

# BrainNNExplainer: An Interpretable Graph Neural Network Framework for Brain Network based Disease Analysis

Hejie Cui<sup>1</sup> Wei Dai<sup>1</sup> Yanqiao Zhu<sup>2,3</sup> Xiaoxiao Li<sup>4</sup> Lifang He<sup>5</sup> Carl Yang<sup>1†</sup>

## Abstract

Interpretable brain network models for disease prediction are of great value for the advancement of neuroscience. GNNs are promising to model complicated network data, but they are prone to overfitting and suffer from poor interpretability, which prevents their usage in decision-critical scenarios like healthcare. To bridge this gap, we propose BrainNNExplainer, an interpretable GNN framework for brain network analysis. It is mainly composed of two jointly learned modules: a backbone prediction model that is specifically designed for brain networks and an explanation generator that highlights disease-specific prominent brain network connections. Extensive experimental results with visualizations on two challenging disease prediction datasets demonstrate the unique interpretability and outstanding performance of BrainNNExplainer.

## 1. Introduction

Brain networks are complex graphs with anatomic regions represented as nodes and connectivities between the brain regions as links (Murugesan et al., 2020). Interpretable models on brain networks for disease prediction play an important role in understanding the biological functions of neural systems, which can be helpful in the early diagnosis of neurological disorders and facilitate neuroscience research (Martensson et al., 2018). Previous models on brain networks have been studied from shallow to deep ones, such as graph kernels (Jie et al., 2016), tensor factorizations (He

et al., 2018; Liu et al., 2018a) and convolutional neural networks (Kawahara et al., 2017; Li et al., 2020).

Recently, Graph Neural Networks (GNNs) attract broad interests due to their established power in different downstream tasks (Kipf & Welling, 2017b; Xu et al., 2019; Velickovic et al., 2018; Yang et al., 2020a). Compared with shallow models, GNNs are promising for brain network analysis with more powerful representation abilities to capture the sophisticated brain network structures (Maron et al., 2018; Yang et al., 2019; 2020b).

However, GNNs as a family of deep learning models are prone to overfitting and lack transparency in their predictions, which prevent their usage in decision-critical applications such as disease diagnosis. Although several approaches have been proposed to explain the predictions of GNNs (Ying et al., 2019; Luo et al., 2020; Vu & Thai, 2020; Yuan et al., 2020), none of them is equipped with a backbone GNN specifically designed for brain networks. Moreover, they do not target at disease prediction, and produce an independent explanation for each instance, whereas for brain networks, we assume that subjects having the same disease may share similar brain network patterns, which means globally shared explanations are needed across instances.

To unleash the power of GNNs in brain network analysis and enable their interpretability, we propose BrainNNExplainer. It is composed of two modules: a backbone BrainNN (Zhu et al., 2021a) which adapts a message-passing GNN for disease prediction on brain networks (Section 2.2), and an explanation generator which learns a globally shared edge mask to highlight the brain network connections that are important for specific diseases (Section 2.3). In order to improve the prediction model and its interpretability, we further propose a three-step training strategy, where the two modules are trained on the original graph or the masked graph iteratively (Section 2.3).

Through experiments on two real-world brain disease datasets (i.e. HIV and Bipolar), we show that BrainNNExplainer can provide explanations that are verifiable based on neuroscience findings. Furthermore, both our backbone model BrainNN and the interpretable version BrainNNExplainer, especially the latter, yield significant improvements

<sup>1</sup>Department of Computer Science, Emory University <sup>2</sup>Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences <sup>3</sup>School of Artificial Intelligence, University of Chinese Academy of Sciences <sup>4</sup>Department of Computer Science, Princeton University <sup>5</sup>Department of Computer Science and Engineering, Lehigh University. Correspondence to: Carl Yang <j.carlyang@emory.edu>.

over the state-of-the-art shallow and deep baselines.

## 2. BrainNNExplainer

### 2.1. Preliminaries

**Problem definition.** Given a weighted brain network  $G = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ , where  $\mathcal{V} = \{v_i\}_{i=1}^n$  is the node set of size  $n$  defined by the regions of interest (ROIs) (same across subjects),  $\mathcal{E} = \mathcal{V} \times \mathcal{V}$  is the edge set, and  $\mathbf{W} \in \mathbb{R}^{n \times n}$  is the weighted adjacency matrix describing connection strengths between ROIs, the model outputs a disease prediction  $y$ . We provide interpretability by learning an edge mask  $\mathbf{M} \in \mathbb{R}^{n \times n}$  that is shared across all subjects to highlight the disease-specific prominent ROI connections.

**Neural system mapping.** One unique property of brain networks is that the ROIs can be partitioned into neural systems according to their structural and functional roles under a specific atlas (Figley et al., 2017; Shirer et al., 2012; Xu et al., 2020), which can facilitate the verification of our generated explanations from the neuroscience perspective. The HIV and BP datasets we use in this paper are based on two different atlases, AAL90 and Brodmann82, respectively. We map the nodes (i.e., ROIs) on both atlases into eight commonly used neural systems, including Visual Network (VN), Auditory Network (AN), Bilateral Limbic Network (BLN), Default Mode Network (DMN), Somato-Motor Network (SMN), Subcortical Network (SN), Memory Network (MN) and Cognitive Control Network (CCN).

### 2.2. The Backbone BrainNN

**Node features construction.** The lack of predictive original ROI features limits the power of GNNs (Cui et al., 2021). To this end, we construct multiple node features based on one-hot ROI identities as well as local statistical measures such as degree profiles (LDP) (Cai & Wang, 2018). In LDP, each feature  $\mathbf{x}_i$  of node  $v_i$  is computed as

$$\mathbf{x}_i = [\deg(v_i); \min(\mathcal{D}_i); \max(\mathcal{D}_i); \text{mean}(\mathcal{D}_i); \text{std}(\mathcal{D}_i)], \quad (1)$$

where  $\mathcal{D}_i = \{\deg(v_j) \mid (v_i, v_j) \in \mathcal{E}\}$  describes the degree statistics of node  $v_i$ 's one-hop neighborhood, and  $[\cdot; \cdot]$  denotes concatenation. Other common artificial node features such as degree, binning degree (Cui et al., 2021) and node2vec (Grover & Leskovec, 2016) are also included as alternatives in our experiments. All of them are consistent across all subjects.

**Edge-weight-aware message passing.** Since the brain region connectivity and correlations are encoded in real-valued edge weights, which can not be handled by existing GNNs, we design an edge-weight-aware message passing mechanism. Specifically, we first construct a message vector  $\mathbf{m}_{ij} \in \mathbb{R}^D$  by concatenating node embeddings of a node  $i$ ,

its neighbor  $j$ , and edge weight  $w_{ij}$ :

$$\mathbf{m}_{ij}^{(l)} = \text{MLP}_{\Theta} \left( \left[ \mathbf{h}_i^{(l)}; \mathbf{h}_j^{(l)}; w_{ij} \right] \right),$$

where  $l$  is the index of the GNN layer. Then we aggregate messages from all neighbors followed by a non-linear transformation (Xie et al., 2020); the propagation rule can be written as:

$$\mathbf{h}_i^{(l)} = \sigma \left( \sum_{j \in \mathcal{N}_i \cup \{i\}} \mathbf{m}_{ij}^{(l-1)} \right),$$

where  $\sigma$  is a non-linear activation function such as ReLU. Finally, another MLP with residual connections is employed (He et al., 2016) for summarizing all node embeddings to compute graph-level embeddings  $\mathbf{z} \in \mathbb{R}^D$ :

$$\mathbf{z}' = \sum_{i \in \mathcal{V}} \mathbf{h}_i^{(k)}, \quad \mathbf{z} = \text{MLP}(\mathbf{z}') + \mathbf{z}'.$$

This GNN can be trained w.r.t. the supervised cross-entropy loss (denoted as  $\mathcal{L}_p$ ) towards disease predictions.

### 2.3. The Explanation Generator

**Shared edge mask as the explanation.** A general approach to generate explanations for GNNs is to find a explanation graph  $G'$  that has the maximum mutual information with the label distribution  $Y$ , where  $G'$  can be a subgraph of  $G$  (Ying et al., 2019) or other alternations of  $G$  (Luo et al., 2020; Schlichtkrull et al., 2020; Yuan et al., 2020). Previous methods usually produce a unique explanation subgraph for each graph subject (e.g. GNNExplainer (Ying et al., 2019), PGExplainer (Luo et al., 2020), and PGM-Explainer (Vu & Thai, 2020)), or through the model-level explanation (e.g. GAT (Veličković et al., 2018)), that cannot drive disease-specific explanation. Considering the unique properties of brain networks (i.e. fixed number and order of nodes under a given atlas) and the characteristics of disease analysis (i.e. subjects with the same disease may share similar brain network connection patterns), a shared explanation graph  $G'$  is feasible in brain networks and can potentially capture more common patterns for disease-specific analysis.

To achieve this, we propose to learn a globally shared edge mask  $\mathbf{M} \in \mathbb{R}^{n \times n}$  and apply it on the individual brain networks across all subjects in a dataset. Specifically, we train  $\mathbf{M}$  by maximizing the mutual information between the BrainNN predictions  $\hat{y}$  on the original graph  $G$  and  $\hat{y}'$  on the masked graph  $G'$ , where  $\mathbf{W}' = \mathbf{W} \odot \sigma(\mathbf{M})$ .  $\odot$  denotes element-wise multiplication, and  $\sigma$  denotes the sigmoid function that maps the mask to  $[0, 1]^{n \times n}$ . Suppose there are

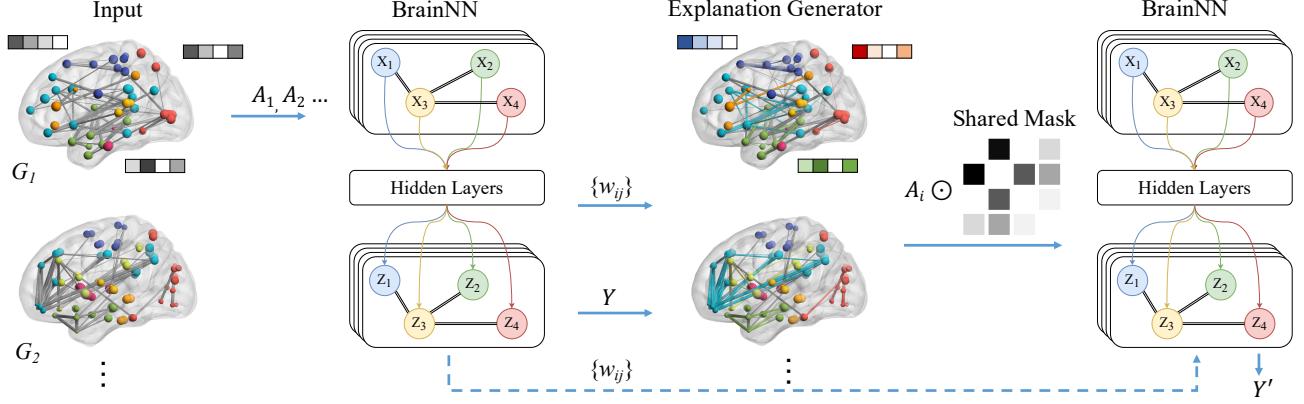


Figure 1: The proposed BrainNNExplainer trained in three-steps: the initial training of BrainNN on the original data, the explanation generation based on trained BrainNN, and the further adjustment of BrainNN based on the explanations.

$C$  classes, mutual information loss can be formulated as:

$$\mathcal{L}_m = - \sum_{c=1}^C \mathbb{1}[y = c] \log P_\Phi(y' = y \mid G = \mathbf{W}').$$

The masked prediction loss  $\mathcal{L}_{p'}$  is the sum of the above mutual information loss  $\mathcal{L}_m$  and the supervised disease prediction loss  $\mathcal{L}_p$  from §2.2,

$$\mathcal{L}_{p'} = \mathcal{L}_m + \mathcal{L}_p.$$

We further apply a sparsity loss  $\mathcal{L}_s$  defined as the sum of all elements of the mask parameters that imposes a regularization on the edge size of  $G'$  to obtain a compact explanation mask, and another element-wise entropy loss  $\mathcal{L}_e$  defined as

$$\mathcal{L}_e = -(M \log(M) + (1 - M) \log(1 - M))$$

from (Ying et al., 2019) to encourage discreteness in mask weight values.

Our final training objective is

$$\mathcal{L} = \mathcal{L}_{p'} + \mathcal{L}_s + \mathcal{L}_e.$$

As a result, our explanation generator will produce a globally shared edge mask  $M$  that can highlight disease-specific prominent brain network connections, and can be further applied on all testing graphs for disease-specific neurological biomarkers and salient ROIs investigation.

**Three-step training strategy.** Overall, BrainNNExplainer is trained in three steps, as shown in Figure 1. In particular, a backbone BrainNN model is first trained on the original data, as described in §2.2. Using this trained prediction model and its prediction as the input, the explanation generator then learns a globally-shared edge mask over all training graphs with other parameters from the prediction model frozen,

as described above. Finally, we apply the learned shared global mask  $M$  on the original training graphs  $G$  to get  $G'$ , then use  $G'$  to train BrainNN backbone model again, where the parameters in backbone model will be further updated with the masked graphs. With this three-step strategy, we further improve the prediction model and obtain a shared explanation mask for model interpretation.

### 3. Experiments

#### 3.1. Datasets and Hyper-parameter Setting

We use two real-world datasets from (Ma et al., 2017) to evaluate the effectiveness of our framework. For each dataset, we randomly divide 80% for training, 10% for validation, and the remaining 10% for testing.

**Human Immunodeficiency Virus Infection (HIV)** is collected from functional magnetic resonance imaging (fMRI), including 35 samples from patients (positive) and 35 healthy controls (negative). Each graph contains 90 nodes (ROIs) and the edge weights corresponding to the adjacency matrix are calculated as the correlations between brain regions.

**Bipolar Disorder (BP)** is also collected from fMRI modality, consisting of 52 bipolar subjects and 45 healthy controls with matched age and gender. It stimulates 82 brain regions, according to Freesurfer-generated cortical/subcortical gray matter regions. Functional brain networks are derived using pairwise BOLD signal correlations.

**Hyper-parameter Setting.** The proposed model is implemented using PyTorch (Paszke et al., 2019) and PyTorch Geometric (Fey & Lenssen, 2019). We use Adam optimizer (Kingma & Ba, 2015) with the initial learning rate setting to 0.001 and a weight decay of 0.00001. The backbone model, composed of three layers of multi-layer perceptron and one layer of edge-weight aware message passing (Section 2.2), is trained for 100 epochs with hidden dimension

setting to 64. Experiments are conducted with multiple common artificial node features (Section 2.2) and different train/test/validation split. The average value of five runs under the optimal hyper-parameter setting is reported for presentation. The implementation will be available after the formal publication of this work.

### 3.2. Interpretability Analysis

**Visualization.** To qualitatively examine the effectiveness of the globally shared mask  $M$ , we follow the similar strategy as the post processings in GNNExplainer (Ying et al., 2019), where a threshold is used to obtain a explanation subgraph  $G'_s$  by removing low-weight edges from  $G'$ .

Figure 2 shows the comparison of connectomes for healthy control and patient groups on two datasets, where edges within the same systems are colored according to the color of nodes it links, while edges across systems are colored gray. The size of an edge is decided by its weight in the explanation graph. We have the following observations.

It can be seen that connectome patterns differ within certain neural systems between the healthy control and patient subject, which could provide potential value for clinical diagnosis. For example, in the HIV dataset, the explanation subgraph of patients excludes many interactions within the Default Mode Network system (DMN), which is colored blue. The connections between superior frontal gyrus (nodes 3, 4) and orbital part (nodes 5, 6, 9, 10) are two examples<sup>1</sup>. Also, interactions within the Visual Network (VN, colored red) system of patients are significantly less than that of healthy controls. For example, connections between cuneus (nodes 45, 46) and lingual gyrus (nodes 47, 48), and those between occipital gyrus (nodes 51, 52, 53, 54) and fusiform gyrus (nodes 55, 56) are found to be missing. These patterns are consistent with the findings in (Herting et al., 2015; Flannery et al., 2021) that alterations in within-and between-network DMN and VN connectivity may relate to known cognitive and visual processing difficulties in HIV.

For the BP dataset, we observe that compared with tight interactions within the Bilateral Limbic Network (BLN, colored green) of the healthy control, the connections within BLN of the patient subject are much more sparse, which may signal pathological changes in this neural system. For instance, it is found that the patient has fewer connections between pyriform cortex (nodes 43, 44) and perirhinal cortex in the rhinal sulcus (nodes 55, 56) than healthy controls, and decreased connections between temporopolar area (nodes 61, 62) and retrosubicular area (nodes 81, 82). These results are in line with previous studies (Das et al., 2020; Ferro et al., 2017). It finds that the parietal lobe, one of the major lobes in the brain roughly located at the upper back area in the

<sup>1</sup>See (Chen et al., 2021) for ROI names and respective node indices.

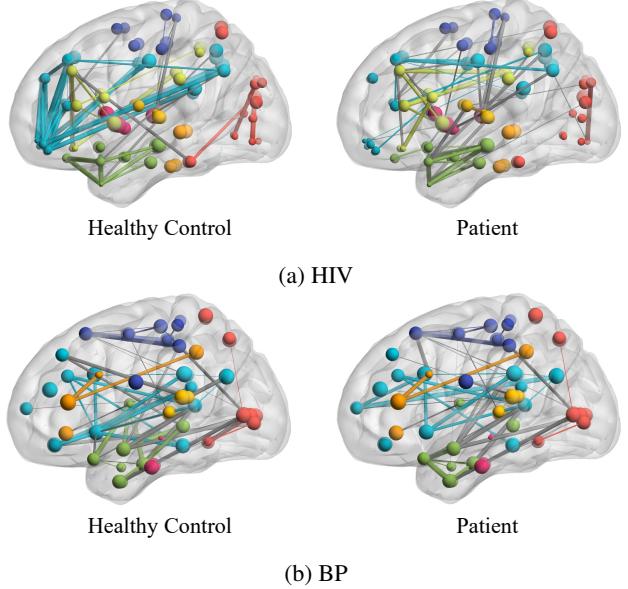


Figure 2: Comparison of explanation graph connectomes in brain networks of a healthy control and a patient on HIV and BP datasets. The colors of neural systems are described as: VN, AN, BLN, DMN, SMN, SN, MN, CCN, respectively.

Table 1: Top ranked neural systems of the explanation subgraph on HIV and BP for both Healthy Control (Normal) and Patient under three comparative measures.

Dataset	Type	Comparative Measures		
		Degree	Strength	Cluster Coefficient
HIV	Normal	DMN BLN CCN	DMN BLN CCN	DMN CCN BLN
	Patient	BLN CCN AN	BLN CCN AN	BLN BLN BLN
BP	Normal	BLN SMN DMN	BLN DMN SMN	SMN VN DMN
	Patient	BLN DMN SMN	BLN DMN SMN	SMN VN

skull and is in charge of processing sensory information it receives from the outside world, is mainly related to Bipolar disease attack. Since parietal lobe ROIs are contained in BLN under our parcellation, the connections missing within the BLN system in our visualization are consistent with existing clinical evidence.

**Interpretation of important brain systems.** To understand which neural systems contribute most to the prediction of a specific disease, we further conduct important brain system interpretation on the explained subgraphs by observing the most manifest nodes with three commonly used measures in brain network analysis: degree, strength, and cluster coefficient (Rubinov & Sporns, 2010). The cluster coefficient of a node in a graph quantifies how close its neighbours are to being a clique (complete graph). Suppose the neighbourhood  $\mathcal{N}_i$  for a node  $v_i$  is its immediately connected neighbours  $\mathcal{N}_i = \{v_j : e_{ij} \in \mathcal{E} \vee e_{ji} \in \mathcal{E}\}$  and  $k_i$  is the number of neighbours of node  $v_i$ , the clustering

coefficient for undirected graphs can be represented as

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in \mathcal{N}_i, e_{jk} \in \mathcal{E}\}|}{k_i(k_i - 1)}.$$

As is shown in Table 1, important neural systems under different metrics show similar characteristics. Specifically, for HIV dataset, both healthy control and patients’ explanation subgraphs reveal the importance of BLN, while DMN is missing from all three metrics in the patient group. This is consistent with our observation on HIV in Figure 2, where the densely connected structure within DMN system degenerated in patient subjects. Regarding BP dataset, BLN, SMN, and DMN are prominent in both patient and healthy controls.

Furthermore, we compare the community structure and modularity (Van Wijk et al., 2010) of our explanation graph  $G'$  against the original graph  $G$  by conducting Newman’s spectral community detection (Newman, 2013). The detected community results are compared with the ground truth neural system partition respectively with different clustering evaluation metrics. Results show that the completeness score of our explained graph achieves about 7.21% improvement over the original graph; the Fowlkes-Mallows score improves over 5.10%; the homogeneity score improves 5.82%; the mutual information score improves 5.12% and the v-measure score improves over 6.44%. The consistent improvements of various clustering evaluation metrics validate the effectiveness of our explanation mask: after the element-wise multiplication with our trained globally-shared explanation mask, the community characteristics are further manifest than the original graphs.

### 3.3. Performance Comparison

We compare our proposed models with baselines from both shallow and deep models for performance evaluation.

**Metrics.** The metrics we used in experiments to evaluate performance are Accuracy and Area Under the ROC Curve (AUC), which are both widely used measures in healthcare domain. Larger values indicate better performance.

**Baselines.** For shallow embeddings methods, we experimented M2E (Liu et al., 2018b), MIC (Shao et al., 2015), MPCA (Lu et al., 2008), and MK-SVM (Dyrba et al., 2015), where the output graph-level embeddings are further processed by logistic regression classifiers to make predictions. We also include three state-of-the-art deep models: GAT (Veličković et al., 2019), GCN (Kipf & Welling, 2017a), and DiffPool (Ying et al., 2018). All the performance of baseline methods are reported under their best settings.

**Results and analysis.** The overall results are presented in Table 2, where our proposed backbone model BrainNN and

Table 2: Performance of different models on HIV and BP datasets. Our methods are colored in gray background and the highest performance is highlighted in boldface.

Method	HIV		BP	
	Accuracy	AUC	Accuracy	AUC
M2E	50.61	51.53	57.78	53.63
MIC	55.63	56.61	51.21	50.12
MPCA	67.24	66.92	56.92	56.86
MK-SVM	65.71	68.89	60.12	56.78
GAT	68.58	67.31	61.31	59.93
GCN	70.16	69.94	64.44	64.24
DiffPool	71.42	71.08	62.22	62.54
BrainNN	74.29	71.67	71.11	64.71
BrainNNExplainer	<b>77.14</b>	<b>75.00</b>	<b>75.56</b>	<b>69.88</b>

the prediction with three-step training from BrainNNExplainer (abbreviated as E-BrainNN in the table) are colored gray. Impressively, both the proposed models yield significant and consistent improvements over all SOTA shallow and deep baselines. Compared with traditional shallow models such as MK-SVM, our backbone BrainNN outperforms them by large margins, with up to 11% absolute improvements on BP, which demonstrates the potential of using deep GNNs on brain networks. The rationale of our edge-weight-aware message passing is supported by the superiority of BrainNN compared with other SOTA deep models such as GAT. Based on this backbone, the performance of three-step training BrainNNExplainer with globally shared mask achieves a further increase of about 5% absolute improvements. This outstanding performance of BrainNNExplainer certifies the unique interpretability of our explanation mask and effectiveness of the proposed framework.

## 4. Conclusion

In this work, we propose BrainNNExplainer, an interpretable GNN framework for brain network based disease analysis, which consists of a brain network oriented GNN predictor and a disease analysis oriented explanation generator. Experimental results with visualizations on two challenging disease prediction datasets validate the unique interpretability and the superior performance of our BrainNNExplainer. Under the framework of BrainNNExplainer, many challenges remain to be solved, such as the lack of supervision and the confinement from small scale datasets for effectively training deep GNN and explanation models. In the near future, we plan to conduct more ablation studies to see how much each component contributes to the system and explore pre-training and transfer learning techniques (Zhu et al., 2021b) based on our current pipeline. Our final aim is to build an interpretable brain analysis system eligible to digest data from different resources and modalities.

## Acknowledgements

Lifang He was supported by NSF OMR N00014-18-1-2009 and Lehigh's accelerator grant S00010293.

## References

- Cai, C. and Wang, Y. A Simple Yet Effective Baseline for Non-Attributed Graph Classification. *arXiv.org*, 2018.
- Chen, N., Shi, J., Li, Y., Ji, S., Zou, Y., Yang, L., Yao, Z., and Hu, B. Decreased dynamism of overlapping brain sub-networks in major depressive disorder. *Journal of psychiatric research*, 133:197–204, 2021.
- Cui, H., Lu, Z., Li, P., and Yang, C. On Positional and Structural Node Features for Graph Neural Networks on Non-attributed Graphs. In *KDD Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.
- Das, T. K., Kumar, J., Francis, S., Liddle, P. F., and Palaniyappan, L. Parietal lobe and disorganisation syndrome in schizophrenia and psychotic bipolar disorder: A bimodal connectivity study. *Psychiatry Research: Neuroimaging*, 303:111139, 2020.
- Dyrba, M., Grothe, M., Kirste, T., and Teipel, S. J. Multi-modal Analysis of Functional and Structural Disconnection in Alzheimer’s Disease Using Multiple Kernel SVM. *Hum. Brain Mapp.*, 36:2118–2131, 2015.
- Ferro, A., Bonivento, C., Delvecchio, G., Bellani, M., Perlini, C., Dusi, N., Marinelli, V., Ruggeri, M., Altamura, A. C., Crespo-Facorro, B., et al. Longitudinal investigation of the parietal lobe anatomy in bipolar disorder and its association with general functioning. *Psychiatry Research: Neuroimaging*, 267:22–31, 2017.
- Fey, M. and Lenssen, J. E. Fast graph representation learning with pytorch geometric. *CoRR*, 2019.
- Figley, T. D., Mortazavi Moghadam, B., Bhullar, N., Kornelsen, J., Courtney, S. M., and Figley, C. R. Probabilistic white matter atlases of human auditory, basal ganglia, language, precuneus, sensorimotor, visual and visuospatial networks. *Frontiers in human neuroscience*, 11:306, 2017.
- Flannery, J. S., Riedel, M. C., Salo, T., Poudel, R., Laird, A. R., Gonzalez, R., and Sutherland, M. T. Hiv infection is linked with reduced error-related default mode network suppression and poorer medication management abilities. *medRxiv*, 2021.
- Grover, A. and Leskovec, J. node2vec: Scalable feature learning for networks. In *KDD*, 2016.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.
- He, L., Chen, K., Xu, W., Zhou, J., and Wang, F. Boosted sparse and low-rank tensor regression. In *NeurIPS*, 2018.
- Herting, M. M., Uban, K. A., Williams, P. L., Gautam, P., Huo, Y., Malee, K., Yogeve, R., Csernansky, J., Wang, L., Nichols, S., et al. Default mode connectivity in youth with perinatally acquired hiv. *Medicine*, 94, 2015.
- Jie, B., Liu, M., Jiang, X., and Zhang, D. Sub-network based kernels for brain network classification. In *International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2016.
- 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. *NeuroImage*, 146:1038–1049, 2017.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- Kipf, T. N. and Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*, 2017a.
- Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017b.
- Li, X., Zhou, Y., Gao, S., Dvornek, N., Zhang, M., Zhuang, J., Gu, S., Scheinost, D., Staib, L., Ventola, P., et al. Braingnn: Interpretable brain graph neural network for fmri analysis. *bioRxiv*, 2020.
- Liu, Y., He, L., Cao, B., Yu, P., Ragin, A., and Leow, A. Multi-view multi-graph embedding for brain network clustering analysis. In *AAAI*, 2018a.
- Liu, Y., He, L., Cao, B., Yu, P. S., Ragin, A. B., and Leow, A. D. Multi-View Multi-Graph Embedding for Brain Network Clustering Analysis. In *AAAI*, 2018b.
- Lu, H., Plataniotis, K. N., and Venetsanopoulos, A. N. MPCA: Multilinear Principal Component Analysis of Tensor Objects. *TNN*, 19:18–39, 2008.
- Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., and Zhang, X. Parameterized explainer for graph neural network. In *NeurIPS*, 2020.
- Ma, G., Lu, C.-T., He, L., Yu, P. S., and Ragin, A. B. Multi-view Graph Embedding with Hub Detection for Brain Network Analysis. In *ICDM*, 2017.
- Maron, H., Ben-Hamu, H., Shamir, N., and Lipman, Y. Invariant and equivariant graph networks. In *ICLR*, 2018.
- Martensson, G., Pereira, J. B., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Lovestone, S.,

- Simmons, A., Volpe, G., et al. Stability of graph theoretical measures in structural brain networks in alzheimer's disease. *Scientific reports*, 8:1–15, 2018.
- Murugesan, G. K., Yogananda, C. G. B., Nalawade, S. S., Davenport, E. M., Wagner, B. C., Kim, W. H., and Maldjian, J. A. Brainnet: Inference of brain network topology using machine learning. *Brain Connect.*, 10:422–435, 2020.
- Newman, M. E. Spectral methods for community detection and graph partitioning. *Physical Review E*, 88:042822, 2013.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.
- Rubinov, M. and Sporns, O. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52:1059–1069, 2010.
- Schlichtkrull, M. S., Cao, N. D., and Titov, I. Interpreting graph neural networks for NLP with differentiable edge masking. *CoRR*, abs/2010.00577, 2020.
- Shao, W., He, L., and Yu, P. S. Clustering on Multi-source Incomplete Data via Tensor Modeling and Factorization. In *PAKDD*, 2015.
- Shirer, W. R., Ryali, S., Rykhlevskaia, E., Menon, V., and Greicius, M. D. Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cerebral cortex*, 22:158–165, 2012.
- Van Wijk, B. C., Stam, C. J., and Daffertshofer, A. Comparing brain networks of different size and connectivity density using graph theory. *PloS one*, 5:e13701, 2010.
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph Attention Networks. In *ICLR*, 2018.
- Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *ICLR*, 2018.
- Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. Deep Graph Infomax. In *ICLR*, 2019.
- Vu, M. N. and Thai, M. T. Pgm-explainer: Probabilistic graphical model explanations for graph neural networks. In *NeurIPS*, 2020.
- Xie, Y., Li, S., Yang, C., Wong, R. C., and Han, J. When do gnns work: Understanding and improving neighborhood aggregation. In *IJCAI*, 2020.
- Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *ICLR*, 2019.
- Xu, M., Wang, Z., Zhang, H., Pantazis, D., Wang, H., and Li, Q. A new graph gaussian embedding method for analyzing the effects of cognitive training. *PLoS computational biology*, 16(9):e1008186, 2020.
- Yang, C., Zhuang, P., Shi, W., Luu, A., and Li, P. Conditional structure generation through graph variational generative adversarial nets. In *NeurIPS*, 2019.
- Yang, C., Xiao, Y., Zhang, Y., Sun, Y., and Han, J. Heterogeneous network representation learning: A unified framework with survey and benchmark. *TKDE*, 2020a.
- Yang, C., Zhang, J., and Han, J. Co-embedding network nodes and hierarchical labels with taxonomy based generative adversarial networks. In *ICDM*, 2020b.
- Ying, R., You, J., Morris, C., Ren, X., Hamilton, W. L., and Leskovec, J. Hierarchical Graph Representation Learning with Differentiable Pooling. In *NeurIPS*, 2018.
- Ying, Z., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. Gnnexplainer: Generating explanations for graph neural networks. In *NeurIPS*, 2019.
- Yuan, H., Yu, H., Gui, S., and Ji, S. Explainability in graph neural networks: A taxonomic survey. *CoRR*, abs/2012.15445, 2020.
- 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 *ICML Workshop on Computational Approaches to Mental Health*, 2021a.
- Zhu, Y., Xu, Y., Liu, Q., and Wu, S. An Empirical Study of Graph Contrastive Learning. *arXiv.org*, 2021b.