NeuroImage 319 (2025) 121422

Contents lists available at ScienceDirect

NeuroImage

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

Review

Discovering robust biomarkers of psychiatric disorders from resting-state functional MRI via graph neural networks: A systematic review

Yi Hao Chan [a,1], Deepank Girish [a,2], Sukrit Gupta [b,3], Jing Xia [a,1], Chockalingam Kasi [a,4], Yinan He [a,4], Conghao Wang [a,2], Jagath C. Rajapakse [a,*,5]

[a] College of Computing and Data Science, Nanyang Technological University, Singapore, 639798, Singapore
[b] Department of Computer Science and Engineering, Indian Institute of Technology, Ropar, 140001, Punjab, India

ARTICLE INFO

Dataset link: https://osf.io/wza6b/

Keywords:
Biomarker discovery
Feature attributions
Graph neural networks
Model explainability
Psychiatric disorders
Resting-state fMRI
Robustness

ABSTRACT

Graph neural networks (GNN) have emerged as a popular tool for modeling functional magnetic resonance imaging (fMRI) datasets. Many recent studies have reported significant improvements in disorder classification performance via more sophisticated GNN designs and highlighted salient features that could be potential biomarkers of the disorder. However, existing methods of evaluating their robustness are often limited to cross-referencing with existing literature, which is a subjective and inconsistent process. In this review, we provide an overview of how GNN and model explainability techniques (specifically, feature attributors) have been applied to fMRI datasets for disorder prediction tasks, with an emphasis on evaluating the robustness of potential biomarkers produced by these feature attributors for psychiatric disorders. Then, 65 studies using GNNs that reported potential fMRI biomarkers for psychiatric disorders (attention-deficit hyperactivity disorder, autism spectrum disorder, major depressive disorder, schizophrenia) published before 9 October 2024 were identified from 2 online databases (Scopus, PubMed). We found that while most studies have performant models, salient features highlighted in these studies (as determined by feature attribution scores) vary greatly across studies on the same disorder. Reproducibility of biomarkers is only limited to a small subset at the level of regions and few transdiagnostic biomarkers were identified. To address these issues, we suggest establishing new standards that are based on objective evaluation metrics to determine the robustness of these potential biomarkers. We further highlight gaps in the existing literature and put together a prediction–attribution–evaluation framework that could set the foundations for future research on discovering robust biomarkers of psychiatric disorders via GNNs.

Contents

* Corresponding author.
E-mail addresses: yihao001@e.ntu.edu.sg (Y.H. Chan), deepank002@e.ntu.edu.sg (D. Girish), sukrit.gupta@iitrpr.ac.in (S. Gupta), jing_xia@ntu.edu.sg (J. Xia), choc0010@e.ntu.edu.sg (C. Kasi), heyi0003@e.ntu.edu.sg (Y. He), conghao001@e.ntu.edu.sg (C. Wang), asjagath@ntu.edu.sg (J.C. Rajapakse).
[1] Research Fellow.
[2] Ph.D. Candidate.
[3] Assistant Professor.
[4] Undergraduate student.
[5] Professor.

https://doi.org/10.1016/j.neuroimage.2025.121422
Received 15 February 2025; Received in revised form 2 August 2025; Accepted 13 August 2025
Available online 29 August 2025

1. Introduction

Psychiatric disorders often manifest as changes in the functional characteristics of the brain. Functional magnetic resonance imaging (fMRI) has been widely used to objectively quantify these functional aberrations and identify the underlying neural substrates in the human brain. This has led to decades of research documenting the associations between psychiatric disorders and disruptions in whole-brain functional connectivity (FC) (Filippi et al., 2023). However, these characterisations have yet to reveal strong and reproducible biomarkers for most disorders (Abi-Dargham et al., 2023; Chollet and Payoux, 2022). The lack of success is commonly attributed to limitations such as disease heterogeneity (Verdi et al., 2021), inter-individual variability (Canario et al., 2021), small effect size (Poldrack et al., 2017; Jia et al., 2018), noise (Liu, 2016), limited dataset sizes (Marek et al., 2022), site effects

(also known as batch effects) when combining data from multiple sites (Bayer et al., 2022) and variability introduced by the choice of pre-processing pipeline (Dadi et al., 2019; Botvinik-Nezer et al., 2020).

Over the past years, larger datasets have emerged through inter-institution collaborations (Laird, 2021) and standardised formats for organising neuroimaging datasets (Gorgolewski et al., 2016), more mature pre-processing pipelines are available (for example, fMRIPrep Esteban et al., 2019), and better harmonisation tools have been developed to reduce batch effects (Hu et al., 2023), alleviating some of the above-mentioned issues in fMRI studies. Coupled with the development of more sophisticated modeling tools, model performances have improved and more potential biomarkers have been discovered in recent years, warranting the need for another review to synthesise their findings.

Many modern modeling tools involve machine learning (ML) algorithms, partly due to the high dimensionality of fMRI datasets. Mass univariate approaches (where each voxel is modeled independently) have traditionally been used, but they do not capture inter-region functional relationships (Habeck and Stern, 2010). To address this limitation, multivariate techniques (e.g. multi-voxel pattern analysis Weaverdyck et al., 2020) have been proposed, involving ML algorithms such as support vector machines (SVM). More recently, deep learning models have been shown to outperform SVM in disease classification tasks (Zhang et al., 2020). While convolutional neural networks (CNN) customised for connectome datasets (Kawahara et al., 2017; Meszlényi et al., 2017) were proposed as an improvement over vanilla deep neural networks (DNN) (Gupta et al., 2021), graph neural networks (GNN) have since emerged as the state-of-the-art deep learning model used in network neuroscience studies (Bessadok et al., 2022). The primary difference between GNNs with other deep learning architectures is the involvement of the adjacency matrix in the learning process. Another difference (between CNN and GNN) is the notion of a neighbor – in the former, this is often limited by spatial Euclidean distance while in the latter, there is no such requirement and it is defined by the adjacency matrix. Instead of learning a separate weight for each neighboring feature, the messaging passing process in GNNs adopts a leaner weight sharing mechanism across neighbors, but also introduces variations in a principled way via the adjacency matrix (often derived from the functional connectome). This also makes it convenient to design techniques that follow functional network organisation (Biswal and Uddin, 2025) (e.g. modularity Wang et al., 2024a), or even capture inter-patient relationships (Parisot et al., 2018).

While vanilla GNNs do not seem to do better than DNNs and CNNs (ElGazzar et al., 2022), more carefully designed GNN architectures have demonstrated significant improvements in disease classification performance (Song et al., 2021; Xiao et al., 2022).

However, disorder prediction is rarely the end goal as clinical adoption of such models is rare (Duda et al., 2023). Instead, these models can be used to provide neurological insights (e.g. nosology, subtyping, etc.). This is most commonly demonstrated via model explainability techniques (henceforth termed as 'attributor', as the majority of them assign an importance score to each feature as a limited form of explanation), ranging from gradient-based methods such as Integrated Gradients (IG) (Chan et al., 2022) or perturbation-based approaches such as GNNExplainer (Gallo et al., 2023). Alternatively, GNN-specific mechanisms such as graph pooling (Li et al., 2021b) can simultaneously train the model and produce feature attribution scores. These attributors typically assign attribution scores to each feature or identify important subgraphs that contribute most to the model's predictions.

While many attributors have been used in existing studies, several GNN-specific attributors remain unexplored. Furthermore, little has been studied about the robustness of these attributions. Existing studies often provide a very limited evaluation of their biomarkers, only reporting the top few features and cross-referencing other studies that had similar findings. Most recently, several studies have started comparing the attributions generated by different feature attributors and found that they could vary across datasets (Gallo et al., 2023), predictors (Zhang et al., 2022b) and even attributors even when the same predictor was used (Hu et al., 2021; Li et al., 2023). To ensure that salient features are truly representative of disorder traits and not mere artifacts, it would be prudent to take a pause and survey the existing literature to identify any convergence in the potential biomarkers that they reported.

In this review, we summarise recent progress on psychiatric disorder prediction via GNNs, with a focus on reviewing attributors used and potential biomarkers discovered by these fMRI studies. Psychiatric disorders included in this review include attention-deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), major depressive disorder (MDD) and schizophrenia (SZ). These disorders were chosen as they generally have more open-source datasets available, as well as more GNN-based studies, as compared to other disorders. We explore the following questions, highlight remaining research gaps in these areas, and suggest pathways for future research on biomarker discovery via GNNs:

1. Are there any graph construction approach and GNN-specific architecture designs that have consistently demonstrated superiority over others (e.g. their proposed GNN-based model works well across multiple datasets and studies)?
2. How do existing studies evaluate the robustness of the potential biomarkers identified by their proposed approach (i.e. combination of predictor and attributor)?
3. For each disorder, is there any convergence of discovered potential biomarkers across multiple studies (that are based on machine learning techniques)? If there are any, are they present in other disorders as well (i.e. potential transdiagnostic biomarkers)?

1.1. Related studies

Several review papers have been written on topics such as the use of GNNs for fMRI data analysis, model explainability in GNNs, and biomarker discovery for various neurodegenerative diseases and neuropsychiatric disorders. However, none has attempted to study these topics in a cohesive manner. Recent reviews on GNN applications in network neuroscience (Bessadok et al., 2022; Zhang et al., 2023c) summarised various graph-based and population-based models that have been proposed. Several benchmarking studies (ElGazzar et al., 2022; Cui et al., 2022; Said et al., 2023) have also been conducted to analyze the efficacy of various GNN designs on fMRI data in a fair and controlled manner that minimises the effect of covariates. Our study builds on top of their findings by synthesising their key takeaways and goes beyond the brain graph–population graph dichotomy to provide a taxonomy of state-of-the-art GNN architectures that are customised for fMRI datasets.

Model explainability methods have been thoroughly reviewed in previous works (Linardatos et al., 2020; Tjoa and Guan, 2020; Marcinkevičs and Vogt, 2023), including methods that are specialised for GNNs (Yuan et al., 2022; Kakkad et al., 2023). Several recent reviews have examined the use of these methods in the medical domain (Munroe et al., 2024; Rahman et al., 2023). However, to the best of our knowledge, no review has been done to assess the relevance of these methods in the context of biomarker discovery from fMRI datasets. Existing reviews on biomarkers (Abi-Dargham et al., 2023; Filippi et al., 2023) tend to summarise various types of biomarkers (often going beyond fMRI) and do not focus on examining and evaluating the reliability of the computational techniques (Nauta et al., 2023; Agarwal et al., 2023) used to derive the biomarkers. In our review paper, we aim to address this gap.

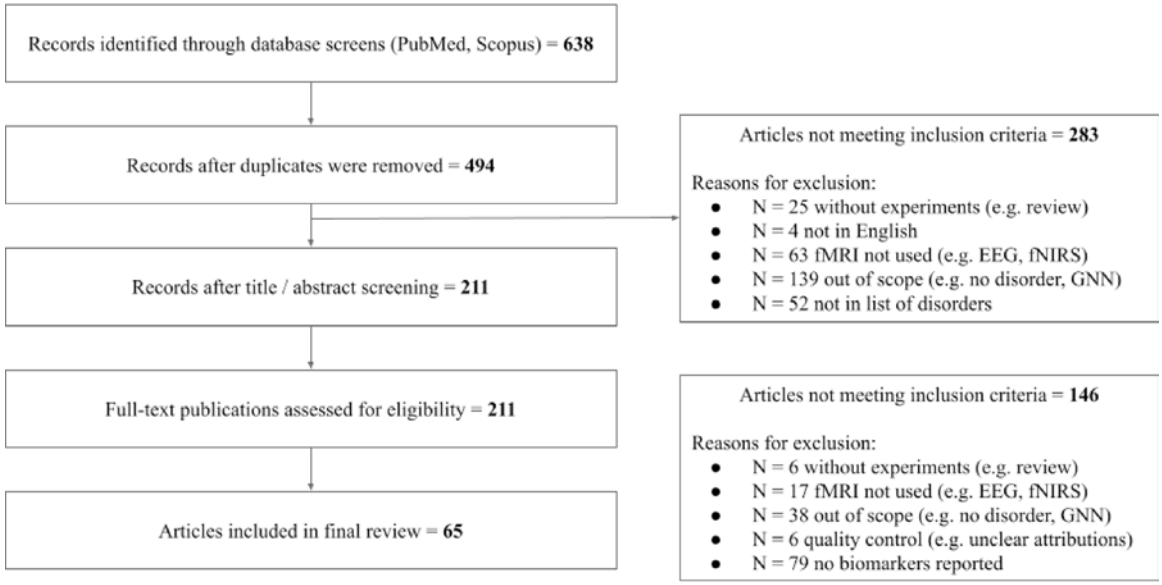

Fig. 1. Flowchart detailing the selection process of this review.

1.2. Review methodology

To identify papers that applied GNNs on fMRI datasets, we performed a search on 9 October 2024 via PubMed and Scopus with the following search query: ("graph neural networks" OR "graph convolutional networks" OR "GNN" OR "GCN") AND ("fmri" OR "functional MRI" OR "functional connectivity"). 158 papers from PubMed and 480 papers from Scopus matched these search terms. Manual filtering was conducted to remove duplicates and remove irrelevant results (e.g. no disorder prediction, resting-state fMRI not used in experiments, no experiments conducted). For papers on disorder prediction, only papers within our defined scope (ADHD, ASD, MDD, SZ) were included and they should report biomarkers clearly (e.g. not just reporting a brain map without any labels) as one major goal of our paper is to consolidate and identify potential biomarkers reported across multiple studies. This was checked manually and independently by 3 reviewers who scanned through the full text. Fig. 1 shows a detailed breakdown of the filtering process. At the end of this filtering process, a total of 65 papers were reviewed. The review and its protocol were not registered as retrospective registration is not feasible. Nevertheless, the PRISMA checklist (Page et al., 2021) and review protocol can be found in the Supplementary materials. Data such as the number of subjects, class splits (disorder, controls), type of GNN, model performance (accuracy), biomarkers reported, etc. were manually extracted from these papers. Variations in labeling techniques (i.e. criteria for diagnosis) across studies were present but minimal as a majority of the studies, for each disorder, relied on the same dataset source. Notably, such variations are more pronounced within each study as datasets are often aggregated from multiple imaging sites which often do not have exactly the same inclusion/exclusion criteria. As an exploratory review, studies were not excluded on the basis of having different labeling techniques. In occasional cases of missing information and there was no feasible way to retrieve them (e.g. model accuracy presented in charts but numbers not reported), these fields were left blank in the tables reported below and not considered in the computation of mean statistics (e.g. mean accuracy). However, studies without any biomarkers clearly reported were excluded. Several studies tested their models across multiple diseases. In such cases, data specific to the disorder (e.g. dataset, class splits) are extracted separated for each study. Potential biomarkers reported were manually collated and matched across studies as there is no standardised reporting format, nor any existing tools that could automate the process.

2. Modeling functional MRI datasets for disorder prediction

Blood-oxygen-level-dependent signals captured in fMRI scans are fundamentally represented as time series data from individual voxels. Even at relatively low resolutions (e.g. 5 mm), the number of voxels ($>10,000$) far outnumbers typical dataset sizes. Coupled with the issue of low signal-to-noise ratio (SNR) in fMRI data (Vizioli et al., 2021), these problems have motivated researchers to group clusters of related voxels together to improve SNR. Examples of such techniques are atlas-based approaches, independent component analysis (ICA) (Calhoun et al., 2009), and functional gradients/manifolds (Hong et al., 2020). The former applies atlases developed by delineating boundaries following anatomical landmarks (such as sulci (Rolls et al., 2020) and gyri (Desikan et al., 2006)) or task-fMRI experiments that identify regions of interest (ROI) that are activated when performing various tasks (Seitzman et al., 2020). This provides an informed way of feature selection to reduce data dimensionality, with the disadvantage of neglecting the voxels that are not part of the atlas' ROIs if a sphere-based approach (i.e. demarcate voxels that are within a certain radius from the ROI's coordinate) is used instead of a parcellation-based approach. On the other hand, ICA and functional gradients take a different approach by learning lower-dimensional representations of the original data. Out of these approaches, atlas-driven dimensionality reduction techniques are most widely used in the studies considered in this review.

2.1. Functional connectivity

Besides modeling the mean time series of each ROI or component directly, another common way to analyze fMRI data is to study the relationship between pairs of time series data. Pearson correlation has been the most common approach to compute such FC matrices. However, it is limited to capturing linear correlations and it could have weaker connections suppressed by noise or imaging artifacts. Alternative metrics introduced to address these limitations include various forms of partial correlation and sparse representation (Yang et al., 2021).

While a majority of existing studies are limited to such pairwise analysis at the level of regions/nodes, several recent studies have experimented with new graph construction techniques to address existing limitations. Going beyond inter-nodal relationships, edge FC was proposed to consider relationships between edges (Faskowitz et al., 2020)

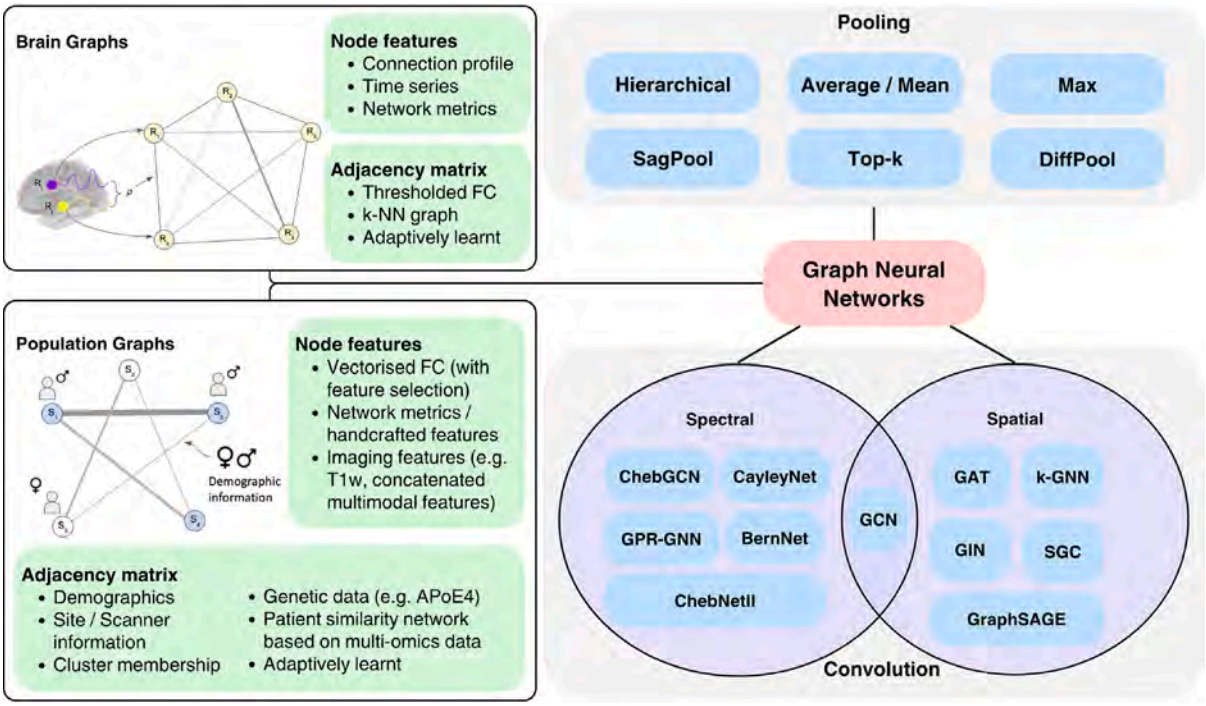

Fig. 2. Summary of the key components of typical GNNs used on fMRI datasets.

(via much larger FC matrices, as all pairwise combinations of nodes are considered). Going beyond pairwise relationships, hypergraphs are higher-order graphs where each hyperedge represents the relationship between two or more nodes. In recent studies, hypergraphs based on connectomes are typically generated dynamically (Wang et al., 2023a; Ji et al., 2022).

Regardless of the choice of analysis type (i.e. region-level or edge-level pairwise connectivity, or higher-order relationships), there are two key paradigms of modeling FC: static FC (sFC) and dynamic FC (dFC). sFC assumes that FC is stable throughout the scan (i.e. Pearson's correlation is computed from the entire time series, without splitting it into parts). On the other hand, dFC does not make such an assumption. The most common approach involves using a sliding window across the time series, generating multiple FC matrices in the process. More sophisticated approaches perform clustering and decomposition on these matrices to produce a single dFC matrix that better captures FC dynamics than the sFC matrix. Du et al. (2018) provides a detailed review of dFC (as well as ICA-based methods). Overall, using sFC at the level of pairwise ROIs remains the most common approach despite the above-mentioned advancements.

2.2. Graph representations of fMRI datasets

fMRI datasets, especially FC matrices, are most naturally represented as graphs. A graph data structure consists of nodes that are linked by edges, which represent the relationship between connected nodes. There are two major paradigms of modeling fMRI datasets: brain graphs (BG) and population graphs (PG). Fig. 2 illustrates the differences between them and lists down several possible options when choosing the node features and adjacency matrix of the BG and PG.

In a BG, each node in the graph represents an ROI and the choice of node features varies across studies — most commonly, the connection profile for that ROI (i.e. the row in the FC matrix that corresponds to that ROI) is used. Other possible node features include regional features that compute various statistics of the voxels within an ROI, such as regional homogeneity (ReHo) and Amplitude of Low Frequency Fluctuations (ALFF) (Arya et al., 2020). Edges store quantitative measures of the relationship between ROI pairs (e.g. Pearson's correlation of the mean time series). Thus, each subject (or scan) is represented as a graph, and graph classification is typically performed. On the other hand, in a PG, each node represents a subject (or scan) and the graph represents the population of interest. Node classification is typically performed. Node features typically store some representation of the imaging data (often after dimensionality reduction via recursive feature elimination (RFE) or principal component analysis (PCA)). Edges store measures of similarity between scans (usually demographic information such as age and gender, or any vector from which distance can be computed). Another alternative is to construct a k-nearest neighbors (k-NN) graph (Wang et al., 2021; Zhang et al., 2021).

PGs allow a much wider variety of data (including metadata) to be incorporated into the analysis. However, such graphs are typically pre-defined rather arbitrarily and tend to be static (Bintsi et al., 2023). Thus, learnable (Cosmo et al., 2020) and adaptive (Song et al., 2021; Park et al., 2023) methods of PG construction have been proposed. Recent studies (Jiang et al., 2020; Xiao et al., 2022; Zhang et al., 2022a; He et al., 2023) have also explored ways to use BG and PG simultaneously. Overall, BG is the dominant approach of input graph construction (used in 82.1% of studies, as compared to 7.7% for PG). Connection profile is by far the most popular node feature (58.9%) and the adjacency matrix is often the thresholded FC matrix (47.4%).

2.3. Encoding fMRI data with graph neural networks

Let $G = (V, A, X)$ represent a graph used by the GNN, where V represents the set of nodes in the graph, $A \in \mathbb{R}^{|V|\times|V|}$ represents the adjacency matrix used and $X \in \mathbb{R}^{|V|\times K}$ represents the node features, each of length K.

Traditionally, network-based analyses have been performed on fMRI datasets for disease studies, revealing insights such as lower clustering coefficient, global efficiency, and node degree for patients with mild cognitive impairment (Filippi et al., 2023). With the advent of ML, researchers started training models that distinguish between healthy subjects and patients. Since graphs cannot be used as input to many of these models, features were handcrafted (e.g. using network metrics Yin et al., 2021) or in the case of FC matrices, vectorised by flattening the lower triangular (Gupta et al., 2021). Doing so loses the graph structure

and leads to a high-dimensional input. Thus, models tend to overfit and feature selection methods (such as two-sample t-test and recursive feature elimination) are often used to address these issues (Teng et al., 2023).

Recently, GNNs – neural networks that are designed to be applied directly to graphs – have been used for encoding fMRI datasets. GNNs provide a parameter efficient means of modeling FC matrices (Li et al., 2021b), reducing the problem of overfitting and allowing for more sophisticated analyses involving modular brain networks (Mei et al., 2022) and multimodal data (He et al., 2023). GNNs can be broadly categorised into spectral GNNs and spatial GNNs.

Spectral GNNs perform convolution by transforming the graph signal and filtering to the spectral domain before convolving. The convolution operation is defined as:

$$g \star x = U(U^T g \circledast U^T x), \tag{1}$$

where $U^T x$ converts a signal x to the spectral domain using graph Fourier transform and Ux transforms the signal x back to the spatial domain using inverse graph Fourier transform. The operation can be simplified as:

$$g_\theta \star x = U g_\theta U^T x, \tag{2}$$

where g_θ denotes a learnable diagonal matrix. Examples of spectral GNN include ChebNet (Defferrard et al., 2016) and a recent improvement of it called ChebNetII (He et al., 2022) which noted that ChebNet can learn inappropriate Chebyshev coefficient which results in overfitting and sub-optimal performance. ChebNetII uses Chebyshev interpolation to overcome this issue, demonstrating better performance.

Spatial GNNs, on the other hand, apply convolutions to the graph based on its topology. Adopting a message-passing paradigm, node features in node i are iteratively updated by aggregating node features from its neighbors $\mathcal{N}(i)$.

$$X_i' = \alpha_{i,i} X_i U + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j} X_j W, \tag{3}$$

where $U, W \in \mathbb{R}^{K \times K'}$ represent learnable weights and α represents coefficients learnt to balance between retaining information from the original node vector and using information gathered from its neighbors. A point to note is that the above formulations present the core convolution operations (without the non-linearity shown) but in actual implementations, normalisation of the adjacency matrix is often performed to prevent the problem of vanishing or exploding gradients when multiple graph convolution layers are used together.

In general, variants of spatial GNN differ in how they weigh and aggregate information. For example, graph attention network (GAT) (Veličković et al., 2018) learns attention scores to choose which neighbors to focus on, while graph isomorphism network (GIN) (Xu et al., 2019) learns to balance between information from neighbors and the node's own node features. Note that graph convolutional network (GCN) (Kipf and Welling, 2017) can be seen as both spatial GNN and spectral GNN, as shown in Fig. 2.

Beyond baseline GNNs, several graph convolution layers and pooling layers have been proposed for use in connectome datasets. Many baseline GNNs focus on node-level aggregation, but FC is often used as the graph of BGs. This motivates the use of edge features as well as edge-based convolutions such as EdgeConv (Wang et al., 2019) or GraphConv (Morris et al., 2019). Fig. 2 also provides a summary of common baseline pooling layers. More details about pooling layers will be discussed in Section 3.2.3 as they are often used as a means to improve model explainability.

Additionally, Fig. S17 summarises the types of GCNs used in these studies. It is evident that GCNs are widely used (present in almost half of the studies) and spatial GCNs (e.g. GIN, GAT, GraphSAGE) are more widely used than spectral GCNs (e.g. ChebNet)

Details about the above GNN architectures have been reviewed by Bessadok et al. (2022) and several benchmarking studies have evaluated the efficacy of GNNs on fMRI datasets. Thus, we will summarise key points from these papers and focus on aspects pertaining to model explainability. BrainGB (Cui et al., 2022) splits the GNN methodology into 4 parts: node feature construction, messaging passing mechanism, attention-enhanced message passing, and pooling strategies. On multiple datasets (both healthy subjects and patients with disorders), they showed that connection profile (i.e. use the corresponding row of the FC matrix as the node features for an ROI), node concat message passing with attention (i.e. multiply learnt attention weights to the neighbor's node feature before concatenating it with the node's feature vector, followed by a multi-layer perceptron), and concat pooling (i.e. concatenate node features from multiple ROIs when performing graph pooling) works best.

Neurograph (Said et al., 2023) applied various spectral and spatial GNNs (GCN, GAT, GIN, etc.) on healthy subjects from the Human Connectome Project dataset for various baseline problems such as task classification, gender classification and age prediction. They showed that their proposed GNN⋆ architecture (which has 3 graph convolution layers with skip connections) works best. They also experimented with various settings such as the number of ROIs used, the sparsity of graphs, and how the node features were created. Model performance was shown to improve when including more ROIs (400 and 1000), using sparser graphs (5%), and using Pearson correlation for node features. It is notable that another benchmarking study (ElGazzar et al., 2022) suggested that GNNs do not even outperform 1D CNN for MDD and ASD classification. However, they opted for a relatively dense FC matrix (50%–90%) and binarised the graph.

Overall, all three benchmarking papers noted that the sparsity of the graph used by the GNN impacts model performance and lower sparsities (below 50%) were shown to be beneficial. Hyperparameter and GNN construction choices that lead to more model parameters (e.g. more ROIs, concatenation of node features) seem to enhance performance, but this should be done with care even though GNNs are parameter efficient (relative to other DNNs). Finally, vanilla GNNs do not seem to clearly outperform non-graph baselines such as SVM and CNN, but more carefully designed GNNs do improve model performance.

2.4. State-of-the-art GNN architectures customised for fMRI datasets

The benchmarking studies discussed above largely involve baseline GNNs and they have not considered state-of-the-art GNN architectures that are customised for fMRI datasets. Such improved architectures have been shown to outperform baselines and generalise better than them. Thus, it would be of interest to use them (instead of baseline models) for biomarker discovery.

These models are often built on top of baseline GNN models, addressing certain limitations by making changes in the (i) input to the GNN, i.e. graph construction process, (ii) formulation of the GNN's message passing and pooling mechanism, (iii) techniques used to train these GNNs. Fig. 3 shows our proposed taxonomy to categorise these customisations of the original baseline GNN.

Most routine applications of GNN on fMRI datasets involve the use of vanilla GNNs (e.g. GCN, ChebNet) constructed using a binarised/thresholded FC matrix as the adjacency matrix and the connection profile as the node features. Numerous state-of-the-art GNN models go beyond this by introducing (i) techniques to learn representations of the original data and use them for graph construction, (ii) higher-order information (such as hypergraphs) to go beyond what a typical FC matrix can represent, and (iii) ways to incorporate multiple views (e.g. sFC and dFC, multiple atlases, etc.). Notably, in the case of hypergraphs, its higher-order relationships require a slightly different GNN formulation (which is actually a generalisation of GNNs (Bai et al., 2021b)). Instead of an adjacency matrix, an incidence matrix $B \in \mathbf{R}^{|V| \times |H|}$ is used, where H represents the set of hyperedges. Each hyperedge $h \in H$ comes along with a diagonal weight matrix $W' \in \mathbf{R}^{|H| \times |H|}$. Another two diagonal matrices are also required: $D_{ii} \in \mathbf{R}^{|V| \times |V|}$ representing vertex degree and $D_{hh} \in \mathbf{R}^{|H| \times |H|}$ representing hyperedge degree. The

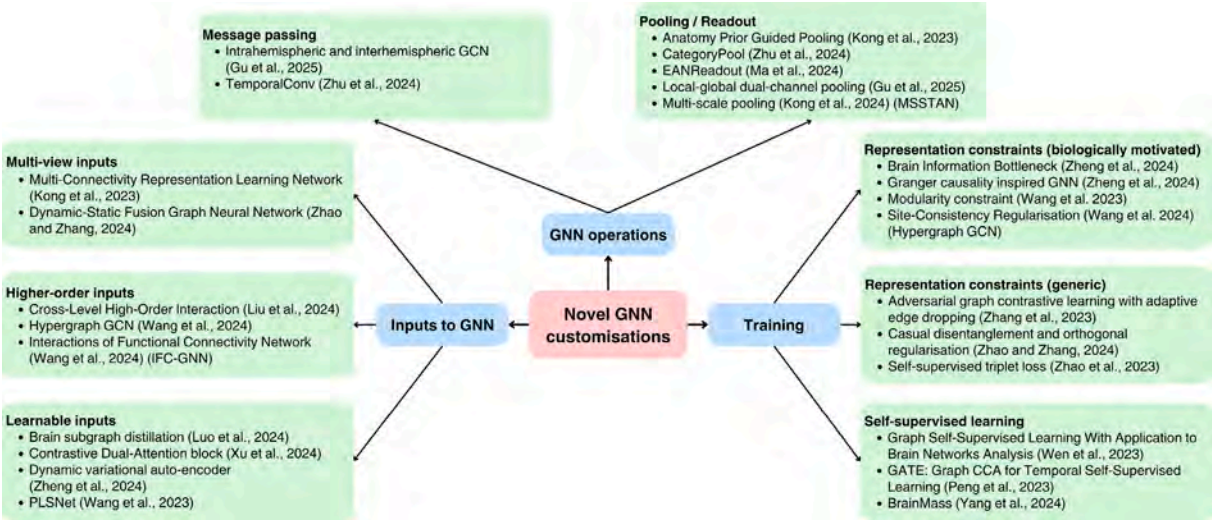

Fig. 3. Our proposed taxonomy of state-of-the-art GNN models customised for fMRI datasets.

message passing mechanism in hypergraph GCNs (HGCN) can then be expressed as:

$$X_i' = \alpha_{iih} X_i H_{ih} H_{ih} W_{hh}' X_i U + \sum_{j \in V \setminus \{i\}} \sum_{h=1}^{H} \alpha_{ijh} H_{ih} H_{jh} W_{hh}' X_j W. \quad (4)$$

Another paradigm of state-of-the-art GNNs modifies the standard message passing and pooling operations so that the model conforms to pre-existing biological knowledge. For instance, instead of aggregating information from all neighbors or relying on attention to learn which neighbors to focus on, one could limit the scope to intra-hemispheric neighbors before inter-hemispheric aggregation (Gu et al., 2025). Pooling is another active area with numerous innovations that make use of our current understanding of the brain to consolidate information across nodes in a biologically meaningful manner (rather than a coarse, one-step pooling operation) (Zhu et al., 2024).

A different class of modifications involves the introduction of constraints in the loss function used to train the GNN. Vanilla GNNs are typically only optimised with respect to the task (e.g. minimise cross-entropy). This gives the neural network too much freedom and it could end up learning representations that are not biologically sound. Adding constraints to the loss function could prevent this. For instance, a modularity constraint (Wang et al., 2023b) was introduced to ensure that embeddings learnt by nodes (ROIs) that originated from the same module were similar, in line with empirical observations of the modular architecture of the functional brain network. Several other studies focus on more generic constraints that aim to improve the robustness of the GNN to minor alterations of the FC matrix. This could have downstream implications on the robustness of biomarkers and would be an interesting area for further research. Finally, self-supervised learning techniques have also been proposed to alleviate the problem of small fMRI datasets. Such techniques collate large datasets for pre-training GNNs, often relying on various novel data augmentation strategies (Peng et al., 2022; Yang et al., 2024). Such advancements enable the construction of foundation models, which have been shown to improve generalisation capabilities.

2.5. Evaluation of strengths and weaknesses

Overall, recent research on using GNNs on fMRI datasets for disorder prediction (as summarised in Fig. 4) has involved much experimentation with a large variety of modeling techniques and numerous customisations on top of baseline models have been proposed. Thus, it would be valuable to evaluate them and suggest future research directions to move the field closer to the goal of producing more robust functional neuromarkers of psychiatric disorders. In this analysis, we focused on studies with over 100 data samples to exclude studies with small datasets that might skew the mean accuracy. Despite the limitations of classification accuracy as a metric (e.g. class imbalance), it was used as that was the most widely available metric and we found that the problem of class imbalance was not very prevalent in the selected studies. Overall, the mean accuracy reported was 75.7%.

Answering the first question mentioned in the introduction (whether any particular graph construction approach and GNN architecture designs are better), the best-performing models across all datasets, based on the reported test accuracies, typically used thresholded FC matrices as the adjacency matrix (i.e. retaining the values above the threshold) and connection profile as node features, concurring with the findings in Cui et al. (2022). However, numerous studies with lower performance also used connection profiles but tend to involve binarised adjacency matrices. The simplicity of this predominant approach, coupled with its high performance, could explain why more advanced methods (e.g. sparse representation, learnable inputs, etc.) have not caught on. Furthermore, in biomarker discovery applications, the current goal is typically to cast a wide net across the whole brain and compute attributions relative to a complete set of features. Thus, sparse representation and learnable inputs might be less suitable in such contexts. Nevertheless, we note that sparsity (as shown in the benchmarking studies) and learning-based approaches are key features in the construction of the adjacency matrix, while preserving the original state of the node features for computing attributions.

Studies that used a mix of both BG and PG tend to perform better than solely using either graphs. Such an approach delivers the best of both worlds (BG allows efficient encoding of input data while incorporating information from the adjacency matrix, while PG allows meta-data such as demographic information to be incorporated easily). However, such an approach inevitably introduces more trainable parameters to the model and without proper benchmarking done, it is uncertain whether the improvements reported are simply a result of larger model capacity (i.e. if the BG-only or PG-only model were to be scaled up to the same number of parameters, they might close up the performance gap).

One major issue with the predominant approach is the choice of brain atlases: in particular, we found that the AAL atlas is a popular choice, being used in 58% of the studies. While the consistent use of the same atlas across studies has the advantage of reducing one source of variability for future meta-studies and reviews, the suitability of the choice of atlas for the target population is often a point of contention. AAL, being derived from a single 'healthy' young adult, might not be

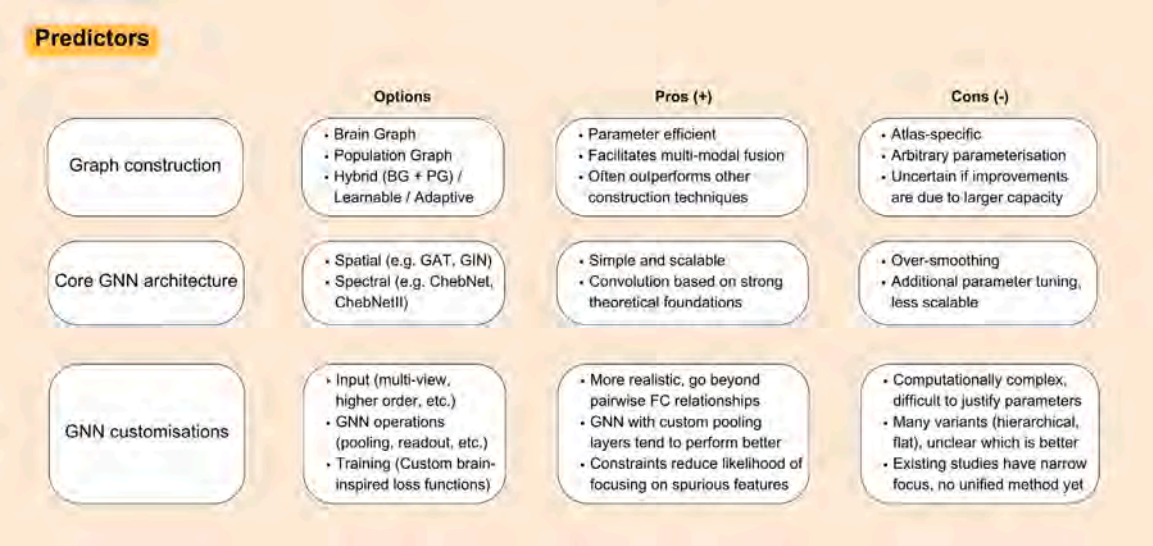

Fig. 4. Key strengths and weaknesses of various aspects of predictors.

the best choice for studies on adolescents or children (e.g. for ASD, ADHD). Additionally, the versions most frequently used (AAL90 and AAL116) have fewer parcels than other atlases, resulting in large ROIs that provide less value as they represent sizable areas of the brain that would be too heterogeneous for any meaningful insight to be drawn. Some studies have proposed multi-atlas strategies, but it is still unclear how best to reconcile disagreements in attributions across atlases (e.g. if there are overlaps in ROIs/voxels across atlases, but they receive very different attributions). Thus, it would be necessary for such issues to be addressed, or alternative dimensionality reduction techniques (e.g. ICA, functional gradients) could be more thoroughly analyzed in future studies.

In terms of the GNN construction, neither spatial nor spectral GCN has a clear lead over the other in terms of model performance. However, out of the classes of GNN architectures, the pooling-based approach featured more frequently among the best-performing models. Most of these advancements provide a clear value over baseline GNNs in terms of their utility for biomarker discovery as the unconstrained nature of baseline GNNs opens it up to the problem of learning spurious relationships. However, many studies did not control nor report the model size (i.e. number of parameters) and do not always compare with the latest models (in part, due to the different focus each model typically has). Thus, it would be prudent to rely on new benchmarking studies, with state-of-the-art GNNs considered, that control for the model size before coming to any conclusions. However, even if significant improvement in model performance is introduced by these novel techniques, another crucial consideration is the impact on the model attributions. For instance, the impact of using synthetic data for graph augmentation in self-supervised pre-training has not been thoroughly evaluated for biomarker discovery purposes (e.g. Does the introduction of synthetic data lead to spurious biomarkers? Do the additional datasets used for augmentation drown out actual signals from the original dataset of interest?).

3. Computing feature attributions via model explainability techniques

Although machine learning models trained on fMRI data can perform disorder classification well, they are rarely adopted clinically (Duda et al., 2023). Thus, studies involving such models often provide additional insights via various model explainability approaches. One type of explanation involves deriving feature attributions from a model trained on disorder classification tasks. These attributions represent the importance of each feature. Features with high attribution scores could be potential biomarkers of the disorder, i.e. identifying characteristics of the disorders (e.g. particularly low connectivity between 2 ROIs, across the population with the disorder and not present in the healthy population).

Early studies analyzed neuroimaging data to study disorders using univariate (Coutanche et al., 2011) or multivariate statistical methods (e.g. multivariate pattern analysis) (Klöppel et al., 2012; Wolfers et al., 2015). These methods provide coefficients of variables showing the significance of biomarkers and enable the generation of statistical models with high diagnostic or predictive potential by focusing on patterns of brain changes that are distributed across multiple regions in disordered states. ML techniques have also been commonly used to select important features for brain disorder classification (Orru et al., 2012; de Filippis et al., 2019). A model that generalises well to the test set and other unseen data points would be expected to have learnt optimal coefficients for each feature and feature importance could be inferred from these values. Features with the highest importance scores could represent potential biomarkers.

With the advent of deep learning techniques, multiple layers of non-linearities are introduced to learn complex relationships between the input and outputs. Furthermore, GNNs make it possible to model FC directly, obviating the need for feature engineering. Although they are often viewed as black boxes, the DNN's decisions could be analyzed via model explainability algorithms such as gradients (Chen et al., 2024), IG, or class activation mapping (CAM) (Qin et al., 2022).

To consolidate the findings from these studies, one could conduct meta-analyses by calculating the contribution of identified biomarkers to specific disorders. For example, BrainMap (Fox et al., 2005) is a database of Montreal Neurological Institute (MNI) coordinates for activation foci consolidated from thousands of experiments. Researchers can extract existing relevant studies from BrainMap using specific keywords, revealing experiments where both activated and non-activated ROIs were pinpointed and indicating their relevance to various disorders. Subsequently, methods such as the Naive Bayes classifier can be used to determine the probability of disorders associated with these ROIs (Yarkoni et al., 2011).

Despite such progress in modeling techniques and efforts to curate larger datasets, biomarkers for brain disorders remain elusive. For instance, efforts to identify diagnostic biomarkers for depression could not arrive at any consistent depression biomarker despite extensive efforts (Winter et al., 2024). Thus, in the next sections, we introduce the literature on model explainability techniques (both model-agnostic

and GNN-specific) with the goal of identifying research gaps (in existing fMRI studies on biomarker discovery) that could be addressed to improve the quality of these potential biomarkers.

3.1. Properties of model explainability techniques

Many model explainability algorithms have been proposed and they have been covered by numerous review papers (Linardatos et al., 2020; Tjoa and Guan, 2020; Marcinkevičs and Vogt, 2023; Yuan et al., 2022). We provide a brief summary of these feature attributors and place greater focus on key insights that are relevant to biomarker discovery. Before delving into the details of each attributor, we note several properties of attributors that are useful to characterise them.

Existing research on model explainability can be separated into three paradigms: (i) 'glass box' (intrinsically interpretable), (ii) 'black box' (reliant on post-hoc explainability methods), (iii) 'gray box' (some interpretation possible, with careful design of the model) (Ali et al., 2023). On one end of the spectrum, 'shallow' models such as linear regression are intrinsically interpretable. In the case where all input features have the same scale, biomarkers can be extracted by identifying features that have the largest coefficients assigned to them by the model fitting process. On the other end of the spectrum, deep learning models learn complex and non-linear relationships that cannot be easily interpreted. They often rely on post-hoc model explainability algorithms that are applied after model training. These algorithms typically generate scores based on some form of gradient computation (with respect to the input) or perturbation. In between these two extremes, some complex models such as fuzzy rule-based systems and Bayesian networks can provide a limited extent of interpretability (Ali et al., 2023). The use of attention scores as well as graph pooling could also be grouped under this category.

While 'glass box' methods are desirable, most of these methods only capture linear relationships, which is likely to be of limited use for disease studies. Unlike how studies on using fMRI data to predict phenotypic information have shown that non-linearities do not give much improvement over linear models, deep learning models have shown better performance for disease classification and prediction of clinical test scores (Zhang et al., 2020). However, many of these models fall under the 'black box' category and more research is needed to create 'gray box', or even 'glass box' alternatives.

All three paradigms of attributors produce attribution scores that can be classified into two categories: local ('instance-level') and global ('model-level'). Local explanations are specific to the input data provided to the model (i.e. each sample has its own attribution scores), while global explanations apply broadly to the entire model (all samples share the same attribution scores). In the context of biomarker discovery, local explanations could be more desirable for clinical use if individual insights are found to be reliable. Additionally, global explanations are unlikely to be helpful for very heterogeneous diseases since heterogeneous diseases would not be fully described by a single set of global attribution scores.

3.2. Model explainability for GNN

GNNs can be difficult to interpret due to the non-Euclidean contextual information in the graph (node features and edge weights) that needs to be taken into account when computing attributions. For BG, attributions can be produced at the level of nodes (i.e. which ROIs contribute the most to the disorder), edges (i.e. which pairs of ROIs contributed most), and node features. For PG, attributions can be produced at such granularity too, but they carry a different meaning: nodes would correspond to patients, edges correspond to pairs of patients and node features would typically correspond to some form of imaging data.

Existing model explainability methods for GNNs can be split into two major groups: self-interpretable and post-hoc. Self-interpretable methods provide explanations simultaneously with the model predictions, while post-hoc methods are only applied after model training is complete. Self-interpretable methods typically include constraints to extract an informative subgraph, or architectural designs where weights that prioritise a subgraph are learnt. Post-hoc techniques can be further split into model-agnostic and GNN-specific techniques. Model agnostic algorithms can be applied to any deep learning model (and for GNNs, regardless of their internal structure), but some of them have been further extended to capture the graph structure (Pope et al., 2019). GNN-specific methods explicitly consider the graph structure when generating the explanations. Model agnostic methods can be categorised into two major groups: gradient-based and perturbation-based. GNN-specific methods follow such a characterisation too but also have unique ones such as techniques based on graph generation. Fig. 5 summarises the taxonomy of GNN explainability methods and in the following subsections, each subcategory will be explained in detail and discussed in the context of biomarker discovery.

3.2.1. Self-interpretable - Structural constraints

Drawing inspiration from CNNs, kernel GNN (**KerGNN**) introduces learnable graph filters (akin to convolutional filters in CNNs) into the messaging passing process in GNNs (Feng et al., 2022). Contrary to the rooted subtree approach used in GNNs that are based on the message-passing paradigm, KerGNN updates each node's embedding based on subgraphs that are centered on the node. This is done via the use of graph kernels (specifically, the random walk kernel) to measure the similarity between subgraphs and the graph filters. Not only does such an approach make GNNs more expressive (going beyond the 1-dimensional Weisfeiler-Leman (1-WL) limit), it also provides additional interpretability via the learnt graph filters, just like how convolutional filters can be visualised.

Such an approach would be useful in the case of disorders that are poorly understood as novel insights could be drawn from these visualisations. However, if there is some existing biological knowledge available, Factor GNN (**FGNN**) (Ma and Zhang, 2019) could be considered. Unlike usual deep learning models where the hidden nodes in deep learning models do not have a physical meaning, nodes in FGNN represent biological units that could be either observable variables or latent variables. Through such representations, FGNN directly incorporates biological knowledge as the inductive bias into the model.

While there have not been any applications of FGNN on fMRI datasets (the original paper applied it on genetic data), one example of how it can be applied is to use functional brain modules (i.e. a group of ROIs that are typically well-connected to each other, yet less connected to ROIs not in the module) as the latent variables and ROIs as the observable variables. This forms a factor graph that can be used to guide the construction of the neural network. Unlike a fully connected layer, input nodes (observable variables) representing ROIs will only be connected to the intermediate layer (latent variables) if the ROI belongs to the module. Then, the hidden layer could be used as input to a fully connected layer to make predictions (e.g. clinical outcomes), or be passed to another stack of factor graphs, forming a deep network.

3.2.2. Self-interpretable - Informational constraints

Instead of introducing architectural designs that guide the learning process (and identify salient subgraphs), methods based on information constraints introduce information bottlenecks such that the mutual information (MI) between the labels and the discovered subgraph is maximised, while keeping the MI between the original graph and subgraph below a predefined threshold. While the former can be approximated via cross-entropy loss, the latter is estimated via techniques such as learnable randomness injection (**LRI**) (Miao et al., 2023) and graph information bottleneck (**GIB**) (Yu et al., 2021). The latter has been further developed in the fMRI biomarker discovery literature by BrainIB (Zheng et al., 2024c). It extends GIB by considering the effects of edges (not just nodes) during subgraph discovery.

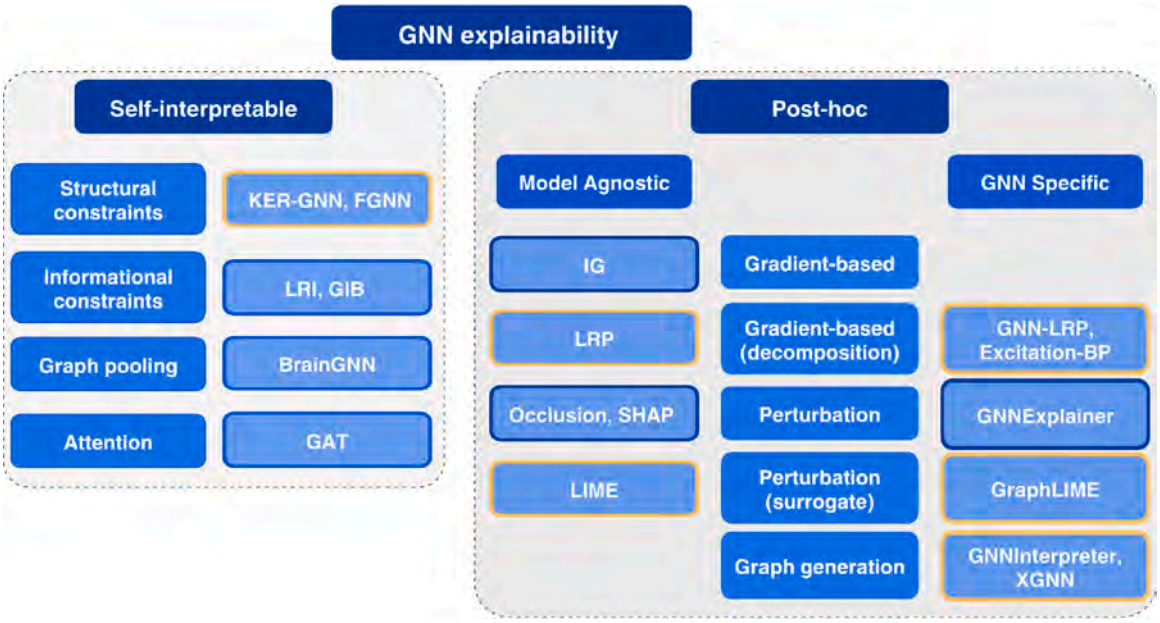

Fig. 5. Taxonomy of attributors applicable to GNNs. Generally, post-hoc methods (typically 'black box') have a larger range of algorithms than self-interpretable (typically 'gray box') ones. Methods that have not been explored in fMRI studies are highlighted in yellow.

3.2.3. Self-interpretable - Graph pooling

Graph pooling is often performed in GNN architectures, especially in graph classification tasks where node features have to be condensed to a lower dimensionality or a single vector. Pooling techniques can be grouped into two categories: (i) flat pooling, where a graph-level representation is generated in one step, and (ii) hierarchical pooling, which gradually coarsens the graph by clustering nodes together or dropping some of them (Liu et al., 2023).

Several pooling techniques customised for FC data have been proposed. Li et al. (2021b) proposed a node/ROI pooling layer (R-pool) in their **BrainGNN** architecture. R-pool projects the node feature embeddings to a learnable weight vector and retains nodes with the highest scores. Hierarchical pooling approaches that consider functional modules have also been proposed (Mei et al., 2022). In this work, three levels of hierarchy were used: (i) ROIs belonging to the same sub-network (e.g. Yeo 7-network parcellation Yeo et al., 2011) and brain hemisphere, (ii) the pair of matching sub-networks from each hemisphere, (iii) combining all sub-networks into a whole brain network. Weights from the final pooling layer were used to identify the sub-networks that contributed most to the model's decision.

3.2.4. Self-interpretable - Attention

The widespread use of attention for model interpretability has also been present in the GNN literature, most popularly via self-attention in **GAT**. Attention scores have been used to identify salient FC features as well (Zhang et al., 2022b; Yu et al., 2022). However, there has been much debate in natural language processing (NLP) research about whether attention scores provide meaningful explanations (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Serrano and Smith, 2019; Bai et al., 2021a). In NLP applications, attention scores were found to not correlate well with multiple gradient-based approaches on recurrent neural networks and different attention distributions can lead to equivalent predictions (Jain and Wallace, 2019). However, Wiegreffe and Pinter (2019) argue that explanations should be further categorised into 'plausible' or 'faithful' explanations. Their results provide additional support that attention does not provide faithful explanations but does not invalidate claims that attention provides plausible explanations. These results suggest that greater care should be taken when using attention scores to discover potential biomarkers. This has been further substantiated in a recent mathematical study of a single-layer multi-head transformer architecture, which showed that post-hoc methods provide more insightful explanations than attention weights (Lopardo et al., 2024). Further research is needed to examine the validity of attention scores for biomarker discovery applications. For instance, Safai et al. (2022) found that the attention weights of a GAT model used for Parkinson's disease classification could be used to identify which brain regions contribute to the classification accuracy, but they vary significantly across attention heads. While they analyzed each head separately, it is currently still unclear how findings from different attention heads of the same GAT model should be reconciled.

3.2.5. Post-hoc - Gradient-based

A large variety of gradient-based approaches have been proposed. Integrated Gradients (**IG**) (Sundararajan et al., 2017) will be discussed here due to its versatility (works on most deep learning models where gradients can be calculated) and widespread use. IG was developed to address the issue of saturation in gradient-based attribution methods. With saturation, the output is no longer sensitive to small changes in the input features, making it difficult to interpret which features are responsible for the prediction of the correct class (Sturmfels et al., 2020). IG avoids this issue by accumulating gradients from interpolated points (controlled by α) between a baseline (x') and the input data (x). In the context of disorder classification, the baseline can be the average data of all healthy subjects. For a given feature k, IG is defined as:

$$IG_k(x) = (x_k - x'_k) \times \int_{\alpha=0}^{1} \frac{\partial f(x'_k + \alpha \times (x - x'_k))}{\partial x_k} d\alpha.$$

Attribution scores from IG provide a local measure of how much the feature contributed to the model's prediction and could be potentially useful for producing personalised insights. At present, its use is still limited to group-level insights, e.g. identifying site-specific and site-invariant biomarkers of SZ (Chan et al., 2022). This is in part due to how IG can also produce noisy pixel attributions in features unrelated to the predicted class. Modified versions of IG, such as GuidedIG (Kapishnikov et al., 2021), have been proposed to address these issues. Features with high gradient scores have a greater impact on predicting the class of the model than features with low gradient scores. Thus, GuidedIG reduces noise in the attribution results by using an adaptive path technique that only incorporates a subset of features with high gradient scores.

3.2.6. Post-hoc - Decomposition

Decomposition-based approaches have some overlaps with gradient-based approaches but differ in the way the scores are computed. Instead of computing the gradients directly with respect to inputs, scores are decomposed starting from the output layer and propagated backwards in a layerwise manner based on pre-defined rules. Layerwise relevance propagation (**LRP**) is one such example. Many forms of LRP exist and the variant called ϵ-LRP will be discussed (Montavon et al., 2019). Starting from the output layer, a score s is assigned to a neuron based on the logit, i.e.

$$s_i^l = \frac{h_{ji}}{\sum_i h_{ji} + \epsilon(\sum_i h_{ji})} s_i^{l+1},$$

where h_{ji} refers to the output from neuron i in layer l to neuron j in layer $l+1$. This is computed layerwise, reallocating the prediction score until the input layer is reached. The total relevance score of s is always preserved for each layer.

LRP does not consider the adjacency matrix in its computations. To address this, **GNN-LRP** (Schnake et al., 2021) distributes scores to different graph walks (and thus have higher computational complexity). **Excitation-BP** is very similar to LRP, but it views the decomposition process from a probability standpoint. Overall, recent decomposition-based approaches like GNN-LRP have not been well-studied in fMRI, with usage mainly limited to LRP (Yan et al., 2017).

3.2.7. Post-hoc - Perturbation

Perturbation-based approaches introduce changes to the input with the motivation that if important features are still retained, the outputs should remain similar. In its simplest implementation (**Occlusion**), this involves masking features one by one and the feature that results in the largest change in output would be deemed to be the most important. **SHAP** (Lundberg and Lee, 2017) takes this idea to completion by considering all feature subsets (i.e. 2^k combinations), so as to compute Shapely scores that have been proven to be the unique solution that fulfills the criteria of local accuracy (model training on best feature subset should have similar predictions with the original model), missingness (features not in the best subset should have no impact on the model output) and completeness (attribution score should not decrease when a different model, where the feature contribution does not decrease, is used). Computing this is too computationally expensive, thus it is achieved via approximation techniques.

In the context of graphs, such perturbations can be performed by discovering subgraphs. **GNNExplainer** (Ying et al., 2019) produces local explanations for GNN predictions by selecting a small subgraph from a given input graph and identifying important node features. Subgraphs are generated by randomly masking nodes in the graph and observing the resulting changes in the model's prediction. A soft mask (i.e. continuous values, not binarised) containing learnable weights is used. Important node features (that are in the nodes within the subgraph) are identified by a binary feature selector. The mask and feature selector are optimised by maximising the mutual information between the original model predictions and the model's predictions given the masked graph. One limitation of GNNExplainer is that the subgraph must be connected, which might not always be applicable to disease biomarkers. Nevertheless, it has been quite popularly used in fMRI studies (Gallo et al., 2023; Hu et al., 2021).

3.2.8. Post-hoc - Surrogate

Surrogate-based approaches have some overlaps with perturbation-based approaches as they tend to rely on perturbations too. However, a distinctive feature is the use of simpler and interpretable models (often linear) to approximate the original complex model. This is possible as it limits the approximation to a local neighborhood and analyze the model predictions of perturbed inputs within this neighborhood. For instance, local interpretable model-agnostic explanations (**LIME**) (Ribeiro et al., 2016) trains a surrogate model on the dataset of perturbed points, weighing them based on their proximity to the chosen data point.

GraphLIME (Huang et al., 2022) is a non-linear version of LIME, a key difference being that it uses Hilbert–Schmidt Independence Criterion (HSIC) lasso, a kernel-based non-linear interpretable feature selection algorithm. The algorithm first computes the importance of each feature in each node by considering the features of the target node and features in the N-hop neighboring nodes. The given target node will aggregate the information from N-hop network neighbors to identify the most significant features. The HSIC lasso method is used to train a linear interpretable model to represent the relationship between the features and target node prediction. Subsequently, the coefficients from the linear interpretable model will be used to identify the top few features that are important for model prediction based on the coefficients. Thus far, no research on FC has used such techniques for biomarker discovery.

3.2.9. Post-hoc - Graph generation

Attributors based on graph generation bear some similarities with GNNExplainer as both identify salient subgraphs. However, generation-based approaches arrive at the subgraph via generative approaches, instead of perturbation. **XGNN** (Yuan et al., 2020) interprets GNNs using a graph generator to identify important graph subgraphs. The generator is trained using reinforcement learning (RL) and validity rules are defined by pre-existing knowledge, making them less suitable for biomarker discovery. On the other hand, **GNNinterpreter** (Wang and Shen, 2023) do not require pre-existing knowledge as it optimises the choice of subgraph by maximising the similarity between embeddings from the important subgraph with that of the average graph embeddings in the target class. Both attributors are different from all other attributors discussed above as they produce global explanations (i.e. one set of explanations for the whole model, across all data points). Thus far, no studies on FC have used these post-hoc attributors based on graph generation for biomarker discovery.

3.2.10. Summary

Fig. S18 provides a visual breakdown of the attributors used in the papers included in this review. It is evident that attributors based on informational constraints, pooling, attention, gradients, and perturbation have been used on FC datasets. Notably, several attributors that are not graph-based have also been tested, e.g. trainable masks, clustering, weights, and variants of Gradient-weighted Class Activation Mapping. However, there remains much room to explore alternative attributors based on graph generation, decomposition, surrogates, and structural constraints.

When choosing an attributor for biomarker discovery, it is crucial to first understand which types of explanations the method can provide (what granularity, what attribution targets are possible, and whether the attributor requires training). As a useful reference for future research, Table 1 summarises these key characteristics of attributors highlighted in Fig. 5. Notably, GNN-specific attributors (e.g. GNNExplainer) have the advantage of being able to identify subgraphs, while generic attributors (e.g. IG) are typically limited to node features.

Overall, there are still several types of attributors not explored in fMRI studies yet. Furthermore, thorough benchmarking studies would be needed to verify the robustness of these attributions, beyond the existing predominant practice of qualitatively comparing the top-k attributions against existing work.

3.3. Evaluation of strengths and weaknesses

Fig. 6 summarises the relative merits of each group of attributors. Self-interpretable attributors have a clear advantage over post-hoc methods as they do not need a separate attributor, which at times requires training yet another deep learning model. However, many of them fundamentally rely on the model learning some form of weight

Table 1
Key characteristics of attributors. Methods above the line are self-interpretable, while those below are post-hoc methods.

Attributor	Approach	Granularity	Target	Trainable
KER-GNN	Structural	Instance	Node features	Yes
FGNN	Structural	Instance	Node features	Yes
LRI	Informational	Instance	Subgraph	Yes
GIB	Informational	Instance	Subgraph	Yes
BrainGNN	Pooling	Instance	Node	Yes
GAT	Attention	Instance	Edge	Yes
IG	Gradients	Instance	Node features	No
LRP	Decomposition	Instance	Node features	No
GNN-LRP	Decomposition	Instance	Edge	No
Excitation BP	Perturbation	Instance	Node features	No
Occlusion	Perturbation	Instance	Node features	No
SHAP	Perturbation	Instance	Node features	No
GNNExplainer	Perturbation	Instance	Subgraph, Node features	Yes
LIME	Surrogate	Instance	Node features	Yes
GraphLIME	Surrogate	Instance	Node features	Yes
GNNInterpreter	Graph generation	Model	Subgraph	Yes
XGNN	Graph generation	Model	Subgraph	Yes

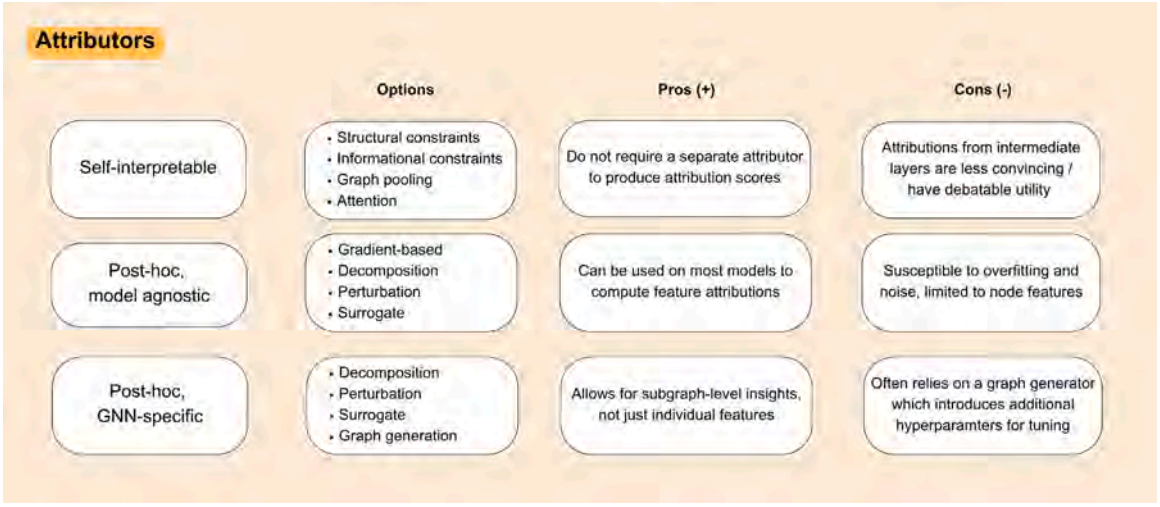

Fig. 6. Summary of the relative merits and limitations of each group of attributors.

matrices, which are designed to be consistent with the input feature dimensions, and thereby seen as a measure of importance. This is sensible for linear models, e.g. the coefficients learnt by a linear regression model with features of the same scale could be interpreted as a measure of feature importance. But when this is applied in the context where non-linear functions are used to transform both the input and outputs of the layer where the weights are extracted from, it might be too simplistic as these representations are not humanly understandable at present, even though they might still seem to correspond to individual ROIs. Given the non-linear transformations, there is little guarantee that a node with higher weight would contribute more greatly to the model's final decision. While existing studies attempt to demonstrate the correctness of their attributions cross-referencing with other research studies, the uncertainties brought about by the lack of understanding of these opaque operations warrant stricter evaluation techniques that go beyond the current practice.

Creating a foolproof technique to evaluate attributors could be challenging due to the lack of ground truth as well as how many of the novel GNNs introduce modifications that are specific to the human brain. The use of other baseline graph datasets with known ground truth, such as MUTAG, could be considered as a proxy for methods that are not tied to neurosciences. In the other scenario, one could attempt to demonstrate the validity of their techniques via synthetic datasets (Agarwal et al., 2023). Post-hoc methods could also be applied on top of these self-interpretable methods to assess the extent of agreement between both methods.

Comparing model agnostic and GNN-specific post-hoc explainability techniques, the most critical factor is often the target that the attributor is able to compute contributions for. In that regard, GNNExplainer provides the most flexibility as it is able to compute both motif-level and node-level contributions. On the other hand, many methods are only limited to a single target type. While several studies have demonstrated how contributions at the level of node features (i.e. edges, when connection profile is used) can be consolidated to nodes, the appropriateness of these aggregations should be studied in a more robust manner, e.g. testing the validity of such operations on synthetic data with known ground truth.

4. Evaluation of attributions

Better biomarker discovery tools are needed as few potential biomarkers turn out to be effective in clinical settings (Parkes et al., 2020). To bring us closer to this goal, one solution could involve developing objective means of assessing the robustness of attribution scores produced by these attributors. Robustness entails the expectation that (i) attributions should capture valid signals (e.g. features that are indeed idiosyncratic characteristics of the disorder) and that (ii) the scores should be reasonably invariant to non-disorder related factors while retaining sensitivity to disorder-related changes. Verification of validity is a challenge for psychiatric disorders as the ground truth is often unavailable. However, it is possible to assess the latter with

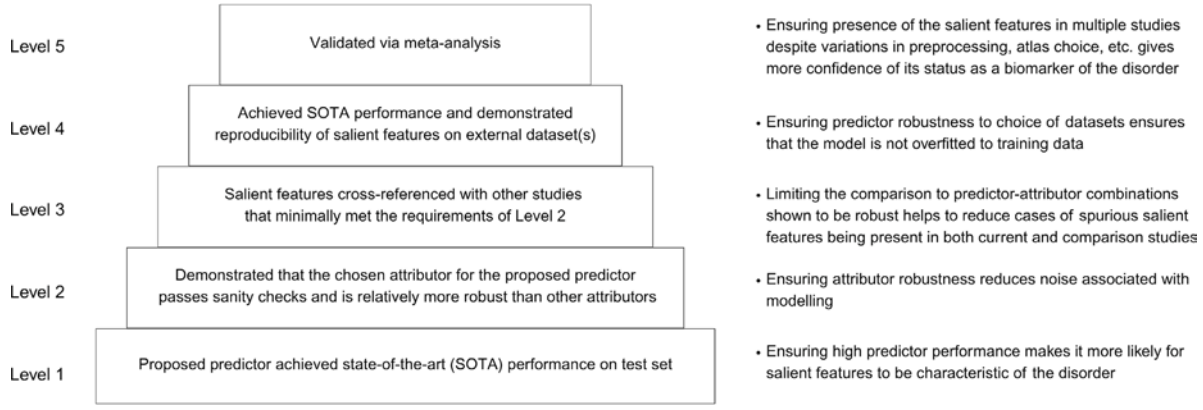

Fig. 7. Proposed framework (Hi-5) for evaluating the robustness of salient features derived via machine learning pipelines.

appropriate metrics defined in the context of deep learning models. For instance, it would be unreasonable for attributions to vary significantly just by changing the size of the hidden channels of a GNN — salient features highlighted by such models would be deemed to be unreliable. The complexity of deep learning models necessitates a comprehensive framework to characterise these factors of variation.

4.1. Hierarchy of robustness for salient features derived via machine learning

In the context of biomarker discovery via ML, there are 3 key factors influencing attribution scores: data, predictors and attributors. Attributors perform the main computation of the scores and predictors have a significant influence in both self-interpretable (directly integrated into the model training process) and post-hoc settings (dependent on the model weights or gradients that are arrived at after model training). However, variations of scores due to the choice of predictors and attributors are artifacts of the modeling process and should be isolated from the biological factors of variation (e.g. datasets covering different subpopulations could have different biomarkers). To disambiguate them, we define a hierarchy of robustness (abbreviated as Hi-5, illustrated in Fig. 7) which consists of the following 5 levels:

1. **State-of-the-art (SOTA) performance achieved on test set**: The proposed predictor is compared against existing state-of-the-art approaches and clearly outperforms them. Achieving the best performance gives greater confidence that the model has learnt meaningful relationships between the data and the labels and thus it would lead to accurate attribution scores. Most existing studies stop at this stage but there is little guarantee that the reported biomarkers are actually reproducible since only 1 dataset and 1 attributor are involved.
2. **Demonstrated that the attributor passes sanity checks and is relatively more robust**: Having high performance on the test set does not guarantee that the model has picked up the correct relationships between data and labels. Attribution scores serve as another layer of checks by providing an alternative method to assess the model performance. However, evident from the results shown in the following sections, the choice of the attributor has a strong influence on the scores. Thus, there is a need to test out multiple attributors. A few papers have done this to a limited extent (Gallo et al., 2023; Li et al., 2023). While performing such evaluations might seem tedious, several metrics have been used as a proxy and this is visited in deeper detail in the next subsection.
3. **Salient features are cross-referenced with other studies that at least passed Level 2**: Although this has the same issues of subjectivity (e.g. which studies to include/exclude) in existing papers that perform such comparisons, the requirement for both attributor robustness and cross-study reproducibility makes it less likely for spurious biomarkers to remain. Salient features that meet this criterion could be considered to be potentially robust, especially if the studies considered are significantly different (e.g. different predictors and attributors used).
4. **SOTA performance achieved on external dataset(s)**: The predictor trained on the original dataset should exhibit good generalisability to a wider variety of data beyond the test set, ideally with no finetuning needed. Furthermore, the final set of reported biomarkers should be present in both the original dataset and the external dataset(s). This has to be carefully evaluated keeping in mind the extent of similarity between the training data and the external dataset, as well as the extent of heterogeneity known to be present in the disorder. Additionally, due to practical constraints in individual studies, it is likely that the comparisons are limited to studies that use the same preprocessing pipeline and atlas as the original study.
5. **Demonstrate through meta-analysis that the biomarker is valid**: At this stage, a wider variety of studies is considered (e.g. different preprocessing pipelines) and biomarkers from studies using different atlases are harmonised. A method to perform this is introduced in Section 4.3 and it serves as a basic demonstration/proof of concept. At a more advanced stage when there is a larger quantity of studies available, meta-analysis should consider heterogeneity of the disorder and carefully select studies included in the analysis.

Put together, Hi-5 provides a framework to assess robustness in 3 key aspects: predictor robustness (model is not overfitted to the training set/does not only capture site-specific salient features), attributor consistency (attributions are not dominated by attributor-specific artifacts that can be manipulated simply by changing attributors) and cross-study reproducibility (salient features picked up by the model are indeed present across multiple datasets). Based on the Hi-5 framework, we define ML-derived salient features to be *potentially robust* if it passes Level 3 (i.e. has predictor robustness and attributor consistency, but not demonstrated to have cross-study reproducibility), and *robust* if they fulfill the criteria in Level 5 (fulfills all 3 above-mentioned aspects). For the avoidance of doubt, this definition is focused on site-invariant biomarkers (i.e. features identified by the model to be important are indeed present in multiple sites/datasets).

Hi-5 is designed with a focus on evaluating salient features derived via ML, but it is not exhaustive. For instance, another desiderata would be for the salient feature to have a significant effect size. A robust biomarker that has a low effect size would not be of much utility as it would require prohibitively many samples to demonstrate the effectiveness of an intervention. Moreover, rare disorders or rarer subtypes of common disorders might not be able to achieve Level 5

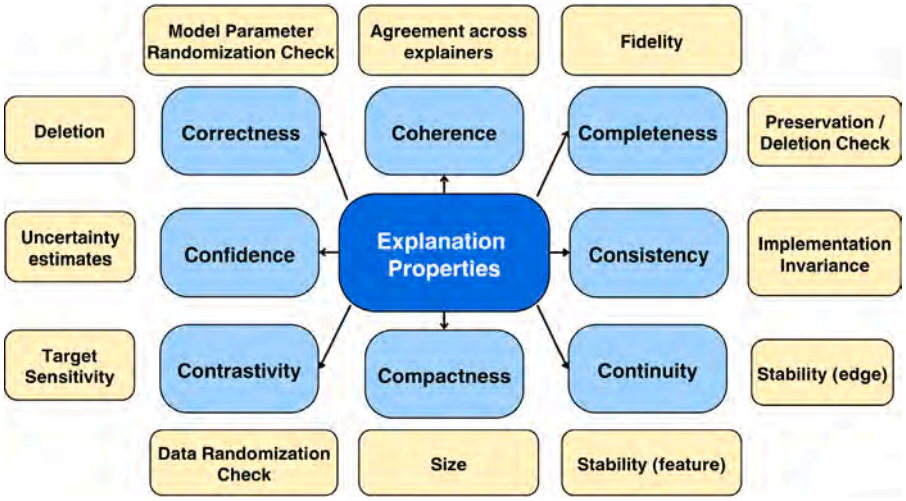

Fig. 8. A subset of the Co-12 properties (Nauta et al., 2023) used to evaluate feature attribution scores produced by attributors, deemed to be relevant for biomarker discovery from fMRI datasets.

due to the lack of datasets or existing studies. Alternative versions of the Hi-5 framework could be designed in future studies for such scenarios, e.g. including carefully designed metrics that considers the impact of dataset distributions on attribution score, such as signal-to-noise ratio maps (Dong et al., 2024). Additionally, Hi-5 in its current form is focused on group-level studies and discussion on how it could be adapted to generate individual-level insights can be found in the Discussion section.

4.2. Existing evaluation techniques

Several methods of evaluating generic attributors (Nauta et al., 2023) as well as GNN-specific attributors (Agarwal et al., 2023; Li et al., 2022c) have been proposed. Nauta et al. (2023) proposed a framework named 'Co-12' which encompasses 12 desired properties of attributors (such as Correctness, Completeness, Consistency, etc.). In this section, we discuss properties that are relevant to biomarker discovery and extend the analysis to GNN-specific attributors. Fig. 8 illustrates the 8 chosen properties and metrics that can be used to measure a model's performance with respect to the corresponding property. An explanation of why the 4 other components are not included can be found in Appendix. Additionally, we note that not all 8 properties have been covered by existing fMRI studies. For those that are covered, the relevant references have been added to the details presented in each subheading below.

Each property below encompasses one or more evaluation metrics. Many metrics involve comparing two distributions (e.g. original attributions versus attributions produced in a different setting) and measuring the distance between them. This can be computed via Hellinger distance, which has an easily interpretable range (0 = perfectly similar, 1 = completely different distributions).

4.2.1. Correctness

Sanity checks on the attributors can be performed via **model parameter randomisation check**, which verifies whether the explanation of a trained model is different from a randomly initialised untrained model. Other forms of evaluating correctness involve **deletion**: changes in model outputs are computed for each feature subset (in the simplest case, independently removing each feature in the node vector) and then correlated with the importance scores. However, this could be too tedious for deep learning models applied on connectome datasets especially if many ROIs are used.

4.2.2. Completeness

Salient subgraphs or feature subsets identified by the attributors should not lose too much information as compared to the original input. This can be assessed via **preservation check** (whether using the selected features as input results in the same model prediction) and **deletion check** (not using the important features results in a different prediction).

Fidelity is based on a similar idea but quantifies the difference by measuring the change in probabilities (as opposed to a simple change in prediction of classes). Note that this can be done both at the level of node features and at the level of subgraphs (in the case of graph datasets like FC). In recent fMRI studies, this has been analyzed at the level of node features (Hu et al., 2021; Safai et al., 2022; Menon et al., 2023). An example of how Fidelity+ (represented by v_{F+}) can be computed is shown below:

$$v_{F+} = \frac{1}{N}\sum_{i}(f(X_i) - f(X_i^+)), \tag{5}$$

where $f(\cdot)$ is the model under study, X_i represents the node features of subject i from a dataset of N subjects whereas X_i^+ represents the subset of node features where features with the highest attribution scores are removed.

4.2.3. Consistency

Explanations should be robust across most variations in model implementations. While deep learning models can involve many hyperparameters that have profound and poorly understood changes to the model, a basic form of assessment for **implementation invariance** could entail checking whether two models with the same architecture, but having different initialisation (i.e. randomly initialised weights), will produce similar outputs and explanations.

4.2.4. Continuity

Given a small change in the input that still produces a similar model output, it would be reasonable to expect that the explanation will remain similar too. **Stability** ensures this by introducing random noise to the input features. For graphs, additional changes can be introduced by randomly switching edges while keeping the number of edges constant. Both explanations (before and after perturbation) can then be compared to determine if the scores have good continuity.

4.2.5. Contrastivity

Explanations produced for data samples belonging to different classes would be expected to differ. This is verified via a **target**

sensitivity check, where the mean explanations for each class can be computed and then compared across classes. Alternatively, a **data randomisation check** can be performed by randomising the labels and verifying that the explanations are changed. Given that biomarker discovery techniques based on ML techniques are often based on disorder classification tasks, it would be important to check if the scores produced have high contrastivity.

4.2.6. Compactness

Explanations should not overwhelm the user and this is especially relevant for FC datasets, where hundreds of ROIs and thousands of connections are analyzed simultaneously in multivariate studies. For instance, while severe diseases might affect widespread areas of the brain, having an explanation that gives similar scores to an overwhelming number of ROIs or edges might not be as useful as explanations that correctly emphasise a small group of features. Computing the **size** of explanations is straightforward for attributors that produce subgraphs (e.g. number of connections divided by the number of edges in a complete graph), but would require thresholding in the case of most gradient-based approaches (since a score is given to each feature, but some scores could be very close to 0). This has been examined in Menon et al. (2023), where SubgraphX was shown to have higher sparsity than GNNExplainer and PGExplainer.

4.2.7. Confidence

While almost all attributors discussed above do not produce uncertainty estimates and many of their scores have no natural variations (e.g. Saliency is based on gradients with respect to inputs, which would be fixed given the same model and input), measures of confidence could be computed based on the logits of the predictor. One example of an implementation is proposed by Atanasova (2024), where class probabilities from the predictor are compared against a predicted confidence value estimated by fitting a logistic regression model to saliency distance values (which computes the distance between saliency across in each class). A similar approach could be used for biomarker discovery applications.

4.2.8. Coherence

Since the ground truth is often unknown in biomarker discovery, coherence would be focused on the extent of agreement between attributors. Biomarkers that are present across attributors based on different approaches could be deemed to be more robust. However, further studies need to be conducted to determine the attributors to use (e.g. attributors with low stability should not be considered). Many existing fMRI studies that analyzed robustness chose to evaluate coherence, revealing that attribution can vary significantly (Li et al., 2023) but there are still instances where there are ample overlaps (Gallo et al., 2023).

4.2.9. Summary of evaluation metrics for assessing attributor robustness

Overall, we found that only 6 out of the 65 fMRI studies covered in this review have used objective metrics to evaluate the robustness of the potential biomarkers highlighted by their model. This is in line with past reviews on the use of interpretable deep learning techniques in more general medical settings which found that only a minority of existing studies (13 out of 75) evaluated robustness (Munroe et al., 2024).

Additionally, most studies are limited to exploring only one aspect of robustness (e.g only using one metric). There is a significant research gap in the area of designing multi-faceted evaluation approaches. An example of how these can be studied is demonstrated in Girish et al. (2024). In essence, metrics that can be customised to focus on top-k features (e.g. Fidelity) can be used to measure the robustness of attributions, while other more generic metrics could serve as a sanity check of the predictor–attributor combination.

4.3. Quantifying cross-study reproducibility

While the above metrics provide sanity checks that make it easier to check for the criteria in Level 2 of the Hi-5 framework (attributor robustness), there remain issues that complicate assessments at higher levels of the hierarchy. At the top (Level 5), studies using different atlases need to be harmonised to derive aggregated insights from them. The difficulty of the task is compounded by the lack of standardisation in the way salient features are reported — many papers only note down a few top features which they can find support from the literature (possibly missing/leaving out some), some studies present a table with the features ranked in order of importance (but the number of features reported varies wildly between 5 to 20) and very few studies share attribution scores in a format usable for thorough analysis.

Depending on the availability of this information, there are 3 scenarios in which insights can be aggregated from these studies: (i) simple frequency counting, (ii) rank-weighted aggregation, (iii) attribution score-weighted aggregation. For each scenario, let S_v represent a reproducibility score for ROI v. Although this can be computed easily, each atlas has its unique set of ROIs that might or might not overlap with those in other atlases. To harmonise them, we create a brain map M, which could be any parcellation map, but ideally should combine both cortical parcellation and subcortical segmentation masks. Let $M_A = \{v_1, v_2, \dots, v_i\}$ represent an area/parcel A of the brain map that contains several salient ROIs. We can then compute the reproducibility score for area A as:

$$S_A = \frac{1}{|M_A|} \sum_{v \in M_A} S_v.$$

Let N represent the total number of studies for a disorder and N_v represent the number of studies ROI v is highlighted as a salient feature. Then, depending on the availability of information, S_v can be defined as:

$$\begin{cases} S_v = \frac{f_{v_i}}{N} & \text{if only counts are available,} \\ S_v = \frac{N-(R[v_i]-1)}{N} = 1 - \frac{1-R[v_i]}{N} & \text{if ranks are available,} \\ S_v = \frac{T_{v_i} - min(T)}{max(T)} & \text{if attribution scores } T \in \mathbb{R}^{\mathbb{K}} \text{ are available.} \end{cases} \tag{6}$$

Based on the studies included in this review paper, the first approach (based on simple frequency counting) would be most feasible for now as there are too few studies that provide ranks or attribution scores. Having this information would have allowed more meaningful meta-analysis. Nevertheless, visualisations of consolidated group-level biomarkers could still be produced via the first approach. These are presented in the next section and implementation details can be found in Appendix.

4.4. Summary

In this section, we have proposed a framework for evaluation of attributions, named Hi-5, and discussed in detail ways to implement Level 2 and Level 5 of this hierarchy of robustnesses. Linking back to the second question mentioned in the introduction (how existing studies evaluate the robustness of their biomarkers), we found that most studies are currently at Level 1 while a few have made initial attempts at Level 2. Such studies that assessed the robustness of attributors were largely limited to aspects such as coherence, completeness, and compactness. Thus, there are still many other aspects such as continuity, contrastivity, and correctness that should be considered in future studies. We hope that this formalisation of a hierarchy of robustness will encourage more studies that achieve higher levels of robustness.

5. Application of GNNs on disorder prediction and biomarker discovery from fMRI data

Biomarkers are biological signatures that can be objectively measured and used to indicate physiological processes, pharmacological responses, or disease status (Jain and Jain, 2010). Diagnostic biomarkers refer to a specific type of biomarker that can be used to identify patients with a particular disorder or condition (Califf, 2018). In this paper, we focus our discussion on potential diagnostic biomarkers of brain disorders derived from fMRI via GNNs.

The salient features detected from trained models are potentially disease biomarkers, but it is uncertain whether they capture true signals of the disorder, or are mere artifacts of noise. To ascertain the usefulness of these potential biomarkers, they need to be evaluated against the following traits: easily accessible, reproducible, specific, and sensitive. Most importantly, it should only change depending on the state of the disease and be unchanged when unrelated factors are varied (Aronson and Ferner, 2017). Most existing studies on neuroimaging biomarkers for brain disorders have not attained such high standards yet and bridging this gap has been a priority for neuroimaging research in recent years (Parkes et al., 2020).

In the following subsections, we summarise key biomarkers identified from the studies included in this review. A key goal of this review is to identify salient features that are reproduced across multiple studies. In view of this, we focus on papers that include resting-state fMRI due to its widespread availability as compared to task fMRI scans. Taking into account that several papers tested their models on multiple datasets and disorders, we found that ASD dominates existing research (53.8% of papers considered in this review) along with MDD (26.9%), while SZ (14.1%) and ADHD (5.1%) are relatively rare.

Salient features present in multiple studies would be more promising biomarkers of the disorder as they exhibit greater robustness than non-reproducible ones. However, this is still a challenging task to identify them as they can vary in granularity: ROI, connections, module, modular connections, and temporal features. Furthermore, the choice of brain atlas and attributor are often different. Thus, to aid readability, the following tables are provided in the Appendix: the names of brain atlases used are abbreviated to make it easy to note how many ROIs are involved and the full names can be found in Table 2. Also, Table 4 provides a mapping of the abbreviations to the full names, while full names for abbreviations of functional modules and ROIs can be found in Tables 3 and 5 respectively.

5.1. Attention deficit hyperactivity disorder

ADHD is a syndrome characterised by the presence of inattentive and/or hyperactive (and impulsive) behavior to an extent that is age-inappropriate and often affects social, academic, and occupational performance. It is widely considered to have 3 broad subtypes based on behavior (combined, inattentive, and hyperactive/impulsive). It typically commences from early childhood years but could persist into adulthood.

Relative to other complex psychiatric disorders, the neural underpinnings of ADHD have been more clearly elucidated: disruption to the response inhibition (dorsomedial frontal cortex, anterior insula/inferior frontal cortex) (Aron et al., 2004) and neural reward processing (ventral striatum) circuits are associated with ADHD (Plichta and Scheres, 2014). Specifically, there is dysregulation of dopaminergic and noradrenergic systems in these regions (Del Campo et al., 2011). Even though there are treatments like psychostimulants and neuromodulation, the cause of ADHD is still poorly understood and diagnosis is complicated by the overlap of symptoms with other related conditions. Thus, it is worth exploring whether fMRI studies could identify regions and functional connections that are implicated in ADHD, beyond our current understanding of the disorder.

In this review, 4 studies on ADHD satisfied our inclusion criteria, as shown in Table 11. They relied on datasets from CUNMET (Spain), ADHD-200 (a mix of sites from the USA and China), and CNI-TLC (with some overlaps with site KKI in ADHD-200). CUNMET contains a mix of combined and inattentive subtypes, ADHD-200 has a mix of all three subtypes and KKI has all three but is dominated by the combined subtype. Hyperactivity/impulsive is under-represented across all datasets. Overall, model performance is moderately high (mean accuracy of 71.8% across the studies). Considering the moderate size of the datasets (mean: 374) used, these results suggest that fMRI has a moderate ability to discern between typical controls and ADHD patients. This remains the case even when including studies that did not perform biomarker analysis.

However, the analysis of salient features was less encouraging. While both Yu et al. and Zhao et al. (trained on different datasets) agree that ADHD subjects have weaker connections between the frontal lobe and temporal lobe, there is no agreement on the level of connections between ROIs. For instance, Yu et al. (2022) highlighted MTG-IFG as a weakened connection, while (Zhao et al., 2022b) highlighted the left temporal pole of STG — right SFG (medial). Nevertheless, as seen in Fig. 9, a majority of studies do implicate the middle frontal gyrus, which could be related to the response inhibition circuit. However, more studies need to be conducted before coming to any conclusion.

Additionally, since ADHD is heterogeneous, it would be more insightful for future studies to analyze salient features separately for each subtype. The presence of heterogeneity in the disorder suggests that biomarker reproducibility might not always be possible (e.g. different datasets are dominated by different subtypes of ADHD) and a subtype-level analysis would have more potential of highlighting replicable biomarkers.

5.2. Autism spectrum disorder

ASD encompasses a wide range of impairments of varying severity, but key characteristics include persistent social impairments and repetitive, restricted behavior including experiencing distress when routines are disrupted. These signs should occur to an extent that affects daily life. As a neurodevelopmental disorder, it often begins in early childhood, often affects intellectual development and it could persist into adulthood. It has been observed to be more prevalent in boys than girls. Existing diagnostic biomarkers are primarily genetic, stemming from de novo mutations (Abi-Dargham et al., 2023). While early studies posited that autistic children have larger head sizes, this was found to be limited to a subgroup (Libero et al., 2016). Findings from functional neuroimaging have yet to converge to a clear consensus (Lord et al., 2020).

In this review, 42 studies on ASD were included, as shown in Table 12. A majority of these studies depend on the ABIDE dataset, which is a multi-site consortium mostly represented by Western data sources (more from the USA than Europe). Despite its prevalent use, major issues with the dataset include variations in exclusion criteria (e.g. some sites do not include females, and have varying fluid intelligence thresholds) and inclusion criteria — most sites relied on the same instruments (Autism Diagnostic Observation Scale/Autism Diagnostic Interview-Revised) but some used it in combination with clinical judgment, which would be subjective by nature. Overall, model performance is rather high with a mean accuracy of 76.1% (mean dataset size of 661). Notably, this remained consistent across dataset sizes. Out of the best-performing studies (across ASD studies, based on the reported test accuracies), studies using connection profiles as features and both BG and PG seem to have contributed to the improved performance.

Given the large number of studies, we will only discuss the most frequently occurring salient found across multiple studies. At the level of regions, precuneus, superior temporal gyrus, thalamus, prefrontal

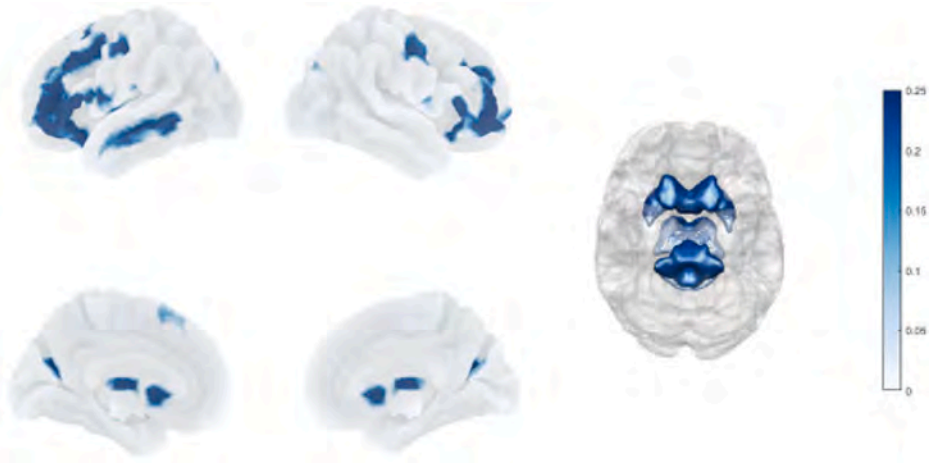

Fig. 9. Surface plot of ROIs highlighted across ADHD studies. Color intensity is based on S_A.

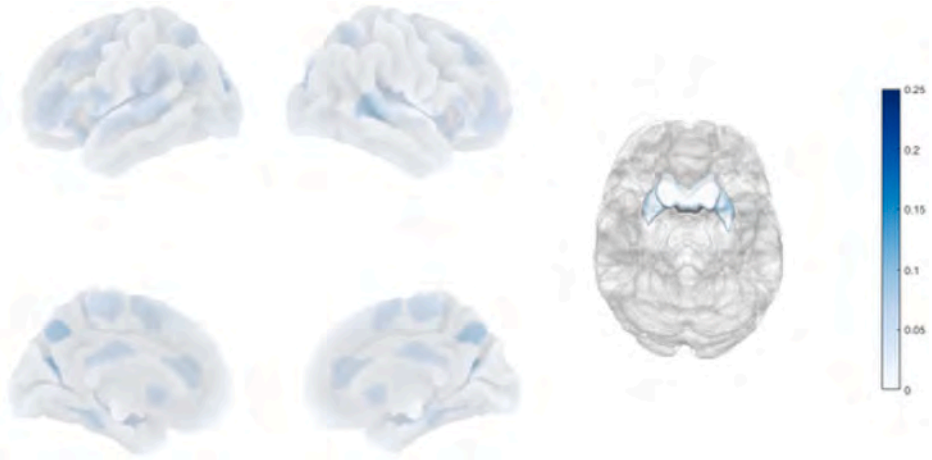

Fig. 10. Surface plot of ROIs highlighted across ASD studies. Color intensity is based on S_A.

cortex, and cingulate gyrus/cortex were implicated in more than 3 studies, as evident in Fig. 10. At the level of connections, no exact matches were found across all studies. At the level of modules, DMN and SN were highlighted in multiple studies but no modular connections were found to be reproducible.

Despite the presence of potential biomarkers that are reproduced across multiple sites, these regions are rather broad, and future research could consider using more fine-grained atlases to arrive at more specific ROIs. Additionally, considering the contrast between the presence of multiple reproducible salient regions and the lack of any reproducible salient connections, future studies could consider identifying salient subgraphs to see if there are any reproducible brain circuits that could reconcile the lack of agreement at the level of connections (Cash et al., 2023).

5.3. Major depressive disorder

MDD presents as a state of persistently low mood (at least two weeks), often accompanied by andohenia. Other symptoms include feelings of worthlessness, suicidal thoughts, and poorer concentration. Significant changes in sleep, appetite, and activity levels often occur too. Existing fMRI studies have revealed significant differences between male and female patients (Tian et al., 2024) and that regions in DMN, SN, and CEN are highly involved (Bondi et al., 2023). Specifically, DMN-linked connections include PCC - precuneus ; SN regions include ACC ; CEN connections include DLPFC - PCC, IFG-DMPFC. However, symptoms of MDD overlap with other disorders, making it difficult to identify biomarkers and traits unique to it.

20 studies on MDD were reviewed, as shown in Table 13. A majority of them (16/20) used subsets of the REST-meta-MDD dataset, which is a collection of over 2000 scans from 17 hospitals in China. Other datasets include psymri (multiple sites, predominantly in Europe) and private datasets collected in China and Korea. Like other multi-site studies, the profile of patients varies widely across aspects such as severity, use of medication, and episodicity. Overall, model performance is moderately high (mean accuracy of 73.6%, mean dataset size of 1003). Notably, smaller datasets (around 500 samples) exhibited greater variability of results. However, mean performance across dataset sizes (500, 1600, 2400) was rather consistent (around 74%).

Potential MDD biomarkers highlighted in these studies were found to have limited robustness. No salient features were consistently represented in the majority of the sites. This could be a result of the heterogeneous nature of the disorder. Some potential biomarkers were present in multiple sites, though they tend to be present in only a few and are mostly limited to ROIs. This includes the pallidum, lingual gyrus, precentral gyrus, SFG, and thalamus, as seen in Fig. 11. At the level of connections, none were found to be present in multiple studies. However, Kong et al. (2025) highlighted several connections that were consistent across datasets (same model architecture, same attributor): left caudate - right caudate, left postcentral gyrus - right postcentral gyrus, left SPG - right SPG, left SOG - right SOG, right precentral gyrus - right postcentral gyrus, left IFG (triangular part) - left IFG (opercular

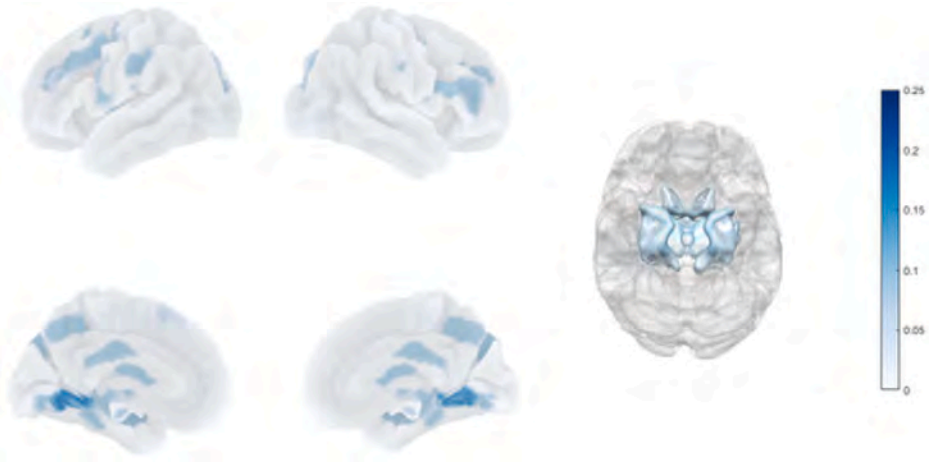

Fig. 11. Surface plot of ROIs highlighted across MDD studies. Color intensity is based on S_A.

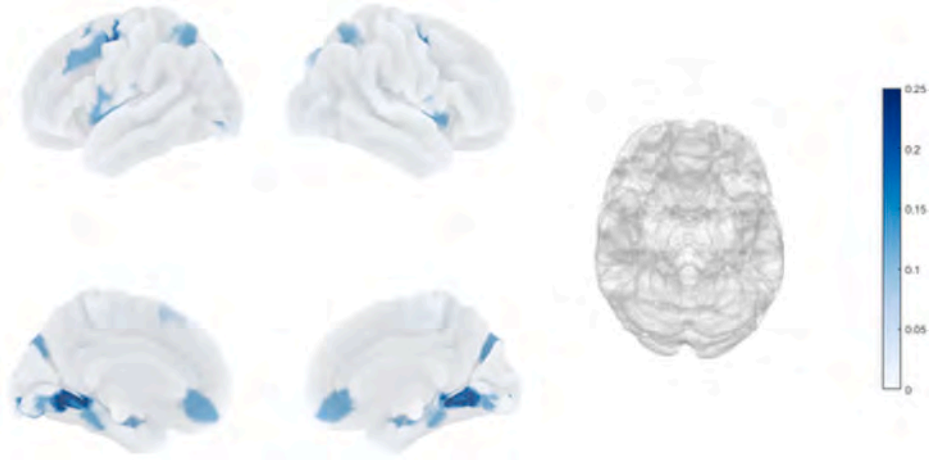

Fig. 12. Surface plot of ROIs highlighted across SZ studies. Color intensity is based on S_A.

part). At the level of modules, the default mode network and limbic network were most consistently highlighted across studies.

In summary, these consolidated findings match past meta-reviews to a limited extent, with the greatest correspondence at the level of modules (DMN, SN), followed by ROIs (IFG). However, no matches at the level of connections could be found.

5.4. Schizophrenia

The definition of schizophrenia has evolved significantly over the past decades (Jablensky, 2010). There is a lack of consensus about how to name it (Lasalvia, 2018) and there is still no clear definition even today. At present, it requires both 'positive' symptoms (such as delusional thinking, and auditory hallucinations) and 'negative' symptoms (such as catatonic behavior, and social withdrawal) to be observed over time (at least a month) ; these symptoms have to be at an extent that affects one's daily life (Tandon et al., 2013). Cognitive impairment is also prevalent in schizophrenia patients (Kahn and Keefe, 2013). Diagnosis is complicated by how symptoms might not show up all at once — it requires a prolonged period of monitoring that involves gradually excluding other related disorders. This motivates the search for alternative ways of diagnosing it that would require less time and therefore allow earlier intervention.

One view of schizophrenia is that of disruptions involving information processing. Functional imaging studies have revealed abnormalities in (i) salience processing (associated with positive symptoms) implicating the ventral striatum and DLPFC ; (ii) reward processing (associated with negative symptoms) implicating the amygdala, medial PFC as well as lowered ventral striatal responses to reward ; (iii) cognitive function (regulation and control of cognitive processes) implicating the DLPFC, rostral ACC and IPL (Kahn et al., 2015) as well as lower connectivity between cortical (PFC) and subcortical regions (basal ganglia, thalamus, cerebellum) (Sheffield and Barch, 2016).

In this review, a wide range of datasets were used by 11 studies on SZ, as shown in Table 14. This includes COBRE (USA), SRPBS (Japan), UCLA (USA), and various sites in China (AMU, FMMU, PKU, UEST). Most sites reported that diagnoses were arrived at via the Structured Clinical Interview used for DSM Disorders (SCID), but inclusion criteria for healthy controls varied. Overall, model performance was high (mean accuracy of 85.4%). We note that 4 sites had a significant class imbalance. However, after excluding them, the performance remained high (84.8%). The average dataset size is 403 but we note that the performance is consistently high across small (around 100) and large datasets (above 1000, combination of multiple sites). This gives greater confidence to the salient features highlighted in these studies.

While there are no salient features that are consistently highlighted across all studies, there are some notable regions that are reported across multiple studies: medial SFG (Chen, Zhu UCLA, and COBRE), left lingual gyrus, left calcarine fissure (Chen, Zhu UCLA), anterior cingulate gyrus (Chen, Zheng), MFG (Zheng, Li). As seen in Fig. 12, the lingual gyrus has particularly higher replicability across studies. Such common findings are notably from different datasets (and countries),

suggesting the possible presence of generalisable biomarkers across geographies.

Additionally, we note that Zhu et al. (2024) found a set of consistent ROIs across 2 datasets (COBRE, UCLA), including left medial SFG, left IOG, and left precentral gyrus. On the other hand, Zheng et al. arrived at quite different biomarkers on the same dataset when different novel attributors were used. This could suggest that attributors have a significant influence and that novel attributors should be carefully tested for robustness before they can be used.

Linking back to the pre-existing knowledge about the neural correlates of Schizophrenia, we found matches for positive symptoms and cognition-related regions but less so for negative symptoms. A group of vision-related substructures such as lingual gyrus, calcarine fissure, and inferior occipital gyrus were found to be implicated, likely related to positive symptoms (e.g. visual hallucinations). For cognition-related symptoms, we found that multiple Chinese sites identified ACC to be important. However, this was not replicated in Western studies (e.g. UCLA, COBRE). Finally, the relatively consistent findings of the medial superior frontal gyrus being implicated might give a more specific localisation of DLPFC-related aberrations in FC. However, it is unclear from the studies whether this is in any way more strongly identified in specific subgroups of schizophrenia symptoms.

5.5. Transdiagnostic biomarkers

Psychiatric disorders often have overlapping symptoms and numerous studies have argued for them to be studied in tandem (van den Heuvel and Sporns, 2019; de Lange et al., 2019). Having summarised the reproducible salient features for each disorder, another round of matching was performed to identify any overlaps of these potential biomarkers across disorders. This results in the following matches:

- Medial frontal gyrus (ADHD and SZ)
- Cingulate gyrus (ASD and SZ)
- Lingual gyrus (MDD and SZ)
- Thalamus (ASD and MDD)
- DMN (ASD and MDD)
- SN (ASD and MDD)

It is clear that SZ has significant overlaps with multiple disorders (ADHD, ASD, MDD), in terms of common brain regions being affected. Additionally, the overlap between ASD and MDD seems stronger than other combinations, with multiple common brain modules and regions being affected in both disorders.

One caveat to this analysis is the possibility that other overlapping brain regions are not considered (e.g. due to the limited number of studies in this review for disorders like ADHD). Nevertheless, this shortlist of regions was found to exhibit a greater degree of robustness across studies (i.e. across pre-processing pipelines, choice of predictor and attributor, etc.) and could be explored in future studies.

Alternatively, transdiagnostic biomarkers can be studied at a higher level (i.e. parcels, instead of ROIs) by leveraging on the framework established in Section 4.3. This has the advantage of being more objective as matching is done based on MNI coordinates, rather than manual matching of salient features (e.g. different atlases might describe the same region in different ways). From Fig. 13, it is evident that at the level of parcels, there are numerous overlaps across disorders. For instance, the 2 parcels highlighted in all disorders are located at the left middle frontal gyrus and the insula/sylvian fissure. While these have been mentioned and discussed in the above sections, the finding here suggests a more general role that these structures play across disorders. While these would not serve as hallmarks of specific disorders, it could be important to study relationships between them and disorder-specific biomarkers. More details about the overlaps highlighted in the UpSet plot can be found in https://osf.io/wza6b/.

5.6. Robustness of attributors

Another analysis that can be conducted is to study the robustness of attributors, based on the salient features reported in existing studies included in this review paper. Specifically, the following questions are explored: (i) How prevalent is it for different attributors to produce different results (while controlling for the variations introduced by choice of dataset and predictors)? (ii) Which attributors produce more reproducible features?

To answer the first question, we explore a few scenarios. First, when other factors are kept constant (i.e. same dataset, same predictor), it would be reasonable to expect that robust attributors should give similar attributions. Two studies (on SZ and MDD, previously discussed in Section 4.2.8 under 'Coherence') demonstrated that when different attributors are used, salient features can vary a lot or have a few overlaps, depending on the choice of attributors. In Li et al. (2023), the use of saliency and GradCAM produced heatmaps with little agreement, while in Gallo et al. (2023), the use of GNNExplainer and Occlusion saw some overlaps of salient features. Second, when broadening the scope to include studies with different predictors and attributors (e.g. Zheng et al. (2024a,b)), we observe that there is practically no agreement between the reported salient features, even though the same dataset was used. This shows how the choice of predictors and attributors has a very significant impact on the attributions. Considering how the dataset is kept constant in these comparisons, these variations clearly do not stem from biological influences, but rather are artifacts produced by the modeling process.

As a contrasting example, when analyzing studies where only the dataset is changed (i.e. same predictor, same attributor), we found considerable overlaps of salient features. However, it is noted that the datasets chosen in these studies have some similarities. For example, in Zhu et al. (2024), COBRE and UCLA are both based in the US, while in Kong et al. (2025), REST-meta-MDD and Zhongda are both based in China. In other related studies when datasets from very different sources were used (e.g. Western and Asian), using the same predictor and attributor resulted in disparate site-specific biomarkers (Chan et al., 2023). With these considered, it could be argued that the choice of predictor and attributor could have an outsized impact relative to the choice of dataset especially when the dataset choices are kept within reasonable bounds (e.g. within geographical and racial boundaries).

Overall, existing studies covered in this review do not give any conclusive evidence of any attributors being superior to others in terms of reproducibility. Based on the above observations, it is very likely that the predictor has a significant role to play too. Considering that existing papers often focus on novel architectures that result in unique predictors being proposed, this would mean that thorough experiments need to be conducted by pairing up the proposed predictor with various attributors as introduced in Section 3.2. This makes it desirable for a means to quickly assess the robustness of attributors to be created. Thereafter, only the most robust attributor for the proposed predictor should then be used for downstream analysis, such as reporting and visualisation of salient features.

5.7. Comparison with neurodegenerative diseases

One challenge with assessing the correctness of salient features highlighted by ML models for psychiatric disorders is the lack of ground truth. A possible heuristic that could be used to verify their correctness is to analyze what these models highlight when applied to neurodegenerative disorders such as dementia. While dementia is not completely solved, it has relatively more consensus in terms of pathophysiology due to the presence of protein aggregates and visible neurodegeneration (Therriault et al., 2024).

Details of the review methodology for this subsection, as well as a summary table of the salient features, are presented in Appendix. From 11 studies (which used GNNs and reported salient features)

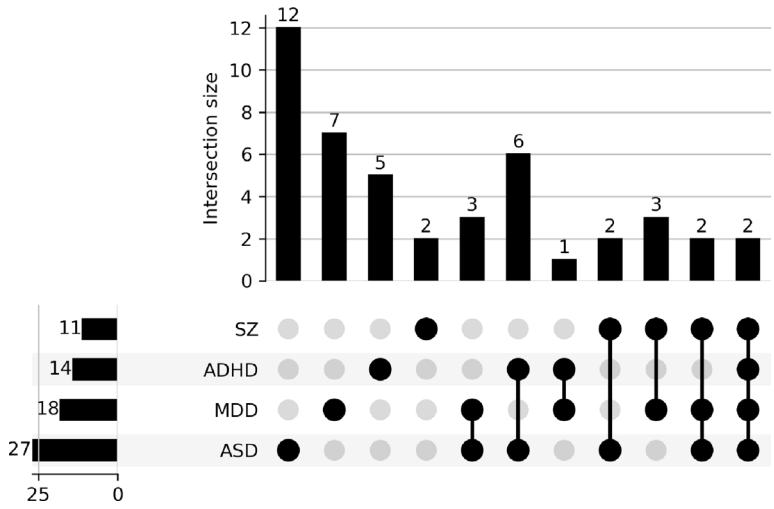

Fig. 13. An UpSet plot showing the overlaps in parcels across disorders, using the DiFuMo atlas with 128 parcels.

spanning various stages and subtypes of dementia, we found that majority of the studies can reasonably capture well-established areas that are affected in dementia, such as the hippocampus and ROIs in the frontal and temporal regions. Default mode network, highlighted in multiple studies, is also strongly associated with the disease. Thus, this gives greater confidence that this framework of using predictors and attributors based on deep learning can produce sensible salient features.

5.8. Summary

Linking back to the third question mentioned in the introduction (whether any salient features are reproducible, whether any transdiagnostic biomarker is present), we have identified a list of potential biomarkers that are present across multiple studies of the same disorder and even identified several potential transdiagnostic biomarkers. However, these are limited to brain regions. Functional connections reported in individual studies exhibited very poor robustness with minimal reproducibility across studies.

Overall, studies on schizophrenia produced the highest classification performances, suggesting that it is possible for GNNs to distinguish SZ patients from healthy controls via fMRI and that the potential biomarkers identified could be more reliable (than those found in studies of other disorders). Research on the other disorders (ASD, ADHD, MDD) could benefit from more thorough benchmarking of state-of-the-art GNN models, or from pivoting to a subtyping approach (i.e. splitting the patient population into subgroups could make it more feasible for disorder classification models to perform better and produce more reliable salient features).

6. Discussion

In order to discover robust biomarkers of psychiatric disorders via machine learning techniques, it is necessary for each component of the biomarker discovery pipeline to work well: (i) a high-performing **predictor** that generalises well out of sample, (ii) an **attributor** that produces stable attribution scores robust to changes unrelated to the disorder, (iii) an **evaluator**, consisting of a set of metrics to evaluate the predictor–attributor combination, that provides objective evidence that the scores corroborate with existing understanding of the disorder. Fig. 14 provides a visualisation of how these 3 components work together.

In this review, we have provided an analysis of the latest research in all 3 components and identified potential biomarkers for several psychiatric disorders, including transdiagnostic ones. Summarising the answers to the questions presented in the introduction, we found that the best predictors tend to use thresholded FC matrices as the adjacency matrix, connection profile as node features and involve both BG and PG. Some graph-based attributors such as GNNExplainer provide greater flexibility of attribution targets than generic ones. The use of evaluators to ascertain the robustness of proposed biomarkers has much room for improvement as they are currently limited to a minority (about 10%) of existing studies. In our own analysis of the reproducibility of these biomarkers, region-level features were found to exhibit greater robustness than edge-level features and this is consistently observed across multiple disorders. In the following discussion, we will provide a deeper evaluation of each component.

We found the reported classification performances of **predictors** to be generally high. However, there are several caveats to this finding. First, most studies do not account for confounds such as age and sex, which could have inflated the model performance (e.g. model captures sex-specific features that will fail to generalise to other datasets). We note that one way to address this is to ensure that these factors are balanced across classes (healthy/disorder), but doing so would often reduce dataset sizes further. On the other hand, the flexibility afforded by GNNs via the use of PGs to incorporate such demographic information could help to alleviate this (e.g. existing BG-only GNNs could add on a PG to account for these factors as covariates). Second, most studies do not demonstrate generalisation to a separate dataset not used during training (i.e. their results merely show their best performance on a test data split from the same dataset, but often do not go further to show generalisation capabilities to a completely different dataset). Third, a deeper analysis reveals that such good performance might be boosted by studies that rely on small datasets. A previous study by Teng et al. (2023) has demonstrated that classification accuracy drops when dataset size increases. This phenomenon is also observed in our analysis, as shown in Fig. 15. This suggests that simply changing the architecture (e.g. to GNN) might not lead to better model performance. Rather, other factors such as dataset size (specifically, the extent of heterogeneity in the sample) have a greater effect on model performance and solutions specific to these issues (rather than just using the latest variant of neural network architectures) need to be developed.

In view of this, we reiterate the need for proper benchmarking. Although several benchmarking studies have been done using GNNs on fMRI datasets, they were focused on baseline GNNs and not state-of-the-art models that have been shown to do better than baselines. Several considerations need to be taken care of in future benchmarking studies: (i) use datasets that are sufficiently large and share precise information about how the dataset is split, so that future studies can follow them, (ii) use of state-of-the-art GNN models, (iii) control for size of model (i.e. number of trainable parameters should be kept similar across models), (iv) report performance metrics beyond just accuracy

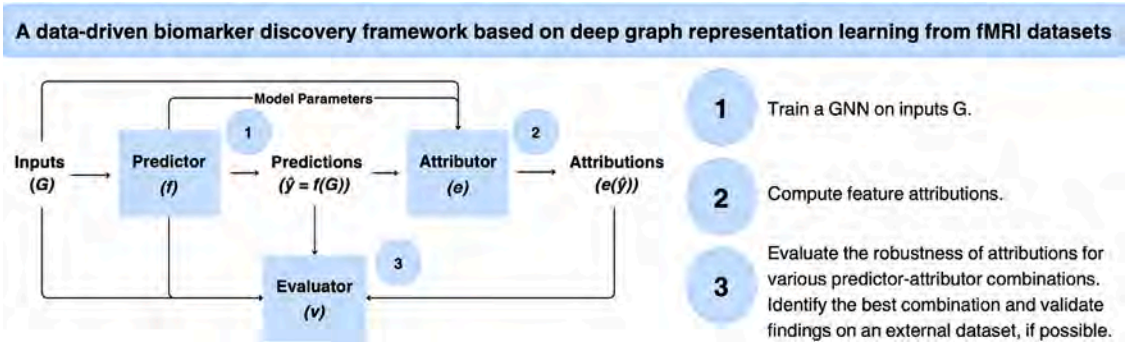

Fig. 14. An illustration showing where the three key stages occur in a typical pipeline that uses a post-hoc attributor. For intrinsically interpretable models, the attributor would become part of the predictor (e.g. pooling layers can be used for producing explanations, but they form part of the GNN architecture).

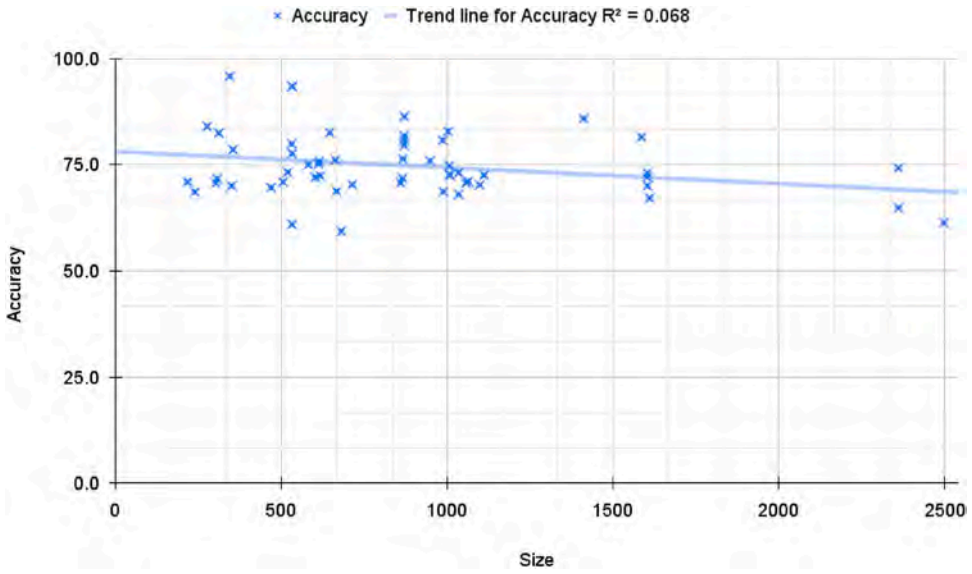

Fig. 15. Plot of model accuracy against dataset size. It is evident that a majority of the studies have dataset sizes below 1000 and classification accuracies generally decrease as the dataset size increases.

(e.g. AUC) so as to account for class imbalances, (v) demonstrate generalisability beyond the dataset used for training (i.e. to a new dataset, not just the test split).

Nevertheless, we caution against dismissing a model just because it generalised to unseen data poorly. Intuitively, a model that has limited performance degradation when tested on unseen datasets would suggest that the features identified by the model to be salient are indeed also important in other datasets of the same disorder. However, more care is needed when choosing these unseen datasets. For instance, datasets aggregated from multiple sites (i.e. data consortiums) are increasingly popular. Such datasets are often gathered from different countries and poor generalisation performance does not immediately mean that fMRI is not useful for identifying patients with the disorder. Rather, it could indicate the presence of heterogeneity in the disorder, especially if the model works well on a test split of the same dataset but performance decays rapidly when using a different dataset (or different imaging site). Such models could still accurately capture site-specific biomarkers that are possibly characteristic of a subtype of the disorder. Thus, more careful evaluation is needed, e.g. using a different site with similar demographics and inclusion/exclusion criteria as the test set. On the other hand, if the goal is to discover biomarkers that are generalisable across the population, then sizable datasets should be expected to be used to justify such claims and demonstration of limited performance degradation in out-of-sample distributions should become a requirement.

On a final note for predictor-related issues, it could be more meaningful to move beyond the use of class labels, especially in view of the litany of studies on disorder classification and their limitations. Binary classification serves as a simplified way to study disorders since patients can be contrasted against 'healthy'/'normal' controls. However, such class labels are a preliminary construct guided by our existing imperfect understanding of these disorders. Furthermore, datasets aggregated from multiple sites are often used without paying heed to variations in inclusion criteria, blurring the lines between what is defined as a control subject and a patient. Thus, it might not be the best approach to constrain our learning algorithms with these labels, especially when the goal is to go beyond our current understanding. For instance, psychiatric disorders have a large number of possible combinations of symptoms that lead to the same diagnosis even though they have different underlying biology (Chen et al., 2023a). In these scenarios, class labels could be unreliable, especially for poorly understood disorders. On the other hand, using a completely data-driven approach is very

challenging. For instance, clustering-based approaches could be applied directly to the data but cluster interpretation is often fraught with subjectivity. Instead, one could predict test scores (such as ADOS, HAM-D, PANSS) or even imaging-guided labels such as PET grading (Li et al., 2022a). This will require the use of regression instead of classification, but the majority of attributors are also able to produce attribution scores in such models.

In terms of the use of **attributors**, it is clear from this review that the choice of attributor restricts the granularity of salient features that can be highlighted. GNN-specific explainers have the distinct advantage of producing subgraph-level attributions. In view of the lack of agreement across edge-level attributions, subgraph/motif-level attributions could be a promising direction as alluded to in Cash et al. (2023), where they showed how a circuit-level analysis can unify seemingly contradictory findings at lower granularities.

Another key insight is that even within the same category, the choice of attributor could introduce significant variability to attribution scores, even when the dataset and predictor remain the same. This warrants a separate benchmarking study to determine the best combination of predictors and attributors for biomarker discovery. In spite of these instabilities, we have identified several consistent region-level biomarkers for each disorder, as well as candidate transdiagnostic biomarkers. However, there are several caveats to these findings.

Firstly, most studies with large datasets almost always rely on data consortiums that collect data from multiple sites (so as to address issues of small datasets, such as inflated accuracies and high standard deviations). While some data consortiums try to standardise scanning protocols to minimise scanner-induced variabilities, the majority of such consortiums pool datasets without standardisation. Existing studies have demonstrated that salient features identified from such aggregated datasets tend to be biased towards the largest site (Chan et al., 2022) and the use of data harmonisation algorithms like ComBat could lead to changes to salient features (Chan et al., 2023). Thus, better data harmonisation techniques would need to be developed and these changes would require further study before they can be used for biomarker discovery.

Secondly, a major limitation of existing studies is that salient features are often aggregated at a class-wide level. However, for them to be clinically useful, personalised insights need to be produced. In such a context, attributors that can produce individualised attributions should be used (e.g. IG ; global attributors such as XGNN would not be useful here). One key concern is that individual heatmaps could be noisy (Kapishnikov et al., 2021). Nevertheless, it is possible to develop techniques to reduce noise even at the individual level. For instance, multiple models (e.g. different random seed) can be used to produce feature attributions for the same individual and these could be averaged to reduce noise. More studies need to be conducted to ascertain the robustness of these individualised insights (e.g. on datasets with multiple scans of the same subject, such as test-retest data). Once this is established, such biomarkers could be used as endpoints to verify the effectiveness of interventions.

Thirdly, one major challenge faced when assessing the robustness of the potential biomarkers brought up by these studies is the lack of consistency in the way the most salient features are reported. This problem is caused by three factors: (i) different atlases used, (ii) unavailability of the raw attribution scores and a lack of clarity on whether biomarkers reported are averaged across test data only, or the entire dataset, (iii) different extent of thoroughness in reporting the salient features (e.g. listing top 10/20/30 features, or just mentioning a few salient features in passing). Furthermore, different granularities of the biomarkers are available due to the design of the GNN architecture and the choice of attributors makes it difficult to harmonise certain features that are less often reported. In this study, we found that a majority of the studies (64.1%) report ROI-level features while another significant portion of studies reports at the level of connections (24.4%). However, functional brain modules (6.4%) and modular connections (5.1%) are less often reported. Overall, harmonising these findings manually is infeasible at a large scale without improvements in reporting standards.

To address this, we suggest the following guidelines as a preliminary step towards establishing a standard that future publications can refer to:

- Attribution scores for each feature should be recorded and shared along with the publication (e.g. as supplementary materials). Minimally, they should contain the ranking of the features (e.g. sorted by importance). Ideally, if scores are available for each individual subject, they should be shared too (e.g. as NumPy files).
- Key metadata about the most salient features (e.g. ROI names as defined by the atlas, along with their MNI coordinates) should be provided in the supplementary materials in the form of a spreadsheet, so as to facilitate future research and meta-analysis.
- Whenever attributors provide information about polarity (i.e. positive and negative scores, reflecting hyper/hypo-connectivity Gupta et al., 2022), these raw values should be reported even if only the absolute value/magnitude is used in the analysis.

Having such guidelines would make it possible to develop computational tools to automate meta-analysis, helping research in this area to progress more quickly.

In the case of **evaluators**, the use of evaluation metrics to measure the robustness of these attribution scores has been an emerging trend in recent years, but little has been done to determine what metrics are appropriate and required for biomarker discovery. In this review, we have attempted to identify the most relevant subset of the Co-12 framework, but much more remains to be done to assess the robustness of the attribution scores produced by attributors. For instance, several unknowns remain, such as an appropriate number of features to keep/remove in checks for Completeness. Once resolved, such scores would provide another set of information, on top of model prediction performances, to determine the robustness of the proposed architecture.

More crucially, the generic evaluation metrics merely serve as a starting point for this research direction. Many novel evaluation metrics could also be proposed in future studies to address other desiderata not covered by these metrics. This includes (i) metrics that are focused on evaluating the consistency of saliency scores across experiments (e.g. repurposing existing measures of reliability such as intraclass correlation coefficient to evaluate cross-cohort consistency), (ii) metrics that consider the relationship between predictions and saliency scores (e.g. in radiology applications, the prediction-saliency correlation metric (Zhang et al., 2023a) computes the correlation between changes in model predictions and the corresponding saliency maps), (iii) metrics that capture brain-specific information (e.g. modularity ratio and hub assortativity coefficient as proposed in Girish et al. (2024)).

These evaluation metrics provide a quick and convenient way to identify the best combination of predictors and attributors. Once this is identified, further validation of the salient features identified by this combination has to be done. For instance, a scatter plot of the FC value against disorder severity (e.g. PANSS for Schizophrenia (Chan et al., 2023)) can be generated and effect sizes can be computed to ascertain the practicality of these biomarkers, e.g. in clinical trials to demonstrate treatment efficacy.

Finally, linking all the above insights to the issues mentioned in the introduction, Fig. 16 highlights the key takeaways from this research. A key end goal of this research is to identify strong and reproducible FC changes that are associated with a disorder. While there are several obstacles that make this challenging, identifying them as intrinsic, data-based and modeling-specific problems allows us to design solutions that specifically target each issue. However, there are still several limitations to the findings we have arrived at. One aspect not covered in our analysis is the effect of variability in pre-processing pipelines (e.g. whether to perform global signal regression, Fisher transform, what threshold to use on the FC matrix etc.) on model performance and

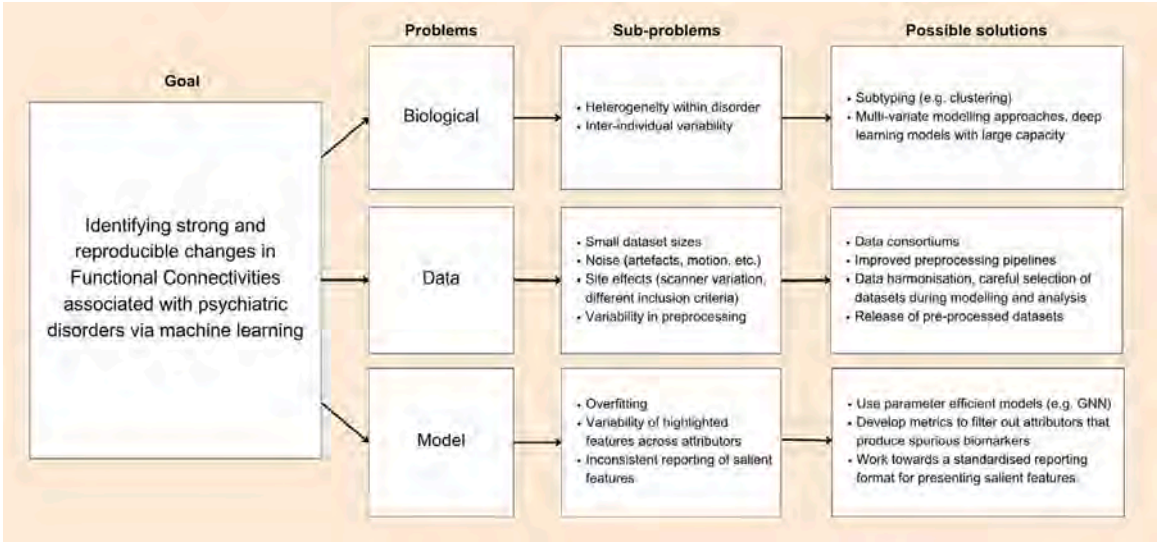

Fig. 16. Overview of key issues faced in ML-driven pipelines for biomarker discovery and possible solutions to remediate them.

biomarker robustness. These issues are complex and should be analyzed in a separate study. One way to reduce the variability of pre-processing pipelines is to release more pre-processed connectome datasets. We note from our analysis that the availability of pre-processed datasets released by data consortiums drives research: about 1/2 of the studies are on ASD and almost all of them used the ABIDE I dataset, especially the preprocessed connectomes with 871 subjects. While the ABIDE II dataset is also available, few studies utilised it due to the lack of pre-processed data. Having pre-processed data made available (on top of the raw brain images) helps to remove one source of variability and also allows results to be compared across studies. Thus, we encourage future data releases to also share pre-processed fMRI datasets or even connectomes.

Additionally, the scope of this review paper is limited to studies that included resting-state fMRI and it is important to keep in mind that it only provides information about brain function at rest. Whenever possible, other modalities (structural, molecular, task-fMRI, etc.) should be studied in tandem to produce a complete picture. In this regard, GNN provides a very flexible architecture to model multi-modal datasets: (i) modalities that are traditionally image-based (e.g. T1w, PET) can still be converted into graphs (Zhang et al., 2023d) such as morphometric similarity networks (Sebenius et al., 2021; Park et al., 2023), (ii) GNN can be used to integrate connectome datasets with multi-omics data (Ghosal, 2023) in a parameter-efficient manner, (iii) population graphs make it possible to integrate non-imaging modalities (Parisot et al., 2018). Multi-modal biomarkers will be another emerging area that would give a more holistic view than what is currently possible with studies only using fMRI (Chen et al., 2023a). While our review paper does include some studies that use multimodal datasets, we note that feature attribution in multimodal/multi-view settings is a bit more complicated. In such architectures, each modality is often taken in as a separate branch and there is a need to consider cross-modal interactions across branches (Niwarthana et al., 2025). Such techniques still have to be extensively validated before biomarkers can be reliably identified from them.

Finally, we note that our review does not include details about temporal GNNs, counterfactual-based explanations (such as Shen et al., 2024) and causal discovery (Rawls et al., 2023). These areas of research are exciting novel directions in biomarker discovery from rs-fMRI data as they go beyond current limitations of static FC, factual explanations and correlation-driven studies, respectively. However, each of these research areas is complex, is relatively less explored within the studies considered in our review, and would require careful exploration to determine their utility for biomarker discovery. For instance, temporal GNNs are essential for dFC-based biomarker discovery, but they face issues such as aliasing effects, determining the appropriate window and step size, etc. Esfahlani et al. (2022). These issues could have significant implications on the robustness of dFC biomarkers and have not been well-studied yet (for instance, evaluation metrics sensitive to temporal attributions are relatively under-explored). We refer interested readers to Kakkad et al. (2023) for a review on temporal GNNs and counterfactual explanations, and Bielczyk et al. (2019) for a review on causal discovery from fMRI.

7. Conclusion

In summary, while there has been an abundance of novel GNN architectures designed for disorder prediction from fMRI data, there remains much room for further research to improve the robustness of the salient features highlighted by these predictors and attributors. Benchmarking studies that involve state-of-the-art GNN predictors customised for fMRI datasets are needed. Studies on optimal choices of predictors and attributors are also required as their robustness on fMRI datasets is still poorly understood. Existing evaluation metrics were designed for generic (graph) datasets and more metrics appropriate for FC datasets are needed to determine whether attributors are sensitive to known properties of FC. The lack of standardised reporting of salient features has made it challenging to consolidate insights from existing studies. Nevertheless, we have proposed a framework, Hi-5, for evaluating them and demonstrated how an analysis based on frequency counts could be done. Eventual improvements in reporting standards, understanding of how attributors interact with predictors and evaluation metrics would allow for a more rigorous meta-analysis to be conducted, potentially revealing new insights for these disorders. Finally, the paucity of reproducible salient features (especially at the level of connections) motivates the search for alternative approaches. Possible directions for future research include looking beyond the use of class labels and pivoting away from the goal of solely chasing for generalisable biomarkers for the entire disorder class, especially for heterogeneous disorders. Moving towards regression tasks, transdiagnostic studies and more fine-grained biomarkers could result in more robust biomarkers of psychiatric (and more generally, neurological) disorders

Table 2
Mapping of abbreviations used for atlases to their full names. More information about the atlases can be found in the corresponding papers that used the atlas in their study. Note that this list is limited to atlases used in studies considered in this review (e.g. it does not include Glasser, or variants of Schaefer's atlas).

Atlas	Full name	ROIs
AAL116	Automated anatomical labeling	116
AAL166	Automated anatomical labeling *v*3	166
AAL90	Automated anatomical labeling	90
BASC325	Bootstrap analysis of stable clusters	325
BM82	Broadmann	82
BN273	Brainnetome	273
BNA246	BrainNet Atlas	246
BV140	BrainVISA Sulci	140
CC200/400	Craddock	200/400
DK308	Desikan–Killiany	308
DK86	Desikan–Killiany, with subcortical ROIs	86
DOS160	Dosenbach	160
DX148	Destrieux	148
EZ115	Eickoff–Zilles	115
HO110/111/112	Harvard–Oxford	110/111/112
JHU81	JHU ICBM-DTI-81	81
MODL128	Dictionaries of functional modes	128
Power264	Power atlas	264
SF100/200	Schaefer	100/200
SHEN268	Shen atlas	268
TT93	Talairach and Tournoux	93
TT97	Talariach	97
YEO114	Yeo 17-network	114

being discovered from fMRI datasets in the near future.

Funding

This work was generously supported by Academic Research Fund Tier-2 grant MOE T2EP20121-0003 and Tier-1 grant RG15/24 of Ministry of Education, Singapore.

CRediT authorship contribution statement

Yi Hao Chan: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Deepank Girish:** Writing – review & editing, Validation, Software, Methodology, Data curation. **Sukrit Gupta:** Writing – review & editing, Methodology, Investigation, Formal analysis, Conceptualization. **Jing Xia:** Writing – review & editing, Methodology, Investigation, Formal analysis. **Chockalingam Kasi:** Software, Methodology. **Yinan He:** Methodology, Formal analysis, Data curation. **Conghao Wang:** Writing – review & editing, Methodology, Investigation, Formal analysis. **Jagath C. Rajapakse:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

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.

Appendix A. Abbreviations used

In this subsection, several tables are presented for a convenient reference of the mapping between abbreviations used in this review paper to their full names.

Table 3
Mapping of abbreviations used for brain networks to their full names.

Abbreviation	Full name
DAN	Dorsal Attention Network
CEN	Central Executive Network
DMN	Default Mode Network
FPN	Frontoparietal Network
LN	Limbic Network
SMN	Somatomotor Network
SN	Salience Network
VAN	Ventral Attention Network

Table 4
Mapping of common abbreviations used to their full names.

Abbreviation	Full name
1-WL	1-dimensional Weisfeiler-Leman
ADHD	Attention Deficit Hyperactivity Disorder
ASD	Autism Spectrum Disorder
BG	Brain Graph
CAM	Class Activation Mapping
CNN	Convolutional Neural Network
dFC	dynamic Functional Connectivity
DNN	Deep Neural Network
FC	Functional Connectivity
fMRI	functional Magnetic Resonance Imaging
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GIN	Graph Isomorphism Network
GNN	Graph Neural Network
HSIC	Hilbert–Schmidt Independence Criterion
ICA	Independent Component Analysis
IG	Integrated Gradients
k-NN	k-Nearest Neighbors
MDD	Major Depressive Disorder
MI	Mutual Information
ML	Machine Learning
MNI	Montreal Neurological Institute
NC	Normal Controls
NLP	Natural Language Processing
PCA	Principal Component Analysis
PG	Population Graph
RF	Random Forest
RFE	Recursive Feature Elimination
RL	Reinforcement Learning
ROI	Region of Interest
sFC	static Functional Connectivity
SNR	Signal-to-Noise Ratio
SVM	Support Vector Machine
SZ	Schizophrenia
TDC	Typically Developing Children

Appendix B. Review protocol

Introduction

Functional magnetic resonance imaging (fMRI) has been widely used to investigate the neural underpinnings of psychiatric disorders. This has been an arduous task due to our limited understanding of their causes as well as the complexity of the data modality. Computational tools are necessary to parse through information stored in millions of voxels, as well as the inter-region relationships measured in functional connectomes that can be constructed from fMRI data, in an attempt to identify differences between normal controls and patients with these disorders. Recently, the rise of deep learning has provided new ways to analyze brain connectomes. Specifically, graph neural networks provide a parameter efficient manner of encoding connectome datasets and

Table 5
Mapping of abbreviations used for brain regions to their full names.

Abbreviation	Full name
ACC	Anterior Cingulate Cortex
ACG	Anterior Cingulate Gyrus
FOG	Frontal Orbital Gyrus
FP	Frontal Pole
IFG	Inferior Frontal Gyrus
IOG	Inferior Occipital Gyrus
IPL	Inferior Parietal Lobule
ITG	Inferior Temporal Gyrus
MFG	Middle Frontal Gyrus
MOG	Middle Occipital Gyrus
MTG	Middle Temporal Gyrus
PARAH	Parahippocampal Gyrus
PoCG	Postcentral Gyrus
PrCG	Precentral Gyrus
PFC	Prefrontal Cortex
SFG	Superior Frontal Gyrus
SMG	Superior Medial Gyrus
SOG	Superior Occipital Gyrus
SPG	Superior Parietal Gyrus
SPL	Superior Parietal Lobule
STG	Superior Temporal Gyrus
TP	Temporal Pole

performing downstream tasks such as disorder prediction. This has led to a steadily increasing number of publications over the past few years. These studies proposed enhanced forms of GNNs that are customised for connectome datasets, surpassing the performance of baseline GNNs.

Several reviews have been conducted to summarise the latest research on using GNNs for disorder prediction. However, these models are rarely used clinically and disorder prediction is not the main goal. Instead, another contribution these models can make is to improve our understanding of these disorders.

A highly performant model would have learnt relationships between the input data and the class labels. While non-linearities obfuscate the processes behind a deep learning model's decision-making process, model explainability techniques have been applied to compute attribution scores that could serve as an indication of feature importance. Driven by this, many recent studies in the field have begun reporting salient features inferred from the model. However, evaluation of these biomarkers is far from robust, often limited to cross-referencing with other studies that report similar biomarkers and neglecting finer details.

Thus, the objective of this review is to compare across multiple studies to determine whether there are any robust potential biomarkers of psychiatric orders. The validity and utility of these biomarkers hinge on the dataset, predictor and explainer. As such, discussions about the predictor need to be in the context of biomarker discovery and its specific needs and this review aims to address this gap. There remains much to explore in terms of evaluating attribution scores — we aim to identify how existing research has achieved this and suggest areas for further investigations. Finally, we will not only highlight potential biomarkers that have been identified across multiple studies of the same disorder, but also attempt to discover trans-diagnostic ones as well.

In summary, the research questions to be answered from the insights gleaned from this review are as follows:

1. Are there any GNN-specific architecture designs that have consistently demonstrated superiority over others (e.g. their proposed GNN-based model works well across multiple datasets and studies)?
2. How do existing studies evaluate the robustness of the potential biomarkers identified by their proposed approach (i.e. combination of predictor and attributor)?
3. For each disorder, is there any convergence of discovered potential biomarkers across multiple studies (that are based on machine learning techniques)? If there are any, are they present in other disorders as well (i.e. potential transdiagnostic biomarkers)?

Methods

In this review, we aim to retrieve studies that used GNNs for disorder prediction and reported potential biomarkers derived from fMRI datasets. Eligibility criteria include (i) within the specified list of psychiatric disorders (attention-deficit hyperactive disorder, autism spectrum disorder, major depressive disorder, schizophrenia), (ii) data modalities must include fMRI, (iii) written in English, (iv) biomarkers must be reported in a clear manner (i.e. not just a brain plot with no corresponding label information), (v) accessible full-text.

Information sources include PubMed and Scopus, which indexes conference proceedings and journal publications where studies on this topic are typically published in. Our search strategy involves the following search terms to be used in both sources: (“graph neural networks” OR “graph convolutional networks” OR “GNN” OR “GCN”) AND (“fmri” OR “functional MRI” OR “functional connectivity”).

This casts a wide net that needs to be narrowed down to exclude irrelevant entries. 3 reviewers will independently parse through the search results over 2 iterations. In the first iteration, irrelevant search results will be filtered out based on the title and abstract. Duplicates are also removed at this stage. In the second iteration, the full text will be retrieved and read through to determine whether they satisfy the eligibility criteria. If relevant, the following fields will be retrieved and stored in a spreadsheet.

- Basic paper information: data source, title, author, year.
- Dataset information: Name of dataset, number and type of modalities, size of dataset, class splits, brain atlas used and number of regions of interests.
- Predictor-related: Base GNN type, type of graph construction (brain graph, population graph), how the adjacency matrix was constructed, what node features were used.
- Performance-related: Best model performance, next best baseline GNN performance, standard deviation of performances (if available).
- Explainer-related: How attribution scores are computed, granularity of attributions (region, connection, module).
- List of biomarkers reported in the paper.

During data synthesis, statistics such as the mean accuracy (overall and for each disorder) will be computed. These are only meant to serve as a gauge of model performance due to their availability — while accuracy has its limitations (e.g. class imbalance), not all papers consistently report alternative metrics. Another set of derived data involves the potential biomarkers that are reproduced across multiple studies (and disorders, for transdiagnostic biomarkers). These will be arrived at via manual matching due to the lack of automated tools that are capable of retrieving such information correctly.

In terms of quality assessment, we note that this systematic review differs from typical ones conducted in medical research — fMRI is not at the stage where it is being used as a test for these disorders, nor are there identified biomarkers that are being used for diagnostic purposes. Rather, the goal is to identify robust biomarkers derived from machine learning models that have been trained on fMRI datasets for disorder prediction tasks (i.e. a preliminary step of biomarker discovery before future prediction models can use these identified biomarkers for diagnostic purposes). Also, a secondary goal of the review is to identify gaps and room for improvement so that future reviews and meta-analyses will be equipped with better data (e.g. existing ways of reporting biomarkers are unstandardised and highly varied, coupled with a lack of access to raw attribution scores, this makes it challenging to perform more sophisticated analysis in this review) Thus, we think that any assessments on risk of bias would be more appropriate and effective in future reviews.

Appendix C. Additional explanations and implementation notes

C.1. Rationale for exclusion of 4 Co-12 factors

1. **Controllability**: This concerns the extent a user has control over an explanation (e.g. ability to correct, or manipulate it within an interface). Being applicable only to interactive explanations, it might not have much relevance at present. However, it could be an important consideration if future clinical use of biomarkers involves clinicians interacting with a graphical user interface that allows them to explore various granularities of biomarkers personalised to the patient.
2. **Context**: This entails creating explanations that consider the users' requirements to ensure that they understand the explanations. It is indeed a relevant consideration, but we do not think that biomarkers for psychiatric disorders are at a state where it is ready for clinical deployment. Once robust biomarkers are established, it would then be necessary to determine exactly which components of the biomarkers to present to address clinicians' needs, as well as the mode of presentation for such information in clinical settings.
3. **Composition**: This is another potentially relevant aspect that concerns the formatting and organisation of explanations. For instance, whether biomarkers should be presented at the level of nodes, edges, motifs/subgraphs, or within the entire connectome. However, at this stage, it is necessary to first ascertain which granularity of biomarkers is most robust and thus this aspect of Co-12 would be a greater concern when functional biomarker discovery for psychiatric disorders is at a more mature stage.
4. **Covariate complexity**: This relates to having human-understandable explanations, but FC studies typically use ROIs as features which are already understandable, unlike individual pixels in images.

C.2. Implementation notes for Section 4.3

In order to aggregate findings from multiple studies, it is necessary to harmonise the various atlases used by them. Direct matching of the ROIs via MNI coordinates is infeasible as it is possible for 2 atlases to have a different center of mass for the same region. While it is possible to set a distance threshold (i.e. if two coordinates are within a certain distance, we consider them to be the same ROI), it is unclear what distance should be used. Instead, we propose to use a brain parcellation (ideally one with multiple scales and appropriately chosen based on the data modality or research question) - if 2 ROIs fall within the same parcel, we consider them to be similar and then conduct downstream analysis at a relatively higher level.

During this atlas harmonisation process, there can be several complications. For instance, some studies use atlases with no publicly available MNI coordinates for each ROI. This will require manual estimation of the center of mass. In the event the ROI consists of multiple blobs, we opted to choose the largest one for the center of mass computation. Other complications include studies that use multiple atlases but do not mention which is used to generate salient features, or studies that use non-standard, customised atlases. In this study, no such scenarios occurred but future reviews that include a broader range of studies will require significant effort to harmonise these atlases, or to exclude studies where it is infeasible to retrieve the information needed.

In this review, we chose the DiFuMo atlas (Dadi et al., 2020) which provides parcellations at different scales (64, 128, 256, 512, 1024). As a proof of concept, we opted for the version with 128 parcels. We have made our analysis code and data produced by it available at the link provided in https://osf.io/wza6b/. In future studies, these could be extended to include other atlases or other ways of computing metrics for cross-study reproducibility.

Appendix D. Summary of predictors and classification performance

D.1. ADHD

See Table 6.

D.2. ASD

See Table 7.

D.3. MDD

See Table 8.

D.4. SZ

See Table 9.

D.5. Additional figure

See Fig. 17.

D.6. Comparative summary of GNN model performance

See Table 10.

Appendix E. Summary of attributors and salient features

E.1. ADHD

See Table 11.

E.2. ASD

See Table 12.

E.3. MDD

See Table 13.

E.4. SZ

See Table 14.

E.5. Additional figure

See Fig. 18

Appendix F. Dementia

The 11 papers on dementia were obtained via the same process as detailed in Section 1.2, with the exception that the search was limited to 15 May 2023 and before. Additionally, some subtypes or stages of dementia (e.g. amnestic mild cognitive impairment, mild cognitive impairment and subjective cognitive decline) had too few studies and thus were not included in Table 15. Most studies relied on the dataset from Alzheimer's Disease Neuroimaging Initiative (ADNI) as it is one of the most established and largest multimodal dataset available.

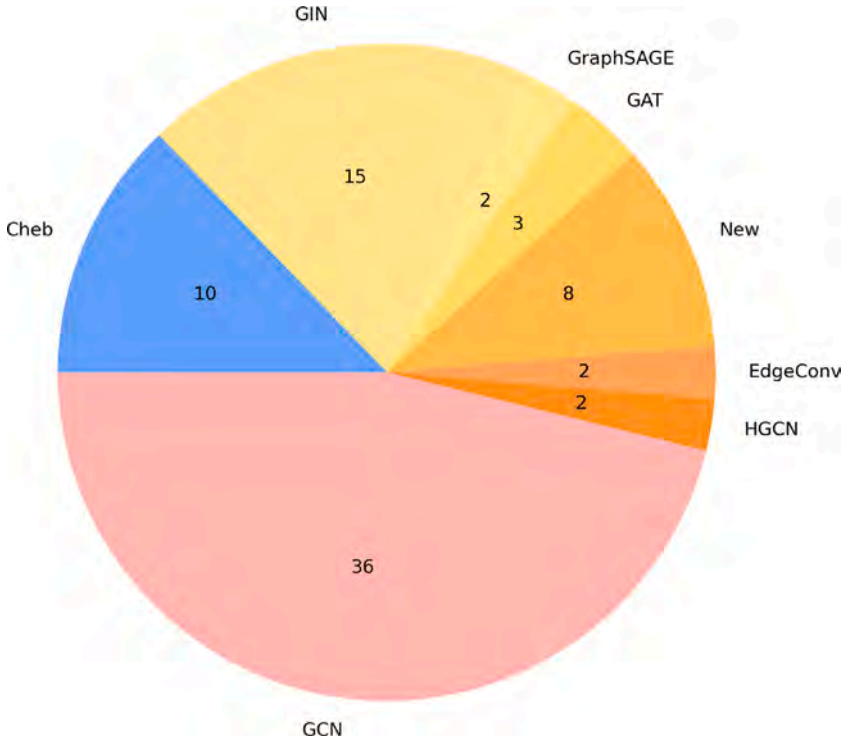

Fig. 17. Choice of GNN used in the papers reviewed. Some studies included more than one type of GNN. 'New' refers to architectures that significantly deviate from baseline GNNs (e.g. customised message passing techniques). Color scheme represents various classes of GNNs - Light red: GCN (spatial/spectral), Shades of yellow: Spatial GNN, Blue: Spectral GNN, Shades of orange: Others.

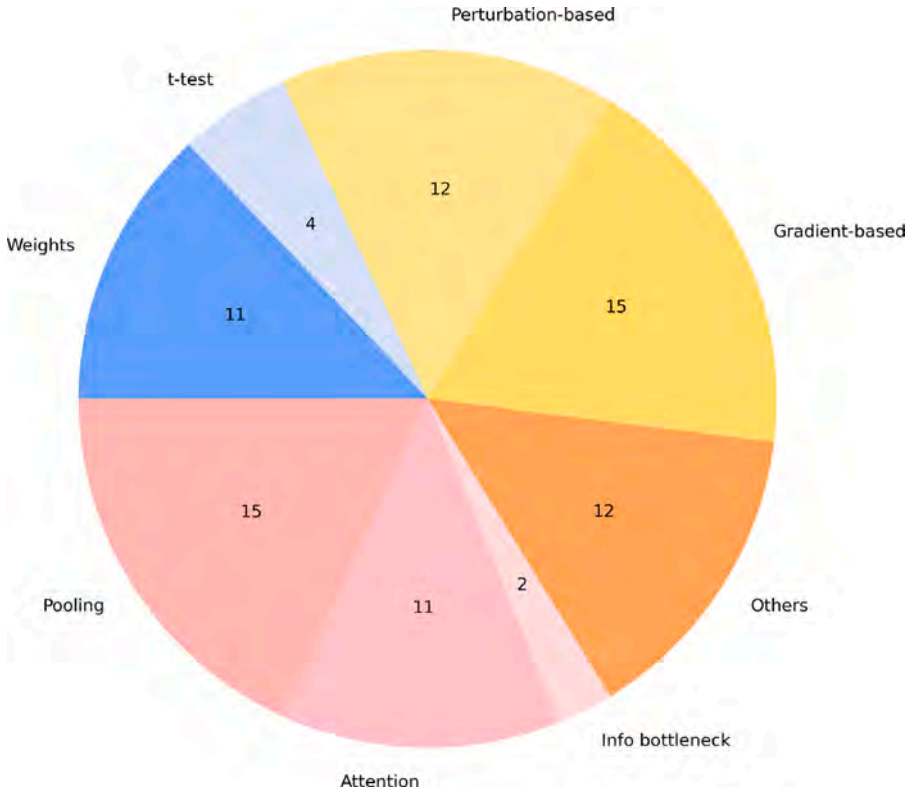

Fig. 18. Distribution of attributors covered in our review. Some studies used more than one type of attributor. Color scheme represents various categories of attributors, following our proposed taxonomy - Shades of red: Self-interpretable, Orange: Others, Shades of yellow: Post-hoc, Shades of blue: 'Traditional' approaches.

Table 6
Summary of findings from ADHD studies. 'Size' refers to the size of the dataset. When modalities beyond sFC are used, they are marked with [d] (dFC) or [M] (multimodal).

Reference	Dataset (Size)	Dataset distribution	Atlas	GNN	Graph	Result	Baseline
Yu et al. (2022)	CNI-TLC (240)	120 ADHD, 120 TDC	CC200	Kipf	BG	68.5%	GAT 58.6%
Zhao et al. (2022b)	ADHD-200 (603)	260 ADHD, 343 TDC	AAL116	New	BG	72.0%	GAT 68.0%
Luo et al. (2024)	ADHD-200 (506)	73 ADHD, 75 TDC	AAL116	GCN	BG	75.8%	GCN 71.7%
Zhang et al. (2023b)	ADHD-200 (506)	215 ADHD, 291 TDC	Multiple	GIN	BG	70.9%	SAGE 62.1%

Table 7
Summary of findings from ASD studies. 'Size' refers to the size of the dataset. When modalities beyond sFC are used, they are marked with [d] (dFC) or [M] (multimodal).

Reference	Dataset (Size)	Dataset distribution	Atlas	GNN	Graph	Result	Baseline
Li et al. (2021a)	ABIDE (866)	402 ASD, 464 TDC	MODL128	Defferrard	BG	71.8%	k-NN 65.8%
Zhang et al. (2022a)	ABIDE (871)	403 ASD, 468 TDC	HO112	Defferrard	Mix	81.8%	pGCN 71.4%
Wang et al. (2022b)	ABIDE (949)	419 ASD, 530 TDC	Multiple	Kipf	PG	75.9%	MLP 75.2%
Zhu et al. (2021b)	ABIDE (1112)	Unclear	AAL116	GraphSAGE	BG	72.5%	pGCN 69.7%
Shao et al. (2021)	ABIDE (871)	403 ASD, 468 TDC	HO111	Kipf	PG	79.5%	MLP 78.1%
Wang et al. (2021)	ABIDE (1057)	525 ASD, 532 TDC	CC200	Edge	BG	70.7%	DNN 70.0%
Li et al. (2022b)	ABIDE (871)	403 ASD, 468 TDC	Multiple	Defferrard	Mix	86.3%	pGCN 69.5%
Zhu et al. (2022)	ABIDE (613) [d]	286 ASD, 327 TDC	HO110	Mix	PG	75.2%	MVGCN 72.0%
Cui et al. (2023)	ABIDE (1035) [d]	505 ASD, 530 TDC	CC200	Edge	Mix	73.1%	SVM 64.0%
Chen et al. (2022)	ABIDE (867) [d]	416 ASD, 451 TDC	HO110	Defferrard	BG	76.3%	GCN 73.2%
Chen et al. (2021)	ABIDE (1007) [M]	481 ASD, 526 TDC	AAL116	New	BG	72.7%	GCN 70.4%
Chen et al. (2024)	ABIDE (1007) [M]	481 ASD, 526 TDC	AAL116	New	BG	74.7%	GCN 70.4%
Li et al. (2020a)	Biopoint (118)	75 ASD, 43 TDC	DX148	GraphSAGE	BG	70.0%	–
Li et al. (2020b)	Biopoint (118)	75 ASD, 43 TDC	DK84	GAT	BG	79.7%	CNN 78.1%
Li et al. (2021b)	Biopoint (118)	75 ASD, 43 TDC	DK84	New	BG	79.8%	GAT 77.4%
Yang et al. (2022)	ABIDE (303)	130 ASD, 173 TDC	SF200	GIN	BG	70.6%	SVM 67.4%
Chu et al. (2022)	ABIDE (351)	155 ASD, 196 TDC	AAL116	Kipf	BG	70.0%	GCN 62.0%
Zhao et al. (2022a)	ABIDE NYU (92) [d]	45 ASD, 47 TDC	AAL116	Kipf	BG	79.9%	FCN 72.6%
Noman et al. (2022)	ABIDE (144) [d]	70 ASD, 74 TDC	Power264	Kipf	BG	66.0%	SVM 63.8%
Zheng et al. (2024a)	ABIDE (1064)	528 ASD, 536 TC	AAL116	GCN	BG	71.0%	GAT 68.0%
Zheng et al. (2024b)	ABIDE (1064)	528 ASD, 536 TC	AAL116	GCN	BG	71.0%	GAT 68.0%
Ma et al. (2024)	ABIDE (714)	334 ASD, 380 TC	AAL116	GCN	BG	70.3%	BrainGNN 67.3%
Zheng et al. (2024c)	ABIDE (1099)	528 ASD, 571 TC	AAL116	GIN	BG	70.2%	GIN 67.9%
Wang et al. (2024d)	ABIDE, 4 sites (355)	167 ASD, 188 TC	AAL116, HO111	HGCN	BG	78.5%	GAT 69.2%
Wang et al. (2024c)	ABIDE (860)	392 ASD, 458 TC	HO110, CC200	GCN	PG	70.7%	GCN 70.1%
Kong et al. (2025)	ABIDE (618) [d]	290 ASD, 328 TC	AAL90	GIN	BG	72.3%	STAGCN 69.1%
Fang et al. (2024)	ABIDE, 3 sites (312) [M]	148 ASD, 164 TC	AAL116	GCN	BG	82.4%	GCN 70.6%
Wang et al. (2024e)	ABIDE (871) [d]	403 ASD, 468 TC	Unclear	Cheb	PG	80.7%	GCN 76.4%
Gu et al. (2025)	ABIDE (871)	403 ASD, 468 HC	HO110	GCN	Both	99.8%	popGCN 96.8%
Wang et al. (2024a)	ABIDE NYU (184) [d]	79 ASD, 105 TC	AAL116	GIN	BG	73.2%	GCN 67.5%
Wei et al. (2023)	Private (138) [M]	61 ASD, 77 TC	AAL90	Cheb	BG	–	–
Bian et al. (2024)	ABIDE (663)	314 ASD, 349 TC	AAL90	New	BG	76.1%	GCN 71.7%
Gu et al. (2024)	ABIDE (871)	403 ASD, 468 TC	HO110	GCN	Both	99.8%	popGCN 96.8%
Wang and Xiao (2023)	ABIDE, 3 sites (307)	145 ASD, 162 TC	AAL116, HO111	HGCN	BG	71.7%	DNN 68.2%
Wang et al. (2023c)	ABIDE (1009)	570 ASD, 539 TC	AAL116, CC200	GCN	BG	72.4%	FBNETGEN 71.3%
Zheng et al. (2023)	ABIDE (582) [d]	289 ASD, 293 TC	Unclear	GCN	BG	75.0%	STAGIN 72.0%
Xu et al. (2024)	ABIDE (989)	455 ASD, 534 TC	SF100	GCN	BG	68.6%	BrainNetCNN 65.9%
Zhang et al. (2023b)	ABIDE (987)	467 ASD, 520 TC	Multiple	GIN	BG	80.7%	SAGE 71.1%
Zhang et al. (2023b)	ABIDE II (532)	243 ASD, 289 TC	Multiple	GIN	BG	79.9%	SAGE 71.2%
Wang et al. (2023b)	ABIDE, NYU (184)	79 ASD, 105 TC	AAL116	GIN	BG	72.6%	GCN 67.5%
Menon et al. (2023)	Private (117)	75 ASD, 43 TC	Unclear	GCN	BG	–	–
Hu et al. (2021)	ABIDE (1035)	505 ASD, 530 TC	HO110	GAT	BG	68.0%	Cheb 63.6%

Table 8
Summary of findings from MDD studies. 'Size' refers to the size of the dataset. When modalities beyond sFC are used, they are marked with [d] (dFC) or [M] (multimodal).

Reference	Dataset (Size)	Dataset distribution	Atlas	GNN	Graph	Result	Baseline
Qin et al. (2022)	REST-meta-MDD (1586)	821 MDD, 765 NC	DOS160	Defferrard	BG	81.5%	–
Gallo et al. (2023)	REST-meta-MDD, psymri (2498)	1249 MDD, 1249 NC	HO112, CC200	Kipf	BG	61.3%	SVM-RBF 61.2%
Kong et al. (2022)	Private (218)	129 MDD, 89 NC	BM82, JHU81	Kipf	BG	70.9%	GAT 68.2%
Jun et al. (2020)	Private (75)	29 MDD, 44 NC	Yeo114	Defferrard	PG	74.1%	SVM 69.8%
Kong et al. (2021)	Private (277) [d]	180 MDD, 97 NC	Unclear	GAT	BG	84.0%	–
Fang et al. (2023)	REST-meta-MDD (681) [d]	356 MDD, 325 NC	AAL116	Kipf	BG	59.3%	STNet 52.0%
Zheng et al. (2024a)	REST-meta-MDD (1604)	828 MDD, 776 HC	AAL116	GCN	BG	72.0%	GAT 63.0%
Zheng et al. (2024b)	REST-meta-MDD (1604)	828 MDD, 776 HC	AAL116	GCN	BG	73.0%	GAT 63.0%

(continued on next page)

Table 8 (*continued*).

Lee et al. (2024)	REST-meta-MDD (470)	245 MDD, 225 HC	Multiple	Cheb	BG	69.6%	MMTGCN 66.9%
Zheng et al. (2024c)	REST-meta-MDD (1604)	828 MDD, 776 HC	AAL 116	GIN	BG	70.0%	GIN 65.4%
Kong et al. (2025)	Zhongda (520) [d]	314 MDD, 206 HC	AAL 90	GIN	BG	73.2%	STAGCN 69.8%
Kong et al. (2025)	REST-meta-MDD (667) [d]	368 MDD, 299 HC	AAL90	GIN	BG	68.7%	STAGCN 66.0%
Liu and Gui (2024)	REST-meta-MDD S20 (533)	282 MDD, 251 HC	AAL116	GCN	BG	77.6%	GCN 75.5%
Gu et al. (2025)	REST-meta-MDD SU (533)	282 MDD, 251 HC	HO110	GCN	Both	93.4%	popGCN 71.6%
Zhao and Zhang (2024)	REST-meta-MDD (1611) [d]	832 MDD, 779 HC	AAL116	GIN	BG	67.1%	STAGIN 64.1%
Gu et al. (2024)	REST-meta-MDD SU (533)	282 MDD, 251 HC	HO110	GCN	Both	93.4%	popGCN 71.6%
Zhang et al. (2023e)	REST-meta-MDD (2361)	1256 MDD, 1105 HC	AAL116	GCN	BG	74.2%	GCN (PG) 61.0%
Kong et al. (2023)	Private (187) [M]	93 MDD, 94 HC	AAL116, Destrieux, HO	GCN	BG	80.2%	MMGNN 73.6%
Dai et al. (2023)	REST-meta-MDD (615)	189 MDD, 426 HC	AAL116, CC200, DOS	GCN	Temporal	75.8%	70.7%
Zhao and Zhang (2022)	REST-meta-MDD (2361)	1256 MDD, 1105 HC	AAL116	GIN	BG	64.8%	SAGE 61.5%
Wang et al. (2023b)	REST-meta-MDD 20 (533) [d]	282 MDD, 251 HC	AAL116	GIN	BG	60.9%	GCN 55.7%

Table 9
Summary of findings from SZ studies. 'Size' refers to the size of the dataset. When modalities beyond sFC are used, they are marked with [d] (dFC) or [M] (multimodal).

Reference	Dataset (Size)	Dataset distribution	Atlas	GNN	Graph	Result	Baseline
Lei et al. (2022)	Private (1412)	505 SZ, 907 NC	AAL116, DOS160	Defferrard	BG	85.8%	SVM 80.9%
Chen et al. (2023b)	Private (345) [M]	140 SZ, 205 NC	AAL90, BNA246	Kipf	BG	95.8%	SVM 81.2%
Sebenius et al. (2021)	COBRE (154) [M]	67 SZ, 87 NC	DK293	Kipf	BG	75.0%	SVM 71.0%
Zheng et al. (2024a)	SRPBS (184)	92 SZ, 92 HC	AAL116	GCN	BG	93.0%	GAT 84.0%
Zheng et al. (2024b)	SRPBS (184)	92 SZ, 92 HC	AAL116	GCN	BG	91.0%	GAT 84.0%
Zhu et al. (2024)	COBRE (112) [d]	48 SZ, 64 HC	AAL90	New	BG	83.6%	BrainGNN 81.3%
Zhu et al. (2024)	UCLA (80) [d]	41 SZ, 39 HC	AAL90	New	BG	89.7%	BrainGNN 84.9%
Wang et al. (2024b)	SRPBS (647)	142 SZ, 505 HC	BV140	GCN	BG	82.5%	GAT 75.4%
Sunil et al. (2024)	UCLA (177)	50 SZ, 122 HC	Custom (AAL3 + HO Cortical)	GCN	BG	82.0%	GCN 77.0%
Li et al. (2023)	In-house (143) [M]	70 SZ, 73 HC	AAL116	GCN	BG	78.4%	SAGE 75.1%
Fan et al. (2023)	Multiple (1003) [d]	411 SZ, 592 HC	AAL116	GIN	BG	82.8%	STAGIN 71.5%

Table 10
Summary of model performance, categorised by type of GNN, across all datasets. Note that GAT, HGCN has been grouped under others and GraphSAGE has been grouped under Spatial GCN.

Core GNN architecture	Mean accuracy	Standard deviation	Strengths	Weaknesses
Spectral GCN	79.7	6.6	Based on strong theoretical foundations based on spectral graph theory.	Higher computational complexity for higher orders, more hyperparameters to tune.
Spatial GCN	77.5	10.1	Simple formulation with better scalability.	Over-smoothing issues if too many layers are added, reducing model performance.
GIN	71.9	5.8	Theoretically more expressive than baseline GNN architectures.	Possible overfitting issues if not properly regularised.
Others	76.6	5.9	Typically customised to ground the GNN construction or learning process to existing biological knowledge.	Typically more complicated and lack of external benchmarking efforts to verify performance improvements.

Table 11
Summary of findings from ADHD studies that identified potential biomarkers. 'Size' refers to the size of the dataset. When modalities beyond sFC are used, they are marked with [d] (dFC) or [M] (multimodal). 'Type' refers to the granularity of attributions.

Reference	Dataset (Size)	Atlas	Attributor (Type)	Salient features
Yu et al. (2022)	CNI-TLC (240)	CC200	Attention (ROI, connection)	**Regions** (−): FP, IFG, MFG, MTG, posterior central gyrus, SOG; **Regions** (+): Brainstem, FOG, precuneus, putamen, TP; **Connections** (−): IFG-MTG, MFG-FP, MFG-posterior central gyrus, posterior central gyrus-SFG; **Connections** (+): Cerebellum-fusiform gyrus, FOG-Precuneus, PARAH-STG
Zhao et al. (2022b)	ADHD-200 (603)	AAL116	Weights (ROI, connection)	**Regions**: Frontal lobe, occipital lobe, subcortical, temporal lobe, posterior-fossa (cerebellum); **Connections** (−): right rolandic operculum - right Heschl gyrus; **Connections** (+): left precuneus - right cerebellum, left STG/TP - right medial SFG
Luo et al. (2024)	ADHD-200 (506)	AAL116	Attention (ROI)	**Regions**: bilateral olfactory cortex, bilateral MFG (orbital part), left gyrus rectus, bilateral posterior cingulum, bilateral putamen
Zhang et al. (2023b)	ADHD-200 (506)	Multiple	Trainable mask (ROI)	**Regions**: right PrCG, right thalamus, and right MFG

Table 12
Summary of findings from ASD studies that identified potential biomarkers. 'Size' refers to the size of the dataset. When modalities beyond sFC are used, they are marked with [d] (dFC) or [M] (multimodal). 'Type' refers to the granularity of attributions.

Reference	Dataset (Size)	Atlas	Attributor (Type)	Salient features
Li et al. (2022b)	ABIDE (866)	Multiple	Finetuning (ROI)	**Regions**: STG ; Cerebellum IV and V, central parieto-occipital sulcus, left superior temporal sulcus, left anterior intraparietal sulcus, cerebrospinal fluid (between superior part of SFG and skull)
Zhang et al. (2022a)	ABIDE (871)	HO112	Attention (ROI)	**Regions**: IFG, PrCG, frontal orbital cortex, PARAH.
Wang et al. (2022b)	ABIDE (949)	Multiple	Unclear (ROI)	**Regions**: angular gyrus, PrCG, precuneus and thalamus
Zhu et al. (2021b)	ABIDE (1112)	AAL116	Pooling (ROI)	**Regions**: lateral PFC, lateral dorsal PFC, SPL
Shao et al. (2021)	ABIDE (871)	HO111	Feature selection (connection)	**Connections**: evenly distributed across the brain, lower in ASD for 25/30 FC e.g. right cuneal cortex and right parietal operculum cortex
Wang et al. (2021)	ABIDE (1057)	CC200	Occlusion (module)	**Modules**: DMN, FPN, VAN
Li et al. (2022b)	ABIDE (871)	Multiple	Clustering (module, modular connection)	**Modules**: CEN, DMN, SN ; Modular connections: CEN-SMN, CEN-SN, DMN-SMN, DMN-CEN, DMN-Visual
Zhu et al. (2022)	ABIDE (613) [d]	HO110	Feature selection (ROI)	**Regions**: lingual gyrus, MFG, SFG
Cui et al. (2023)	ABIDE (1035) [d]	CC200	Gradients (ROI)	**Regions**: FP, precuneus, brain stem, paracingulate gyrus, lingual, OP, lateral occipital cortex, frontal orbital cortex
Chen et al. (2022)	ABIDE (867) [d]	HO110	Feature selection (connection)	**Connections**: right pallidum-right IFG, left frontal orbital cortex-left central opercular cortex, left supramarginal gyrus - right ITG
Chen et al. (2021)	ABIDE (1007) [M]	AAL116	Gradients (ROI, connection)	**Regions**: (higher T1w): DMN, reward, memory and motor ; (higher ALFF) reward and motor ; **Connections**: inter >intra, low homotopic interhemispheric connection in limbic regions
Chen et al. (2024)	ABIDE (1007) [M]	AAL116	Gradients (connection)	**Connections** (−): right MTG and multiple ROIs in the frontal, parietal, and occipital lobes; mix of higher and lower FCs between ROIs in the limbic regions to multiple other regions
Li et al. (2020a)	Biopoint (118)	DX148	Clustering (ROI)	**Regions**: PFC, cingulate cortex
Li et al. (2020b)	Biopoint (118)	DK84	Pooling (ROI)	**Regions**: dorsal striatum, thalamus, frontal gyrus
Li et al. (2021b)	Biopoint (118)	DK84	Pooling (ROI)	**Regions**: frontal gyrus, temporal lobe, cingulate gyrus, occipital pole, angular gyrus
Yang et al. (2022)	ABIDE (303)	SF200	Pooling (ROI)	**Regions**: right parietal cortex, left visual cortex, right lateral PFC, left PFC, left cingulate cortex
Chu et al. (2022)	ABIDE (351)	AAL116	Attention (ROI)	**Regions**: hippocampus, PARAH, putamen, thalamus
Zhao et al. (2022a)	ABIDE (92) [d]	AAL116	Pooling (ROI)	**Regions** (Lo): bilateral cerebellum and right hippocampus ; **Regions** (Ho): left insula, left putamen, medial aspect of right SFG

(*continued on next page*)

Table 12 (*continued*).

Noman et al. (2022)	ABIDE (144) [d]	Power264	Clustering (module)	**Modules**: ASD stronger FC in visual, DMN and SN ; TDC higher FC in sensory and auditory networks.
Zheng et al. (2024a)	ABIDE (1064)	AAL116	Info bottleneck (connection, module)	**Connections**: right STG - (left PARAH, right PrCG, right PoCG), **Modules**: SMN, SMN-LN
Zheng et al. (2024b)	ABIDE (1064)	AAL116	Prototype (connection)	**Connections**: right SFG (orbital part) - right orbitofrontal cortex, right PoCG - (left PoCG, right PrCG)
Ma et al. (2024)	ABIDE (714)	AAL116	Ablation (module)	**Modules**: DMN
Zheng et al. (2024c)	ABIDE (1099)	AAL116	Info bottleneck (connection)	**Connections**: right superior parietal gyrus - (vermis, left MOG, right orbitofrontal cortex, cerebellum)
Wang et al. (2024d)	ABIDE (355)	AAL116	GradCAM (ROI)	**Regions**: MTG, MFG, Precuneus, STG, Lingual gyrus
Wang et al. (2024c)	ABIDE (860)	CC200	Weights (ROI)	**Regions**: STG, Precuneus, supramarginal gyrus, SPL, MFG, parahippocampal gyrus, angular gyrus, putamen, left pallidum
Kong et al. (2025)	ABIDE (618) [d]	AAL90	Attention (ROI, connection)	**Regions**: left putamen, bilateral insula; **Connections**: left medial SFG - right medial STF, left orbitofrontal cortex (superior medial) - right orbitofrontal cortex (superior medial) ; left IFG (triangular) - left IFG (operculum), right IFG (triangular) - right IFG (operculum)
Fang et al. (2024)	ABIDE (312) [M]	AAL116	Attention (ROI)	**Regions**: right posterior cingulate gyrus, right ACG, right paracentral lobule, vermis
Wang et al. (2024e)	ABIDE (871) [d]	Unclear	Elastic Net (Connection)	**Connections**: right rectus gyrus - right cerebellum, left paracentral lobule - right middle occipital cortex, left precuneus - left cerebellum
Gu et al. (2025)	ABIDE (871)	HO110	Pooling (ROI)	**Regions**: temporal gyrus, PARAH, frontal cortex, accumbens, central cortex, thalamus, brainstem, and caudate nucleus
Wang et al. (2024a)	ABIDE (184) [d]	AAL116	Lasso (connection)	**Connections**: left SMG, right ACG ; left angular gyrus - left middle orbitofrontal cortex
Wei et al. (2023)	Private (138) [M]	AAL90	Eigenvalues (ROI)	**Regions**: left middle orbitofrontal cortex, bilateral supplementary motor area, left PrCG, bilateral rolandic operculum
Bian et al. (2024)	ABIDE (663)	AAL90	Graph filtration (connection)	**Connections**: left amygdala - left ITG, left hippocampus - right PoCG, left thalamus - right MTG, left dorsal SFG - left insula
Gu et al. (2024)	ABIDE (871)	HO110	Pooling (ROI)	**Regions**: Temporal gyrus, PARAH, frontal cortex, accumbens, central cortex, thalamus, brainstem, caudate nucleus
Wang and Xiao (2023)	ABIDE (307)	AAL116	GradCAM (ROI)	**Regions**: Precuneus
Wang et al. (2023c)	ABIDE (1009)	AAL116	Pooling (ROI)	**Regions**: Temporal pole (MTG, STG), cerebellum
Zheng et al. (2023)	ABIDE (582) [d]	Unclear	Attention (modular connection)	**Modular connection**: LN - SMN
Xu et al. (2024)	ABIDE (989)	SF100	Attention (connection)	**Connections**: connections between PFC, parietal lobe, cingulate cortex
Zhang et al. (2023b)	ABIDE (987)	AAL90	Trainable mask (ROI)	**Regions**: right MTG, left fusiform gyrus, left MOG, right linugal gryus, left MTG, right precuneus
Zhang et al. (2023b)	ABIDE II (532)	AAL90	Trainable mask (ROI)	**Regions**: right MTG, left cerebellum, left MOG, right SFG, right STG, right precuneus
Wang et al. (2023b)	ABIDE NYU (184)	AAL116	Lasso	**Connections**: thalamus, MTG - cerebellum
Menon et al. (2023)	Private (114)	Unclear	SubgraphX	**Regions**: right frontal orbital cortex
Hu et al. (2021)	ABIDE (1035)	HO110	Multiple (connection)	**Connections**: right hippocampus - right frontal medial cortex, left frontal pole - right IFG, left PrCG

Table 13
Summary of findings from MDD studies that identified potential biomarkers. 'Size' refers to the size of the dataset. When modalities beyond sFC are used, they are marked with [d] (dFC) or [M] (multimodal). 'Type' refers to the granularity of attributions.

Reference	Dataset (Size)	Atlas	Attributor (Type)	Salient features
Qin et al. (2022)	REST-meta-MDD (1586)	DOS160	CAM (ROI, module)	**Regions**: right dorsal ACC, right ventrolateral PFC, left IPL, left posterior insula. **Modules**: DMN, FPN, Cingulo-Opercular network.
Gallo et al. (2023)	Multiple (2498)	Multiple	Multiple (connection)	**Connections**: stronger inter-hemispheric thalamic connection across most datasets and attributors
Kong et al. (2022)	Private (218)	Multiple	Weights (connection)	**Connections**: right anterior corona radiata - left dorsolateral PFC

(*continued on next page*)

Table 13 (*continued*).

Jun et al. (2020)	Private (75)	YEO114	Gradients (connection)	**Connections** (reduced in EC): left dorsal PFC - left precentral ventral region, striate cortex, parietal medial region, IPL, PARAH
Kong et al. (2021)	Private (277) [d]	Unclear	Unclear (ROI)	**Regions**: bilateral pallidum, right putamen, bilateral MFG, right PoCG, right Heschl gyrus, right caudate, right olfactory cortex, right IFG, triangular part.
Fang et al. (2023)	REST-meta-MDD (681) [d]	AAL116	Attention (connection, temporal)	**Connections**: cross-hemisphere connections within the insula and lingual gyrus ; lingual gyrus - calcarine sulcus ; Temporal: middle & end of time series
Zheng et al. (2024a)	REST-meta-MDD (1604)	AAL116	Info bottleneck (connection)	**Connections**: left rectus - (left cerebellum, right cuneus, right paracingulate gyri, left lingual gyrus, left SFG (medial orbital))
Zheng et al. (2024b)	REST-meta-MDD (1604)	AAL116	Prototype (connection)	**Connections**: left fusiform gyrus - right fusiform gyrus, left SPG - right SPG
Lee et al. (2024)	REST-meta-MDD (470)	HO112	t-test (ROI)	**Region**: ITG, frontal medial cortex, subcallosal cortex, PHG, temporal fusiform cortex, occipital gyrus, brainstem
Zheng et al. (2024c)	REST-meta-MDD (1604)	AAL116	Info bottleneck (connection)	**Connections**: left MTG - (right rolandic operculum, left fusiform gyrus, left PARAH, right SFG (medial orbital))
Kong et al. (2025)	Zhongda (520) [d]	AAL90	Attention (connection, ROI)	**Connections**: bilateral caudate, thalamus, paracentral lobule, posterior cingulate gyrus, cuneus, SOG ; left IFG triangular part - opercular part, right PrCG - right PoCG ; **Regions**: bilateral lingual gyrus, PoCG, PrCG, SOG ; right cuneus, left calcarine fissure
Kong et al. (2025)	REST-meta-MDD (667) [d]	AAL90	Attention (connection, ROI)	**Connections**: bilateral caudate, insula, PoCG, SMG, SPG, IOG, SOG, left IFG triangular part - opercular part, right IFG triangular part - opercular part ; **Regions**: left posterior cingulate gyrus, right fusiform gyrus, right ITG, left MOG, left lingual gyrus, left calcarine fissure
Liu and Gui (2024)	REST-meta-MDD (533)	AAL116	Pooling (ROI)	**Regions**: Precentral gyrus, SFG, Cuneus, Lingual gyrus, and Fusiform gyrus
Gu et al. (2025)	REST-meta-MDD (533)	HO110	Pooling (ROI)	**Regions**: left pallidum, right anterior ITG, left frontal orbital, right PARAH, left thalamus
Zhao and Zhang (2024)	REST-meta-MDD (1611) [d]	AAL116	Attention (ROI, modules)	**Regions**: bilateral rectus, MTG, right middle frontal gyrus (orbital part) ; **Modules**: DAN, FPN ; LN, DMN
Gu et al. (2024)	REST-meta-MDD (533)	HO110	Pooling (ROI)	**Regions**: pallidum, ITG, frontal operculum cortex, PARAH, thalamus, temporal fusiform cortex, amygdala, accumbens and orbital cortex
Zhang et al. (2023e)	REST-meta-MDD (2361)	AAL116	Subgraph (ROI)	**Regions**: Thalamus
Kong et al. (2023)	Private (187) [M]	AAL116	Weights (ROI)	**Regions**: bilateral medial SFG, bilateral dorsolateral SFG, bilateral caudate nucleus, bilateral precuneus, right hippocampus, and left lenticular nucleus
Dai et al. (2023)	REST-meta-MDD (615)	AAL116	GradCAM (ROI)	**Regions**: left insula, right amygdala, left SPG
Zhao and Zhang (2022)	REST-meta-MDD (2361)	AAL116	Pooling (ROI, module)	**Regions**: left inferior parietal lobe, right MFG, left PrCG ; **Modules**: DMN, LN
Wang et al. (2023b)	REST-meta-MDD (533) [d]	AAL116	Lasso (connection)	**Connections**: hippocampus-IFG, cerebellum-support motor, thalamus-temporal pole/MTG

Table 14
Summary of findings from SZ studies that identified potential biomarkers. 'Size' refers to the size of the dataset. When modalities beyond sFC are used, they are marked with [d] (dFC) or [M] (multimodal). 'Type' refers to the granularity of attributions.

Reference	Dataset (Size)	Atlas	Attributor (Type)	Salient features
Lei et al. (2022)	Multiple (1412)	Multiple	CAM (ROI)	**Regions**: decreased nodal efficiency in bilateral putamen and pallidum in SZ, across both atlases
Chen et al. (2023b)	Private (345) [M]	Multiple	Pooling (ROI)	**Regions**: bilateral rectus gyrus, bilateral lingual gyrus, bilateral cuneus; right medial orbitofrontal cortex, medial SFG, calcarine cortex, ACG
Sebenius et al. (2021)	COBRE (154) [M]	DK293	Pooling (ROI)	**Regions**: ROI relevance scores had a stronger correlation with SC than FC (specific regions unclear from plot)
Zheng et al. (2024a)	SRPBS (184)	AAL116	Information bottleneck (connection)	**Connections**: left hippocampus - left PARAH, left hippocampus - right hippocampus
Zheng et al. (2024b)	SRPBS (184)	AAL116	Prototype (connection)	**Connections**: right ACG - right orbital part of MFG, left rectus - right orbital part of MFG, left orbital part of MFG - right orbital part of MFG ; left STG - left insula, left SMG - left insula, left putamen - left insula
Zhu et al. (2024)	COBRE (112) [d]	AAL90	Pooling (ROI, module)	**Regions**: left SOG, left fusiform gyrus, left medial SFG, left calcarine fissure, left IOG, left PrCG ; **Modules**: VN, DMN, DAN
Zhu et al. (2024)	UCLA (80) [d]	AAL90	Pooling (ROI, module)	**Regions**: left lingual gyrus, left medial SFG, left IOG, left PrCG, left calcarine fissure ; **Modules**: DMN, VN, DAN

(*continued on next page*)

Table 14 (*continued*).

Wang et al. (2024b)	SRPBS (647)	BV140	GNNExplainer (ROI)	**Regions**: ventricle, temporal gyrus
Sunil et al. (2024)	UCLA (177)	Custom AAL164	GNNExplainer (ROI)	**Regions**: Supramarginal Gyrus (anterior division), ITG (posterior, temporooccipital, anterior division), STG (posterior division, left), right SPL, MTG (temporooccipital part, right)
Li et al. (2023)	In-house (143) [M]	AAL116	GradCAM, Saliency (ROI)	**Regions**: bilateral insula, MFG, lower FG, MOG, angular gyrus, superior marginal gyrus
Fan et al. (2023)	Multiple (1034) [d]	AAL116	Attention (module)	**Module**: subcortical-cerebellar circuit

Table 15
Summary of findings from dementia studies that identified potential biomarkers. 'Size' refers to the size of the dataset. When modalities beyond sFC are used, they are marked with [d] (dFC) or [M] (multimodal). 'Type' refers to the granularity of attributions.

Reference	Dataset (Size)	Atlas	Attributor (Type)	Salient features
Significant memory concern				
Zuo et al. (2022)	ADNI (168) [d]	AAL90	Occlusion (ROI)	**Regions**: orbital part of the SFG, anterior cingulate, paracingulate gyri, calcarine fissure, lingual gyrus, precuneus, paracentral lobule, caudate nucleus, lenticular nucleus putamen
Zhu et al. (2021a)	ADNI (138) [d]	AAL90	Unclear (ROI)	**Regions**: ITG, MFG, IFG, left hippocampus
Song et al. (2022)	Multiple (207) [M]	AAL90	Unclear (ROI)	**Regions**: right ITG, right insula, left olfactory cortex, left angular gyrus, right amygdala, right precuneus
Song et al. (2021)	ADNI (88) [M]	AAL90	RFE (connection)	**Connections**: function: precentral gyrus (left and right) - left superior TP ; structure: left precentral gyrus - left putamen
Early mild cognitive impairment				
Mei et al. (2022)	ADNI (910)	SF200	Pooling (module)	**Modules**: DAN, VAN, DMN
Lee et al. (2021)	ADNI (101) [d]	YEO114	RL (ROI)	**Regions**: ROIs in DMN, including bilateral parahippocampal cortices; right insula, involving both SMN and SN/VAN
Yu et al. (2019)	ADNI (88) [d]	AAL90	Weights (ROI)	**Regions**: medial part of left SFG, right putamen
Zhu et al. (2021a)	ADNI (180) [d]	AAL90	Unclear (ROI)	**Regions**: ITG, MFG, left hippocampus and IFG
Lei et al. (2023)	ADNI (154) [M]	AAL90	Occlusion (ROI)	**Regions**: IFG, olfactory cortex, PARAH, MOG, ITG
Song et al. (2022)	Multiple (249) [M]	AAL90	Unclear (ROI)	**Regions**: right ITG, right insula, left olfactory cortex, left angular gyrus, right amygdala, right precuneus
Song et al. (2021)	ADNI (88) [M]	AAL90	RFE (connection)	**Connections**: function: right precentral gyrus - right putamen ; structure: left superior frontal lobe - left thalamus
Yu et al. (2019)	ADNI (82) [d]	AAL90	Weights (ROI)	**Regions**: medial part of left superior frontal gyrus and right putamen
Lei et al. (2023)	ADNI (107) [M]	AAL90	Occlusion (ROI)	**Regions**: IFG, olfactory cortex, PARAH, MOG, ITG
Song et al. (2022)	Multiple (329) [M]	AAL90	Unclear (ROI)	**Regions**: right ITG, right insula, left olfactory cortex, left angular gyrus, right amygdala, right precuneus
Song et al. (2021)	ADNI (82) [M]	AAL90	RFE (connection)	**Connections**: function: orbital part of the MFG/SFG - right ITG ; structure: left superior frontal lobe - left thalamus
Alzheimer's disease				
Alorf and Khan (2022)	ADNI (83)	AAL90	Weights (ROI)	**Regions**: left MFG, left orbital SFG, right precentral gyrus
Xing et al. (2021)	ADNI (292) [d]	AAL116	GradCAM (ROI)	**Regions**: bilateral hippocampus, right precuneus, right frontal mid-cortex, left precentral cortex
Wang et al. (2022a)	ADNI (107) [d]	AAL90	Attention (ROI)	**Regions**: frontal and temporal regions

Data availability

Data generated and code used for the review paper can be found in https://osf.io/wza6b/.

References

Abi-Dargham, A., Moeller, S.J., Ali, F., DeLorenzo, C., Domschke, K., Horga, G., Jutla, A., Kotov, R., Paulus, M.P., Rubio, J.M., et al., 2023. Candidate biomarkers in psychiatric disorders: state of the field. World Psychiatry 22 (2), 236–262.

Agarwal, C., Queen, O., Lakkaraju, H., Zitnik, M., 2023. Evaluating explainability for graph neural networks. Sci. Data 10 (1), 144.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F., 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Inf. Fusion 99, 101805.

Alorf, A., Khan, M.U.G., 2022. Multi-label classification of Alzheimer's disease stages from resting-state fMRI-based correlation connectivity data and deep learning. Comput. Biol. Med. 151, 106240.

Aron, A.R., Robbins, T.W., Poldrack, R.A., 2004. Inhibition and the right inferior frontal cortex. Trends Cogn. Sci. 8 (4), 170–177.

Aronson, J.K., Ferner, R.E., 2017. Biomarkers—a general review. Curr. Protoc. Pharmacol. 76 (1), 9–23.

Arya, D., Olij, R., Gupta, D.K., El Gazzar, A., Wingen, G., Worring, M., Thomas, R.M., 2020. Fusing structural and functional MRIs using graph convolutional networks for autism classification. In: Medical Imaging with Deep Learning. PMLR, pp. 44–61.

Atanasova, P., 2024. A diagnostic study of explainability techniques for text classification. In: Accountable and Explainable Methods for Complex Reasoning over Text. Springer, pp. 155–187.

Bai, B., Liang, J., Zhang, G., Li, H., Bai, K., Wang, F., 2021a. Why attentions may not be interpretable? In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 25–34.

Bai, S., Zhang, F., Torr, P.H., 2021b. Hypergraph convolution and hypergraph attention. Pattern Recognit. 110, 107637.

Bayer, J.M., Thompson, P.M., Ching, C.R., Liu, M., Chen, A., Panzenhagen, A.C., Jahanshad, N., Marquand, A., Schmaal, L., Sämann, P.G., 2022. Site effects how-to and when: An overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses. Front. Neurol. 13, 923988.

Bessadok, A., Mahjoub, M.A., Rekik, I., 2022. Graph neural networks in network neuroscience. IEEE Trans. Pattern Anal. Mach. Intell. 45 (5), 5833–5848.

Bian, C., Xia, N., Xie, A., Cong, S., Dong, Q., 2024. Adversarially trained persistent homology based graph convolutional network for disease identification using brain connectivity. IEEE Trans. Med. Imaging 43 (1), 503–516.

Bielczyk, N.Z., Uithol, S., van Mourik, T., Anderson, P., Glennon, J.C., Buitelaar, J.K., 2019. Disentangling causal webs in the brain using functional magnetic resonance imaging: A review of current approaches. Netw. Neurosci. 3 (2), 237–273.

Bintsi, K.-M., Mueller, T.T., Starck, S., Baltatzis, V., Hammers, A., Rueckert, D., 2023. A comparative study of population-graph construction methods and graph neural networks for brain age regression. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 64–73.

Biswal, B.B., Uddin, L.Q., 2025. The history and future of resting-state functional magnetic resonance imaging. Nature 641 (8065), 1121–1131.

Bondi, E., Maggioni, E., Brambilla, P., Delvecchio, G., 2023. A systematic review on the potential use of machine learning to classify major depressive disorder from healthy controls using resting state fMRI measures. Neurosci. Biobehav. Rev. 144, 104972.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J.A., Adcock, R.A., et al., 2020. Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582 (7810), 84–88.

Calhoun, V.D., Liu, J., Adalı, T., 2009. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. Neuroimage 45 (1), S163–S172.

Califf, R.M., 2018. Biomarker definitions and their applications. Exp. Biol. Med. 243 (3), 213–221.

Canario, E., Chen, D., Biswal, B., 2021. A review of resting-state fMRI and its use to examine psychiatric disorders. Psychoradiology 1 (1), 42–53.

Cash, R.F., Müller, V.I., Fitzgerald, P.B., Eickhoff, S.B., Zalesky, A., 2023. Altered brain activity in unipolar depression unveiled using connectomics. Nat. Ment. Heal. 1 (3), 174–185.

Chan, Y.H., Yew, W.C., Chew, Q.H., Sim, K., Rajapakse, J.C., 2023. Elucidating salient site-specific functional connectivity features and site-invariant biomarkers in schizophrenia via deep neural networks. Sci. Rep. 13 (1), 21047.

Chan, Y.H., Yew, W.C., Rajapakse, J.C., 2022. Semi-supervised learning with data harmonisation for biomarker discovery from resting state fMRI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 441–451.

Chen, X., Ke, P., Huang, Y., Zhou, J., Li, H., Peng, R., Huang, J., Liang, L., Ma, G., Li, X., et al., 2023b. Discriminative analysis of schizophrenia patients using graph convolutional networks: A combined multimodal MRI and connectomics analysis. Front. Neurosci. 17, 1140801.

Chen, Y., Liu, A., Fu, X., Wen, J., Chen, X., 2022. An invertible dynamic graph convolutional network for multi-Center ASD classification. Front. Neurosci. 15, 828512.

Chen, J., Patil, K.R., Yeo, B.T., Eickhoff, S.B., 2023a. Leveraging machine learning for gaining neurobiological and nosological insights in psychiatric research. Biol. Psychiatry 93 (1), 18–28.

Chen, Y., Yan, J., Jiang, M., Zhang, T., Zhao, Z., Zhao, W., Zheng, J., Yao, D., Zhang, R., Kendrick, K.M., Jiang, X., 2024. Adversarial learning based node-edge graph attention networks for autism spectrum disorder identification. IEEE Trans. Neural Netw. Learn. Syst. 35 (6), 7275–7286. http://dx.doi.org/10.1109/TNNLS.2022.3154755.

Chen, Y., Yan, J., Jiang, M., Zhao, Z., Zhao, W., Zhang, R., Kendrick, K.M., Jiang, X., 2021. Attention-based node-edge graph convolutional networks for identification of autism spectrum disorder using multi-modal mri data. In: Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part III 4. Springer, pp. 374–385.

Chollet, F., Payoux, P., 2022. Functional imaging for neurodegenerative diseases. Press. Méd. 51 (2), 104121.

Chu, Y., Ren, H., Qiao, L., Liu, M., 2022. Resting-state functional MRI adaptation with attention graph convolution network for brain disorder identification. Brain Sci. 12 (10), 1413.

Cosmo, L., Kazi, A., Ahmadi, S.-A., Navab, N., Bronstein, M., 2020. Latent-graph learning for disease prediction. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23. Springer, pp. 643–653.

Coutanche, M.N., Thompson-Schill, S.L., Schultz, R.T., 2011. Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. Neuroimage 57 (1), 113–123.

Cui, H., Dai, W., Zhu, Y., Kan, X., Gu, A.A.C., Lukemire, J., Zhan, L., He, L., Guo, Y., Yang, C., 2022. Braingb: A benchmark for brain network analysis with graph neural networks. IEEE Trans. Med. Imaging 42 (2), 493–506.

Cui, W., Du, J., Sun, M., Zhu, S., Zhao, S., Peng, Z., Tan, L., Li, Y., 2023. Dynamic multi-site graph convolutional network for autism spectrum disorder identification. Comput. Biol. Med. 157, 106749.

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., Varoquaux, G., Initiative, A.D.N., et al., 2019. Benchmarking functional connectome-based predictive models for resting-state fMRI. NeuroImage 192, 115–134.

Dadi, K., Varoquaux, G., Machlouzarides-Shalit, A., Gorgolewski, K.J., Wassermann, D., Thirion, B., Mensch, A., 2020. Fine-grain atlases of functional modes for fMRI analysis. NeuroImage 221, 117126.

Dai, P., Lu, D., Shi, Y., Zhou, Y., Xiong, T., Zhou, X., Chen, Z., Zou, B., Tang, H., Huang, Z., et al., 2023. Classification of recurrent major depressive disorder using a new time series feature extraction method through multisite rs-fMRI data. J. Affect. Disord. 339, 511–519.

Defferrard, M., Bresson, X., Vandergheynst, P., 2016. Convolutional neural networks on graphs with fast localized spectral filtering. Adv. Neural Inf. Process. Syst. 29.

Del Campo, N., Chamberlain, S.R., Sahakian, B.J., Robbins, T.W., 2011. The roles of dopamine and noradrenaline in the pathophysiology and treatment of attention-deficit/hyperactivity disorder. Biol. Psychiatry 69 (12), e145–e157.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., 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.

Dong, Y., Batalle, D., Deprez, M., 2024. Reproducible comparison and interpretation of machine learning classifiers to predict autism on the ABIDE multimodal dataset. MedRxiv 2024–09.

Du, Y., Fu, Z., Calhoun, V.D., 2018. Classification and prediction of brain disorders using functional connectivity: promising but challenging. Front. Neurosci. 12, 525.

Duda, M., Iraji, A., Ford, J.M., Lim, K.O., Mathalon, D.H., Mueller, B.A., Potkin, S.G., Preda, A., Van Erp, T.G., Calhoun, V.D., 2023. Reliability and clinical utility of spatially constrained estimates of intrinsic functional networks from very short fMRI scans. Hum. Brain Mapp. 44 (6), 2620–2635.

ElGazzar, A., Thomas, R., Van Wingen, G., 2022. Benchmarking graph neural networks for FMRI analysis. arXiv preprint arXiv:2211.08927.

Esfahlani, F.Z., Byrge, L., Tanner, J., Sporns, O., Kennedy, D.P., Betzel, R.F., 2022. Edge-centric analysis of time-varying functional brain networks with applications in autism spectrum disorder. NeuroImage 263, 119591.

Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., et al., 2019. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nature Methods 16 (1), 111–116.

Fan, Z., Su, J., Xu, J., Hu, D., Zeng, L.-L., 2023. DGST-former: Dynamic graph with spatio-temporal transpose-transformer for brain fMRI data. In: 2023 7th Asian Conference on Artificial Intelligence Technology. ACAIT, IEEE, pp. 530–535.

Fang, Y., Wang, M., Potter, G.G., Liu, M., 2023. Unsupervised cross-domain functional MRI adaptation for automated major depressive disorder identification. Med. Image Anal. 84, 102707.

Fang, J., Zhang, D.-f., Xie, K., Xu, L., Bi, X.-a., 2024. Bilinear perceptual fusion algorithm based on brain functional and structural data for ASD diagnosis and regions of interest identification. Interdiscip. Sci.: Comput. Life Sci. 1–15.

Faskowitz, J., Esfahlani, F.Z., Jo, Y., Sporns, O., Betzel, R.F., 2020. Edge-centric functional network representations of human cerebral cortex reveal overlapping system-level architecture. Nature Neurosci. 23 (12), 1644–1654.

Feng, A., You, C., Wang, S., Tassiulas, L., 2022. Kergnns: Interpretable graph neural networks with graph kernels. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 6614–6622.

Filippi, M., Spinelli, E.G., Cividini, C., Ghirelli, A., Basaia, S., Agosta, F., 2023. The human functional connectome in neurodegenerative diseases: relationship to pathology and clinical progression. Expert. Rev. Neurother. 23 (1), 59–73.

de Filippis, R., Carbone, E.A., Gaetano, R., Bruni, A., Pugliese, V., Segura-Garcia, C., De Fazio, P., 2019. Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: a systematic review. Neuropsychiatr. Dis. Treat. 1605–1627.

Fox, P.T., Laird, A.R., Fox, S.P., Fox, P.M., Uecker, A.M., Crank, M., Koenig, S.F., Lancaster, J.L., 2005. BrainMap taxonomy of experimental design: description and evaluation. Hum. Brain Mapp. 25 (1), 185–198.

Gallo, S., El-Gazzar, A., Zhutovsky, P., Thomas, R.M., Javaheripour, N., Li, M., Bartova, L., Bathula, D., Dannlowski, U., Davey, C., et al., 2023. Functional connectivity signatures of major depressive disorder: machine learning analysis of two multicenter neuroimaging studies. Mol. Psychiatry 28 (7), 3013–3022.

Ghosal, S., 2023. Interpretable Machine Learning and Deep Learning Frameworks for Predictive Analytics and Biomarker Discovery from Multimodal Imaging Genetics Data (Ph.D. thesis). Johns Hopkins University.

Girish, D., Chan, Y.H., Gupta, S., Xia, J., Rajapakse, J., 2024. Robustness of explainable AI algorithms for disease biomarker discovery from functional connectivity datasets. In: IEEE-EMBS International Conference on Biomedical and Health Informatics.

Gorgolewski, K.J., Auer, T., Calhoun, V.D., Craddock, R.C., Das, S., Duff, E.P., Flandin, G., Ghosh, S.S., Glatard, T., Halchenko, Y.O., et al., 2016. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci. Data 3 (1), 1–9.

Gu, Y., Peng, S., Li, Y., Gao, L., Dong, Y., 2024. A novel population graph neural network based on functional connectivity for mental disorders detection. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, pp. 221–233.

Gu, Y., Peng, S., Li, Y., Gao, L., Dong, Y., 2025. FC-HGNN: A heterogeneous graph neural network based on brain functional connectivity for mental disorder identification. Inf. Fusion 113, 102619.

Gupta, S., Chan, Y.H., Rajapakse, J.C., 2021. Obtaining leaner deep neural networks for decoding brain functional connectome in a single shot. Neurocomputing 453, 326–336.

Gupta, S., Lim, M., Rajapakse, J.C., 2022. Decoding task specific and task general functional architectures of the brain. Hum. Brain Mapp. 43 (9), 2801–2816.

Habeck, C., Stern, Y., 2010. Multivariate data analysis for neuroimaging data: overview and application to Alzheimer's disease. Cell Biochem. Biophys. 58 (2), 53–67.

He, Y., Chan, Y.H., Rajapakse, J.C., 2023. Predicting gender from structural and functional connectomes via brain and population graph neural networks. BioRxiv 2023–11.

He, M., Wei, Z., Wen, J.-R., 2022. Convolutional neural networks on graphs with chebyshev approximation, revisited. Adv. Neural Inf. Process. Syst. 35, 7264–7276.

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

Hong, S.-J., Xu, T., Nikolaidis, A., Smallwood, J., Margulies, D.S., Bernhardt, B., Vogelstein, J., Milham, M.P., 2020. Toward a connectivity gradient-based framework for reproducible biomarker discovery. NeuroImage 223, 117322.

Hu, J., Cao, L., Li, T., Dong, S., Li, P., 2021. GAT-LI: a graph attention network based learning and interpreting method for functional brain network classification. BMC Bioinformatics 22, 1–20.

Hu, F., Chen, A.A., Horng, H., Bashyam, V., Davatzikos, C., Alexander-Bloch, A., Li, M., Shou, H., Satterthwaite, T.D., Yu, M., et al., 2023. Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. NeuroImage 120125.

Huang, Q., Yamada, M., Tian, Y., Singh, D., Chang, Y., 2022. Graphlime: Local interpretable model explanations for graph neural networks. IEEE Trans. Knowl. Data Eng. 35 (7), 6968–6972.

Jablensky, A., 2010. The diagnostic concept of schizophrenia: its history, evolution, and future prospects. Dialogues Clin. Neurosci. 12 (3), 271–287.

Jain, K.K., Jain, K.K., 2010. The Handbook of Biomarkers, vol. 6, Springer.

Jain, S., Wallace, B.C., 2019. Attention is not explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 3543–3556.

Ji, J., Ren, Y., Lei, M., 2022. FC–HAT: hypergraph attention network for functional brain network classification. Inform. Sci. 608, 1301–1316.

Jia, X.-Z., Zhao, N., Barton, B., Burciu, R., Carrière, N., Cerasa, A., Chen, B.-Y., Chen, J., Coombes, S., Defebvre, L., et al., 2018. Small effect size leads to reproducibility failure in resting-state fMRI studies. BioRxiv 285171.

Jiang, H., Cao, P., Xu, M., Yang, J., Zaiane, O., 2020. Hi-GCN: A hierarchical graph convolution network for graph embedding learning of brain network and brain disorders prediction. Comput. Biol. Med. 127, 104096.

Jun, E., Na, K.-S., Kang, W., Lee, J., Suk, H.-I., Ham, B.-J., 2020. Identifying resting-state effective connectivity abnormalities in drug-naïve major depressive disorder diagnosis via graph convolutional networks. Hum. Brain Mapp. 41 (17), 4997–5014.

Kahn, R.S., Keefe, R.S., 2013. Schizophrenia is a cognitive illness: time for a change in focus. JAMA Psychiatry 70 (10), 1107–1112.

Kahn, R.S., et al., 2015. Schizophrenia. Nat. Rev. Dis. Prim. 1 (1), 15067. http://dx.doi.org/10.1038/nrdp.2015.67.

Kakkad, J., Jannu, J., Sharma, K., Aggarwal, C., Medya, S., 2023. A survey on explainability of graph neural networks. arXiv preprint arXiv:2306.01958.

Kapishnikov, A., Venugopalan, S., Avci, B., Wedin, B., Terry, M., Bolukbasi, T., 2021. Guided integrated gradients: An adaptive path method for removing noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5050–5058.

Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G., 2017. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage 146, 1038–1049.

Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=SJU4ayYgl.

Klöppel, S., Abdulkadir, A., Jack Jr., C.R., Koutsouleris, N., Mourão-Miranda, J., Vemuri, P., 2012. Diagnostic neuroimaging across diseases. Neuroimage 61 (2), 457–463.

Kong, Y., Gao, S., Yue, Y., Hou, Z., Shu, H., Xie, C., Zhang, Z., Yuan, Y., 2021. Spatio-temporal graph convolutional network for diagnosis and treatment response prediction of major depressive disorder from functional connectivity. Hum. Brain Mapp. 42 (12), 3922–3933.

Kong, Y., Niu, S., Gao, H., Yue, Y., Shu, H., Xie, C., Zhang, Z., Yuan, Y., 2022. Multi-stage graph fusion networks for major depressive disorder diagnosis. IEEE Trans. Affect. Comput. 13 (4), 1917–1928.

Kong, Y., Wang, W., Liu, X., Gao, S., Hou, Z., Xie, C., Zhang, Z., Yuan, Y., 2023. Multi-connectivity representation learning network for major depressive disorder diagnosis. IEEE Trans. Med. Imaging 42 (10), 3012–3024.

Kong, Y., Zhang, X., Wang, W., Zhou, Y., Li, Y., Yuan, Y., 2025. Multi-scale spatial-temporal attention networks for functional connectome classification. IEEE Trans. Med. Imaging 44 (1), 475–488. http://dx.doi.org/10.1109/TMI.2024.3448214.

Laird, A.R., 2021. Large, open datasets for human connectomics research: Considerations for reproducible and responsible data use. NeuroImage 244, 118579.

de Lange, S.C., Scholtens, L.H., Initiative, A.D.N., van den Berg, L.H., Boks, M.P., Bozzali, M., Cahn, W., Dannlowski, U., Durston, S., Geuze, E., et al., 2019. Shared vulnerability for connectome alterations across psychiatric and neurological brain disorders. Nat. Hum. Behav. 3 (9), 988–998.

Lasalvia, A., 2018. Words matter: after more than a century 'schizophrenia'needs rebranding. BJPsych Adv. 24 (1), 33–36.

Lee, J., Ko, W., Kang, E., Suk, H.-I., Initiative, A.D.N., et al., 2021. A unified framework for personalized regions selection and functional relation modeling for early MCI identification. NeuroImage 236, 118048.

Lee, D.-J., Shin, D.-H., Son, Y.-H., Han, J.-W., Oh, J.-H., Kim, D.-H., Jeong, J.-H., Kam, T.-E., 2024. Spectral graph neural network-based multi-atlas brain network fusion for major depressive disorder diagnosis. IEEE J. Biomed. Heal. Inform. 28 (5), 2967–2978. http://dx.doi.org/10.1109/JBHI.2024.3366662.

Lei, D., Qin, K., Pinaya, W.H., Young, J., Van Amelsvoort, T., Marcelis, M., Donohoe, G., Mothersill, D.O., Corvin, A., Vieira, S., et al., 2022. Graph convolutional networks reveal network-level functional dysconnectivity in schizophrenia. Schizophr. Bull. 48 (4), 881–892.

Lei, B., Zhu, Y., Yu, S., Hu, H., Xu, Y., Yue, G., Wang, T., Zhao, C., Chen, S., Yang, P., et al., 2023. Multi-scale enhanced graph convolutional network for mild cognitive impairment detection. Pattern Recognit. 134, 109106.

Li, X., Dvornek, N.C., Zhuang, J., Ventola, P., Duncan, J., 2020a. Graph embedding using infomax for ASD classification and brain functional difference detection. In: Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging. SPIE, 1131702.

Li, L., Gong, J., Yan, W., Feng, X., Yang, Z., Wen, F., Luo, C., 2023. Classification of schizophrenia based on graph product depth neural network fusion of fMRI and dMRI multidimensional information. In: Second International Conference on Biomedical and Intelligent Systems. IC-BIS 2023, SPIE, pp. 479–484.

Li, L., Jiang, H., Wen, G., Cao, P., Xu, M., Liu, X., Yang, J., Zaiane, O., 2022b. TE-HI-GCN: An ensemble of transfer hierarchical graph convolutional networks for disorder diagnosis. Neuroinformatics 1–23.

Li, C., Liu, M., Xia, J., Mei, L., Yang, Q., Shi, F., Zhang, H., Shen, D., 2022a. Predicting brain amyloid-β PET grades with graph convolutional networks based on functional MRI and multi-level functional connectivity. J. Alzheimer's Dis. 86 (4), 1679–1693.

Li, J., Wang, F., Pan, J., Wen, Z., 2021a. Identification of autism spectrum disorder with functional graph discriminative network. Front. Neurosci. 15, 729937.

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

Li, X., Zhou, Y., Dvornek, N.C., Zhang, M., Zhuang, J., Ventola, P., Duncan, J.S., 2020b. Pooling regularized graph neural network for fmri biomarker analysis. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23. Springer, pp. 625–635.

Li, Y., Zhou, J., Verma, S., Chen, F., 2022c. A survey of explainable graph neural networks: Taxonomy and evaluation metrics. arXiv preprint arXiv:2207.12599.

Libero, L.E., Nordahl, C.W., Li, D.D., Ferrer, E., Rogers, S.J., Amaral, D.G., 2016. Persistence of megalencephaly in a subgroup of young boys with autism spectrum disorder. Autism Res. 9 (11), 1169–1182.

Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2020. Explainable ai: A review of machine learning interpretability methods. Entropy 23 (1), 18.

Liu, T.T., 2016. Noise contributions to the fMRI signal: An overview. NeuroImage 143, 141–151.

Liu, S., Gui, R., 2024. Fusing multi-scale fMRI features using a brain-inspired multi-channel graph neural network for major depressive disorder diagnosis. Biomed. Signal Process. Control. 90, 105837.

Liu, C., Zhan, Y., Wu, J., Li, C., Du, B., Hu, W., Liu, T., Tao, D., 2023. Graph pooling for graph neural networks: progress, challenges, and opportunities. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. pp. 6712–6722.

Lopardo, G., Precioso, F., Garreau, D., 2024. Attention meets post-hoc interpretability: A mathematical perspective. In: Forty-First International Conference on Machine Learning. URL: https://openreview.net/forum?id=wnkC5T11Z9.

Lord, C., Brugha, T.S., Charman, T., Cusack, J., Dumas, G., Frazier, T., Jones, E.J., Jones, R.M., Pickles, A., State, M.W., et al., 2020. Autism spectrum disorder. Nat. Rev. Dis. Prim. 6 (1), 1–23.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

Luo, X., Wu, J., Yang, J., Chen, H., Li, Z., Peng, H., Zhou, C., 2024. Knowledge distillation guided interpretable brain subgraph neural networks for brain disorder exploration. IEEE Trans. Neural Netw. Learn. Syst..

Ma, C., Li, W., Ke, S., Lv, J., Zhou, T., Zou, L., 2024. Identification of autism spectrum disorder using multiple functional connectivity-based graph convolutional network. Med. Biol. Eng. Comput. 1–12.

Ma, T., Zhang, A., 2019. Incorporating biological knowledge with factor graph neural network for interpretable deep learning. arXiv preprint arXiv:1906.00537.

Marcinkevičs, R., Vogt, J.E., 2023. Interpretable and explainable machine learning: A methods-centric overview with concrete examples. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. e1493.

Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., Donohue, M.R., Foran, W., Miller, R.L., Hendrickson, T.J., et al., 2022. Reproducible brain-wide association studies require thousands of individuals. Nature 603 (7902), 654–660.

Mei, L., Liu, M., Bian, L., Zhang, Y., Shi, F., Zhang, H., Shen, D., 2022. Modular graph encoding and hierarchical readout for functional brain network based eMCI diagnosis. In: MICCAI Workshop on Imaging Systems for GI Endoscopy. Springer, pp. 69–78.

Menon, G.S., Poornima, J., Sojan, A.T., Mridhula, P., Mohan, A., 2023. ASDEXPLAINER: An interpretable graph neural network framework for brain network based autism spectrum disorder analysis. In: 2023 14th International Conference on Computing Communication and Networking Technologies. ICCCNT, IEEE, pp. 1–7.

Meszlényi, R.J., Buza, K., Vidnyánszky, Z., 2017. Resting state fMRI functional connectivity-based classification using a convolutional neural network architecture. Front. Neuroinformatics 11, 61.

Miao, S., Luo, Y., Liu, M., Li, P., 2023. Interpretable geometric deep learning via learnable randomness injection. In: The Eleventh International Conference on Learning Representations. URL: https://openreview.net/forum?id=6u7mf9s2A9.

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.-R., 2019. Layer-wise relevance propagation: an overview. Explain. AI: Interpret. Explain. Vis. Deep Learn. 193–209.

Morris, C., Ritzert, M., Fey, M., Hamilton, W.L., Lenssen, J.E., Rattan, G., Grohe, M., 2019. Weisfeiler and leman go neural: Higher-order graph neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 4602–4609.

Munroe, L., da Silva, M., Heidari, F., Grigorescu, I., Dahan, S., Robinson, E.C., Deprez, M., So, P.-W., 2024. Applications of interpretable deep learning in neuroimaging: a comprehensive review. Imaging Neurosci. 2, 1–37.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C., 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Comput. Surv. 55 (13s), 1–42.

Niwarthana, A., Kasi, C., Chan, Y.H., Wang, C., Rajapakse, J.C., 2025. Decoding brain structure and gene expression interactions in Alzheimer's disease pathology. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1–5.

Noman, F., Yap, S.-Y., Phan, R.C.-W., Ombao, H., Ting, C.-M., 2022. Graph autoencoder-based embedded learning in dynamic brain networks for autism spectrum disorder identification. In: 2022 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 2891–2895.

Orru, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G., Mechelli, A., 2012. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci. Biobehav. Rev. 36 (4), 1140–1152.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Bmj 372.

Parisot, S., Ktena, S.I., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., Rueckert, D., 2018. Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. Med. Image Anal. 48, 117–130.

Park, H.W., Kim, S.Y., Lee, W.H., 2023. Graph convolutional network with morphometric similarity networks for schizophrenia classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 626–636.

Parkes, L., Satterthwaite, T.D., Bassett, D.S., 2020. Towards precise resting-state fMRI biomarkers in psychiatry: synthesizing developments in transdiagnostic research, dimensional models of psychopathology, and normative neurodevelopment. Curr. Opin. Neurobiol. 65, 120–128.

Peng, L., Wang, N., Xu, J., Zhu, X., Li, X., 2022. GATE: Graph CCA for temporal self-supervised learning for label-efficient fMRI analysis. IEEE Trans. Med. Imaging 42 (2), 391–402.

Plichta, M.M., Scheres, A., 2014. Ventral–striatal responsiveness during reward anticipation in ADHD and its relation to trait impulsivity in the healthy population: A meta-analytic review of the fMRI literature. Neurosci. Biobehav. Rev. 38, 125–134.

Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, M.R., Nichols, T.E., Poline, J.-B., Vul, E., Yarkoni, T., 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nature Rev. Neurosci. 18 (2), 115–126.

Pope, P.E., Kolouri, S., Rostami, M., Martin, C.E., Hoffmann, H., 2019. Explainability methods for graph convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10772–10781.

Qin, K., Lei, D., Pinaya, W.H., Pan, N., Li, W., Zhu, Z., Sweeney, J.A., Mechelli, A., Gong, Q., 2022. Using graph convolutional network to characterize individuals with major depressive disorder across multiple imaging sites. EBioMedicine 78.

Rahman, M.M., Calhoun, V.D., Plis, S.M., 2023. Looking deeper into interpretable deep learning in neuroimaging: a comprehensive survey. arXiv preprint arXiv: 2307.09615.

Rawls, E., Andrews, B., Lim, K., Kummerfeld, E., 2023. Causal discovery for fMRI data: Challenges, solutions, and a case study. ArXiv.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. " why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144.

Rolls, E.T., Huang, C.-C., Lin, C.-P., Feng, J., Joliot, M., 2020. Automated anatomical labelling atlas 3. Neuroimage 206, 116189.

Safai, A., Vakharia, N., Prasad, S., Saini, J., Shah, A., Lenka, A., Pal, P.K., Ingalhalikar, M., 2022. Multimodal brain connectomics-based prediction of Parkinson's disease using graph attention networks. Front. Neurosci. 15, 741489.

Said, A., Bayrak, R.G., Derr, T., Shabbir, M., Moyer, D., Chang, C., Koutsoukos, X.D., 2023. NeuroGraph: Benchmarks for graph machine learning in brain connectomics. In: Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track. URL: https://openreview.net/forum?id=MEa0cQeURw.

Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schütt, K.T., Müller, K.-R., Montavon, G., 2021. Higher-order explanations of graph neural networks via relevant walks. IEEE Trans. Pattern Anal. Mach. Intell. 44 (11), 7581–7596.

Sebenius, I., Campbell, A., Morgan, S.E., Bullmore, E.T., Liò, P., 2021. Multimodal graph coarsening for interpretable, MRI-based brain graph neural network. In: 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing. MLSP, IEEE, pp. 1–6.

Seitzman, B.A., Gratton, C., Marek, S., Raut, R.V., Dosenbach, N.U., Schlaggar, B.L., Petersen, S.E., Greene, D.J., 2020. A set of functionally-defined brain regions with improved representation of the subcortex and cerebellum. Neuroimage 206, 116290.

Serrano, S., Smith, N.A., 2019. Is attention interpretable? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2931–2951.

Shao, L., Fu, C., You, Y., Fu, D., 2021. Classification of ASD based on fMRI data with deep learning. Cogn. Neurodynamics 15 (6), 961–974.

Sheffield, J.M., Barch, D.M., 2016. Cognition and resting-state functional connectivity in schizophrenia. Neurosci. Biobehav. Rev. 61, 108–120.

Shen, X., Song, Z., Zhang, Z., 2024. GCAN: Generative counterfactual attention-guided network for explainable cognitive decline diagnostics based on fMRI functional connectivity. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 416–426.

Song, X., Zhou, F., Frangi, A.F., Cao, J., Xiao, X., Lei, Y., Wang, T., Lei, B., 2021. Graph convolution network with similarity awareness and adaptive calibration for disease-induced deterioration prediction. Med. Image Anal. 69, 101947.

Song, X., Zhou, F., Frangi, A.F., Cao, J., Xiao, X., Lei, Y., Wang, T., Lei, B., 2022. Multicenter and multichannel pooling GCN for early AD diagnosis based on dual-modality fused brain network. IEEE Trans. Med. Imaging 42 (2), 354–367.

Sturmfels, P., Lundberg, S., Lee, S.-I., 2020. Visualizing the impact of feature attribution baselines. Distill 5 (1), e22.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. In: International Conference on Machine Learning. PMLR, pp. 3319–3328.

Sunil, G., Gowtham, S., Bose, A., Harish, S., Srinivasa, G., 2024. Graph neural network and machine learning analysis of functional neuroimaging for understanding schizophrenia. BMC Neurosci. 25 (1), 2.

Tandon, R., Gaebel, W., Barch, D.M., Bustillo, J., Gur, R.E., Heckers, S., Malaspina, D., Owen, M.J., Schultz, S., Tsuang, M., et al., 2013. Definition and description of schizophrenia in the DSM-5. Schizophr. Res. 150 (1), 3–10.

Teng, J., Mi, C., Shi, J., Li, N., 2023. Brain disease research based on functional magnetic resonance imaging data and machine learning: a review. Front. Neurosci. 17, 1227491.

Therriault, J., Schindler, S.E., Salvadó, G., Pascoal, T.A., Benedet, A.L., Ashton, N.J., Karikari, T.K., Apostolova, L., Murray, M.E., Verberk, I., et al., 2024. Biomarker-based staging of Alzheimer disease: rationale and clinical applications. Nat. Rev. Neurol. 20 (4), 232–244.

Tian, X., Hu, N., Lu, L., Tan, L., Li, P., 2024. Gender differences in major depressive disorder at different ages: a REST-meta-MDD project-based study. BMC Psychiatry 24 (1), 1–9.

Tjoa, E., Guan, C., 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Trans. Neural Netw. Learn. Syst. 32 (11), 4793–4813.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y., 2018. Graph attention networks. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=rJXMpikCZ.

Verdi, S., Marquand, A.F., Schott, J.M., Cole, J.H., 2021. Beyond the average patient: how neuroimaging models can address heterogeneity in dementia. Brain 144 (10), 2946–2953.

Vizioli, L., Moeller, S., Dowdle, L., Akçakaya, M., De Martino, F., Yacoub, E., Uğurbil, K., 2021. Lowering the thermal noise barrier in functional brain mapping with magnetic resonance imaging. Nat. Commun. 12 (1), 5181.

Wang, W., Hu, X., Xiao, L., Wang, Y.-P., 2024c. Adaptive multiview community-preserved graph convolutional network for multiatlas-based functional connectivity analysis. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 2056–2060.

Wang, L., Li, K., Hu, X.P., 2021. Graph convolutional network for fMRI analysis based on connectivity neighborhood. Netw. Neurosci. 5 (1), 83–95.

Wang, J., Li, H., Qu, G., Cecil, K.M., Dillman, J.R., Parikh, N.A., He, L., 2023a. Dynamic weighted hypergraph convolutional network for brain functional connectome analysis. Med. Image Anal. 87, 102828.

Wang, Y., Liu, J., Xiang, Y., Wang, J., Chen, Q., Chong, J., 2022b. MAGE: Automatic diagnosis of autism spectrum disorders using multi-atlas graph convolutional networks and ensemble learning. Neurocomputing 469, 346–353.

Wang, Y., Long, H., Zhou, Q., Bo, T., Zheng, J., 2023c. Plsnet: Position-aware gcn-based autism spectrum disorder diagnosis via fc learning and rois sifting. Comput. Biol. Med. 163, 107184.

Wang, X., Shen, H.W., 2023. GNNInterpreter: A probabilistic generative model-level explanation for graph neural networks. In: The Eleventh International Conference on Learning Representations. URL: https://openreview.net/forum?id=rqq6Dh8t4d.

Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M., 2019. Dynamic graph cnn for learning on point clouds. ACM Trans. Graph. (Tog) 38 (5), 1–12.
Wang, S., Tang, H., Himeno, R., Solé-Casals, J., Caiafa, C.F., Han, S., Aoki, S., Sun, Z., 2024b. Optimizing graph neural network architectures for schizophrenia spectrum disorder prediction using evolutionary algorithms. Comput. Methods Programs Biomed. 257, 108419.
Wang, Q., Wang, W., Fang, Y., Yap, P.-T., Zhu, H., Li, H.-J., Qiao, L., Liu, M., 2024a. Leveraging brain modularity prior for interpretable representation learning of fMRI. IEEE Trans. Biomed. Eng. 71 (8), 2391–2401. http://dx.doi.org/10.1109/TBME.2024.3370415.
Wang, Q., Wu, M., Fang, Y., Wang, W., Qiao, L., Liu, M., 2023b. Modularity-constrained dynamic representation learning for interpretable brain disorder analysis with functional MRI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 46–56.
Wang, W., Xiao, L., 2023. Consistency guided multiview hypergraph embedding learning with multiatlas-based functional connectivity networks using resting-state fMRI. In: Chinese Conference on Pattern Recognition and Computer Vision. PRCV, Springer, pp. 170–181.
Wang, W., Xiao, L., Qu, G., Calhoun, V.D., Wang, Y.-P., Sun, X., 2024d. Multiview hyperedge-aware hypergraph embedding learning for multisite, multiatlas fMRI based functional connectivity network analysis. Med. Image Anal. 94, 103144.
Wang, L., Yuan, W., Zeng, L., Xu, J., Mo, Y., Zhao, X., Peng, L., 2022a. Dementia analysis from functional connectivity network with graph neural networks. Inf. Process. Manage. 59 (3), 102901.
Wang, X., Zhang, X., Chen, Y., Yang, X., 2024e. IFC-GNN: Combining interactions of functional connectivity with multimodal graph neural networks for ASD brain disorder analysis. Alex. Eng. J. 98, 44–55.
Weaverdyck, M.E., Lieberman, M.D., Parkinson, C., 2020. Tools of the Trade Multivoxel pattern analysis in fMRI: a practical introduction for social and affective neuroscientists. Soc. Cogn. Affect. Neurosci. 15 (4), 487–509.
Wei, L., Liu, B., He, J., Zhang, M., Huang, Y., 2023. Autistic spectrum disorders diagnose with graph neural networks. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8819–8827.
Wiegreffe, S., Pinter, Y., 2019. Attention is not not explanation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP, pp. 11–20.
Winter, N.R., Blanke, J., Leenings, R., Ernsting, J., Fisch, L., Sarink, K., Barkhau, C., Emden, D., Thiel, K., Flinkenflügel, K., et al., 2024. A systematic evaluation of machine learning–based biomarkers for major depressive disorder. JAMA Psychiatry.
Wolfers, T., Buitelaar, J.K., Beckmann, C.F., Franke, B., Marquand, A.F., 2015. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci. Biobehav. Rev. 57, 328–349.
Xiao, T., Zeng, L., Shi, X., Zhu, X., Wu, G., 2022. Dual-graph learning convolutional networks for interpretable Alzheimer's disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 406–415.
Xing, X., Li, Q., Yuan, M., Wei, H., Xue, Z., Wang, T., Shi, F., Shen, D., 2021. DS-GCNs: Connectome classification using dynamic spectral graph convolution networks with assistant task training. Cerebral Cortex 31 (2), 1259–1269.
Xu, J., Bian, Q., Li, X., Zhang, A., Ke, Y., Qiao, M., Zhang, W., Sim, W.K.J., Gulyás, B., 2024. Contrastive graph pooling for explainable classification of brain networks. IEEE Trans. Med. Imaging.
Xu, K., Hu, W., Leskovec, J., Jegelka, S., 2019. How powerful are graph neural networks? In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=ryGs6iA5Km.
Yan, W., Plis, S., Calhoun, V.D., Liu, S., Jiang, R., Jiang, T.-Z., Sui, J., 2017. Discriminating schizophrenia from normal controls using resting state functional network connectivity: A deep neural network and layer-wise relevance propagation method. In: 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing. MLSP, IEEE, pp. 1–6.
Yang, S., Jin, D., Liu, J., He, Y., 2022. Identification of young high-functioning autism individuals based on functional connectome using graph isomorphism network: A pilot study. Brain Sci. 12 (7), 883.
Yang, C., Wang, P., Tan, J., Liu, Q., Li, X., 2021. Autism spectrum disorder diagnosis using graph attention network based on spatial-constrained sparse functional brain networks. Comput. Biol. Med. 139, 104963.
Yang, Y., Ye, C., Su, G., Zhang, Z., Chang, Z., Chen, H., Chan, P., Yu, Y., Ma, T., 2024. BrainMass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. IEEE Trans. Med. Imaging.
Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D., 2011. Large-scale automated synthesis of human functional neuroimaging data. Nature Methods 8 (8), 665–670.
Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., et al., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol..
Yin, W., Li, L., Wu, F.-X., 2021. A graph attention neural network for diagnosing ASD with fMRI data. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine. BIBM, IEEE, pp. 1131–1136.
Ying, Z., Bourgeois, D., You, J., Zitnik, M., Leskovec, J., 2019. Gnnexplainer: Generating explanations for graph neural networks. Adv. Neural Inf. Process. Syst. 32.
Yu, A., Chen, L., Qiao, C., 2022. Graph convolutional network with attention mechanism for discovering the brain's abnormal activity of attention deficit hyperactivity disorder. In: 2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. CISP-BMEI, IEEE, pp. 1–5.
Yu, J., Xu, T., Rong, Y., Bian, Y., Huang, J., He, R., 2021. Recognizing predictive substructures with subgraph information bottleneck. IEEE Trans. Pattern Anal. Mach. Intell. 46 (3), 1650–1663.
Yu, S., Yue, G., Elazab, A., Song, X., Wang, T., Lei, B., 2019. Multi-scale graph convolutional network for mild cognitive impairment detection. In: Graph Learning in Medical Imaging: First International Workshop, GLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 1. Springer, pp. 79–87.
Yuan, H., Tang, J., Hu, X., Ji, S., 2020. Xgnn: Towards model-level explanations of graph neural networks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 430–438.
Yuan, H., Yu, H., Gui, S., Ji, S., 2022. Explainability in graph neural networks: A taxonomic survey. IEEE Trans. Pattern Anal. Mach. Intell. 45 (5), 5782–5799.
Zhang, J., Chao, H., Dasegowda, G., Wang, G., Kalra, M.K., Yan, P., 2023a. Revisiting the trustworthiness of saliency methods in radiology AI. Radiol.: Artif. Intell. 6 (1), e220221.
Zhang, S., Chen, X., Shen, X., Ren, B., Yu, Z., Yang, H., Jiang, X., Shen, D., Zhou, Y., Zhang, X.-Y., 2023b. A-GCL: Adversarial graph contrastive learning for fMRI analysis to diagnose neurodevelopmental disorders. Med. Image Anal. 90, 102932.
Zhang, Y., He, X., Chan, Y.H., Teng, Q., Rajapakse, J.C., 2023d. Multi-modal graph neural network for early diagnosis of Alzheimer's disease from sMRI and PET scans. Comput. Biol. Med. 164, 107328.
Zhang, Z., Li, G., Niu, J., Du, S., Gao, T., Liu, W., Jiang, Z., Tang, X., Xu, Y., 2022b. Identifying biomarkers of subjective cognitive decline using graph convolutional neural network for fMRI analysis. In: 2022 IEEE International Conference on Mechatronics and Automation. ICMA, IEEE, pp. 1306–1311.
Zhang, Y., Liu, X., Tang, P., Zhang, Z., 2023e. SLG-NET: Subgraph neural network with local-global braingraph feature extraction modules and a novel subgraph generation algorithm for automated identification of major depressive disorder. In: International Conference on Neural Information Processing. Springer, pp. 31–42.
Zhang, H., Song, R., Wang, L., Zhang, L., Wang, D., Wang, C., Zhang, W., 2022a. Classification of brain disorders in rs-fMRI via local-to-global graph neural networks. IEEE Trans. Med. Imaging 42 (2), 444–455.
Zhang, L., Wang, M., Liu, M., Zhang, D., 2020. A survey on deep learning for neuroimaging-based brain disorder analysis. Front. Neurosci. 14, 779.
Zhang, L., Wang, J.-R., Ma, Y., 2021. Graph convolutional networks via low-rank subspace for multi-site rs-fMRI ASD diagnosis. In: 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. CISP-BMEI, IEEE, pp. 1–6.
Zhang, S., Yang, J., Zhang, Y., Zhong, J., Hu, W., Li, C., Jiang, J., 2023c. The combination of a graph neural network technique and brain imaging to diagnose neurological disorders: A review and outlook. Brain Sci. 13 (10), 1462.
Zhao, K., Duka, B., Xie, H., Oathes, D.J., Calhoun, V., Zhang, Y., 2022b. A dynamic graph convolutional neural network framework reveals new insights into connectome dysfunctions in ADHD. NeuroImage 246, 118774.
Zhao, F., Li, N., Pan, H., Chen, X., Li, Y., Zhang, H., Mao, N., Cheng, D., 2022a. Multi-view feature enhancement based on self-attention mechanism graph convolutional network for autism spectrum disorder diagnosis. Front. Hum. Neurosci. 16, 918969.
Zhao, T., Zhang, G., 2022. Detecting major depressive disorder by graph neural network exploiting resting-state functional MRI. In: International Conference on Neural Information Processing. Springer, pp. 255–266.
Zhao, T., Zhang, G., 2024. Enhancing major depressive disorder diagnosis with dynamic-static fusion graph neural networks. IEEE J. Biomed. Heal. Inform. 28 (8), 4701–4710. http://dx.doi.org/10.1109/JBHI.2024.3395611.
Zheng, K., Ma, B., Chen, B., 2023. DynBrainGNN: Towards spatio-temporal interpretable graph neural network based on dynamic brain connectome for psychiatric diagnosis. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 164–173.
Zheng, K., Yu, S., Chen, B., 2024a. Ci-gnn: A granger causality-inspired graph neural network for interpretable brain network-based psychiatric diagnosis. Neural Netw. 172, 106147.
Zheng, K., Yu, S., Chen, L., Dang, L., Chen, B., 2024b. BPI-GNN: Interpretable brain network-based psychiatric diagnosis and subtyping. NeuroImage 292, 120594.
Zheng, K., Yu, S., Li, B., Jenssen, R., Chen, B., 2024c. Brainib: Interpretable brain network-based psychiatric diagnosis with graph information bottleneck. IEEE Trans. Neural Netw. Learn. Syst..
Zhu, Y., Song, X., Qiu, Y., Zhao, C., Lei, B., 2021a. Structure and feature based graph U-net for early Alzheimer's disease prediction. In: Multimodal Learning for Clinical Decision Support: 11th International Workshop, ML-CDS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 11. Springer, pp. 93–104.
Zhu, C., Tan, Y., Yang, S., Miao, J., Zhu, J., Huang, H., Yao, D., Luo, C., 2024. Temporal dynamic synchronous functional brain network for schizophrenia classification and lateralization analysis. IEEE Trans. Med. Imaging 43 (12), 4307–4318. http://dx.doi.org/10.1109/TMI.2024.3419041.

Zhu, Z., Wang, B., Li, S., 2021b. A triple-pooling graph neural network for multi-scale topological learning of brain functional connectivity: Application to ASD diagnosis. In: Artificial Intelligence: First CAAI International Conference, CICAI 2021, Hangzhou, China, June 5–6, 2021, Proceedings, Part II 1. Springer, pp. 359–370.

Zhu, H., Wang, J., Zhao, Y.-P., Lu, M., Shi, J., 2022. Contrastive multi-view composite graph convolutional networks based on contribution learning for autism spectrum disorder classification. IEEE Trans. Biomed. Eng. 70 (6), 1943–1954.

Zuo, Q., Lu, L., Wang, L., Zuo, J., Ouyang, T., 2022. Constructing brain functional network by adversarial temporal-spatial aligned transformer for early AD analysis. Front. Neurosci. 16, 1087176.