frontiers | Frontiers in Big Data

TYPE Methods
PUBLISHED 01 July 2024
DOI 10.3389/fdata.2024.1410424



OPEN ACCESS

EDITED BY
Rex Ying,
Yale University, United States

REVIEWED BY
Ninghao Liu,
University of Georgia, United States
Xiangyu Zhao,
City University of Hong Kong, Hong Kong SAR, China

*CORRESPONDENCE
Liang Zhao
liang.zhao@emory.edu

RECEIVED 01 April 2024
ACCEPTED 04 June 2024
PUBLISHED 01 July 2024

CITATION
Etemadyrad N, Gao Y, Manoj Pudukotai Dinakarrao S and Zhao L (2024) Global explanation supervision for Graph Neural Networks. *Front. Big Data* 7:1410424. doi: 10.3389/fdata.2024.1410424



# Global explanation supervision for Graph Neural Networks

Negar Etemadyrad[1], Yuyang Gao[2], Sai Manoj Pudukotai Dinakarrao[1] and Liang Zhao[2]*

[1]Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA, United States, [2]Department of Computer Science, Emory University, Atlanta, GA, United States

With the increasing popularity of Graph Neural Networks (GNNs) for predictive tasks on graph structured data, research on their explainability is becoming more critical and achieving significant progress. Although many methods are proposed to explain the predictions of GNNs, their focus is mainly on "how to generate explanations." However, other important research questions like "whether the GNN explanations are inaccurate," "what if the explanations are inaccurate," and "how to adjust the model to generate more accurate explanations" have gained little attention. Our previous GNN Explanation Supervision (GNES) framework demonstrated effectiveness on improving the reasonability of the local explanation while still keep or even improve the backbone GNNs model performance. In many applications instead of per sample explanations, we need to find global explanations which are reasonable and faithful to the domain data. Simply learning to explain GNNs locally is not an optimal solution to a global understanding of the model. To improve the explainability power of the GNES framework, we propose the Global GNN Explanation Supervision (GGNES) technique which uses a basic trained GNN and a global extension of the loss function used in the GNES framework. This GNN creates local explanations which are fed to a Global Logic-based GNN Explainer, an existing technique that can learn the global Explanation in terms of a logic formula. These two frameworks are then trained iteratively to generate reasonable global explanations. Extensive experiments demonstrate the effectiveness of the proposed model on improving the global explanations while keeping the performance similar or even increase the model prediction power.

KEYWORDS

**graph, Graph Neural Networks, global explainability, human-in-the-loop, graphical concepts**

## 1 Introduction

As Deep Neural Networks (DNNs) are widely deployed in sensitive application areas, recent years have seen an explosion of research in understanding how DNNs work under the hood (e.g., explainable AI, or XAI; Adadi and Berrada, 2018; Arrieta et al., 2020) and more importantly, how to improve DNNs using human knowledge (Hong et al., 2020). In particular, Graph Neural Networks (GNNs) have been increasingly grabbed attention in several research fields, including computer vision (Fukui et al., 2019; Pope et al., 2019), natural language processing (Annervaz et al., 2018), medical domain (De Haan et al., 2009), and beyond. Such trend is attributed to the practical implication of graph data—many real-world data, such as social networks (Fan et al., 2019), chemical molecules (Scarselli et al., 2008), and financial data (Matsunaga et al., 2019), are represented as graphs.

However, similar to other DNNs' architectures, GNNs also offer only limited transparency, imposing significant challenges in observing when GNNs make successful/unsuccessful predictions (Hong et al., 2020; Wu et al., 2021). Some real world examples for adjusting model explanation to improve Graph Neural Networks (GNNs) can be seen in Figure 1. This issue motivates a surge of recent research on GNN explanation techniques, including gradients-based methods, where the gradients are used to indicate the importance of different input features (Baldassarre and Azizpour, 2019; Pope et al., 2019); perturbation-based methods, where an additional optimization step is typically used to find the important input that influences the model output the most with input perturbations (Ying et al., 2019; Luo et al., 2020; Schlichtkrull et al., 2020); response-based methods, where the output response signal is backpropagated as an importance score layer by layer until the input space (Baldassarre and Azizpour, 2019; Pope et al., 2019; Schnake et al., 2020); surrogate-based methods, where the explanation obtained from an interpretable surrogate model that is trained to fit the original prediction is used to explain the original model (Huang et al., 2020; Vu and Thai, 2020; Zhang et al., 2020); and global explanation methods, where graph patterns are generated to maximize the predicted probability for a certain class and use such graph patterns to explain the class (Yuan et al., 2020a). Unlike local explanation models which explain the model prediction per input sample, global explanation techniques aim at providing the general insights and high-level understanding of the predictions of a deep graph model. Specifically, they investigate what input graph patterns can lead to a certain GNN behavior or maximize a certain prediction. This is essential in many real-world critical applications and can substantially increase human trust in GNNs' prediction ability. As an example, consider classifying graph molecules as either having a mutagenic effect or not (Azzolin et al., 2022). The mutagenicity of a molecule is correlated with the presence of electron-attracting elements conjugated with nitro groups (e.g., $NO_2$). Accordingly, designing an explanation model that can provide a global understanding of the GNN classification is an urgent need. This could be achieved by designing an explainer that manages to recover all the existing well-known $NO_2$ motifs as an indicator of mutagenicity. Additional examples include gender (male vs. female) or age (young vs. old) classification of human subjects based on structural or functional connectivity matrices, obtained through magnetic resonance imaging of the corresponding subjects. In this case, rather than a per sample explanation, we need a per class explanation in form of high-level, generic insight on differences in the input connectivity matrices of these subjects.

Despite the recent fast progress on GNN explanation techniques, the existing research body focuses on "how to generate GNN explanations" instead of "whether the GNN explanations are inaccurate," "what if the explanations are inaccurate," and "how to adjust the model to generate more accurate explanations." Answering the above questions is highly beneficial to the model developers and the users of GNN explanation techniques but is also extremely difficult due to several challenges: **1) Lack of an automatic learning framework for identifying and adjusting unreasonable explanations on GNNs.** Although there are plenty of existing works on GNN explanations, they are not able to ensure the correctness of explanations, not able to identify the incorrect explanations, nor able to adjust the unreasonable explanations. The technique that can enable this has not been well-explored yet and is technically challenging due to the additional involvement of another backpropagation originated from explanation error. **2) Difficulty in aligning the node and edge explanations.** Existing GNN explanation works usually focus on either node and edge explanation, while the interplay and consistency between the explanations of nodes and edges are extremely challenging to maintain and jointly adjusted. **3) Difficulty in jointly improving model performance and explainability with limited explanation supervision.** Due to the high cost for human annotation, it can be impractical to assume the full accessibility to the human explanation label during model training. Thus, designing an effective framework that can best leverage a partially labeled dataset is on-demand yet challenging. **4) Lack of a learning framework that can employ global explanation of a GNN model to improve its performance and global explainability through global explanation supervision.** In many applications, we have access to the ground-truth explanations annotated by domain experts that can demonstrate the behavior of the data as a whole, and hence, we are motivated to employ that as the supervision signal to improve performance and global-level explainability. Designing a learning framework that utilizes this type of information is an interesting line of research which has yet remained unexplored.

To address the above challenges, beyond merely finding a solution to produce global GNN explanations, this study focuses on a global GNN explanation supervision framework for correcting the unreasonable explanations and learning how to explain GNNs from a global aspect correctly. Although the previously proposed Graph Neural Network Explanation Supervision (GNES) framework (Gao et al., 2021) has proved effective on improving the reasonability of the model explanation per local samples, while still keep or even improve the backbone GNN model performance, it still lacks the ability to guide the global model explanation generation. In many real-world decision-critical application, the ability to explain the reason for each class prediction through a single robust overview of the model is a critical requirement. To address the inefficacy of the existing GNES model in improving the global explainability through global explanation supervision, in this study, we extend the GNES model by proposing the Global GNN Explanation Supervision (GGNES), whose effectiveness is similar to the GNES model but can improve the GNN model global explanation generation (and potentially its prediction) through guiding the global explanations generated while training the model. The major contributions of this study are summarized as follows: **(1) Develop a generic framework for training GNNs while improving the reasonability and faithfulness of the global explanations generated for the model.** We propose the GGNES model built upon concept-based explainability and our previously proposed GNES model. GGNES enables learning reasonable and faithful global explanation, in terms of logic formulas, while training a GNN model. These formulas are constructed from a combination of learned graphical concepts which are derived from local explanations. **(2) Develop the formulation that can take the**

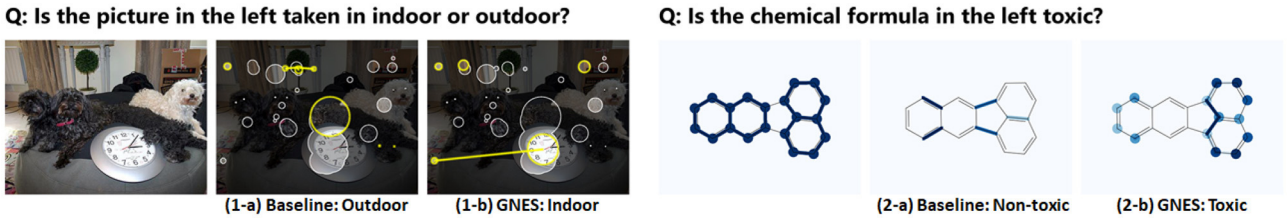


FIGURE 1
Cases for adjusting model explanation to improve Graph Neural Networks (GNNs). Scene graph **(left three)**: from the left, an input image, explanation before adjustment (1-a, inaccurate), and explanation after the adjustment (1-b, accurate). Note that the model explanation has been shifted from puppy eyes and back, rods, and an artificial tree to curtains, a clock, and a rug. Molecular formula **(right three)**: from the left, an input formula, explanation before the adjustment (2-a, inaccurate), and explanation after the adjustment (2-b, accurate). Reactivity for this molecule is mostly affected by benzene ring sub-components in the overall molecular structure. 2-b highlights the main benzene rings of the molecule more effectively than 2-a.

**model-generated global node (or edge) level explanation of a GNN, and use that as additional supervision to train the GNN model.** The explanations generated by the GNN model remain differentiable to the backbone model's parameters. This makes the global explanation supervision feasible as the model parameters can be affected and tuned during training. **(3) Conduct comprehensive experiments to evaluate the effectiveness of the proposed model.** Extensive experiments on three real-world datasets demonstrate that the proposed model improved the backbone GNN model both in terms of prediction power and global explainability across different application domains. In addition, qualitative analyzes, including case studies, are provided to demonstrate the effectiveness of the proposed framework.

## 2 Related work

In this section, we first introduce our previously proposed GNES framework. Then, we note that our work draws inspiration from the research fields of graph neural network explanations that provide the model generated explanations, and explanation supervision on DNNs which enables the design of pipelines for the human-in-the-loop adjustment on the DNNs based on their explanations.

### 2.1 Our previously proposed GNES framework

In our previous study (Gao et al., 2021), we proposed a framework that learns how to jointly optimize both model prediction and model explanation by enforcing both whole graph regularization and weak supervision on model explanations. For the graph regularization, we proposed a unified explanation formulation for both node-level and edge-level explanations by enforcing the consistency between them. The node- and edge-level explanation techniques we proposed are also generic and rigorously demonstrated to cover several existing major explainers as special cases. However, in some applications, the ground truth explanations demonstrate the behavior of the data as a whole instead of each individual sample. Accordingly, we need a learning framework that utilizes this type of information through global explanation supervision and hence improves both model prediction and global model explanation.

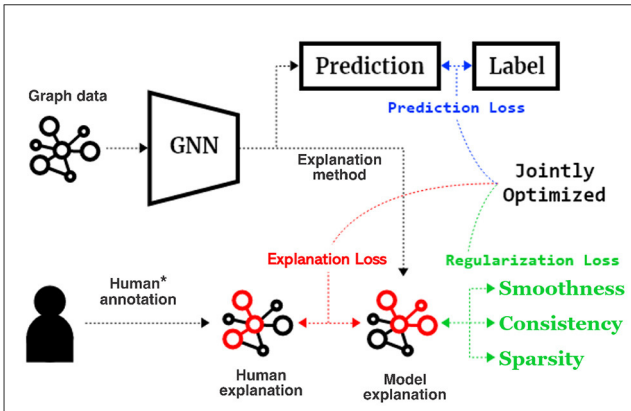


FIGURE 2
Proposed GNN Explanation Supervision (GNES) framework that jointly optimized the GNN models based on (1) a prediction loss, (2) an explanation loss on the human annotation and model explanation, and (3) a graph regularization loss to inject high-level principles of the graph-structured explanation. Notice that we only assume limited accessibility to the human annotation for only a small set of samples (10% in our experiments).

### 2.2 Graph Neural Networks explanations

Most of the existing GNN explanation methods are instance-level methods, where the methods explain the models by identifying important input features for its prediction (Yuan et al., 2020b). The first category is gradients-based methods, where the gradients are used to indicate the importance of different input features. Existing methods are SA (Baldassarre and Azizpour, 2019), Guided BP (Baldassarre and Azizpour, 2019), CAM (Pope et al., 2019), and GradCAM (Pope et al., 2019). In Etemadyrad et al. (2022), the authors propose a novel post hoc explanation technique to find the subgraphs in input that majorly influence one or more subgraphs in the output domain by using gradient information and solving a classical community detection objective (De Domenico et al., 2015). The second category is perturbation-based methods, where an additional optimization step is typically used to find the important input that influences the model output the most with input perturbations. Existing methods are GNNExplainer (Ying et al., 2019), PGExplainer (Luo et al., 2020), and GraphMask (Schlichtkrull et al., 2020). The third category is the response-based method, where the output response signal is backpropagated

as an importance score layer by layer until the input space. Existing methods in this category include LRP (Baldassarre and Azizpour, 2019), Excitation BP (Pope et al., 2019), and GNN-LRP (Schnake et al., 2020). The last category is surrogate-based methods, where the explanation obtained from an interpretable surrogate model that is trained to fit the original prediction is used to explain the original model. The surrogate methods include GraphLime (Huang et al., 2020), RelEx (Zhang et al., 2020), and PGM-Explainer (Vu and Thai, 2020). In addition to instance-level explanation methods, very recently, the global explanation of the GNN model has also been explored by XGNN (Yuan et al., 2020a). Please see Yuan et al. (2020b) for a survey of explainability in Graph Neural Networks. Even though there are plenty of existing explanation methods for GNNs, most of the methods above can not be applied to explanation supervision mechanism as the goal is to apply supervision on the generated explanation such that the backbone GNN model itself can be fine-tuned accordingly to generate better explanations as well as keep or even improve the model performance. To enable this fine-tuning process over the explanation, the explanation itself needs to be differentiable to the backbone GNN model's parameters. In other words, only the explanation that is directly calculated from the computational pipeline (such as gradients-based and response-based methods) can be used to apply this additional explanation supervision to fine-tune the backbone GNN models explanation. The perturbation-based and surrogate-based methods all require additional optimization steps to obtain the explanation and thus are unable to be end-to-end trained with the explanation supervision on the backbone GNNs.

## 2.3 Explanation supervision on DNNs

The potential of using *explanation*–methods devised for understanding which sub-parts in an instance are important for making a prediction–in improving DNNs has been studied in many domains across different applications. In fact, explanation supervision has been widely studied on image data by the computer vision community (Das et al., 2017; Linsley et al., 2018; Qiao et al., 2018; Mitsuhara et al., 2019; Zhang et al., 2019; Chen et al., 2020; Patro et al., 2020). Linsley et al. (2018) have demonstrated that the benefit of using stronger supervisory signals by teaching networks where to attend, which looks similar to the proposed approach. Moreover, Mitsuhara et al. (2019) have proposed a *post-hoc* fine-tuning strategy where an end-user is asked to manually edit the model's explanation to interactively adjust its output. Such edited explanations are then used as ground-truth explanations (from humans) to further fine-tune the model. In addition, several works in the Visual Question Answering (VQA) domain have proposed to use explanation supervision to obtain improved explanation on both the text data and the image data (Das et al., 2017; Qiao et al., 2018; Zhang et al., 2019; Patro et al., 2020). In addition to image data, the explanation supervision has also been studied on other data types, such as texts (Ross et al., 2017; Jacovi and Goldberg, 2020), attributed data (Visotsky et al., 2019), and more.

Gao et al. (2024) provide a systematic survey on Explanation-Guided Learning (EGL), a line of research that focuses on leveraging additional supervision signals or prior knowledge obtained from human explanations into machine learning models' reasoning process. According to Gao et al. (2024), EGL methods provide either global (Weinberger et al., 2020; Erion et al., 2021) or local guidance (Gao et al., 2022a,b; Shi et al., 2023)) by injecting prior knowledge or adding supervision signals to improve the model's global (or local) explanation. In Erion et al. (2021), the authors introduce attribution priors to optimize for higher-level properties of explanations, such as smoothness and sparsity. Lee et al. (2022) illustrate how to upgrade a deep model to its self explainable version that can predict and explain with logic rules learned with widely-used deep learning modules. Gupta et al. (2024) introduce Concept Distillation to create richer concepts using a pre-trained teacher model. They demonstrate how concept-sensitive training can improve model interpretability, reduce biases, and induce prior knowledge. Sha et al. (2023) propose a rational extraction technique built based on an adversarial approach that calibrates the information between a guider, a typical neural model that does the prediction, and a selector-predictor model that additionally produces a rationale for the guider' prediction. Shi et al. (2023) develop the ENGAGE framework as a local guidance EGL, built upon Explanation Guidance Data Augmentation, which leverages explanation to inform graph augmentation, and uses contrastive learning for training representations to preserve the key parts in graphs while removing uninformative artifacts.

However, to our best knowledge, explanation supervision on graph-structured data with graph neural networks through learning logic-based concepts has not been explored before, and we are the first to propose a framework to handle this open research problem.

# 3 Model

In this section, we introduce our proposed Global Explanation Supervision framework for GNNs. First, we briefly summarize the explanation regularizations (i.e., explanation consistency and sparsity) proposed by Gao et al. (2021) and how these components enhance the quality of model explanations in a global level. Then, we will introduce the proposed Global node-level, in addition to the Global edge-level explanation supervision definition and formulation.

**Formal definition of the problem:** Let $\mathcal{G} = (X, A)$ denotes an attributed graph with $N$ nodes be defined with its node attributes $X \in \mathbb{R}^{N \times d_{in}}$ and its adjacency matrix $A \in \mathbb{R}^{N \times N}$ (weighted or binary), where $d_{in}$ denotes the dimension of input feature. Let $y$ be the class label for graph $\mathcal{G}$. The general goal for a GNN model is to learn the mapping function $f$ for each graph $\mathcal{G}$ to its corresponding label $y$,

$$\mathcal{F} : \mathcal{G} \longrightarrow y$$

Following Kipf and Welling (2016) and similar to Gao et al. (2021), we employ the basic definition of Graph Convolutional Networks (GCN; Kipf and Welling, 2016), for an attributed graph $\mathcal{G} = (X, A)$ with $y$ as the class label for graph $\mathcal{G}$, where a graph convolutional layer can be defined as Equation (1).

$$F^{(l)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}F^{(l-1)}W^{(l)}) \tag{1}$$

where $F^{(l)}$ denotes the activations at layer $l$, and $F^{(0)} = X$; $\tilde{A} = A + I_N$ is the adjacency matrix with added self connections where $I_N \in \mathbb{R}^{N \times N}$ is the identity matrix; $\tilde{D}$ is the degree matrix of $\tilde{A}$, where

TABLE 1 Raw formulas as extracted by the Entropy Layer.

| Dataset | Task | Raw formulas |
|---|---|---|
| HCP structural | Gender prediction | Female $\Longleftrightarrow P_0 \vee P_1 \vee P_2$ |
| | | Male $\Longleftrightarrow P_3$ |
| HCP functional | Age prediction | Old $\Longleftrightarrow P_0 \vee (P_1 \wedge P_2)$ |
| | | Young $\Longleftrightarrow P_2 \vee P_3$ |
| ABIDE | ASD classification | Typical $\Longleftrightarrow P_0 \wedge P_1$ |
| | | Control $\Longleftrightarrow P_2$ |

$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$; The trainable weight matrix for layer $l$ is denoted as $W^{(l)} \in \mathbb{R}^{d^{(l)} \times d^{(l+1)}}$; $\sigma(\cdot)$ is the element-wise non-linear activation function. Additionally, a similar design as in Pope et al. (2019) is employed to this backbone GNN model in which using several layers of graph convolutional layers followed by a global average pooling (GAP) layer over the graph nodes can address any concerns when working with variable input graph size.

## 3.1 GGNES framework

The goal here would be to design a framework that can generate global explanations which are closer to the human annotations through global explanation supervision. The prediction performance is expected to stay the same or possibly also improve. The global explanation supervision is possible via defining the learning objective of the proposed framework as a joint optimization. As shown in Equation (2) and following the framework in Figure (2), the objective function is a combination of model prediction loss (e.g., the cross-entropy loss), the global explanation loss (which is a function of the absolute or squared difference between class level human and model explanations), and global model explanation regularizations (graph regularizations that follow high-level graph-structured rules to the explanation). These three terms are computed per class and combined thereafter to form the global explanation supervision framework. Concretely, we employ the objective function as

$$\min \mathcal{L}_{\text{Pred,c}} + \underbrace{\mathcal{L}_{\text{Att,c}}(\langle M_c, M_c'\rangle, \langle E_c, E_c'\rangle)}_{\text{global explanation loss}} + \underbrace{\Omega_c(M_c, E_c)}_{\text{regularization}} \quad (2)$$

where $M_c \in \mathbb{R}^{N\times 1}$ and $E_c \in \mathbb{R}^{N\times N}$ denote the model-generated node-level and edge-level explanations of class $c$ using a given explanation method. And $M_c'$, $E_c'$ are the corresponding ground-truth explanations of class $c$, marked by the human annotators. The human annotations are provided globally for all samples and are unique per class, but equal for the samples of each class. These are used as additional guidance to make the explanation supervision possible. $\mathcal{L}_{\text{Pred,c}}$ is the typical prediction loss (such as the cross-entropy loss) on the training set. The proposed explanation loss $\mathcal{L}_{\text{Att,c}}$ measures the discrepancies between model and human explanations globally both on node level and edge level, as Equation (3)

$$\mathcal{L}_{\text{Att}}(\langle M_c, M_c'\rangle, \langle E_c, E_c'\rangle) = \underbrace{\alpha_n \text{dist}(M_c, M_c')}_{\text{global node loss}} + \underbrace{\alpha_e \text{dist}(E_c, E_c')}_{\text{global edge loss}} \quad (3)$$

where $\alpha_n$ and $\alpha_e$ are the scale factors for balancing global node-level and global edge-level loss; the function dist$(x, y)$ measures the mean element-wise distance between the inputs $x$ and $y$, a common choose can be absolute difference or squared difference.

However, in practice, in many applications, it is not feasible to obtain the human explanations for the whole dataset. As a remedy, we only apply the global explanation loss to the classes that have the ground-truth labels for the human explanations and apply the high-level graph rules to regulate the model explanation for each class even if the human annotation is unavailable (Gao et al., 2021). Specifically, we employ the global explanation consistency, in addition to the global sparsity regularization. The former can regulate the global node and edge explanation simultaneously so that the model is more likely to generate a globally consistent and smooth explanation over nodes and edges. The global sparsity regularization is designed to regulate the model to only focus on a few important nodes and edges for the explanations. Thus, we propose Equation (4) for global graph regularizations to obtain more reasonable model explanations:

$$\Omega_c(M_c, E_c) = \underbrace{\beta \Omega_c^{con}(M_c, E_c)}_{\text{explanation consistency}} + \underbrace{\gamma \Omega_c^{s}(M_c, E_c)}_{\text{sparsity}} \quad (4)$$

where $\beta$ is the scaling factor for the global explanation consistency between node and edge explanations, $\gamma$ is the scaling factor for the sparsity constraints on both node and edge explanations. These regularizations are described in more detail below:

### 3.1.1 Global explanation consistency regularization

The global node explanation and edge explanation are not independent, but rather highly correlated with each other. One natural assumption about the global node explanation smoothness is that the adjacent nodes should share similar importance. However, this assumption can be too strong and sometimes lead to over-smoothing of the node explanation and tend to yield indistinguishable patterns for the explanation. In addition, it ignores the connection between the node and edge explanations, which can be a crucial factor for the explanation model to generate a global consistent explanation.

Here, we propose to take one step further regarding the smoothness assumption about the explanation by considering both node and edge explanations and making them more consistent with each other. Concretely, instead of treating all pairs of adjacent nodes equally important when enforcing the smoothness constraint, we propose to weight them by the corresponding edge importance such that the explanation consistency is better enforced on those nodes and edges that are deemed important. Mathematically, the global explanation consistency can be measured by Equation (5)

$$\Omega_c^{con}(M_c, E_c) = \frac{1}{T_c} \sum_k \frac{1}{2N^2} \sum_{i,j} E_{c,i,j} A_{i,j}^k \| M_{c,i} - M_{c,j} \|^2 \quad (5)$$

where $k$ is the index of sample belonging to class $c$, $A_{i,j}^k$ is the adjacency matrix for sample $k$, and $T_c$ is the total number of samples in class $c$. The above regularization can be interpreted as

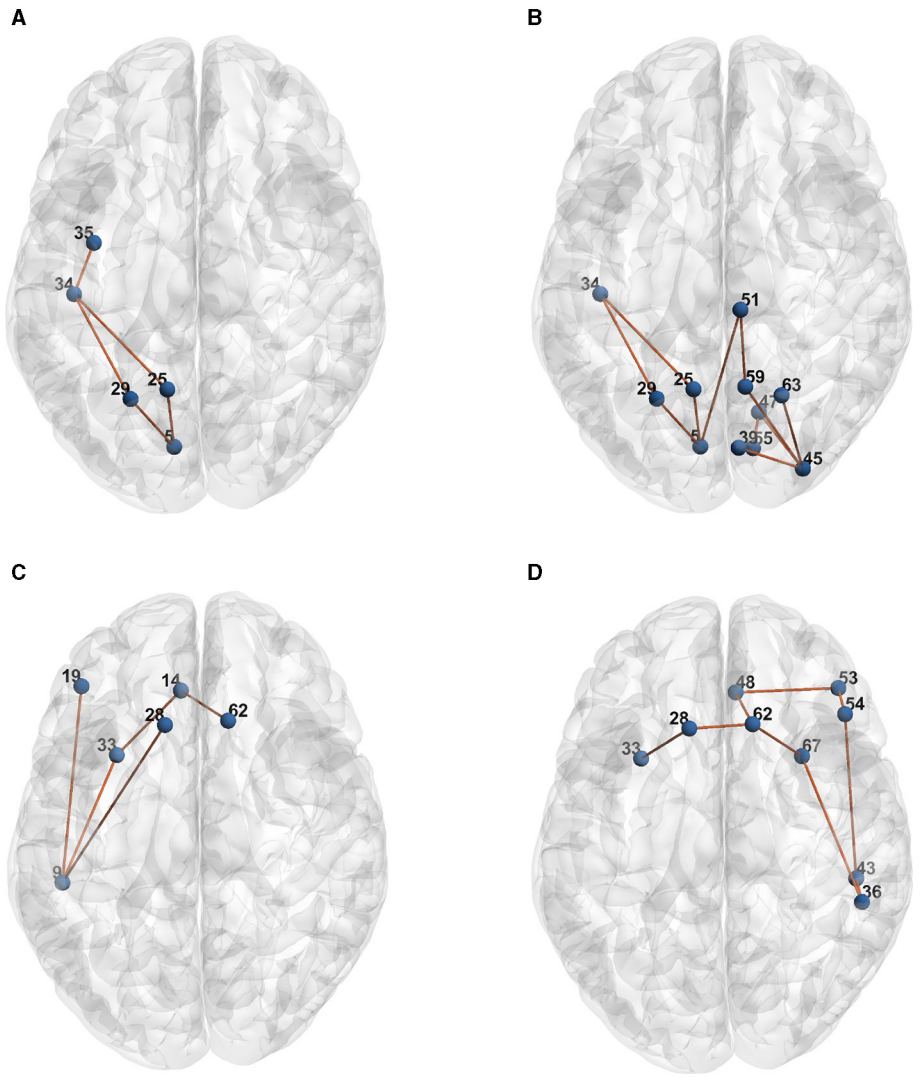


FIGURE 3
Representative element of each learned concept for HCP Structural dataset. **(A)** $P_0$. **(B)** $P_1$. **(C)** $P_2$. **(D)** $P_3$.

follows: given a pair of nodes $i$ and $j$ that is adjacent (i.e., $A_{i,j} = 1$), if the edge that connects the two nodes is important (i.e., $E_{i,j}$ is high), then the nodes it connects also tend to be consistent.

### 3.1.2 Sparsity regularization

As sparsity is a common practice for the model explanation, we apply the $\ell_1$ norm to regulate both the node-level and the edge-level explanations, as Equation (6)

$$\Omega_c^s(M_c, E_c) = \frac{1}{N}\|M_c\|_1 + \frac{1}{N^2}\|E_c\|_1 \tag{6}$$

Overall, the benefits of applying the proposed regularization terms are 3-fold. First, the regularization terms do not rely on the specific human labels on the explanation, which can be very limited and hard to acquire in practice. Thus, they can be very crucial in the scenarios where the explanation labels are scarce. Second, since the explanation for the node and edge can be highly relevant, the proposed explanation consistency regularization can be critical for enforcing the model to generate more reasonable and consistent results that better align with the human explanation. Lastly, our overall framework is very flexible such that the regularization terms are not affected by changing the specification of the node and edge explanation formulation in Equations (7, 12), respectively, making the proposed framework easily applicable to give explanation and apply explanation supervision on any downstream applications with little to no overhead.

The regularization term in Equation (2) is employed to first regulate the node and edge explanation and make them consistent and smooth through considering the dependence of node and edge explanations. Additionally, to lead the model to generate more realistic explanations, the sparsity regularization is also applied which can regulate the model to only focus on a few important nodes and edges for the explanations.

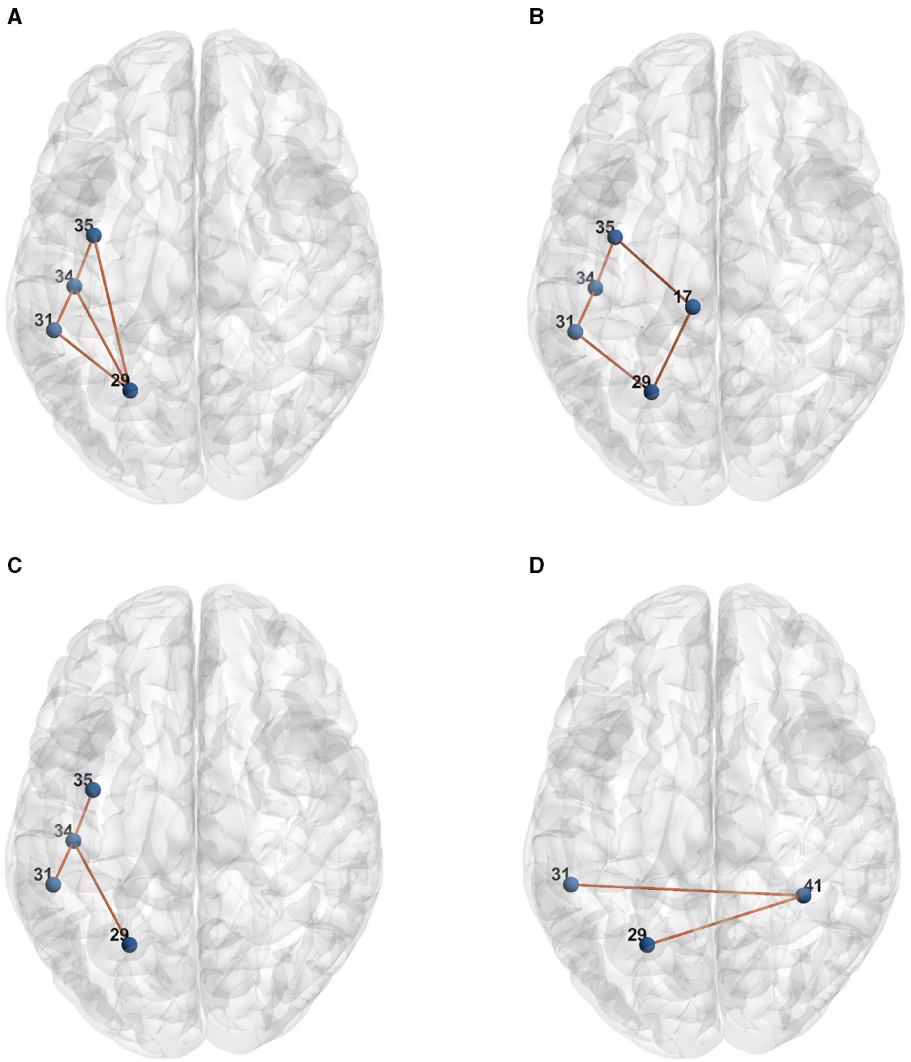


FIGURE 4
Representative element of each learned concept for HCP Functional dataset. **(A)** $P_0$. **(B)** $P_1$. **(C)** $P_2$. **(D)** $P_3$.

## 3.2 Global node explanation formulation for global explanation supervision

In many applications, the ground-truth explanations (on synthetic data) or the domain knowledge (on real-world data) provides node-level explanation of the data in a global manner rather than per sample/instance. In this case, we need to provide a single robust overview of the model predictions. Accordingly, we aim to propose a framework that can both generate the global node explanation by capturing the behavior of the GNN model as a whole (rather than providing instance-specific explanations which could be noisy or not faithful to the model predictions) and also employ it as a supervision signal to further improve the global node-level explanations generated by the model.

To this end, we employ the gradient and the response/activation information which are also the main components for local node explanation supervision as described in Gao et al. (2021). We then aggregate this information over all instances so we can produce a model-generated global explanation that remains differentiable to the backbone GNN model's parameters. This makes the global explanation supervision feasible as the model parameters can be affected and tuned during training. Mathematically, given the output $y_c^i$ on class $c$ and sample $i$, the global explanation for node $n$ at layer $l$ can be computed as follows:

$$M_{n,c}^{(l)} = \Psi\left(\frac{\partial y_c^1}{\partial F_n^{(l)}}, ..., \frac{\partial y_c^i}{\partial F_n^{(l)}}, ..., \frac{\partial y_c^Z}{\partial F_n^{(l)}}, F_n^{(l)}\right) \tag{7}$$

where $\frac{\partial y_c^i}{\partial F_n^{(l)}}$ represents the gradient of the features of node $n$ at layer $l$ given class $c$ and sample $i$, $Z$ is the total number of samples, and $F_n^{(l)}$ denotes the node activation at layer $l$. The function $\Psi$ in Equation (7) can generate any simple to more complicated computations over the input gradients and the activation. Two simple examples are shown in Equations (8, 9), where the gradients are employed to generate simple gradient-based local explanation

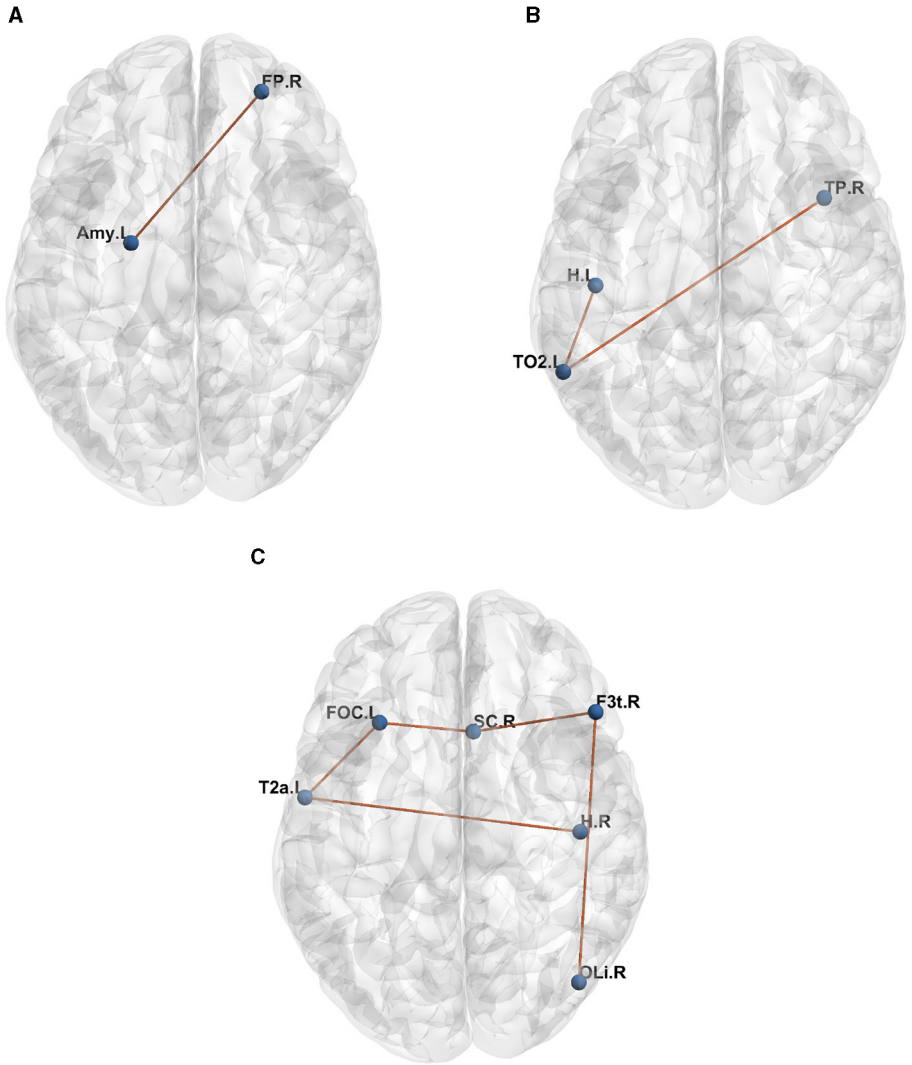


FIGURE 5
Representative element of each learned concept for ABIDE dataset. **(A)** $P_0$. **(B)** $P_1$. **(C)** $P_2$.

for each sample, which are then aggregated using the min or max function to form the final global explanation:

$$M_{n,c}^{(l)} = \min(\|\text{ReLU}(\frac{\partial y_c^1}{\partial F_n^{(l)}})\|, \|\text{ReLU}(\frac{\partial y_c^2}{\partial F_n^{(l)}})\|, ..., \|\text{ReLU}(\frac{\partial y_c^Z}{\partial F_n^{(l)}})\|) \tag{8}$$

$$M_{n,c}^{(l)} = \max(\|\text{ReLU}(\frac{\partial y_c^1}{\partial F_n^{(l)}})\|, \|\text{ReLU}(\frac{\partial y_c^2}{\partial F_n^{(l)}})\|, ..., \|\text{ReLU}(\frac{\partial y_c^Z}{\partial F_n^{(l)}})\|) \tag{9}$$

The other form of aggregation is to average over the local explanations to get the global-level explanation:

$$M_{n,c}^{(l)} = \frac{1}{Z}\sum_{i=1}^{Z} \|\text{ReLU}(\frac{\partial y_c^i}{\partial F_n^{(l)}})\| \tag{10}$$

More complicated technique, described as concept-based global explainer in Azzolin et al. (2022), with some variations, can be used and formulated as below:

$$M_{n,c}^{(l)} = \Lambda(P_1, P_2, ..., P_m) \tag{11}$$

where $P_i$ is the $i-th$ learned prototype which is initialized randomly from a uniform distribution and learned through training the GLGExplainer framework described in Azzolin et al. (2022) and $m$ is the total number of prototypes which is a hyperparameter and tuned separately for each dataset. $\Lambda$ is also a learnable Boolean function that generates a logical combination of the learned prototypes following (Azzolin et al., 2022). In this setting, Equation (11) is a logic formula constructed using graphical concepts derived from local explanations. Concepts can be described as intermediate, high-level and semantically meaningful units of information commonly used by humans to

TABLE 2 Performance and model-generated explanation evaluation among the proposed models and the baseline on two HCP, in addition to one ABIDE graph classification tasks.

| Dataset | Global_exp _method | ACC | AUC | Node MSE | Node MAE | Edge MSE | Edge MAE |
|---|---|---|---|---|---|---|---|
| HCP functional | None | 0.736 | 0.843 | 0.392 | 0.436 | – | – |
| | Avg | 0.741 | 0.854 | **0.311** | 0.394 | – | – |
| | max | 0.736 | 0.843 | 0.324 | 0.418 | – | – |
| | min | 0.738 | 0.845 | 0.321 | 0.414 | – | – |
| | concept_based | **0.759** | **0.899** | **0.311** | **0.372** | – | – |
| HCP structural | None | 0.829 | 0.961 | 0.238 | 0.436 | – | – |
| | Avg | 0.833 | 0.965 | 0.224 | 0.322 | – | – |
| | max | 0.830 | **0.971** | 0.220 | 0.397 | – | – |
| | min | 0.833 | **0.971** | 0.217 | 0.323 | – | – |
| | concept_based | **0.838** | **0.971** | **0.101** | **0.223** | – | – |
| ABIDE | None | 0.730 | 0.868 | 0.237 | 0.437 | 0.065 | 0.033 |
| | Avg | 0.735 | 0.870 | 0.218 | 0.416 | 0.051 | 0.031 |
| | max | 0.732 | 0.871 | 0.215 | 0.406 | 0.055 | 0.025 |
| | min | 0.730 | 0.868 | 0.222 | 0.413 | 0.061 | **0.021** |
| | concept_based | **0.744** | **0.885** | **0.191** | **0.331** | **0.043** | 0.024 |

The results are obtained from five individual runs for every setting. The best results for each task are highlighted with boldface font, and the second bests are underlined.

explain their decisions. More details for GLGExplainer are given in Azzolin et al. (2022).

The training process based on Equation (11) consists of three steps. First, a basic GCN is trained by optimizing only the first term in Equation (3). Second, the local explanations generated by this trained GNN are fed as inputs to the GLGExplainer which can construct the logic formula of Equation (11). Last, the original GCN is re-trained through the full loss function in Equation (3). For the third or last step, we only employ the logic formula from step 2 and discard the prototypes generated. Instead we randomly initialize the values of prototypes from a uniform distribution. Accordingly, the GCN and GLGExplainer are trained iteratively until the value of prototypes would converge. Note that all the parameters of GLGExplainer in step three are exactly equal to those in step 2, except for the prototypes that remain learnable and are updated at each iteration.

For all the functions in Equations (8–11), the results are computed and included in the Experiments section with further discussions.

## 3.3 Global edge explanation formulation for global explanation supervision

While several works have studied global node-level explanation topic, little to no work has explored the global edge-level explanation and its applications. However, in many scenarios, the latter can be more crucial and meaningful than the former as the domain knowledge or human annotations describe the relationship between nodes rather than the nodes in particular.

Similar to the global node explanation supervision, we need to propose a unified edge-level explanation formulation which generates explanations that are differentiable to the backbone model's parameters. Taking the gradient of each edge in the input adjacency matrix, as well as the response/activation of the pairs of nodes that are associated with that edge, and using the chain rule, we can define suitable model generated explanations for each instance. Concretely, given the output $y_c^i$ on class $c$ and sample $i$, the global edge explanation between node $n$ and node $m$ at layer $l$ can be computed as the aggregation of all edge explanations for single instances. More precisely, this is a function of the edge gradients for all samples, in addition to node activations:

$$E^{(l)}_{n,m,c} = \Phi\left(\frac{\partial y_c^1}{\partial F^{(l)}} \cdot \frac{\partial F^{(l)}}{\partial A^1_{n,m}}, \ldots, \frac{\partial y_c^i}{\partial F^{(l)}} \cdot \frac{\partial F^{(l)}}{\partial A^i_{n,m}}, \ldots, \frac{\partial y_c^N}{\partial F^{(l)}} \cdot \frac{\partial F^{(l)}}{\partial A^N_{n,m}}, F^{(l)}_n, F^{(l)}_m\right) \tag{12}$$

where $\frac{\partial y_c^i}{\partial F^{(l)}} \cdot \frac{\partial F^{(l)}}{\partial A^i_{n,m}}$ represents the gradient of the edge that connects nodes $n$ and $m$ at layer $l$ given class $c$ and sample $i$; $F^{(l)}_n$ and $F^{(l)}_m$ denote the activation of node $n$ and node $m$ at layer $l$, respectively, and $N$ is the total number of instances. Similar to previous formulation in Equation (7), $\Phi$ can combine the local explanations of all samples, providing a global explanation for the overall behavior of the GNN. A simple example is the min or max value among all the gradient-based

TABLE 3 Fidelity, accuracy, and concept purity computed over test sets for all datasets.

| Dataset | Fidelity | Accuracy | Concept purity |
|---|---|---|---|
| HCP structural | 0.91 | 0.89 | 0.82 |
| HCP functional | 0.81 | 0.83 | 0.85 |
| ABIDE | 0.78 | 0.79 | 0.85 |

edge-level local explanations, which can be formulated as Equation (13)

$$E^{(l)}_{n,m,c} = \min(\|\mathrm{ReLU}(\frac{\partial y^1_c}{\partial F^{(l)}} \cdot \frac{\partial F^{(l)}}{\partial A^1_{n,m}})\|, ..., \|\mathrm{ReLU}(\frac{\partial y^i_c}{\partial F^{(l)}} \cdot \frac{\partial F^{(l)}}{\partial A^i_{n,m}})\|, ..., \|\mathrm{ReLU}(\frac{\partial y^Z_c}{\partial F^{(l)}} \cdot \frac{\partial F^{(l)}}{\partial A^Z_{n,m}})\|) \quad (13)$$

where $Z$ is the total number of samples, and a similar formulation can be used to find the max value of the local explanations. Averaging over the local explanations can also be another aggregator to generate the global .explanation and can be shown by Equation (14)

$$E^{(l)}_{n,m,c} = \frac{1}{Z}\sum_{i=1}^{Z} \|\mathrm{ReLU}(\frac{\partial y^i_c}{\partial F^{(l)}} \cdot \frac{\partial F^{(l)}}{\partial A^i_{n,m}})\| \quad (14)$$

Similar to Equation (11), the global edge explanation can also be represented as a learnable logic combination of concepts. As long as the GNN model can generate local explanations that are a subgraph of the input data, these can be fed into the GLGExplainer in Azzolin et al. (2022) which can learn the formula, and parameters in Equation (11) and generate the global explanation per class.

These various functions for $\Phi$ are investigated in detail in the Experiment section.

# 4 Experiments

We test our Global GNN Explanation Supervision framework on the datasets extracted from two publicly available sources including HCP (Human Connectome Project) and the ABIDE (Autism Brain Imaging Data Exchange) database. These datasets, in addition to the implementation details, evaluation metrics, and comparison methods are described in turn below.

## 4.1 Datasets

### 4.1.1 Magnetic resonance imaging data

The (structural, diffusion, and functional) MRI data were extracted from the Human Connectome Project website (https://db.humanconnectome.org/), specifically, the 1,200 Subjects Release, February 2017 (Van Essen et al., 2013), which provided (MRI) data from 1,200 young adult (ages 22–35) subjects. Here, two tasks are defined as binary classification of a given subject as Female vs. Male, in addition to Young (22–29) vs. Old (29–35). The age and gender labels were provided as additional meta features. For the ground-truth explanations of each class, we refer to Gong et al. (2009), which has investigated age and sex effects on the anatomical connectivity patterns of 95 normal subjects ranging in age from 19 to 85 years. Accordingly, cortical regions which show significant effect for young, old, male, or female subjects were separately identified for each group for Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002). To use these as annotations for HCP dataset, these regions were then mapped to Desikan-Killiany (DK) atlas (Desikan et al., 2006), by finding the closest node (Euclidean distance) in DK to each identified node in AAL atlas. The resulting DK nodes are provided in Supplementary material for each class under study.

The raw MRI data were then preprocessed using the HCP pipeline (WU-Minn, 2017). For the diffusion MRI, this was followed by the BEDPOSTX (Bayesian Estimation of Diffusion Parameters Obtained using Sampling Techniques, modeling crossing X fibers) algorithm in the FMRIB Software Library (Jenkinson et al., 2012, FSL), which models white matter fiber orientations and crossing fibers for probabilistic tractography. The resting state blood-oxygen-level-dependent functional MRI (r-fMRI) time series data were acquired from participants, in four runs of ∼15 min for each participant, including two runs on two different days (Day 1 and Day 2). These measurements were collected with the subject supine and still, with eyes open, to track physiological changes in the brain (i.e., changes in blood flow and oxygen levels) that occur in resting state, when an explicit task is not being performed (Biswal, 2012; Buckner et al., 2013).

**Extracting SC and FC:** To construct the SC matrix for each subject, we ran Probtrackx in FSL with 68 regions of interest (ROIs) obtained from the the DK atlas. For the remaining parameter setting in Probtractx, we followed the recommendations of the tutorial (in St.Louis, 2020) provided by HCP. Finally, the resulting SC matrices were normalized by dividing the respective row sum from each non-zero value.

Three steps were followed to extract the functional connectivity from the r-fMRI time series data, for each day: 1. Concatenate the time series for the two runs together; 2. For each of the 68 ROIs defined by the Desikan-Killiany atlas, average all the time series to create a single ROI time series; and 3. obtain the functional connectivities by either (a) Computing the pairwise ROI time series' Pearson correlations using FSLNets (of Heidelberg Department of Neuroradiology, 2014) with the full correlation option, thus generating **Dataset 1**; or following similar three steps as mentioned for Dataset 1, except that we Concatenate the time series for the two runs performed in day 2 together, thus generating **Dataset 2**. In this study, we followed with the experiments only using **Dataset 1** due to the high amount of computation and resources required for each Dataset.

### 4.1.2 ABIDE dataset

We analyzed r-fMRI in the Autism Brain Imaging Data Exchange (ABIDE; Di Martino et al., 2014). It compiles a dataset of 1,112 r-fMRI participants by gathering data from 16

TABLE 4 Performance and model-generated explanation evaluation for two additional local explainers with GCN as the backbone model.

| Dataset | Local_exp _method | Global_exp _method | ACC | AUC | Node MSE | Node MAE | Edge MSE | Edge MAE |
|---|---|---|---|---|---|---|---|---|
| HCP functional | Guided BP | None | 0.736 | 0.843 | 0.392 | 0.436 | – | – |
| | | Avg | 0.736 | 0.845 | 0.341 | 0.381 | – | – |
| | | max | 0.736 | 0.843 | 0.355 | 0.410 | – | – |
| | | min | 0.739 | 0.845 | 0.321 | 0.382 | – | – |
| | | concept_based | **0.752** | **0.870** | **0.311** | **0.377** | – | – |
| | Grad-CAM | None | 0.736 | 0.843 | 0.392 | 0.436 | – | – |
| | | Avg | 0.738 | 0.854 | **0.311** | 0.384 | – | – |
| | | max | 0.737 | 0.845 | 0.338 | 0.422 | – | – |
| | | min | 0.736 | 0.845 | 0.348 | 0.417 | – | – |
| | | concept_based | **0.749** | **0.893** | **0.311** | **0.362** | – | – |
| HCP structural | Guided BP | None | 0.829 | 0.961 | 0.238 | 0.436 | – | – |
| | | Avg | 0.831 | 0.965 | 0.211 | 0.289 | – | – |
| | | max | 0.833 | **0.970** | 0.224 | 0.318 | – | – |
| | | min | 0.835 | 0.966 | 0.221 | 0.314 | – | – |
| | | concept_based | **0.840** | **0.970** | **0.118** | **0.233** | – | – |
| | Grad-CAM | None | 0.829 | 0.961 | 0.238 | 0.436 | – | – |
| | | Avg | 0.835 | 0.966 | 0.188 | 0.239 | – | – |
| | | max | 0.833 | **0.968** | 0.124 | 0.228 | – | – |
| | | min | 0.829 | 0.965 | 0.201 | 0.314 | – | – |
| | | concept_based | **0.838** | **0.968** | **0.111** | **0.225** | – | – |
| ABIDE | Guided BP | None | 0.730 | 0.868 | 0.237 | 0.437 | 0.065 | 0.033 |
| | | Avg | 0.735 | 0.873 | 0.230 | 0.416 | 0.055 | 0.030 |
| | | max | 0.732 | 0.874 | 0.227 | 0.416 | 0.053 | 0.025 |
| | | min | 0.730 | 0.869 | 0.222 | 0.403 | 0.061 | **0.027** |
| | | concept_based | **0.744** | **0.883** | **0.200** | **0.355** | **0.045** | **0.027** |
| | Grad-CAM | None | 0.730 | 0.868 | 0.237 | 0.437 | 0.065 | 0.033 |
| | | Avg | 0.735 | 0.878 | 0.218 | 0.403 | **0.045** | **0.023** |
| | | max | 0.730 | 0.869 | 0.218 | 0.392 | 0.055 | 0.026 |
| | | min | 0.730 | 0.868 | 0.212 | 0.412 | 0.061 | 0.026 |
| | | concept_based | **0.741** | **0.885** | **0.195** | **0.337** | **0.045** | 0.025 |

The results are obtained from five individual runs for every setting. The best results for each task are highlighted with boldface font, and the second bests are underlined.

international imaging sites that have aggregated and are openly sharing neuroimaging data from 539 individuals suffering from ASD and 573 typical controls (TCs). The task is to classify a subject as either belonging to ASD or the control group, based on their r-fMRI data. Since there was no prior coordination between sites, the scan and diagnostic/assessment protocols vary across sites. Accordingly, we rely on a publicly available preprocessed version of this dataset provided by the Preprocessed Connectome Project (PCP) initiative. PCP preprocessed the data using four different pipelines, all of which implemented fairly similar steps, but varied in the algorithms used for each step and the parameters. We specifically used the data processed with the Configurable Pipeline for the Analysis of Connectomes, C-PAC (Craddock et al., 2013), which provides further minimally preprocessed data through the python package, cpac. C-PAC comes pre-packaged with a default pipeline, as well as a growing library of pre-configured pipelines. These pipelines could be edited or built from scratch, using the provided pipeline builder. For our experiments, we used the default processing pipeline. For more details, please see Craddock et al. (2013) on how we extracted time series for the Harvard-Oxford atlas. We finally used the same steps as HCP dataset, to compute the functional connectivity matrices. Additionally, we used the biomarkers extracted by Kunda et al. (2020), as the ground-truth labels for explanation supervision. These include the top five most contributing FC edges for ASD and TC classification, respectively (10 overall connections), built using

TABLE 5 Performance and model-generated explanation evaluation for all three local explainers with DGCNN as the backbone model.

| Dataset | Local_exp_method | Global_exp_method | ACC | AUC | Node MSE | Node MAE | Edge MSE | Edge MAE |
|---|---|---|---|---|---|---|---|---|
| HCP functional | Gradient based | None | 0.708 | 0.791 | 0.394 | 0.438 | – | – |
| | | Avg | 0.715 | 0.809 | 0.343 | 0.398 | – | – |
| | | max | 0.708 | 0.792 | 0.344 | 0.402 | – | – |
| | | min | 0.712 | 0.812 | 0.332 | 0.422 | – | – |
| | | concept_based | **0.718** | **0.855** | **0.301** | **0.328** | – | – |
| | Guided BP | None | 0.708 | 0.791 | 0.394 | 0.438 | – | – |
| | | Avg | 0.712 | 0.832 | 0.342 | 0.400 | – | – |
| | | max | 0.712 | 0.838 | 0.339 | 0.400 | – | – |
| | | min | 0.708 | 0.805 | 0.337 | 0.391 | – | – |
| | | concept_based | **0.715** | **0.861** | **0.319** | **0.370** | – | – |
| | Grad-CAM | None | 0.708 | 0.791 | 0.394 | 0.438 | – | – |
| | | Avg | 0.721 | 0.844 | 0.330 | **0.375** | – | – |
| | | max | 0.708 | 0.843 | **0.308** | 0.390 | – | – |
| | | min | 0.708 | 0.835 | 0.320 | 0.385 | – | – |
| | | concept_based | **0.725** | **0.857** | 0.315 | **0.375** | – | – |
| HCP structural | Gradient based | None | 0.803 | 0.941 | 0.279 | 0.446 | – | – |
| | | Avg | 0.812 | 0.954 | 0.189 | **0.297** | – | – |
| | | max | 0.803 | 0.943 | 0.224 | 0.318 | – | – |
| | | min | 0.808 | 0.945 | 0.185 | 0.314 | – | – |
| | | concept_based | **0.812** | **0.958** | **0.145** | 0.298 | – | – |
| | Guided BP | None | 0.803 | 0.941 | 0.279 | 0.446 | – | – |
| | | Avg | 0.808 | 0.953 | 0.231 | 0.294 | – | – |
| | | max | 0.808 | 0.954 | 0.198 | 0.318 | – | – |
| | | min | 0.803 | 0.945 | 0.221 | 0.314 | – | – |
| | | concept_based | **0.814** | **0.961** | **0.161** | **0.286** | – | – |
| | Grad-CAM | None | 0.803 | 0.941 | 0.279 | 0.446 | – | – |
| | | Avg | 0.811 | 0.962 | 0.201 | 0.304 | – | – |
| | | max | 0.806 | 0.953 | 0.228 | 0.388 | – | – |
| | | min | 0.803 | 0.952 | 0.220 | 0.401 | – | – |
| | | concept_based | **0.815** | **0.963** | **0.183** | **0.272** | – | – |
| ABIDE | Gradient based | None | 0.730 | 0.860 | 0.292 | 0.446 | 0.065 | 0.033 |
| | | Avg | 0.741 | 0.868 | 0.221 | 0.394 | 0.045 | **0.026** |
| | | max | 0.741 | 0.867 | 0.254 | 0.418 | 0.045 | **0.026** |
| | | min | 0.732 | 0.865 | 0.271 | 0.392 | 0.054 | 0.029 |
| | | concept_based | **0.744** | **0.874** | **0.211** | **0.372** | **0.043** | **0.026** |
| | Guided BP | None | 0.730 | 0.860 | 0.292 | 0.446 | 0.065 | 0.033 |
| | | Avg | 0.737 | 0.873 | 0.199 | 0.362 | **0.041** | 0.026 |
| | | max | 0.732 | 0.865 | 0.198 | 0.400 | 0.051 | **0.023** |
| | | min | 0.732 | 0.867 | 0.198 | 0.403 | 0.053 | 0.027 |
| | | concept_based | **0.742** | **0.880** | **0.197** | **0.343** | 0.048 | 0.026 |

*(Continued)*

TABLE 5 (Continued)

| Dataset | Local_exp _method | Global_exp _method | ACC | AUC | Node MSE | Node MAE | Edge MSE | Edge MAE |
|---|---|---|---|---|---|---|---|---|
| | Grad-CAM | None | 0.730 | 0.860 | 0.292 | 0.446 | 0.065 | 0.033 |
| | | Avg | 0.738 | 0.873 | 0.221 | 0.401 | 0.051 | 0.027 |
| | | max | 0.734 | 0.872 | 0.221 | 0.398 | 0.056 | 0.027 |
| | | min | 0.735 | 0.870 | 0.219 | 0.415 | 0.060 | 0.028 |
| | | concept_based | **0.744** | **0.883** | **0.193** | **0.335** | **0.043** | **0.025** |

The results are obtained from five individual runs for every setting. The best results for each task are highlighted with boldface font, and the second bests are underlined.

the Harvard-Oxford (HO) brain atlas (Jenkinson et al., 2012) as a point of reference.

## 4.2 Implementation details

Following the previous work on the explanation supervision for GNNs, we used a 3 layer GCN as our backbone GNN model. The hidden dimension size for the three graph convolutional layers is tuned separately for each dataset/task. We used 2 for Gender prediction and 3 for age prediction tasks. For the ASD classification task, we found 3 to best classify the dataset. These hidden layers are followed by a global average pooling (GAP) layer, and a softmax classifier. Models were trained for 200, 300, and 260 epochs using the ADAM optimizer (Kingma and Ba, 2014), respectively, with a learning rate of 0.001 in all three cases. For the remaining details of implementation and parameters, we followed all the settings in Gao et al. (2021), unless otherwise specified.

For the GLGExplainer, we prepared the input using the simple gradient-based local explainer in the backbone GNN. The number of prototypes was set to 4 and 2 for the HCP dataset and the ABIDE data, respectively. This explainer was trained using all the remaining settings and parameters including the optimizer, learning rate, batch size, focusing parameter, and auxiliary loss coefficients, in addition to the E-LEN, from the original proposed model (Azzolin et al., 2022).

### 4.2.1 Evaluation metrics

We evaluate the effectiveness of the proposed GGNES model in terms of prediction performance as well as in terms of global explainability. Specifically, for model performance assessment, we use accuracy (ACC) and Area Under the Curve (AUC) scores to measure the prediction power of the GNNs on the prediction tasks for all the datasets. In addition, we leverage the human/domain-labeled explanation on the test set to quantitatively assess the goodness of the model explanation. Specifically, for both node-level and edge-level global explanations, we treat the human explanation as the gold standard and compute the distance between human and global model explanation via Mean Square Error (MSE) and Mean Absolute Error (MAE). Additionally, we evaluate our model on: (i) FIDELITY, which represents the accuracy of the E-LEN in matching the predictions of the GNN model to explain; (ii) ACCURACY, which represents the accuracy of the formulas in matching the ground-truth labels of the graphs; (iii) CONCEPT PURITY, which is computed for every cluster independently and measures how good the embedding is at clustering the local explanations (Azzolin et al., 2022), and is computed through Equation (15)

$$\text{ConceptPurity}(C_i) = \frac{\text{count\_most\_frequent\_label}(C_i)}{|C_i|} \tag{15}$$

where $C_i$ corresponds to the cluster having $p_i$ as the learned prototype, and $count\_most\_frequent\_label(C_i)$ returns the number of local explanations annotated with the most present label in cluster $C_i$. The Concept Purity results are reported by computing the mean and the standard deviation across all clusters. For a more detailed description of these metrics, see Azzolin et al. (2022).

### 4.2.2 Comparison methods

Since there is no existing work on global explanation supervision on GNNs, we demonstrate the effectiveness of our model by comparing the evaluation metrics in the following scenarios:

- No explanation supervision technique is used.
- min, max, or Average functions are used to generate global explanation and perform explanation supervision.
- Concept-based global explanation is constructed and further used for supervision.

## 4.3 Experimental results

Table 1 presents the raw formulas extracted by the Entropy Layer. Those formulas can be further described in a more human-understandable format after finding the representative elements of each cluster as shown in Figures 3–5, which correspond to HCP structural, HCP functional, and ABIDE datasets, respectively. Each of these Figures contains a number of sub-Figures that show the learned prototypes described in Section 3.2. Specifically, for each prototype $p_j$, the local explanation $\bar{G}$ such that $\bar{G} = \text{argmax}_{\bar{G}' \in D}\, d(p_j, h(\bar{G}'))$ is reported. Here, $D$ is a list of local explanations obtained after the binarization step in GLGExplainer. For details on this step, in addition to the definition for distance function $d()$, please see Azzolin et al. (2022). The nodes in Figures 3, 4 refer to DK atlas and are labeled with numbers for better readability, while the nodes in Figure 5 correspond to HO atlas. See Supplementary material for the label names corresponding to the labels we used in Figures 3, 4. For a list of HO atlas labels used in Figure 5, see Atlas (2023).

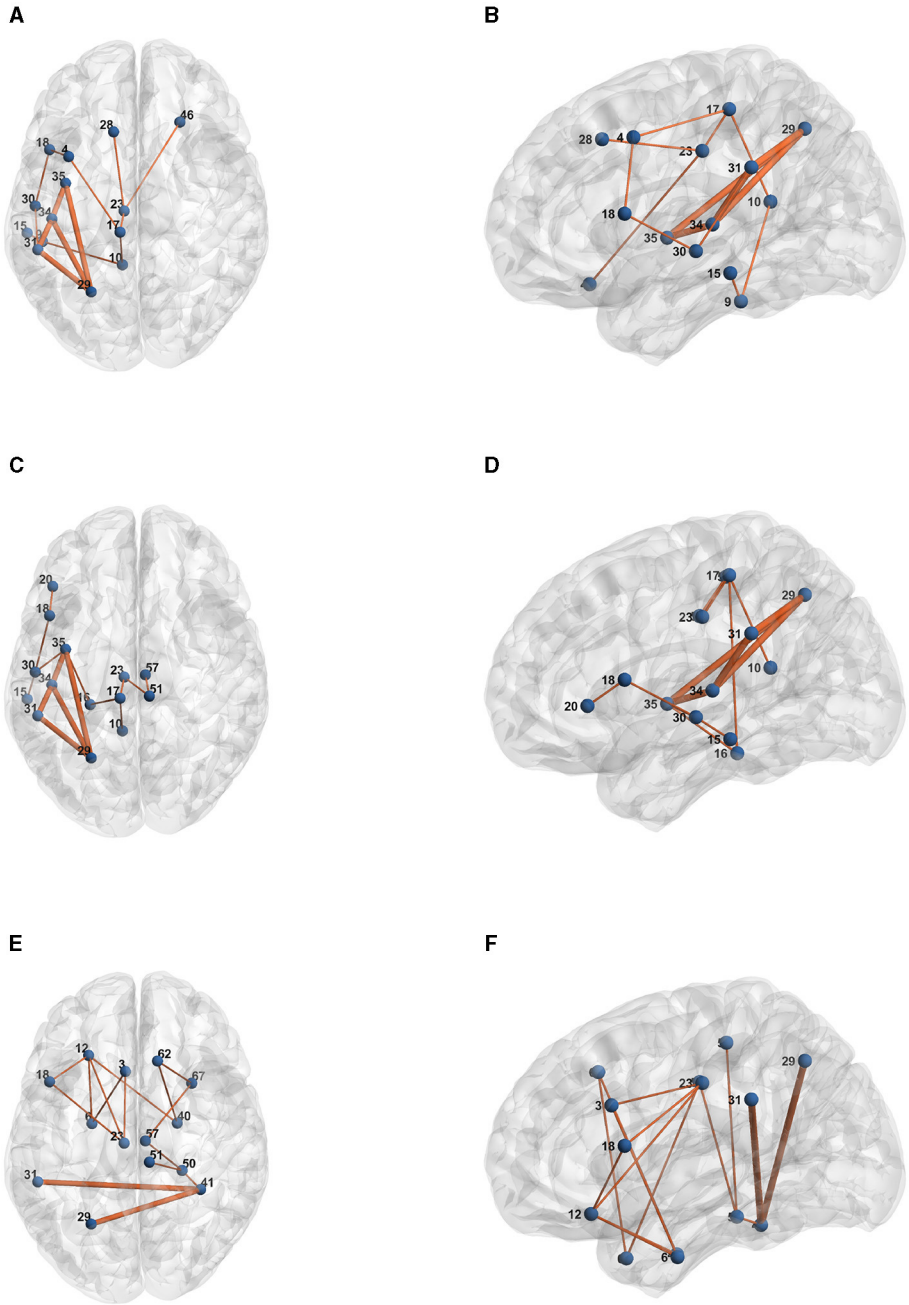


FIGURE 6
Examples of input graphs with their explanations in bold as extracted by Gradient-based edge explanation technique, for HCP structural connectivity dataset. **(A)** Subject 286-Female-Axial view. **(B)** Subject 286-Female-Sagittal view. **(C)** Subject 543-Female-Axial view. **(D)** Subject 543-Female-Sagittal view. **(E)** Subject 56-Male-Axial view. **(F)** Subject 56-Male-Sagittal view.

### 4.3.1 Performance

Table 2 shows the model performance and model-generated explanation quality for the three described datasets. The results are obtained from 5 individual runs for every setting. The best results for each dataset are highlighted with boldface font, and the second bests are underlined. For the HCP datasets, for both Age and Gender prediction tasks, the human annotations contain only node-level explanations, but for the ABIDE dataset we have both ground-truth (domain-labeled) node-level and edge-level explanations available for all samples. In general, our proposed Global Explanation Supervision model variations outperformed the non-explanations supervision GNN model in terms of both prediction power as well as explainability on all three datasets. More specifically, the performance results for different variations suggested that global explanation supervison can have positive effects in all scenarios on both prediction power, in addition to the explanation correctness. The most complicated model (i.e., the concept-based supervision model) achieved the best performance, out-performing baseline GNN by 1–6% and 1–3% on AUC and ACC scores, respectively. In addition, in terms of

explainability, there is significant improvement in both node and edge-level explanations, when comparing the backbone GNN and the concept-based supervision models. In particular, we observed between 19–57% increase in node MSE and 14–48% in node MAE, and more than 33% improvement for edge MSE and MAE explanations.

These results demonstrate the general effectiveness of the proposed framework both on largely correcting the model-generated global explanation, in addition to improving the model performance and prediction power. In addition, among different variations used for global explanation generation, we observe constant superiority of the more sophisticated concept-based technique compared to the others, while no clear excellence of Avg, max, or min methods when comparing one to the other was remarkable.

To further evaluate the extracted global explanation formulas presented in Table 1, we computed Fidelity, Accuracy, and Concept Purity over the test set. The results are reported in Table 3 for the three datasets. As it can be seen, on average, the clusters are quite homogeneous, which means the model has learned a good mapping from the local explanations to the concepts space. Also the concept purity is at its lowest for HCP structural dataset while has the highest value for the same set. The accuracy results demonstrate that the formula in Table 1 can correctly match the behavior of the model in most samples. Additionally, it is important to note that by looking at the fidelity results, it is clear that the explainer is generating an explanation for the ground-truth labeling of the dataset, while capturing the underlying predictive behavior of the GNN it is supposed to explain.

### 4.3.2 Effect of choice of local explainer and backbone GNN model

To evaluate the proposed model more comprehensively, we repeated experiments for the model performance and model-generated explanation quality for all datasets for two additional local explanation techniques, Guided BP and GradCAM, and one other backbone GNN model, DGCNN (Zhang et al., 2018). The results are shown in Tables 4, 5. As these results show, we continue to see superiority of our proposed Global Explanation Supervision model variations compared to non-explanation supervised scenarios. The concept-based supervision model again achieved the best performance, out-performing baseline GNN, for the two backbone GNNs and all three local explanation techniques. Additionally, we observe significant improvement in both node, and edge-level MSE and MAE, when using the concept-based supervision models. These improvements, both in explanation quality and model performance, for the concept-based technique largely exceed other simple aggregation methods (e.g., Averaging) in almost all settings as well.

### 4.3.3 Qualitative analysis: case studies

Here, we provide some case studies of the input data and the model explanation derived from gradient based explanation technique and binarized following (Azzolin et al., 2022). We report some random examples for each dataset, with their extracted explanation in bold, as illustrated in Figure 6.

## 5 Conclusion

In this study, we address an existing challenge for explainability in GNNs, by proposing the Global GNN Explanation Supervision (GGNES) technique which uses a basic trained GNN and a global extension of the loss function used in the GNES framework. This GNN creates local explanations which are fed to a Global Logic-based GNN Explainer, an existing technique that can learn the global Explanation in terms of a logic formula. These two frameworks are then trained iteratively to generate reasonable global explanations. Extensive experiments demonstrate the effectiveness of the proposed model on improving the global explanations while keeping the performance similar or even increase the model prediction power.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

## Ethics statement

The studies involving humans were approved by the Autism Brain Imaging Data Exchange, the Human Connectome Project. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## Author contributions

NE: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing—original draft, Writing—review & editing. YG: Methodology, Writing—original draft, Writing—review & editing. SM: Resources, Writing—original draft, Writing—review & editing. LZ: Conceptualization, Formal analysis, Methodology, Writing—original draft, Writing—review & editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the NSF Grant Nos. 2432418, 2414115, 2007716, 2007976, 1942594, 1907805, and 2318831, Cisco Faculty Research Award, Amazon Research Award.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships

that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2024.1410424/full#supplementary-material

## References


Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052

Annervaz, K., Chowdhury, S. B. R., and Dukkipati, A. (2018). Learning beyond datasets: knowledge graph augmented neural networks for natural language processing. *arXiv preprint arXiv:1802.05930*. doi: 10.18653/v1/N18-1029

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Informa. Fus.* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Atlas, T. S. B. (2023). *Brain Atlas*. Available online at: https://scalablebrainatlas.incf.org/human/HOA06 (accessed March 15, 2024).

Azzolin, S., Longa, A., Barbiero, P., Lió, P., and Passerini, A. (2022). Global explainability of GNNs via logic combination of learned concepts. *arXiv preprint arXiv:2210.07147*. doi: 10.48550/arXiv.2210.07147

Baldassarre, F., and Azizpour, H. (2019). Explainability techniques for graph convolutional networks. *arXiv preprint arXiv:1905.13686*. doi: 10.48550/arXiv.1905.13686

Biswal, B. B. (2012). Resting state fMRI: a personal history. *Neuroimage* 62, 938–944. doi: 10.1016/j.neuroimage.2012.01.090

Buckner, R. L., Krienen, F. M., and Yeo, B. T. (2013). Opportunities and limitations of intrinsic functional connectivity MRI. *Nat. Neurosci.* 16, 832–837. doi: 10.1038/nn.3423

Chen, S., Jiang, M., Yang, J., and Zhao, Q. (2020). "Air: attention with reasoning capability," in *European Conference on Computer Vision* (Berlin: Springer), 91–107.

Craddock, C., Sikka, S., Cheung, B., Khanuja, R., Ghosh, S. S., Yan, C., et al. (2013). Towards automated analysis of connectomes: the configurable pipeline for the analysis of connectomes (C-PAC). *Front. Neuroinform.* 9:42. doi: 10.3389/conf.fninf.2013.09.00042

Das, A., Agrawal, H., Zitnick, L., Parikh, D., and Batra, D. (2017). Human attention in visual question answering: do humans and deep networks look at the same regions? *Comput. Vis. Image Underst.* 163, 90–100. doi: 10.1016/j.cviu.2017.10.001

De Domenico, M., Lancichinetti, A., Arenas, A., and Rosvall, M. (2015). Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Phys. Rev. X* 5:011027. doi: 10.1103/PhysRevX.5.011027

De Haan, W., Pijnenburg, Y. A., Strijers, R. L., van der Made, Y., van der Flier, W. M., Scheltens, P., et al. (2009). Functional neural network analysis in frontotemporal dementia and Alzheimer's disease using EEG and graph theory. *BMC Neurosci.* 10, 1–12. doi: 10.1186/1471-2202-10-101

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31, 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiat.* 19, 659–667. doi: 10.1038/mp.2013.78

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., and Lee, S. I. (2021). Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nat. Machine Intell.* 3, 620–631. doi: 10.1038/s42256-021-00343-w

Etemadyrad, N., Gao, Y., Li, Q., Guo, X., Krueger, F., Lin, Q., et al. (2022). Functional connectivity prediction with deep learning for graph transformation. *IEEE Trans. Neural Netw. Learn. Syst.* 35, 4862–4875. doi: 10.1109/TNNLS.2022.3197337

Fan, W., Ma, Y., Li, Q., He, Y., Zhao, E., Tang, J., et al. (2019). "Graph neural networks for social recommendation," in *The World Wide Web Conference* (New York, NY: ACM), 417–426.

Fukui, H., Hirakawa, T., Yamashita, T., and Fujiyoshi, H. (2019). "Attention branch network: learning of attention mechanism for visual explanation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New York, NY: Computer Vision Foundation/IEEE), 10705–10714.

Gao, Y., Gu, S., Jiang, J., Hong, S. R., Yu, D., and Zhao, L. (2024). Going beyond XAI: a systematic survey for explanation-guided learning. *ACM Comput. Surv.* 56, 1–39. doi: 10.1145/3644073

Gao, Y., Sun, T., Bhatt, R., Yu, D., Hong, S., and Zhao, L. (2021). "GNES: learning to explain graph neural networks," in *2021 IEEE International Conference on Data Mining (ICDM)* (New York, NY: IEEE), 131–140.

Gao, Y., Sun, T. S., Bai, G., Gu, S., Hong, S. R., and Liang, Z. (2022a). "RES: a robust framework for guiding visual explanation," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (New York, NY: ACM), 432–442.

Gao, Y., Sun, T. S., Zhao, L., and Hong, S. R. (2022b). Aligning eyes between humans and deep neural network through interactive attention alignment. *Proc. ACM Hum. Comput. Interact.* 6, 1–28. doi: 10.1145/3555590

Gong, G., Rosa-Neto, P., Carbonell, F., Chen, Z. J., He, Y., and Evans, A. C. (2009). Age-and gender-related differences in the cortical anatomical network. *J. Neurosci.* 29, 15684–15693. doi: 10.1523/JNEUROSCI.2308-09.2009

Gupta, A., Saini, S., and Narayanan, P. (2024). Concept distillation: leveraging human-centered explanations for model improvement. *Adv. Neural Inform. Process. Syst.* 36:15303. doi: 10.48550/arXiv.2311.15303

Hong, S. R., Hullman, J., and Bertini, E. (2020). Human factors in model interpretability: industry practices, challenges, and needs. *Proc. ACM Hum. Comput. Interact.* 4, 1–26. doi: 10.1145/3392878

Huang, Q., Yamada, M., Tian, Y., Singh, D., Yin, D., and Chang, Y. (2020). Graphlime: Local interpretable model explanations for graph neural networks. *arXiv preprint arXiv:2001.06216*. doi: 10.48550/arXiv.2001.06216

in St.Louis, W. U. (2020). Available online at: https://wustl.app.box.com/s/wna2cu94pqgt8zskg687mj8z-lmfj1pq7 (accessed February 15, 2022).

Jacovi, A., and Goldberg, Y. (2020). Aligning faithful interpretations with their social attribution. *arXiv preprint arXiv:2006.01067*. doi: 10.48550/arXiv.2006.01067

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012a). Fsl. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. doi: 10.48550/arXiv.1412.6980

Kipf, T. N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*. doi: 10.48550/arXiv.1609.02907

Kunda, M., Zhou, S., Gong, G., and Lu, H. (2020). Improving multi-site autism classification based on site-dependence minimisation and second-order functional connectivity. *IEEE Trans. Med. Imag.* 42, 55–65. doi: 10.1109/TMI.2022.3203899

Lee, S., Wang, X., Han, S., Yi, X., Xie, X., and Cha, M. (2022). Self-explaining deep models with logic rule reasoning. *Adv. Neural Inform. Process. Syst.* 35, 3203–3216. doi: 10.48550/arXiv.2210.07024

Linsley, D., Shiebler, D., Eberhardt, S., and Serre, T. (2018). Learning what and where to attend. *arXiv preprint arXiv:1805.08819*. doi: 10.48550/arXiv.1805.08819

Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., et al. (2020). Parameterized explainer for graph neural network. *arXiv preprint arXiv:2011.04573*. doi: 10.48550/arXiv.2011.04573

Matsunaga, D., Suzumura, T., and Takahashi, T. (2019). Exploring graph neural networks for stock market predictions with rolling window analysis. *arXiv preprint arXiv:1909.10660*. doi: 10.48550/arXiv.1909.10660

Mitsuhara, M., Fukui, H., Sakashita, Y., Ogata, T., Hirakawa, T., Yamashita, T., et al. (2019). Embedding human knowledge into deep neural network via attention map. *arXiv preprint arXiv:1905.03540*. doi: 10.48550/arXiv.1905.03540

of Heidelberg Department of Neuroradiology, D. A. H. M. U. (2014). *Fsl-Scripts*. Available online at: https://github.com/ahheckel/FSL-scripts (accessed January 10, 2024).

Patro, B., Namboodiri, V., and Anupriy, A. (2020). "Explanation vs. attention: a two-player game to obtain attention for VQA," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34* (Palo Alto, CA: AAAI Press), 11848–11855.

Pope, P. E., Kolouri, S., Rostami, M., Martin, C. E., and Hoffmann, H. (2019). "Explainability methods for graph convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New York, NY: ComputerVisionFoundation/IEEE), 10772–10781.

Qiao, T., Dong, J., and Xu, D. (2018). "Exploring human-like attention supervision in visual question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32* (Palo Alto, CA: AAAI Press).

Ross, A. S., Hughes, M. C., and Doshi-Velez, F. (2017). Right for the right reasons: training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*. doi: 10.48550/arXiv.1703.03717

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE Trans. Neural Netw.* 20, 61–80. doi: 10.1109/TNN.2008.2005605

Schlichtkrull, M. S., De Cao, N., and Titov, I. (2020). Interpreting graph neural networks for NLP with differentiable edge masking. *arXiv preprint arXiv:2010.00577*. doi: 10.48550/arXiv.2010.00577

Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schütt, K. T., Müller, K. R., et al. (2020). *Higher-Order Explanations of Graph Neural Networks via Relevant Walks*. New York, NY: IEEE.

Sha, L., Camburu, O. M., and Lukasiewicz, T. (2023). Rationalizing predictions by adversarial information calibration. *Artif. Intell.* 315:103828. doi: 10.1016/j.artint.2022.103828

Shi, Y., Zhou, K., and Liu, N. (2023). "Engage: explanation guided data augmentation for graph representation learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Berlin: Springer), 104–121.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., et al. (2013). The wu-minn human connectome project: an overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041

Visotsky, R., Atzmon, Y., and Chechik, G. (2019). Few-shot learning with per-sample rich supervision. *arXiv preprint arXiv:1906.03859*. doi: 10.48550/arXiv.1906.03859

Vu, M. N., and Thai, M. T. (2020). PGM-explainer: probabilistic graphical model explanations for graph neural networks. *arXiv preprint arXiv:2010.05788*. doi: 10.48550/arXiv.2010.05788

Weinberger, E., Janizek, J., and Lee, S.-I. (2020). Learning deep attribution priors based on prior knowledge. *Adv. Neural Inform. Process. Syst.* 33, 14034–14045. doi: 10.48550/arXiv.1912.10065

Wu, L., Cui, P., Pei, J., and Zhao, L. (2021). *Graph Neural Networks: Foundations, Frontiers, and Applications*. Singapore: Springer.

WU-Minn (2017). *1200 Subjects Data Release Reference Manual*. H. C. P. Available online at: https://www.humanconnectome.org (accessed March 3, 2024).

Ying, R., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019). GNNExplainer: generating explanations for graph neural networks. *Adv. Neural Inform. Process. Syst.* 32:9240. doi: 10.48550/arXiv.1903.03894

Yuan, H., Tang, J., Hu, X., and Ji, S. (2020a). "XGNN: towards model-level explanations of graph neural networks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY: ACM), 430–438.

Yuan, H., Yu, H., Gui, S., and Ji, S. (2020b). Explainability in graph neural networks: a taxonomic survey. *arXiv preprint arXiv:2012.15445*. doi: 10.48550/arXiv.2012.15445

Zhang, M., Cui, Z., Neumann, M., and Chen, Y. (2018). "An end-to-end deep learning architecture for graph classification," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32* (Palo Alto, CA: AAAI Press).

Zhang, Y., Defazio, D., and Ramesh, A. (2020). RelEx: a model-agnostic relational model explainer. *arXiv preprint arXiv:2006.00305*. doi: 10.48550/arXiv.2006.00305

Zhang, Y., Niebles, J. C., and Soto, A. (2019). "Interpretable visual question answering by visual grounding from attention supervision mining," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (Waikoloa, HI: IEEE), 349–357.