

We propose an interpretable graph neural network (GNN) framework for ASD classification using rs-fMRI data from the Autism Brain Imaging Data Exchange (ABIDE) [6]. 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 [17].

Incorporating attention mechanisms and hierarchical pooling within the GNN is motivated by the multi-scale structure of the brain, which spans local circuits, intermediate subsystems, and large-scale functional networks. Attention enables the model to assign dynamic importance to inter-regional connections, while hierarchical pooling reflects the brain's nested structure by learning coarser graph representations that preserve key topological features. GNNExplainer [19], a post-hoc GNN-based interpretability method, highlights influential subgraphs and node features, 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, although their emphasis on local or modular structure may overlook higher-order dependencies or distributed patterns across non-contiguous brain regions [3, 14]. For instance, subtle but coordinated disruptions spanning distant areas may not be captured by any single metric yet could emerge through models that learn from the global graph structure. Graph Convolutional Networks (GCNs) [9] introduced a learnable mechanism for node feature aggregation across the network, improving performance on various classification tasks, including neurological disorders [10].

Attention and Pooling in Graph Neural Networks. Uniform neighborhood aggregation in standard GCN layers can dilute critical edges. Graph Attention Networks (GATs) [18] 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 [11] and DiffPool [20] 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 [15]. GNNExplainer [19] 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 $k \in \{1, \dots, M\}$ 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. 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 fMRI volumes, each corresponding to a brain-wide signal.

To define edge connectivity, we compute the partial correlation matrix by inverting the empirical covariance matrix of the ROIs' time series. The partial correlation between ROI i and j is given by:

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

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

$$A_{ij}^{(k)} = \begin{cases} |\pi_{ij}^{(k)}| & \text{if } |\pi_{ij}^{(k)}| > \tau^{(k)} \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where $\tau^{(k)}$ is the threshold corresponding to the 90th percentile of the absolute non-zero partial correlations. This approach emphasizes the magnitude of interaction between regions while discarding the direction under the assumption that both excitatory and inhibitory relationships may reflect meaningful coordination. However, we note that the treatment of negative correlations, particularly whether to include, exclude, or interpret them separately, remains an area of active debate in the literature.

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) [22], 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) [21], which measures the similarity of a voxel's time series to those of its immediate neighbors and reflects local synchronization (a voxel represents a cube of brain tissue); 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:

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

Additionally, we include graph-level features that capture phenotypic information that may influence neuroimaging patterns. Specifically, we include four categorical features: acquisition site, sex, eye status at scan, and handedness, and four numerical features: age of the patient, full-scale IQ (FIQ), verbal IQ (VIQ), and performance IQ (PIQ).

The resulting graph is passed through a graph neural network composed of graph attention layers and hierarchical pooling modules. A readout layer produces a graph-level embedding, which is concatenated with the phenotypic features projected onto the latent space, producing an embedding used to predict the diagnostic label.

4 Dataset Description

The Autism Brain Imaging Data Exchange (ABIDE) [6] is a large multi-site repository containing over one thousand rs-fMRI scans from individuals with ASD and matched TD controls. We utilize the ABIDE Preprocessed repository [4] 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 resulting dataset is 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.

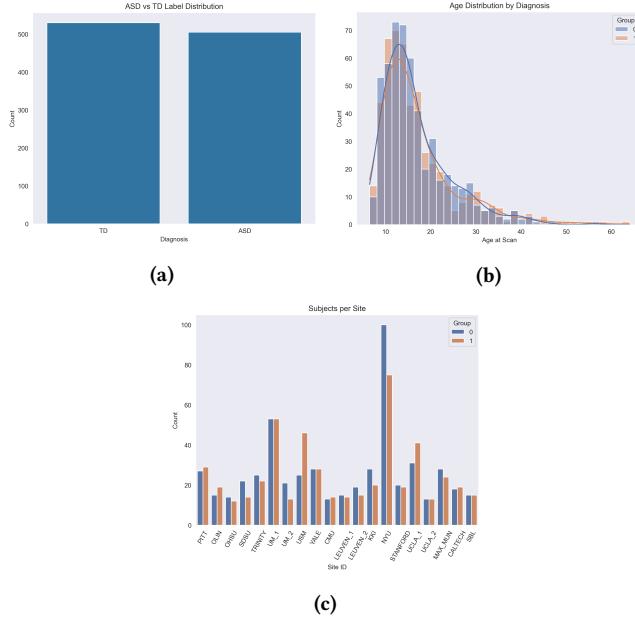


Figure 2: Distribution of patients. (a) Per diagnosis. (b) Per age. (c) Per site.

5 Methodology

The brain is parcellated using the Craddock 200 (CC200) functional atlas [5], which divides the brain into $N = 200$ spatially coherent ROIs. Figure 3 shows the CC200 atlas overlaid on a brain template, illustrating the spatial distribution of ROIs used in all subsequent analyses. The overall model pipeline is illustrated in Figure 4.

We present visualizations from a single subject (Subject 0) to further illustrate our graph construction framework. Figure 5 displays five mean time series from randomly selected ROIs. Node features are 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 7.a and 7.b, respectively. Functional connectivity is quantified using both Pearson and partial correlation. Figure 8 shows the corresponding matrices and the binarized adjacency matrix formed by thresholding the absolute partial correlations at the 90th percentile. The subject-specific connectome is visualized in Figure 9.

The model of interest, which combines attention layers and hierarchical pooling, is illustrated in Figure 6.

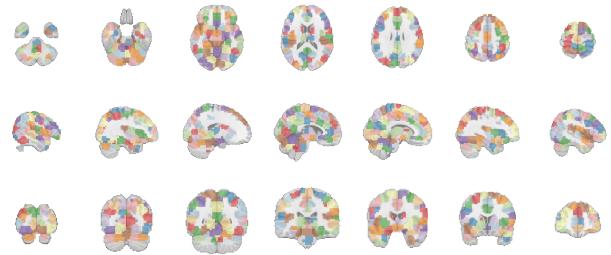


Figure 3: Slices of the Craddock 200 atlas. (Top) Along the x -axis. (Middle) Along the y -axis. (Bottom) Along the z -axis.

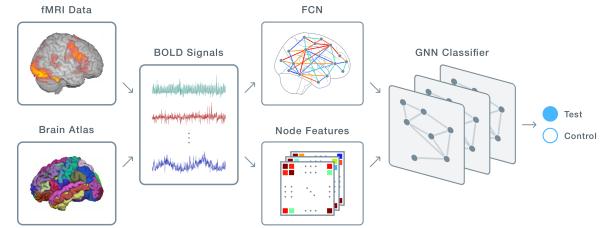


Figure 4: Model pipeline. Adapted from [16].

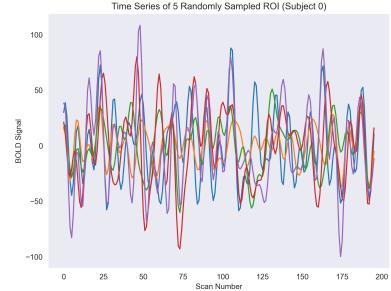
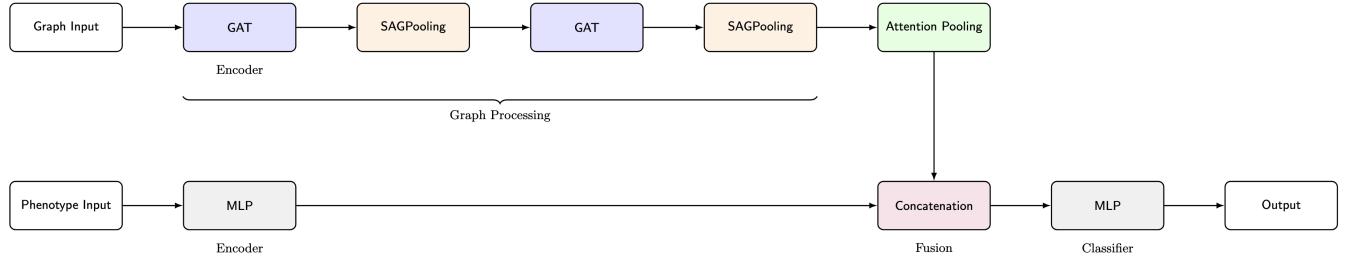
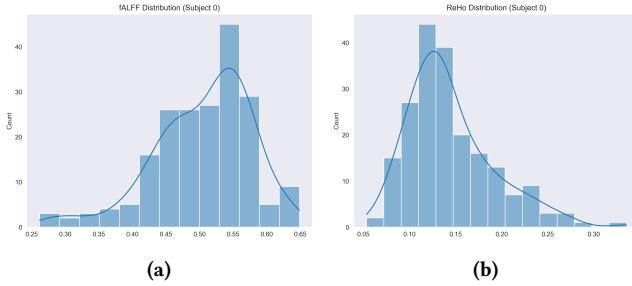
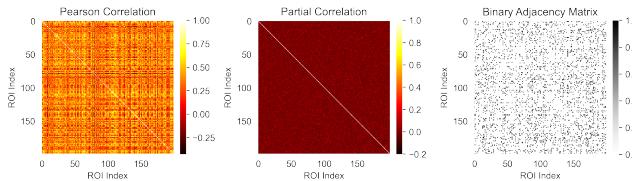
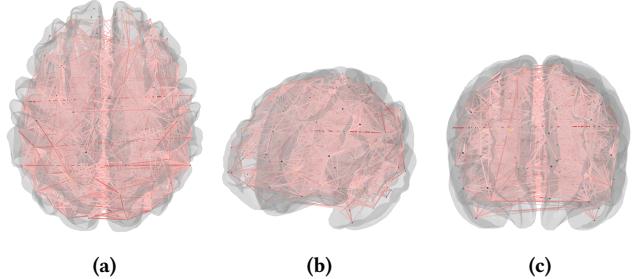


Figure 5: Time series of five randomly sampled ROIs for Subject 0.

**Figure 6: Model Architecture.****Figure 7: Distribution of node features for Subject 0. (a) fALFF Distribution. (b) ReHo Distribution.****Figure 8: Functional Connectivity Matrices for Subject 0.****Figure 9: Connectome of Subject 0. The redness of the links represent partial correlations. (a) Top. (b) Front/Left. (c) Front.**

6 Experimental Setup

Baseline model. We implement a two-layer graph convolutional network (GCN) using PyTorch Geometric [7]. Each layer has 64 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.

Model of interest. We implement a graph neural network using PyTorch Geometric based on GATv2 layers [2]—an improved variant of GAT that replaces its static attention mechanism with dynamic attention, allowing attention scores to be conditioned on both source and target nodes and thereby increasing expressiveness—and incorporate hierarchical pooling. The architecture consists of two graph attention layers. The first employs four attention heads with output features averaged across heads (rather than concatenated), yielding a 64-dimensional node embedding. The second uses a single-head attention layer, maintaining the same dimensionality. Each graph convolution is followed by an ELU activation and dropout regularization with probability $p = 0.3$. To progressively reduce graph resolution while preserving informative substructures, we apply SAGPooling after each convolution, with pooling ratios of 0.8 and 0.6, respectively. To obtain a graph-level representation, we use a learnable attention-based aggregation mechanism: a two-layer multilayer perceptron (MLP) with ReLU activation and dropout regularization computes scalar attention weights over node embeddings, producing a weighted sum of node features.

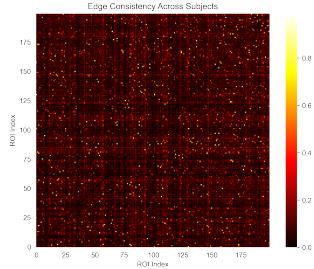
Phenotypic features. For both models, the 8-dimensional phenotypic features are projected into the latent space via a linear transformation followed by a ReLU activation. The graph and phenotype embeddings are concatenated and passed through a final two-layer MLP with dropout regularization for binary classification.

Model training and evaluation. The two models are trained using the AdamW optimizer [12]—an optimization algorithm that modifies the implementation of weight decay in Adam [8] by decoupling weight decay from the gradient update—with a learning rate of 0.001 and a weight decay of 0.0005, and binary cross-entropy loss with logits that account for class imbalance through a positive class weight. 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 graphs across all subjects. As shown in Figure 10, most edges occur infrequently, with a small subset of pairs of ROIs appearing consistently across individuals.

The baseline model, when evaluated under an 80/20 random split repeated over five runs, achieved an average accuracy of about 61%.

**Figure 10: Edge connectivity across subjects.**

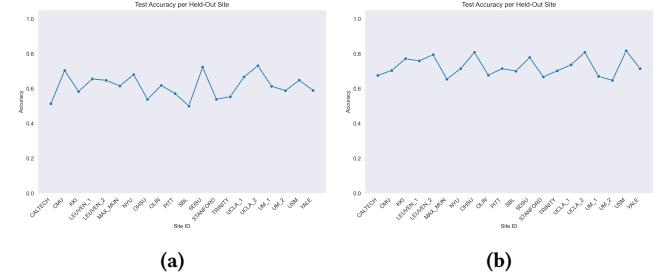
In contrast, the proposed model, incorporating graph attention mechanisms and hierarchical pooling, reached an average accuracy of 68.2%. Classification metrics are reported in Table 1. Under a LOSO evaluation, performance slightly varies across sites (Table 2, Figure 11), e.g., accuracy varies by about 5% across sites for the GAT and 7% for the GCN (standard deviations). Notably, the GAT consistently outperforms the GCN on nearly all sites, often by a substantial margin. For instance, at OHSU, the GAT achieves an accuracy of over 25 percentage points higher than the GCN. This suggests that the GAT generalizes more effectively to out-of-distribution data, while the GCN appears to overfit residual noise linked to site-specific variability despite extensive preprocessing of the time series.

Table 1: Classification metrics, averaged over 5 runs.

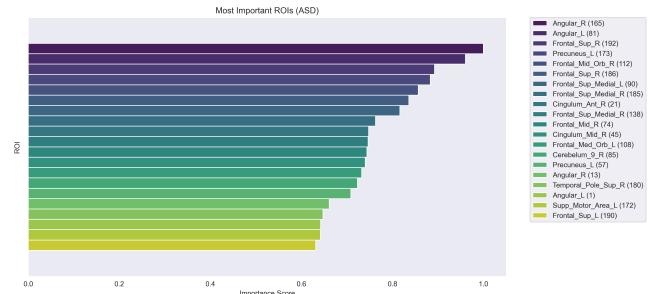
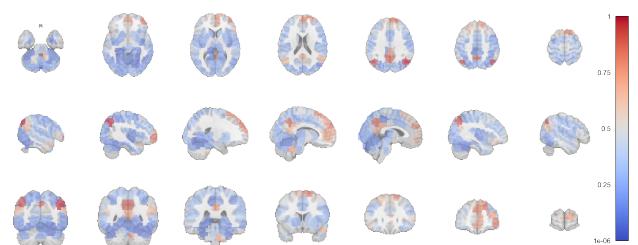
Metric	GCN	GAT
Accuracy	0.610 ± 0.041	0.682 ± 0.020
Precision	0.588 ± 0.041	0.680 ± 0.011
Recall	0.755 ± 0.098	0.680 ± 0.074
F1 Score	0.656 ± 0.033	0.678 ± 0.035
ROC-AUC	0.675 ± 0.0165	0.714 ± 0.022

Table 2: LOSO Classification Metrics.

Site	Accuracy		Precision		Recall		F1 Score		ROC-AUC	
	GCN	GAT	GCN	GAT	GCN	GAT	GCN	GAT	GCN	GAT
CALTECH	0.51	0.68	0.52	0.64	0.84	0.84	0.64	0.73	0.65	0.70
CMU	0.70	0.70	0.71	0.80	0.71	0.57	0.71	0.67	0.74	0.78
KKI	0.58	0.77	0.50	0.76	0.90	0.65	0.64	0.70	0.75	0.71
LEUVEN_1	0.66	0.76	0.75	0.89	0.43	0.57	0.55	0.70	0.72	0.81
LEUVEN_2	0.65	0.79	0.71	0.90	0.33	0.60	0.45	0.72	0.79	0.84
MAX_MUN	0.62	0.65	0.57	0.67	0.71	0.50	0.63	0.57	0.59	0.65
NYU	0.68	0.71	0.60	0.72	0.76	0.55	0.67	0.62	0.72	0.77
OHSU	0.54	0.81	0.50	0.82	0.75	0.75	0.60	0.78	0.64	0.79
OLIN	0.62	0.68	0.62	0.79	0.84	0.58	0.71	0.67	0.60	0.68
PITT	0.57	0.71	0.61	0.78	0.48	0.62	0.54	0.69	0.70	0.75
SBL	0.50	0.70	0.50	0.80	0.87	0.53	0.63	0.64	0.58	0.70
SDSU	0.72	0.78	0.62	0.71	0.71	0.71	0.67	0.71	0.75	0.75
STANFORD	0.54	0.67	0.51	0.62	0.95	0.84	0.67	0.71	0.70	0.71
TRINITY	0.55	0.70	0.52	0.62	0.77	0.91	0.62	0.74	0.65	0.69
UCLA_1	0.67	0.74	0.81	0.78	0.54	0.76	0.65	0.77	0.72	0.75
UCLA_2	0.73	0.81	0.69	0.79	0.85	0.85	0.76	0.81	0.78	0.84
UM_1	0.61	0.67	0.57	0.61	0.87	0.94	0.69	0.74	0.72	0.61
UM_2	0.59	0.65	0.47	0.57	0.69	0.31	0.56	0.40	0.71	0.56
USM	0.65	0.82	0.92	0.85	0.50	0.87	0.65	0.86	0.79	0.88
YALE	0.59	0.71	0.56	0.75	0.79	0.64	0.66	0.69	0.70	0.73

**Figure 11: Accuracy per Held-Out Site. (a) GAT. (b) GCN.**

The regions identified as most influential for distinguishing individuals with ASD from TD controls closely align with functional domains consistently implicated in autism research, validating established findings and reinforcing the plausibility and interpretability of our model. The angular gyrus (bilaterally) and the right superior frontal and middle orbitofrontal cortices emerged as top predictors for ASD. These areas are involved in social cognition, attention control, and reward processing—domains frequently altered in ASD. For TD individuals, the most predictive regions included the left superior medial frontal gyrus, left medial orbitofrontal cortex, and right anterior cingulate cortex—regions critical to emotion regulation, executive control, and adaptive decision-making, functions often altered in ASD. The left precuneus ranked highly for both ASD and TD individuals, underscoring its central role in self-referential processing and default mode network dynamics.

**Figure 12: Top 20 ROIs (ASD).****Figure 13: Influential ROIs (ASD).**

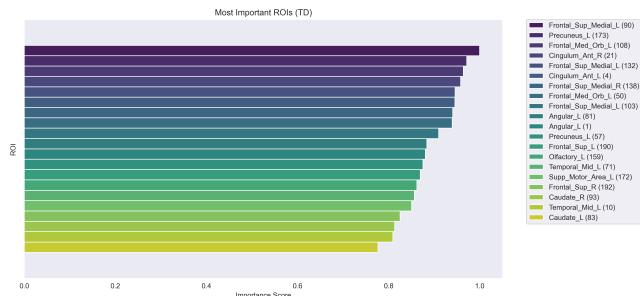


Figure 14: Top 20 ROIs (TD).

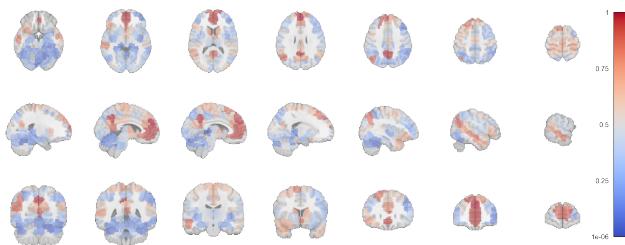


Figure 15: Influential ROIs (TD).

8 Discussion

Our study demonstrates that a GNN incorporating attention mechanisms and hierarchical pooling effectively classifies ASD from rs-fMRI data. The superior performance of the proposed GAT compared to the baseline GCN highlights the value of attention-based architectures combined with hierarchical pooling modules. Specifically, graph attention enables the model to weigh edges dynamically, thereby prioritizing critical connections that differentiate ASD from typical development. Hierarchical pooling further captures the multi-scale nature of brain connectivity.

The interpretability analysis using GNNExplainer provided meaningful insights, identifying several influential ROIs known from the literature to be associated with ASD. For example, the angular gyrus (bilaterally), the right superior frontal cortex, and the middle orbitofrontal cortex were among the top predictors for ASD, aligning well with existing findings implicating these regions in social cognition, attention regulation, and reward processing abnormalities characteristic of ASD.

Despite the interpretability benefits demonstrated, the limitations of the current approach must be acknowledged. The model's classification accuracy, though competitive and aligned with previous GNN-based approaches, remains moderate. Recent work employing dynamic networks or transformer-based architectures has achieved higher accuracy, suggesting that incorporating temporal modeling could further enhance performance [13]. Additionally, the multi-site nature of the ABIDE dataset presents challenges due to variability in scanning protocols and participant demographics. To address these challenges, future research could explore domain adaptation techniques or transfer learning to enhance robustness across heterogeneous datasets.

Moreover, the binary classification framework simplifies the rich clinical spectrum of ASD. Future studies could expand this analysis to multi-class frameworks, capturing finer variations of symptom severity and variability within ASD. Additionally, while GNNExplainer provides valuable insights consistent with the neuroimaging literature, it remains associative; experimental validation through targeted clinical studies is essential to confirm causal relationships between identified neural markers and ASD symptomatology.

Exploring contrastive learning approaches may uncover more discriminative connectivity patterns within brain networks, potentially improving performance and interpretability. Further investigations into multimodal fusion techniques—integrating structural, functional, and phenotypic data—could yield more comprehensive representations.

9 Conclusion

This study demonstrates the potential of graph neural networks (GNNs) for classifying Autism Spectrum Disorder (ASD) from resting-state fMRI data. By leveraging complementary connectivity measures, local neural dynamics, and phenotypic features, we effectively classify ASD using a GNN that employs graph attention layers and hierarchical pooling modules. Leveraging GNNExplainer, we identify meaningful brain regions consistent with findings in the ASD neuroimaging literature. Overall, this work underscores the importance of hierarchical pooling in capturing the brain's multi-scale structure and validates the utility of GNN-based interpretability methods in advancing neuroimaging research.

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] Shaked Brody, Uri Alon, and Eran Yahav. 2021. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491* (2021).
- [3] 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.
- [4] 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.
- [5] 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.
- [6] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Eri 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.
- [7] Matthias Fey and Jan Eric Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. *arXiv preprint arXiv:1903.02428* (2019).
- [8] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [9] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [10] 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.
- [11] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. 2019. Self-attention graph pooling. In *International conference on machine learning*. pmlr, 3734–3743.
- [12] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [13] Zhihao Peng, Zhibin He, Yu Jiang, Pengyu Wang, and Yixuan Yuan. 2024. GBT: Geometric-oriented Brain Transformer for Autism Diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 142–152.
- [14] Mikail Rubinov and Olaf Sporns. 2010. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 3 (2010), 1059–1069.
- [15] 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.
- [16] Jonathan Shabrook and Paul C. Bogdan. 2020. DeepFCN. <https://github.com/shabrook/DeepFCN>. Created as part of research in the Dolcos Lab at the Beckman Institute for Advanced Science and Technology.
- [17] 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.
- [18] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
- [19] 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).
- [20] 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).
- [21] 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.
- [22] 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.