Neural Networks 148 (2022) 254–265

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Quantifying the reproducibility of graph neural networks using multigraph data representation

Ahmed Nebli[a,b], Mohammed Amine Gharsallaoui[a], Zeynep Gürler[a], Islem Rekik[a,c,1,*],
for the Alzheimer's Disease Neuroimaging Initiative[2]

[a] *BASIRA lab, Faculty of Computer and Informatics Engineering, Istanbul Technical University, Istanbul, Turkey*
[b] *National School for Computer Science, University of Manouba, Tunisia*
[c] *School of Science and Engineering, Computing, University of Dundee, UK*

ARTICLE INFO

Article history:
Received 15 July 2021
Received in revised form 10 January 2022
Accepted 26 January 2022
Available online 3 February 2022

Keywords:
Reproducibility
Graph neural networks
Brain connectivity multigraphs
Brain biomarkers

ABSTRACT

Graph neural networks (GNNs) have witnessed an unprecedented proliferation in tackling several problems in computer vision, computer-aided diagnosis and related fields. While prior studies have focused on boosting the model accuracy, quantifying the reproducibility of the most discriminative features identified by GNNs is still an intact problem that yields concerns about their reliability in clinical applications in particular. Specifically, the reproducibility of biological markers across clinical datasets and distribution shifts across classes (e.g., healthy and disordered brains) is of paramount importance in revealing the underpinning mechanisms of diseases as well as propelling the development of personalized treatment. Motivated by these issues, we propose, for the first time, reproducibility-based GNN selection (RG-Select), a framework for GNN reproducibility assessment via the quantification of the most discriminative features (i.e., biomarkers) shared between different models. To ascertain the soundness of our framework, the reproducibility assessment embraces variations of different factors such as training strategies and data perturbations. Despite these challenges, our framework successfully yielded replicable conclusions across different training strategies and various clinical datasets. Our findings could thus pave the way for the development of biomarker trustworthiness and reliability assessment methods for computer-aided diagnosis and prognosis tasks. RG-Select code is available on GitHub at https://github.com/basiralab/RG-Select.

1. Introduction

The scope of deep learning (DL) application in neuroscience is marking an exponential growth in many directions thanks to its proven efficiency in tackling many problems such as classification (Richards et al., 2019) or regression (Smith, Ganesh, & Liu, 2013). The abundance of non-invasive neuroimaging datasets acquired from different modalities (e.g., structural and functional MRI) and the availability of new computational frameworks are indubitably pushing the boundaries of research towards deepening our understanding of brain connectivity (Bassett & Sporns, 2017). In network neuroscience, in particular, the graph structure is considered as a powerful data representation thanks to its capacity in encoding connections between different brain regions (delEtoile & Adeli, 2017; Farahani, Karwowski, & Lighthall, 2019; He & Evans, 2010; van den Heuvel & Sporns, 2019). In fact, a brain connectome is a map of connections in the brain wiring different anatomical regions of interest (ROIs), providing a comprehensive map of the network structure of the brain. Thus, it helps better understand the anatomically based interactions between different ROIs (Toga, Clark, Thompson, Shattuck, & Van Horn, 2012) in a non-invasive manner. A brain connectome can be modeled as a graph where each node denotes an ROI and an edge connects two ROIs quantifying their interaction. Applying traditional DL frameworks to graphs does not lead to satisfactory results due to their incapacity in exploiting the topological properties of such non-Euclidian data (Bronstein, Bruna, LeCun, Szlam, & Vandergheynst, 2017; Henaff, Bruna, & LeCun, 2015). To mitigate this limitation, graph neural networks (GNNs), an extended family of DL methods dealing with non-Euclidian data,

[*] Corresponding author at: BASIRA lab, Faculty of Computer and Informatics Engineering, Istanbul Technical University, Istanbul, Turkey.
E-mail address: irekik@itu.edu.tr (I. Rekik).

[1] http://basira-lab.com/.

[2] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/howtoapply/ADNIAcknowledgementList.pdf.

https://doi.org/10.1016/j.neunet.2022.01.018
0893-6080/

have been proposed as an alternative to traditional DL algorithms in many fields (Monti et al., 2017; Wang et al., 2019; Zhang, Cui, & Zhu, 2020) including the field of network neuroscience (Bessadok, Mahjoub, & Rekik, 2021; Wang, Sapra, George and Silva, 2021). GNNs have demonstrated a promising potential in capturing the topological features of graphs to perform a given task such as classification or regression (Wu et al., 2020; Xu, Hu, Leskovec, & Jegelka, 2018; Zhou et al., 2020).

Thus far, most DL and GNN classification models applied in network neuroscience have focused on increasing the accuracy in discriminating between two neurological states (e.g., healthy and neurologically disordered) (Alper, Bach, Henry Riche, Isenberg, & Fekete, 2013; Bessadok et al., 2021; Rashid et al., 2016; Shirer, Ryali, Rykhlevskaia, Menon, & Greicius, 2012). Notably, instead of evaluating the efficiency in discriminating between two classes, GNNs can be evaluated in their capacity to reproduce the most reliable set of discriminative ROIs in a given learning task. Specifically, if two models have consensus over the most important features/biomarkers, this shows that those features are reproducible across models. Hence, since different models end up finding the same top discriminative features, this shows that such models are reproducible. Adding to that, we evaluate if such consensus holds up when we change the training and testing data distributions using various cross-validation strategies. The ability of a particular GNN model to consistently reproduce the same top features in concordance with the majority of other models across various cross-validation strategies demonstrates the high reproducibility of such a model. Hence the most reproducible model acts as a central node in the GNN-to-GNN reproducibility matrix (Fig. 2). However, accuracy-based GNN comparison, which focuses only on the end classification results, overlooks the actual biomarker reliability (i.e., the neuroscientific meaning behind the identified biomarkers). Yet, unlike the accuracy-based GNN assessment, under the reproducibility definition, a reproducible biomarker can be reliably investigated in clinical treatments, where patients with the same brain disorder show a higher disease–biomarker overlap (Povero et al., 2020) (e.g., decreased cortical thickness in Alzheimer patients).

Although being subject to confusion with "interpretability/explainability", "reproducibility" has been suggested to investigate the model's ability in reproducing the same most discriminative features (e.g., biomarkers) between two classes across data distribution perturbations (Georges, Mhiri, & Rekik, 2020). While "interpretability" focuses on debunking how different layers and weights contribute to GNN's decision making (i.e., classification, segmentation) (Li et al., 2021), "reproducibility" studies evaluate the ability of a given GNN in producing and *reproducing* consistent findings across multiple data perturbations. Here, we are interested in the latter goal with the aim to study and quantify a given GNN's reproducibility. Specifically, in our case, the predictions made by GNNs are learned by identifying the different brain connectivity alterations between brain regions that mark a particular disorder. To deepen our understanding of brain connectivity, quantifying the reproducibility of GNNs in terms of biomarkers becomes crucial to investigate their reliability more rigorously. In this context, the reproducibility of a model can be looked at as how likely it is *congruent with other models*. Specifically, here we define the reproducibility score of a given GNN model based on the intersection of its most relevant features with feature sets identified by other GNN models. As such, a GNN's reproducibility assessment has to be generalized across various perturbations of the training and testing data distributions.

A few studies have been proposed to tackle the problem of biomarker or feature reproducibility. Jin et al. (2020) worked on reproducibility across datasets collected from different sites to evaluate the generalizability of a given model. Du et al. (2020) investigated the reproducibility of biomarkers across datasets to extract the most reproducible brain alterations responsible for a neurological abnormality. Although they have generated robust conclusions, such methods do not study the reproducibility across brain connectivity multigraph datasets (i.e., graphs with different connectivity measures for the same pair of nodes). Another line of works has focused on reproducibility across models (Georges et al., 2020). This approach reflects more consistency since it considers multiple models at once and takes into consideration datasets containing brain multigraphs (Dhifallah & Rekik, 2020; Lisowska & Rekik, 2017; Mahjoub, Mahjoub, & Rekik, 2018). However, the proposed framework in Georges et al. (2020) focused only on traditional feature selection (FS) methods and cannot be applied directly to GNNs due to their complexity. In fact, extracting the top biomarkers in FS methods is inherently straightforward, unlike other frameworks. Most GNNs include graph embeddings or graph reshaping operations which alter the original dimensions in the input space (Bessadok et al., 2021). To overcome these limitations, in analogy with reproducibility of FS methods where the most discriminative features from the input space are selected, we can look at the weights learned in a deep learning model as an indicator of the discriminativeness for biomarkers (i.e., sample features in general). For this purpose, we extract the weights of a given GNN model in its last layer preserving the original graph dimensions. Consequently, we build a feature map for each GNN characterizing the discriminativeness assigned to the neurological biomarkers, which denote brain ROIs in our case. This choice is also justified by the fact that the last layer can be looked at as a weighted combination of all the previous neurons in a given neural network. Consequently, we analyze the intersection of the different GNN specific feature maps using different strategies with the aim of selecting the *most reproducible GNN* as illustrated in Fig. 1.

Throughout this paper, we use the term "reproducibility" to describe how well a GNN can reproduce the same findings given several perturbation techniques. More importantly, in this study, we propose the concept of reproducibility as a criterion for best GNN model selection. Notably, the feature map of weights respective to biomarkers reflects the importance accorded by a given GNN projected in the input domain. By conceptualizing the reproducibility in feature selection case as the consensus in terms of selected biomarkers across different models, we can extend this approach to the GNN models by regarding the learned weights as an importance factor. We can then pick the top-weighted biomarkers to investigate the overlap across GNNs from different angles. To ensure generalizability, our study implicates variations in multiple factors such as brain connectivity measures, training data distribution perturbation strategies and the number of top biomarkers to be considered for a given dataset of two different neurological states (e.g., healthy *vs.* disordered). Depending on these factors, we target GNN reproducibility by using different techniques with the aim of establishing a generalizable and trustworthy clinical interpretation.

In view of these aims, we propose reproducibility based GNN selection (RG-Select),[3] a novel framework that investigates the reproducibility of GNN classifiers in datasets of brain connectivity multigraphs, where two nodes are connected by multiple edges, each capturing a particular facet of the brain interactivity. Specifically, we aim to rigorously assess our framework with different settings in order to provide generalizable results. In this context, our study incorporates the variations of the following factors: (1) GNNs, (2) brain connectivity measures per dataset, (3) training strategies, (4) number of top biomarkers to be selected, and (5) connectivity measures (e.g., cortical thickness and sulcal depth).

[3] https://github.com/basiralab/RG-Select.

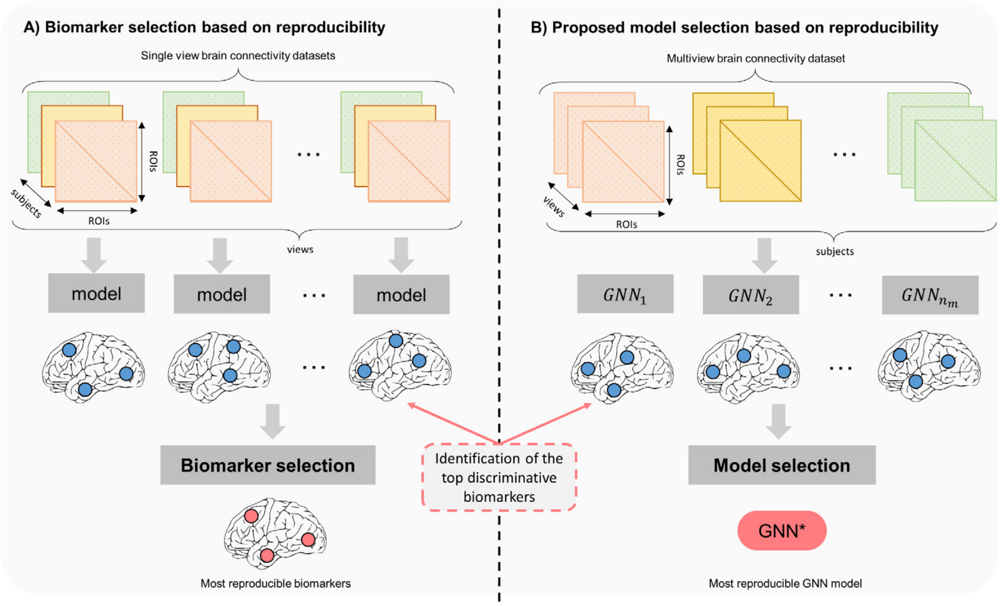
A) Biomarker selection based on reproducibility
Single view brain connectivity datasets
ROIs
subjects
ROIs
views
model
model
model
Biomarker selection
Most reproducible biomarkers
Identification of the top discriminative biomarkers
B) Proposed model selection based on reproducibility
Multiview brain connectivity dataset
ROIs
views
ROIs
subjects
GNN_1
GNN_2
GNN_{n_m}
Model selection
GNN*
Most reproducible GNN model

Fig. 1. Proposed reproducibility-based model selection. **(A)** A model selection based on the ability of the model to reproduce biomarkers using a single view data set. **(B)** Model selection based on the ability of the model to reproduce biomarkers given multi-view dataset.

Taking into account those factors and given a pool of GNN models and a particular dataset of interest, our RG-Select identifies the most reproducible GNN model.

2. Proposed reproducibility based graph neural network selection (RG-Select)

In this section, we present in detail the proposed framework RG-Select for quantifying the reproducibility of GNNs as illustrated in Fig. 2. First, we construct single view datasets by separating the views in each multigraph. We train a set of GNNs on each dataset, separately. Following training, we extract a set of top discriminative biomarkers (i.e., ROIs) based on the ranking of their respective learned weights. In detail, we extract sets of top biomarkers with different sizes for generalizability purpose. Finally, we assign scores to each pair of models based on the inter-model discriminative biomarker overlap. Inter-model reproducibility scores will be used eventually to build the overall reproducibility matrix which incorporates variations of the different factors.

2.1. Problem statement

We denote $\mathcal{D} = (\mathcal{G}, \mathcal{Y})$ as the dataset containing brain connectivity multigraphs with a set of classes respective to different brain neurological states to classify. Let $\mathcal{G} = \{\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_n\}$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$ denote the set of the brain connectivity multigraphs and their labels, respectively. Each connectivity multigraph $\mathbf{G}_i$ is obtained by stacking (i.e., concatenating) a set of n_v views (also referred to as edge types). Each view is a single connectivity matrix representing a distinct cortical measurement (e.g., cortical thickness). We formulate a view as $\mathbf{X}_i^j \in \mathbb{R}^{n_r \times n_r}$, where $j \in \{1, \ldots, n_v\}$ is the view index in the multigraph. As such, a brain connectivity multigraph can be represented as a tensor $\mathbf{X}_i \in \mathbb{R}^{n_r \times n_r \times n_v}$ and a label $y_i \in \{0, 1\}$.

Let $\mathcal{D}^j = (\mathcal{G}^j, \mathcal{Y})$ be the dataset constructed from the jth view. Given a pool of n_m GNNs $\{GNN_1, GNN_2, \ldots GNN_{n_m}\}$, we are interested in training a GNN model $GNN_i : \mathcal{G} \rightarrow \mathcal{Y}$ on the separate single view dataset $\{\mathcal{D}^j\}_{j=1}^{n_v}$. We aim to identify the best GNN that reproduces the same biomarkers differentiating between two brain states against different data perturbation strategies. Thus, we extract the weight vector $\mathbf{w}_i \in \mathbb{R}^{n_r}$ learned by the ith GNN model, where $i \in \{1, 2, \ldots, n_m\}$ in each experiment. For a given dataset, we extract the weights for all the views and GNNs. Next, we rank the biomarkers based on the absolute value of their respective weights. Finally, we compute the reproducibility scores as detailed in what follows.

2.2. Model selection and evaluation

Consistent with previous machine learning practices, we conducted separate model selection and evaluation steps to ensure fairness in the assessment of the models following the protocol detailed in Errica, Podda, Bacciu, and Micheli (2019). To do so, we partition our training set into an inner training set and holdout subset. Next, we train the GNN on the inner training set and validate it on the holdout subset for the model selection. The model selection aims to tune the hyperparameters based on the performance on the validation set. Next, we select the optimal hyperparameters combination that brought the best results on the validation set. We then use the optimal hyperparameters in the model evaluation step where each model is assessed on a separate test set depending on the different k-fold cross-validation (CV). k-fold CV consists of k different training/test splits used to evaluate the performance of the model. In each iteration, the model is tested on a subset of samples never used in the model selection step where the model is trained. We also ensure label stratification in the different data partitions so that the class proportions are preserved across all the training/test/validation splits. This protocol is motivated by the fact that there are some issues about the separation between model *selection* and *assessment* in different state-of-the-art GNN official implementations leading to unfair and biased comparisons (Errica et al., 2019).

2.3. GNN training modes

We used different modes of training for our GNNs to ensure the generalizability of the results. We conducted resourceful training which is based on the conventional k-fold cross-validation protocol. It trains the models on the training set following the same fairness diagram detailed in Errica et al. (2019).

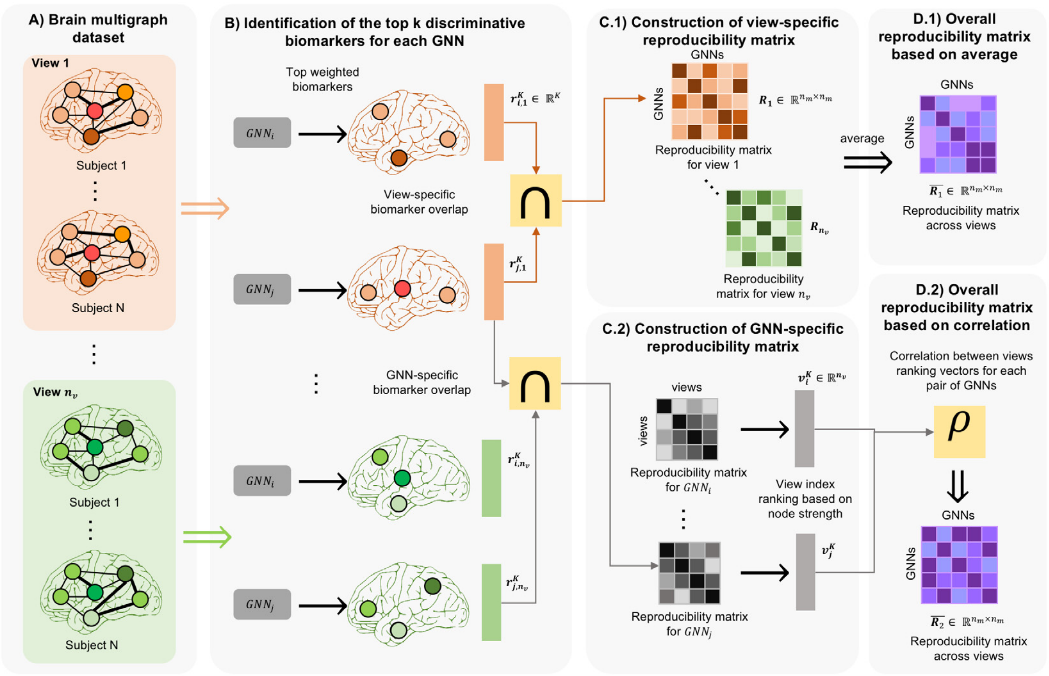
A) Brain multigraph dataset
View 1
Subject 1
Subject N
View n_v
Subject 1
Subject N
B) Identification of the top k discriminative biomarkers for each GNN
Top weighted biomarkers
GNN_i
GNN_j
GNN_i
GNN_j
$r^K_{i,1} \in \mathbb{R}^K$
$r^K_{j,1}$
r^K_{i,n_v}
r^K_{j,n_v}
View-specific biomarker overlap
GNN-specific biomarker overlap
C.1) Construction of view-specific reproducibility matrix
GNNs
GNNs
$R_1 \in \mathbb{R}^{n_m \times n_m}$
Reproducibility matrix for view 1
R_{n_v}
Reproducibility matrix for view n_v
average
D.1) Overall reproducibility matrix based on average
GNNs
GNNs
$\overline{R_1} \in \mathbb{R}^{n_m \times n_m}$
Reproducibility matrix across views
C.2) Construction of GNN-specific reproducibility matrix
views
views
$v^K_i \in \mathbb{R}^{n_v}$
Reproducibility matrix for GNN_i
View index ranking based on node strength
v^K_j
Reproducibility matrix for GNN_j
D.2) Overall reproducibility matrix based on correlation
Correlation between views ranking vectors for each pair of GNNs
ρ
GNNs
GNNs
$\overline{R_2} \in \mathbb{R}^{n_m \times n_m}$
Reproducibility matrix across views

Fig. 2. Illustration of the proposed framework for GNN reproducibility assessment. (A) We start with datasets of single-view graphs. (B) We train different GNN models on the datasets. For each model-dataset combination, we extract the biomarkers absolute value weights vector. Then, we rank the resulting vectors to identify the top discriminative brain regions. (C.1) We calculate the overlap ratio between resulting vectors from different GNNs and the same view. For each view, we build a reproducibility matrix. (C.2) We compute the overlap ratio between resulting vectors from different views and the same GNN. Consequently, we obtain a matrix for each GNN. Next, we extract ranking vectors based on the node strength vectors resulting from these matrices. This step results in a ranking vector for each GNN. (D.1) We compute the average of the resulting matrices to obtain the overall reproducibility matrix. (D.2) We calculate the correlation between pairs of GNN resulting ranking vectors.

Moreover, we used frugal training which follows a few-shot learning approach. This mode performs the model training on only a few samples and the evaluation on all the remaining subjects in the dataset. Using both methods is important to ensure that the results of our framework are agnostic to data perturbations and training strategies.

2.4. Biomarker selection

In contrast with the conventional approach which focuses on accuracy evaluation of a given classifier, here, we focus on model reproducibility in top discriminative features (i.e., biomarkers). Typically, the extraction of the most discriminative biomarkers for FS methods is straightforward. However, GNN methods have different architectures which makes it hard to implement a generalized way to extract the most important biomarkers. To circumvent this issue, we extract the weights of the last layer preserving the dimensionality of the input data (i.e., having the same number of features/ROIs). Specifically, given n_r ROIs for each brain connectome, we rank these biomarkers based on their learned weights by the selected GNN. Based on that ranking, we extract $\mathbf{r}^{K_h}_{i,j} \in \mathbb{R}^{K_h}$ the vector containing the top K_h biomarkers based on the weights learned by the ith GNN trained on the jth view of the input multigraph dataset.

Definition 1. Let $\mathbf{r}^k_{i,v}$, $\mathbf{r}^k_{j,v} \in \mathbb{R}^{n_r}$ denote two the vectors containing the top k biomarkers learned on the same view v by GNN_i and GNN_j, respectively. We denote $r^k_{i,v}$ and $r^k_{j,v}$ as the two sets containing the regions included in $\mathbf{r}^k_{i,v}$, $\mathbf{r}^k_{j,v}$, respectively. We define the view-specific reproducibility on the view v at threshold k between models i and j as: $p_v\,(\mathbf{r}^k_{i,v}, \mathbf{r}^k_{j,v}) = \frac{|r^k_{i,v} \cap r^k_{j,v}|}{k}$

Definition 2. Let $\mathbf{r}^k_{g,i}$, $\mathbf{r}^k_{g,j} \in \mathbb{R}^{n_r}$ denote two the vectors containing the top k biomarkers learned by the same GNN_g on the views i and j, respectively. We denote $r^k_{g,i}$ and $r^k_{g,j}$ as the two sets containing the regions included in $\mathbf{r}^k_{g,i}$, $\mathbf{r}^k_{g,j}$, respectively. We define the GNN-specific reproducibility by GNN_g at threshold k between views i and j as: $p_g\,(\mathbf{r}^k_{g,i}, \mathbf{r}^k_{g,j}) = \frac{|r^k_{g,i} \cap r^k_{g,j}|}{k}$

2.5. View-specific reproducibility matrix

For a pool containing n_m GNNs, we aim to quantify the reproducibility between each pair of models. Since reproducibility reflects the commonalities between two sets of biomarkers, we propose to compute the ratio of the overlapping ROIs. First, we need to quantify the reproducibilities in the same domain (i.e., the same view). In other terms, for a given view v and a threshold K_h we calculate the ratio $p_v(r^{K_h}_{i,v}, r^{K_h}_{j,v})$ for each pair of models GNN_i and GNN_j. Having the reproducibility calculated for each pair of GNNs, we construct the reproducibility matrix $\mathbf{R}^{K_h}_v \in \mathbb{R}^{n_m \times n_m}$ where $\mathbf{R}^{K_h}_v(i,j) = p_v(r^{K_h}_{i,v} r^{K_h}_{j,v})$. Next, we generate the average reproducibility matrix by merging all the reproducibility matrices across the different p thresholds $\mathbf{R}_v(i,j) = \frac{\sum_{h=1}^{n_k} \mathbf{R}^{K_h}_v(i,j)}{n_k}$, where n_k is the number of threshold values. Finally, after calculating the reproducibility locally (i.e., for each view) we need to get a general overview of the reproducibility across all views. Therefore, we average the resulting matrix over all the views and training modes (i.e., perturbation strategy).

2.6. GNN-specific reproducibility matrix

Another way to quantify reproducibility is to start from quantifying the commonalities across views for the same GNN. This is motivated by the fact that GNNs might have varying behaviors (i.e., different learned weight distributions) across different data views. For the same model, we measure the GNN-specific reproducibility between the different views of the dataset (see Section 3.4.2). For a given GNN_g, we construct the matrix $\mathbf{R}_g^{K_h} \in \mathbb{R}^{n_v \times n_v}$ where $\mathbf{R}_g^{K_h}(i,j) = p_g(r_{g,i}^{K_h}, r_{g,j}^{K_h})$. Then, we average over the thresholds, $\mathbf{R}_g(i,j) = \frac{\sum_{h=1}^{n_k} \mathbf{R}_g^{K_h}(i,j)}{n_k}$. Finally, we calculate the average of the GNN-specific reproducibility matrix for each model across all the different training modes.

3. Results

3.1. Evaluation datasets

We evaluated our reproducibility framework on a small-scale and a large-scale brain connectivity datasets. The first dataset (AD/LMCI) contains 77 subjects (41 subjects are diagnosed with Alzheimer's diseases (AD) (average age 70.4 ± 7.5) and 36 diagnosed with Late Mild Cognitive Impairment (LMCI) (average age 74.1±6.7)) from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database GO public dataset) (Weiner et al., 2010). The second dataset (ASD/NC) includes 300 subjects (all under 15 years old) equally partitioned between autism spectral disorder (ASD), and normal control (NC) states extracted from Autism Brain Imaging Data Exchange ABIDE I public dataset (Di Martino et al., 2014).

For both datasets, the connectivities (i.e., views or measurements) were obtained using FreeSurfer (Fischl, 2012) by constructing cortical morphological networks for each subject from structural T1-w MRI (Dhifallah & Rekik, 2020; Mahjoub et al., 2018). Next, both left and right cortical hemispheres (LH and RH) are parcellated into 35 cortical ROIs using Desikan–Killiany cortical atlas, respectively (Desikan et al., 2006; Lisowska & Rekik, 2017). Both AD/LMCI (RH and LH) brain multigraphs are constructed using 4 cortical measures: maximum principal curvature, cortical thickness, sulcal depth and average curvature. As for the ASD/NC dataset, brain multigraphs are generated from six cortical attributes which are the same attributes used for AD/LMCI datasets in addition to cortical surface area and minimum principle area. Specifically, for each node ROI_i and for each cortical attribute, we calculate the average cortical measurement $\overline{a}_i$ across all its vertices. The weight of the connectivity linking ROI_i and ROI_j is the absolute distance between their average cortical attributes: $|\overline{a}_i - \overline{a}_j|$.

3.2. GNN models

For our reproducibility framework, we used 5 state-of-the-art GNN architectures: DiffPool (Ying et al., 2018), GAT (Veličković et al., 2017), GCN (Kipf & Welling, 2016), SAGPool (Lee, Lee, & Kang, 2019) and g-U-Nets (Gao & Ji, 2019). DiffPool performs a differential pooling to generate a learned hierarchical representation of an input graph. At each pooling layer, DiffPool learns how to make soft assignment of nodes into clusters which will be the nodes of the following layer (Ying et al., 2018). GAT and GCN are originally designed for node classification. Here, we adapt them to perform the graph classification task. Therefore, before a final linear layer, we insert a global mean pooling layer projecting the node scores into a global score for the whole graph. GAT learns different weights to the neighborhood to perform the aggregation in the following layer. GCN sequentially learns convolution weights that encode neighborhood features and local graph structure (Kipf & Welling, 2016). SAGPool performs graph convolutions to learn pooling and unpooling of graphs based on self-attention (Lee et al., 2019). g-U-Nets is a U-shape based GNN combining multiple encoders and decoders that perform pooling and unpooling of the graph, respectively (Gao & Ji, 2019).

3.3. Training settings and hyperparameters

We have used two different types of training in our experiments: resourceful and frugal. For the resourceful training, we trained our models in the conventional train/test approach. To do so, we have made 3-fold and 5-fold cross-validation strategies. In addition to the resourceful training approach based on the *k*-fold cross-validation, we also evaluated our experiments with a frugal training approach based on few-shot learning. Here, we only trained the model on 2 samples per class for each dataset. To limit any major intervention of parameters/sample selection with respect to our findings, we run our experiments for 100 times, each with different randomizations. We also used four thresholds for the top biomarkers extraction, which are 5, 10, 15, and 20. All the hyperparameters were selected using grid search. For all models, the learning rates ranged between 0.0001 and 0.001. For DiffPool, the hidden dimension, the output dimension, the assignment ratio and the number of convolution layers were equal to 256, 512, 0.1 and 3, respectively. For GAT, the numbers of hidden units and head attentions were equal to 8. For GCN, the number of hidden units is equal to 64. For g-U-Nets, the number of layers, hidden and convolution layer dimensions were equal to 3, 512 and 48, respectively. For SAGPool, the hidden dimension and the pooling ratio were equal to 256 and 0.5, respectively.

3.4. Overall reproducibility matrices

3.4.1. View-specific matrix based reproducibility

To quantify the reproducibility across GNN models we used 4 different methods. The first method consists of calculating the average between the view-specific reproducibility matrices over all the views of the selected dataset. This method is intuitive in order to combine the information calculated for each view.

3.4.2. GNN-specific matrix based reproducibility

We rank the views based on the GNN-specific reproducibility matrix. For each GNN, we extract a vector indicating the ranks of the views. Next, we calculate the correlation coefficient across pairs of GNNs based on their respective reproducibility matrices. Consequently, we construct a reproducibility matrix containing pairwise relations between GNNs. This method is useful to reflect how the GNN are likely to have the same behavior across views.

3.5. Most reproducible GNN selection

In what follows, we define the reproducibility matrix as the summation of the two matrices detailed above. To take advantage of both GNN-specific and view-specific matrices, we sum the two reproducibility matrices detailed above. As such, we can look at the overall reproducibility matrix as a graph where the nodes represent the GNN models. Consequently, we use the node strength to quantify the reproducibility scores of the GNN models. This is conceptualized based on the intuition that the model reproducibility reflects the consensus in biomarkers with other models. Projecting this idea on the graph topology, node strength is a topological measure that encodes the magnitude of the connections with remaining entities in the graph. We define the most reproducible model as the node having the highest strength in the reproducibility graph.

For each neurological dataset, we trained GNN models using two different training modes: CV and FS. For CV, we average 3-fold, 5-fold and 10-fold results. For FS, we average the 100 different randomizations that we have conducted to select the few training samples. Figs. 3 and 4 illustrate the reproducibility matrices for AD/LMCI RH and LH datasets, respectively. For these datasets, the most reproducible GNNs are DiffPool and SAGPool,

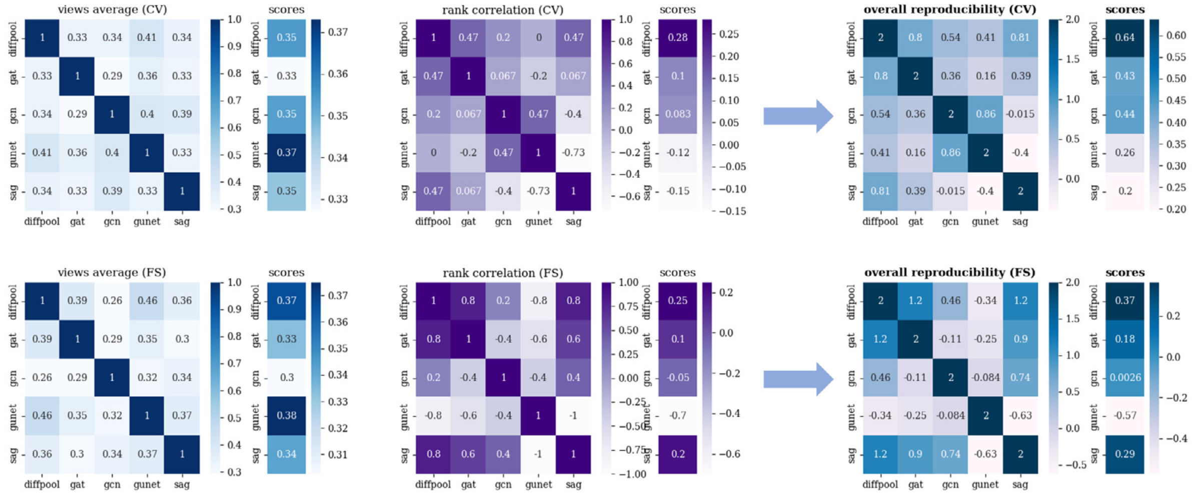

Fig. 3. *Heatmaps of reproducibility matrices of AD/LMCI LH dataset.* The matrices were computed on cross-validation and few-shot training strategies, separately. For each heatmap, we associate a score vector where each value represents the average of its corresponding row (in the heatmap). AD: Alzheimer's disease. LMCI: late mild cognitive impairment. LH: left hemisphere. CV: cross-validation. FS: few-shot.

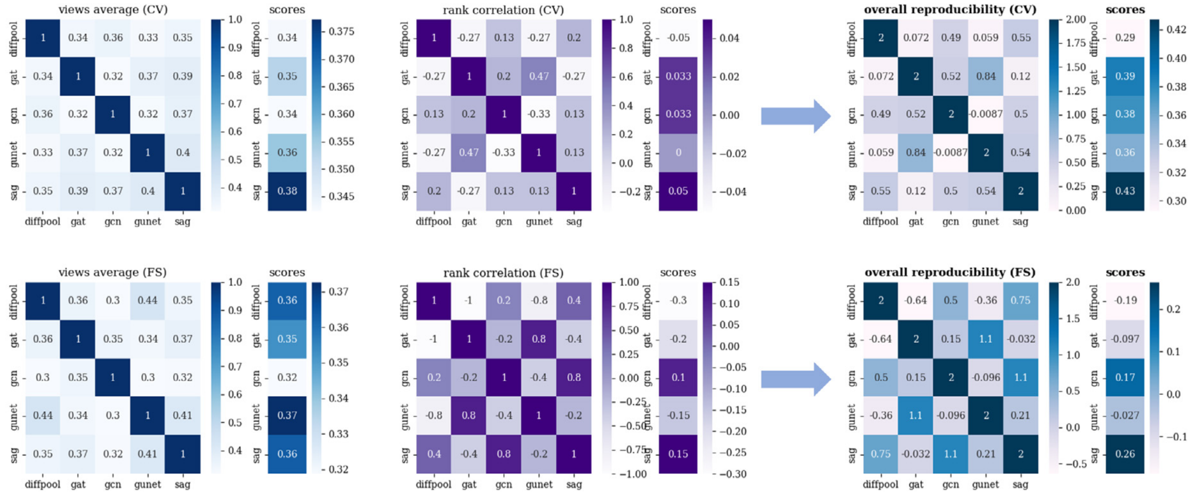

Fig. 4. *Heatmaps of reproducibility matrices of AD/LMCI RH dataset*. The matrices were computed on cross-validation and few-shot training strategies, separately. For each heatmap, we associate a score vector where each value represents the average of its corresponding row (in the heatmap). AD: Alzheimer's disease. LMCI: late mild cognitive impairment. RH: right hemisphere. CV: cross-validation. FS: few-shot.

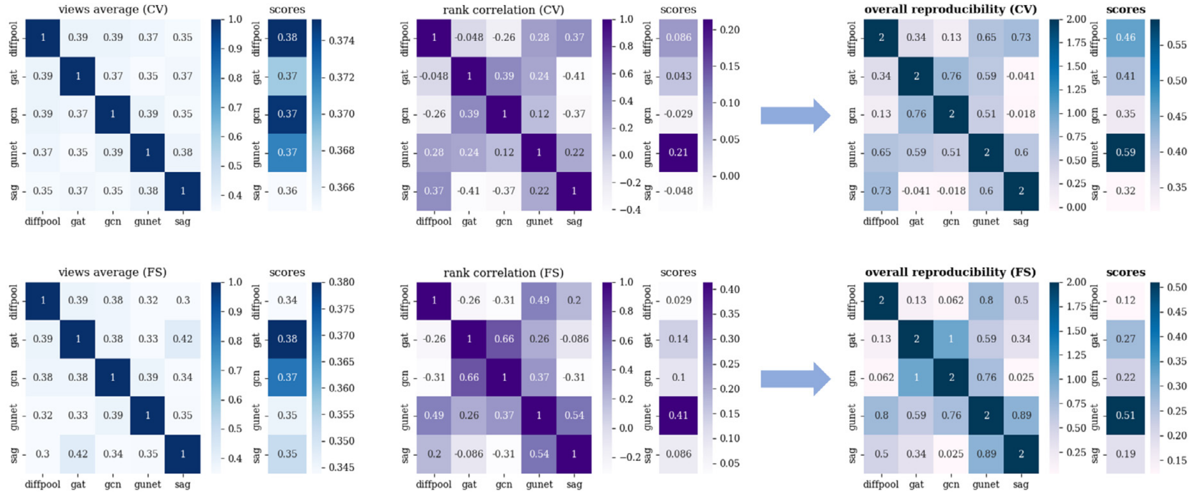

Fig. 5. *Heatmaps of reproducibility matrices of ASD/NC LH dataset.* The matrices were computed on cross-validation and few-shot training strategies, separately. For each heatmap, we associate a score vector where each value represents the average of its corresponding row (in the heatmap). ASD: autism spectrum disorder. NC: normal control. LH: left hemisphere. CV: cross-validation. FS: few-shot.

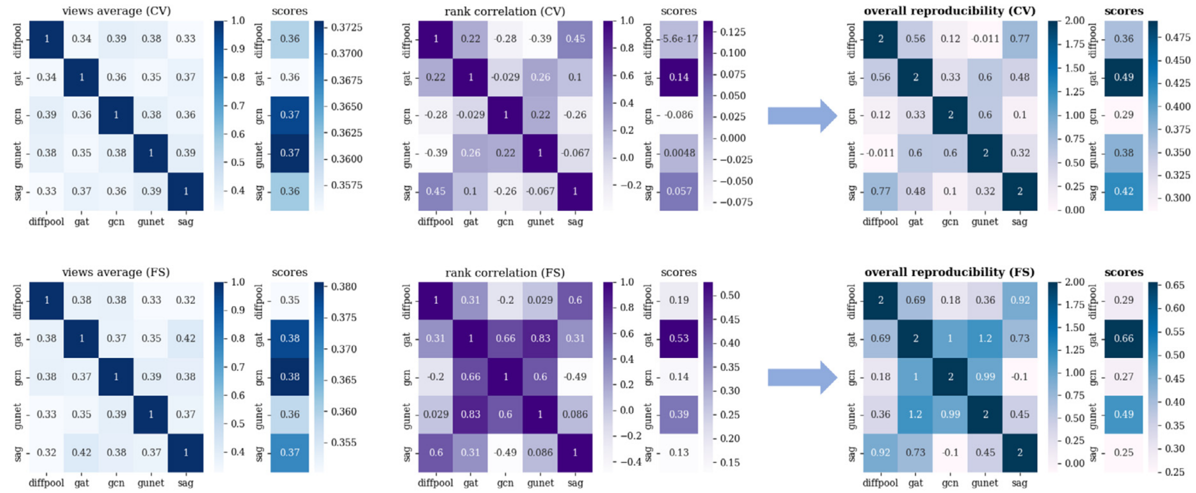

Fig. 6. *Heatmaps of reproducibility matrices of ASD/NC RH dataset.* The matrices were computed on cross-validation and few-shot training strategies, separately. For each heatmap, we associate a score vector where each value represents the average of its corresponding row (in the heatmap). ASD: autism spectrum disorder. NC: normal control. RH: right hemisphere. CV: cross-validation. FS: few-shot.

respectively. We also note that the most reproducible method is the same across training modes. Figs. 6 and 5 illustrate the reproducibility matrices for ASD/NC datasets. Based on the overall matrices, g-U-Nets and GAT are the most reproducible models on LH and RH, respectively. For all datasets, the results show that the most reproducible model selection is generalized over different training modes. This emphasizes our framework's ability to assess reproducibility across different data distribution perturbation strategies. In addition, the model having the highest node strength might not be the same across reproducibility scores (correlation-based and average-based). This reflects that the GNN selection highly depends on the reproducibility score. However, the summation of the resulting matrices gave consistent conclusions regarding the most reproducible model selection. Once the most reproducible GNN model is selected, we extract its learned weights as in Fig. 7. The most discriminative biomarkers will be further detailed in the discussion section.

Table 1
Average classification accuracy across views by different GNN models using 5-fold cross-validation. ASD: autism spectrum disorder. NC: normal control. AD: Alzheimer's disease. LMCI: late mild cognitive impairment. LH: left hemisphere. RH: right hemisphere.

Datasets	ASD/NC		AD/LMCI	
	LH	RH	LH	RH
GCN (Kipf & Welling, 2016)	52.5	57.7	52.46	50.46
GAT (Veličković et al., 2017)	56.66	57.5	53.84	48.52
DiffPool (Ying et al., 2018)	53.3	54.17	59.84	47.53
g-U-Nets (Gao & Ji, 2019)	49.59	51.17	57.23	37.23
SAGPool (Lee et al., 2019)	52.16	52.58	53.50	55.10

4. Discussion

To the best of our knowledge, here, we proposed the first study to evaluate the reproducibility of GNN models. Our method, RG-Select, quantifies the reproducibility of a model based on the consensus of its most discriminative biomarkers across other models in a given pool of GNNs. Based on the node strength concept from graph theory, our framework quantifies the reproducibility score of GNNs. In contrast with other methods, RG-Select is applicable to datasets of multigraphs which indicates the challenging level of heterogeneity that can be handled by our framework. This also reflects the generalizability of the settings that we considered in our framework to determine the most reproducible model. Our framework succeeded in producing replicable results across different training modes in all the datasets. In more detail, for each dataset, the most reproducible method was the same across different training settings. Consequently, it identifies the most reproducible biomarkers as the most congruent features across models.

4.1. Most reproducible biomarkers

Fig. 7 displays the absolute value of weights respective to biomarkers learned by the most reproducible model for datasets AD/LMCI RH, AD/LMCI LH, ASD/NC RH and ASD/NC LH, respectively. Our framework identified DiffPool as the most reproducible method for AD/LMCI RH dataset as illustrated in Fig. 7. The most discriminative ROIs are the lingual gyrus and pericalcarine cortex. Supporting our findings, Shi et al. (2020) found that lingual gyrus plays an important role in the neuropathophysiology of depression in AD. Furthermore, Yang et al. (2019) found that pericalcarine cortex was among the brain regions which have a significant reduction in cortical thickness with AD patients. For AD/LMCI LH dataset, SAGPool was selected as the most reproducible model as illustrated in Fig. 3. The most important biomarkers in this dataset as illustrated in 7 are insula cortex and transverse temporal cortex. Lou et al. (2021) found that the insula cortex has a significant variation in T1 values with AD patients. In addition, Barnes et al. (1991) showed that the density of the selective angiotensin converting enzyme in temporal cortex is significantly higher with AD patients. For ASD/NC RH dataset, GAT was the most reproducible model. Following the model training, the most two discriminative regions were precentral gyrus and bank of the superior temporal sulcus. Nebel, Eloyan, Barber, and Mostofsky (2014) found that the precentral gyrus is highly related to the severity of ASD traits in brain connectivity network. Moreover, Zilbovicius et al. (2006) stated different experiments showing abnormal or absent superior temporal sulcus activation in patients with ASD during tasks involving social cognition. For the ASD/NC LH dataset, g-U-Nets was the most reproducible GNN model. The experiments have shown that posterior-cingulate cortex and insula cortex are the most discriminative biomarkers. Gogolla (2017) confirmed that irregularities in insula cortex connectivities are linked to autistic symptom severity. In addition, different studies showed that Abnormalities in posterior-cingulate cortex responses during interpersonal interaction highly correlate with the severity of patients' autistic symptoms (Chiu et al., 2008; Leech & Sharp, 2014).

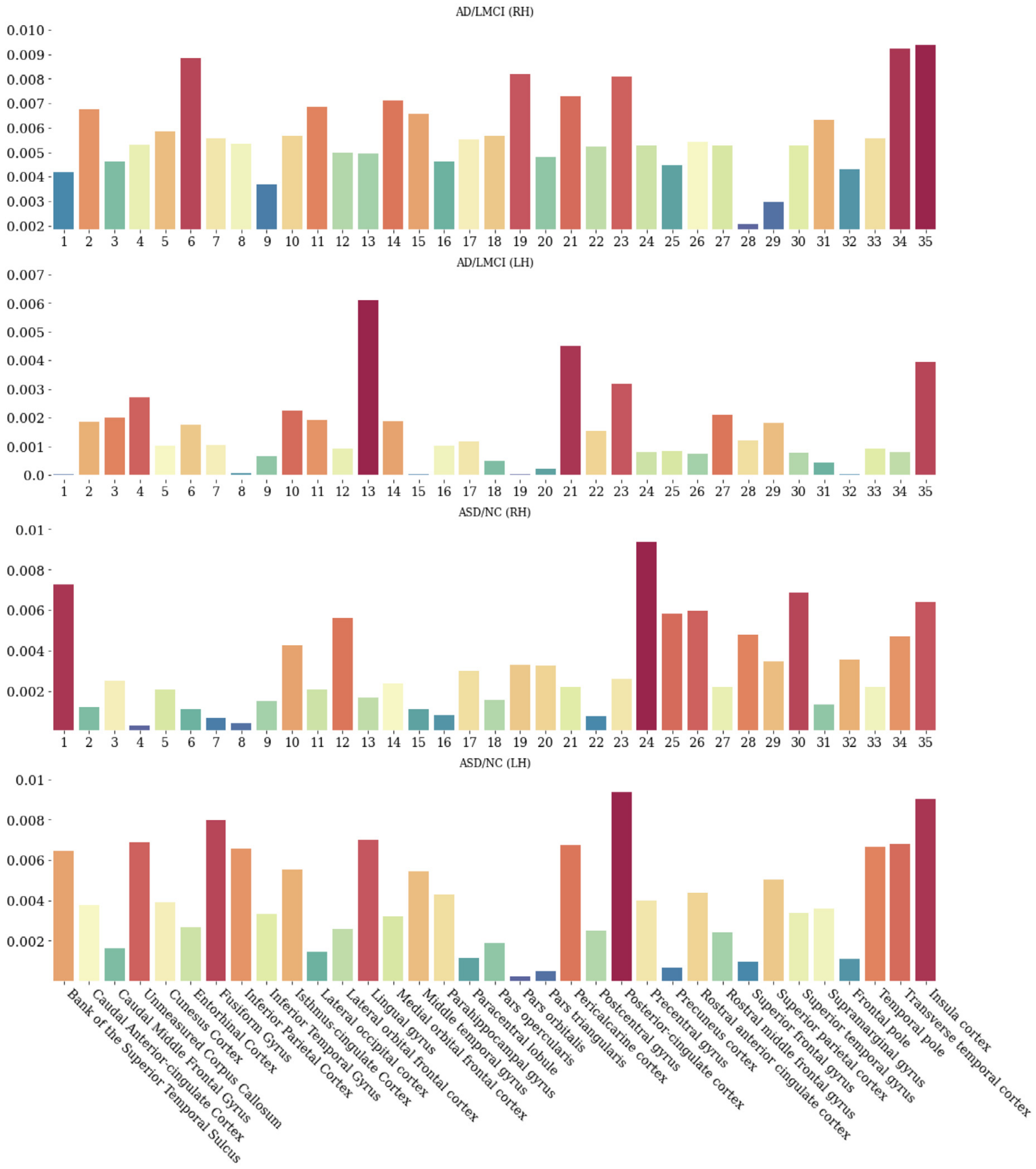

Fig. 7. *The learned weights for the cortical regions of the brain by the most reproducible model using the four datasets.* AD: Alzheimer's disease. LMCI: late mild cognitive impairment. ASD: autism spectrum disorder. NC: normal control. RH: right hemisphere. LH: left hemisphere. CV: cross-validation. FS: few-shot.

4.2. Reproducibility evaluation scores

Tables 2–5 contain all the reproducibility scores for all the models on each dataset. We denote the views average and rank correlation based reproducibility scores as v.a and r.c, respectively. Here, we detail the other reproducibility scores not mentioned in the methods section. The first score is the strength correlation (s.c). We extract the node weights of the GNN-specific reproducibility matrix for each model. Then, we compute their average over all the thresholds. The dimension of the resulting vector is equal to n_v. Finally, we calculate the correlation score of each pair of GNN resulting vectors. The second score is the accumulated weights correlation (a.w.c). instead of averaging over the thresholds, here, we accumulate them within one vector. The dimension of the resulting vector is $n_v \times n_k$. Next, for each pair of GNNs, we compute the correlation score between their respective resulting vectors. The third score is the accumulated weighted intersection (a.w.i): We calculate the accumulated vectors for each GNN. Next, we implement a weighted intersection of the resulting vectors. The weighted intersection takes into consideration of two vectors: accumulated strengths and accumulated rankings. The accumulated strengths vector is the same as the previous method. The accumulated ranking vector contains the rankings of the views (e.g., node strengths of the GNN-specific reproducibility graph). This vector mainly represents the similarities between the strengths and weighted by the similarities in the rankings. In other terms, it gives high scores to elements having close rankings and close strengths. It would also penalize the pair of vectors

Table 2
Reproducibility scores for AD/LMCI left hemisphere dataset using different GNN models. AD: Alzheimer's disease. LMCI: late mild cognitive impairment. CV: cross-validation. FS: few-shot. v.a: views average. r.c: rank correlation. a.w.i: accumulated weighted intersection. a.w.c: accumulated weights correlation. s.c: strength correlation. a.r.i: accumulated rank intersection. KL: KL divergence. L2: L_2 distance between vectors of scores.

Method	Training	v.a	r.c	a.w.i	a.w.c	s.c	a.r.i	KL	L2
DiffPool (Ying et al., 2018)	CV	0.354	**0.283**	0.314	0.264	**0.238**	0.255	**1.575**	**2.624**
	FS	0.366	**0.25**	**0.314**	0.103	**0.32**	**0.266**	**1.782**	2.809
GAT (Veličković et al., 2017)	CV	0.327	0.1	0.284	0.397	0.042	0.24	3.259	3.819
	FS	0.334	0.1	0.247	**0.422**	−0.013	0.156	4.109	4.119
GCN (Kipf & Welling, 2016)	CV	0.353	0.083	**0.329**	0.302	0.04	0.219	1.695	**2.647**
	FS	0.303	−0.05	0.302	0.205	0.027	0.188	2.345	3.017
g-U-Nets (Gao & Ji, 2019)	CV	**0.373**	−0.117	0.272	0.262	−0.008	**0.286**	2.728	3.45
	FS	**0.375**	−0.7	0.253	0.223	−0.496	0.25	3.244	3.746
SAGPool (Lee et al., 2019)	CV	0.345	−0.15	0.274	**0.482**	0.087	0.26	2.749	3.524
	FS	0.342	0.2	0.299	0.396	0.298	0.234	2.926	3.58

Table 3
Reproducibility scores for AD/LMCI right hemisphere dataset using different GNN models. AD: Alzheimer's disease. LMCI: late mild cognitive impairment. CV: cross-validation. FS: few-shot. v.a: views average. r.c: rank correlation. a.w.i: accumulated weighted intersection. a.w.c: accumulated weights correlation. s.c: strength correlation. a.r.i: accumulated rank intersection. KL: KL divergence. L2: L_2 distance between vectors of scores.

Method	Training	v.a	r.c	a.w.i	a.w.c	s.c	a.r.i	KL	L2
DiffPool (Ying et al., 2018)	CV	0.343	−0.05	0.345	**0.109**	−0.013	0.281	**1.499**	**2.484**
	FS	0.363	−0.3	0.297	0.145	−0.257	0.219	**1.023**	**2.39**
GAT (Veličković et al., 2017)	CV	0.353	0.033	0.265	0.047	−0.049	0.172	2.703	3.672
	FS	0.353	−0.2	0.246	0.176	−0.4	0.234	1.961	3.238
GCN (Kipf & Welling, 2016)	CV	0.342	0.033	**0.364**	−0.215	−0.004	**0.292**	1.986	2.769
	FS	0.319	0.1	**0.347**	−0.59	0.113	**0.266**	1.777	2.848
g-U-Nets (Gao & Ji, 2019)	CV	0.356	0	0.303	0.027	−0.113	0.224	2.017	2.822
	FS	**0.373**	−0.15	0.261	0.161	−0.181	0.234	1.748	2.804
SAGPool (Lee et al., 2019)	CV	**0.377**	**0.05**	0.298	0.079	−**0.001**	0.271	3.129	3.944
	FS	0.363	**0.15**	0.28	**0.246**	**0.095**	0.266	2.581	3.597

Table 4
Reproducibility scores for ASD/NC left hemisphere dataset using different GNN models. ASD: autism spectrum disorder. NC: normal control. CV: cross-validation. FS: few-shot. v.a: views average. r.c: rank correlation. a.w.i: accumulated weighted intersection. a.w.c: accumulated weights correlation. s.c: strength correlation. a.r.i: accumulated rank intersection. KL: KL divergence. L2: L_2 distance between vectors of scores.

Method	Training	v.a	r.c	a.w.i	a.w.c	s.c	a.r.i	KL	L2
DiffPool (Ying et al., 2018)	CV	**0.375**	0.086	**0.346**	0.415	0.01	0.194	6.901	5.779
	FS	0.344	0.029	**0.383**	0.52	0.002	**0.24**	4.511	**4.622**
GAT (Veličković et al., 2017)	CV	0.369	0.043	0.277	**0.627**	0.095	0.208	12.879	8.605
	FS	**0.38**	0.143	0.264	**0.63**	0.225	0.188	12.574	8.853
GCN (Kipf & Welling, 2016)	CV	**0.375**	−0.029	0.334	0.361	0.028	0.17	**5.618**	**5.381**
	FS	0.37	0.1	0.375	0.27	0.393	0.219	4.609	4.693
g-U-Nets (Gao & Ji, 2019)	CV	0.372	**0.214**	0.315	0.416	**0.216**	**0.222**	8.232	6.456
	FS	0.346	**0.414**	0.339	0.476	**0.436**	**0.24**	5.632	5.306
SAGPool (Lee et al., 2019)	CV	0.365	−0.048	0.31	0.559	−0.13	0.17	8.239	6.929
	FS	0.353	0.086	0.361	0.572	0.043	0.177	**4.136**	5.384

Table 5
Reproducibility scores for ASD/NC right hemisphere dataset using different GNN models. ASD: autism spectrum disorder. NC: normal control. CV: cross-validation. FS: few-shot. v.a: views average. r.c: rank correlation. a.w.i: accumulated weighted intersection. a.w.c: accumulated weights correlation. s.c: strength correlation. a.r.i: accumulated rank intersection. KL: KL divergence. L2: L_2 distance between vectors of scores.

Method	Training	v.a	r.c	a.w.i	a.w.c	s.c	a.r.i	KL	L2
DiffPool (Ying et al., 2018)	CV	0.361	0	**0.341**	0.37	0.151	0.198	6.93	**5.788**
	FS	0.35	0.186	**0.377**	0.52	0.234	0.177	4.441	**4.52**
GAT (Veličković et al., 2017)	CV	0.355	**0.138**	0.281	**0.624**	**0.2**	**0.208**	15.025	9.07
	FS	0.378	**0.529**	0.244	**0.627**	**0.464**	**0.229**	15.764	9.645
GCN (Kipf & Welling, 2016)	CV	0.372	−0.086	0.337	0.356	−0.078	0.153	**6.513**	5.777
	FS	**0.381**	0.143	0.372	0.099	0.22	0.156	6.583	5.559
g-U-Nets (Gao & Ji, 2019)	CV	**0.373**	0.005	0.294	0.342	−0.036	0.132	8.931	6.694
	FS	0.358	0.386	0.333	0.401	0.147	0.146	5.779	5.142
SAGPool (Lee et al., 2019)	CV	0.362	0.057	0.303	0.562	0.045	0.177	7.726	6.666
	FS	0.37	0.129	0.321	0.516	0.091	0.188	**3.957**	5.034

elements if the elements have close strengths but different ranks. The fourth score is the accumulated rank intersection (a.r.i). It accumulates the vectors of biomarkers at different thresholding based on the GNN-specific reproducibility matrix. Next, we rank the views for each GNN. Eventually, we calculate the correlation between each pair of GNN resulting ranking vectors. The fifth score is the KL divergence (KL). We calculate the ranking vectors of each GNNs as the previous methods. Next, we calculate the KL divergence of the resulting vectors. Unlike previous methods, this score reflects the dissimilarity between both distributions. Therefore, we are interested in identifying the model having the smallest score. Finally, we have the $L2$ distance score (L2) which is constructed by calculating the L_2 distance between pairs of the resulting vectors. This score reflects the dissimilarity between GNNs. Hence, based on this score the model having the smallest value is identified as the most reproducible GNN.

We note that for some metrics such as r.c, s.c., and a.w.c, it is possible to obtain negative values since these metrics result in the range [1, 1]. More importantly, if we consider each reproducibility score independently, the model selection conclusions will be divergent. This also emphasizes the importance of the two reproducibility scores that we have chosen in our framework. For instance, if we focus on Table 2, each score identifies a different GNN as the most reproducible method. However, the majority of scores indicate that DiffPool is the most reproducible method which confirms our findings for AD/LMCI left hemisphere dataset in Fig. 3. For the right hemisphere of the same dataset, the majority of scores in Table 3 indicates that SAGPool is the most reproducible model which is the same method selected by our framework for this dataset as detailed in Fig. 4. In Fig. 5, our framework selected g-U-Nets as the most reproducible model for ASD/NC left hemisphere dataset. The same result is confirmed by the majority of the reproducibility scores illustrated in Table 4. Finally, the majority of reproducibility scores in Table 5 reflects that GAT is the most reproducible model which is correlated with our findings in Fig. 6.

At the time of writing this paper, the reported GNNs construct the state-of-the-art for geometric deep learning models. A plethora of studies (Gao & Xu, 2020; Wang et al., 2021; Wieder et al., 2020) has already shown the outperformance of these GNNs in classification tasks for many datasets. We have carefully inspected the accuracy and training loss of all presented GNN models prior to including them in our study for reproducibility check. As shown in Table 1, the majority of GNN models displayed an average classification accuracy higher than 0.5 across all data views using 5-fold cross-validation — with very minor exceptions (e.g., g-U-Nets for RH AD/LMCI dataset). Fluctuation in accuracy is notable across models; however, the main focus of this paper is model reproducibility rather than model accuracy.

4.3. Limitations and future directions

Although our RG-Select successfully identifies the *most reproducible* graph neural architecture in a given pool of GNNs for a target multigraph classification task, it has a few limitations. First, our model does not identify the most reproducible view-specific biomarkers since we conceptualized the reproducibility paradigm as finding the model that produces the same biomarkers across different data views *and* various perturbation strategies. However, since a graph is regarded as a special instance of a multigraph where the number of node-to-node edges is equal to 1, one can directly use the GNN-specific reproducibility matrix to first identify the most reproducible GNN then the most reproducible biomarkers. Second, this study only focuses on GNN reproducibility while somewhat overlooking its learning performance in terms of classification accuracy. In our future work, we will investigate the trade-off between GNN reproducibility and performance in different classification tasks. Finally, although we used 5 models for the classification, we have only trained our GNN-based models in a fully supervised manner. As an extension of our RG-Select, we intend to encompass other families of classification methods including semi-supervised and weakly deep learning models.

5. Conclusion

While the majority of classification models have focused on boosting the accuracy of a given model, in this study, we address the problem of feature reproducibility. To the best of our knowledge, this is the first work investigating the reproducibility of GNNs in biomarkers using multigraph brain connectivity datasets. Our RG-Select demonstrated consistent results against different training strategies: cross-validation and a few-shot learning. Moreover, we evaluated our framework on both small-scale and large-scale datasets. This work presents a big stride in precision medicine since it incorporates the reproducibility of neurological biomarkers against different perturbations of multi-view clinical datasets. We believe that reproducibility frameworks can make major contributions towards unifying clinical interpretation by enhancing the extraction of the set of biomarkers responsible for brain connectivity alterations in neurologically disordered populations. One major drawback of our framework is the computational time consumed to run all the experiments. To circumvent this issue, we aim, in the foreseeable future, to predict the influence of different perturbations on the overall reproducibility of a given model instead of running it on all datasets.

Code availability

An open-source Python implementation of RG-Select is available on GitHub at https://github.com/basiralab/RG-Select. The release includes a tutorial, notes regarding Python packages, which need to be installed. Information regarding input format can be also found in the same repository. Input files contain the learned weights by different GNNs. However, the framework works with any data respecting the same shape of the weights vectors extracted from the GNNs.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that support the findings of this study are publicly available from ADNI data (http://adni.loni.usc.edu/). For reproducibility and comparability, the authors will make available upon request all morphological networks generated based on the four cortical attributes (maximum principal curvature, cortical thickness, sulcal depth, and average curvature) for the 77 subjects (41 AD and 36 LMCI) following the approval by ADNI Consortium. Our large-scale dataset is also available from the public ABIDE initiative (http://fcon_1000.projects.nitrc.org/indi/abide/). Following the approval by the ABIDE initiative, all morphological networks generated from the six cortical attributes (cortical surface area and minimum principle area in addition to 4 aforementioned measures) for the 300 subjects (150 NC and 150 ASD) are also accessible from the authors upon request.

Acknowledgments

This work was funded by generous grants from the European H2020 Marie Sklodowska-Curie action (Grant No. 101003403, http://basira-lab.com/normnets/) to I.R. and the Scientific and Technological Research Council of Turkey to I.R. under the TUBITAK 2232 Fellowship for Outstanding Researchers, Turkey (no. 118C288, http://basira-lab.com/reprime/). However, all scientific contributions made in this project are owned and approved solely by the authors. M.A.G. is supported by TUBITAK, Turkey.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro-Imaging at the University of Southern California.

References

Alper, B., Bach, B., Henry Riche, N., Isenberg, T., & Fekete, J.-D. (2013). Weighted graph comparison techniques for brain connectivity analysis. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 483–492).

Barnes, N. M., Cheng, C. H., Costall, B., Naylor, R. J., Williams, T. J., & Wischik, C. M. (1991). Angiofensin converting enzyme density is increased in temporal cortex from patients with Alzheimer's disease. *European Journal of Pharmacology*, *200*(2–3), 289–292.

Bassett, D. S., & Sporns, O. (2017). Network neuroscience. *Nature Neuroscience*, *20*(3), 353–364.

Bessadok, A., Mahjoub, M. A., & Rekik, I. (2021). Graph neural networks in network neuroscience. arXiv preprint arXiv:2106.03535.

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

Chiu, P. H., Kayali, M. A., Kishida, K. T., Tomlin, D., Klinger, L. G., Klinger, M. R., et al. (2008). Self responses along cingulate cortex reveal quantitative neural phenotype for high-functioning autism. *Neuron*, *57*(3), 463–473.

delEtoile, J., & Adeli, H. (2017). Graph theory and brain connectivity in Alzheimer's disease. *The Neuroscientist*, *23*(6), 616–626.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968–980.

Dhifallah, S., Rekik, I., & Alzheimer's Disease Neuroimaging Initiative, et al. (2020). Estimation of connectional brain templates using selective multi-view network normalization. *Medical Image Analysis*, *59*, Article 101567.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., 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), 659–667.

Du, Y., Fu, Z., Sui, J., Gao, S., Xing, Y., Lin, D., et al. (2020). NeuroMark: An automated and adaptive ICA based pipeline to identify reproducible fMRI markers of brain disorders. *NeuroImage: Clinical*, *28*, Article 102375.

Errica, F., Podda, M., Bacciu, D., & Micheli, A. (2019). A fair comparison of graph neural networks for graph classification. arXiv preprint arXiv:1912.09893.

Farahani, F. V., Karwowski, W., & Lighthall, N. R. (2019). Application of graph theory for identifying connectivity patterns in human brain networks: a systematic review. *Frontiers in Neuroscience*, *13*, 585.

Fischl, B. (2012). FreeSurfer. *Neuroimage*, *62*(2), 774–781.

Gao, H., & Ji, S. (2019). Graph u-nets. In *International conference on machine learning* (pp. 2083–2092). PMLR.

Gao, J., & Xu, C. (2020). Ci-gnn: Building a category-instance graph for zero-shot video classification. *IEEE Transactions on Multimedia*, *22*(12), 3088–3100.

Georges, N., Mhiri, I., Rekik, I., & Alzheimer's Disease Neuroimaging Initiative, et al. (2020). Identifying the best data-driven feature selection method for boosting reproducibility in classification tasks. *Pattern Recognition*, *101*, Article 107183.

Gogolla, N. (2017). The insular cortex. *Current Biology*, *27*(12), R580–R586.

He, Y., & Evans, A. (2010). Graph theoretical modeling of brain connectivity. *Current Opinion in Neurology*, *23*(4), 341–350.

Henaff, M., Bruna, J., & LeCun, Y. (2015). Deep convolutional networks on graph-structured data. arXiv preprint arXiv:1506.05163.

van den Heuvel, M. P., & Sporns, O. (2019). A cross-disorder connectome landscape of brain dysconnectivity. *Nature Reviews Neuroscience*, *20*(7), 435–446.

Jin, D., Zhou, B., Han, Y., Ren, J., Han, T., Liu, B., et al. (2020). Generalizable, reproducible, and neuroscientifically interpretable imaging biomarkers for Alzheimer's disease. *Advanced Science*, *7*(14), Article 2000675.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Lee, J., Lee, I., & Kang, J. (2019). Self-attention graph pooling. In *International conference on machine learning* (pp. 3734–3743). PMLR.

Leech, R., & Sharp, D. J. (2014). The role of the posterior cingulate cortex in cognition and disease. *Brain*, *137*(1), 12–32.

Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., et al. (2021). Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, *74*, Article 102233.

Lisowska, A., Rekik, I., & Alzheimers Disease Neuroimaging Initiative, et al. (2017). Pairing-based ensemble classifier learning using convolutional brain multiplexes and multi-view brain networks for early dementia diagnosis. In *International workshop on connectomics in neuroimaging* (pp. 42–50). Springer.

Lou, B., Jiang, Y., Li, C., Wu, P.-Y., Li, S., Qin, B., et al. (2021). Quantitative analysis of synthetic magnetic resonance imaging in alzheimer's disease. *Frontiers in Aging Neuroscience*.

Mahjoub, I., Mahjoub, M. A., & Rekik, I. (2018). Brain multiplexes reveal morphological connectional biomarkers fingerprinting late brain dementia states. *Scientific Reports*, *8*(1), 1–14.

Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., & Bronstein, M. M. (2017). Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5115–5124).

Nebel, M. B., Eloyan, A., Barber, A. D., & Mostofsky, S. H. (2014). Precentral gyrus functional connectivity signatures of autism. *Frontiers in Systems Neuroscience*, *8*, 80.

Povero, D., Yamashita, H., Ren, W., Subramanian, M. G., Myers, R. P., Eguchi, A., et al. (2020). Characterization and proteome of circulating extracellular vesicles as potential biomarkers for NASH. *Hepatology Communications*, *4*(9), 1263–1278.

Rashid, B., Arbabshirani, M. R., Damaraju, E., Cetin, M. S., Miller, R., Pearlson, G. D., et al. (2016). Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity. *Neuroimage*, *134*, 645–657.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770.

Shi, Z., Cao, X., Hu, J., Jiang, L., Mei, X., Zheng, H., et al. (2020). Retinal nerve fiber layer thickness is associated with hippocampus and lingual gyrus volumes in nondemented older adults. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *99*, Article 109824.

Shirer, W. R., Ryali, S., Rykhlevskaia, E., Menon, V., & Greicius, M. D. (2012). Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cerebral Cortex*, *22*(1), 158–165.

Smith, P. F., Ganesh, S., & Liu, P. (2013). A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of Neuroscience Methods*, *220*(1), 85–91.

Toga, A. W., Clark, K. A., Thompson, P. M., Shattuck, D. W., & Van Horn, J. D. (2012). Mapping the human connectome. *Neurosurgery*, *71*(1), 1–5.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. arXiv preprint arXiv:1710.10903.

Wang, P. Y., Sapra, S., George, V. K., & Silva, G. A. (2021). Generalizable machine learning in neuroscience using graph neural networks. *Frontiers in Artificial Intelligence*, *4*, 4.

Wang, M., Yu, L., Zheng, D., Gan, Q., Gai, Y., Ye, Z., et al. (2019). Deep graph library: Towards efficient and scalable deep learning on graphs.

Wang, Y., Zhang, J., Guo, S., Yin, H., Li, C., & Chen, H. (2021). Decoupling representation learning and classification for GNN-based anomaly detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 1239–1248).

Weiner, M. W., Aisen, P. S., Jack, C. R., Jr., Jagust, W. J., Trojanowski, J. Q., Shaw, L., et al. (2010). The Alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimer's & Dementia*, *6*(3), 202–211.

Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., et al. (2020). A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How powerful are graph neural networks? arXiv preprint arXiv:1810.00826.

Yang, H., Xu, H., Li, Q., Jin, Y., Jiang, W., Wang, J., et al. (2019). Study of brain morphology change in Alzheimer's disease and amnestic mild cognitive impairment compared with normal controls. *General Psychiatry*, *32*(2).

Ying, R., You, J., Morris, C., Ren, X., Hamilton, W. L., & Leskovec, J. (2018). Hierarchical graph representation learning with differentiable pooling. arXiv preprint arXiv:1806.08804.

Zhang, Z., Cui, P., & Zhu, W. (2020). Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., et al. (2020). Graph neural networks: A review of methods and applications. *AI Open*, *1*, 57–81.

Zilbovicius, M., Meresse, I., Chabane, N., Brunelle, F., Samson, Y., & Boddaert, N. (2006). Autism, the superior temporal sulcus and social perception. *Trends in Neurosciences*, *29*(7), 359–366.