Make Model Transparent: Brain Network Analysis via Causal and Knowledge Graph Learning

Lingyuan Meng[1], Ke Liang[1†], Hao Yu[1], Haotian Wang[1], Miaomiao Li[2†], Xinwang Liu[1]

[1]The College of Computer Science and Technology, National University of Defense Technology, Changsha, 410073, China.

[2]The School of Electronic Information and Electrical Engineering, Changsha College, Changsha, 410022, China.

Abstract

Brain network analysis technology reveals the organizational mechanism and information processing mode by constructing the structural connection network between brain regions. It has achieved satisfactory results in brain disease prediction tasks, promoting the progress of neuroscience. In recent years, graph transformer has become the most mainstream method for brain analysis with its powerful feature extraction ability and attention mechanism. However, these methods face two challenges, i.e., lack of interpretability, and neglect of semantic associations among brain regions. To solve these problems, we proposed a large language model (LLM)-driven causal knowledge brain network transformer framework, termed BrainCKT, which is plug-and-play, and can adapt to most of the existing mainstream graph transformer-based methods. Specifically, we constructed a brain region causal graph and used its adjacency matrix to guide the learning process of the self-attention mechanism. In addition, we constructed a brain science knowledge graph and encoded it through a pre-trained model to enhance the original brain region features. Finally, we integrated BrainCKT into four mainstream graph transformer baselines for verification. Experimental results on two brain imaging datasets proved the effectiveness of BrainCKT.

Introduction

The brain is the most vital organ of the human body, which serves as the central hub for orchestrating life activities (Genon et al. 2018). Therefore, analyzing the function of the brain has become an important goal of human development. In practice, fully analyzing the brain can help people effectively predict neurological diseases and human physical signs (Luo et al. 2024; Guan et al. 2025; Liu et al. 2024). For example, Figure 1 illustrates that fluctuations in the blood oxygen level-dependent (BOLD) signal of the hippocampus can reveal the progression of Alzheimer's disease. Although modern medicine has anatomically divided the brain into various regions, each of which has been extensively studied, most existing approaches only focus on a few regions with a high technique cost (Su et al. 2021; Hu et al. 2024; Feng et al. 2025).

As a relational data structure, graphs are able to model complex relationships in the human world. In recent years, modeling the brain as a graph structure and leveraging deep learning techniques to extract neural signal features has achieved promising results in various brain analysis tasks (Mohammadi and Karwowski 2024). In such models, nodes represent brain regions of interest (ROIs), and edges denote the functional connectivity strength between them. Such graph-based modeling strategy is known as brain network analysis, which has garnered significant attention.

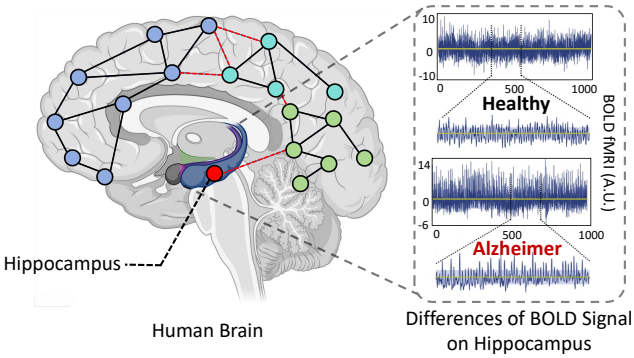


Figure 1: Illustration of the significance and application value of brain network analysis. Specifically, there are significant differences in blood level signals in the hippocampus region between healthy people and patients with Alzheimer's disease.

The emergence of transformer models in the field of graph representation learning has led to remarkable progress in brain network analysis tasks. For instance, BrainNETTF (Kan et al. 2022b) utilized efficient attention weights to learns pairwise connection strengths. Furthermore, BrainNETTF proposed an Orthonormal Clustering Readout operation, generating cluster-aware and informative graph embeddings. ALTER (Yu et al. 2024) solves the long-distance path dependency in brain networks and proposes a long-distance-aware brain network transformer framework. However, these methods still face two critical challenges: (1) **Neglect of semantic associations among ROIs**, and (2) **Lack of interpretability**. Specifically, for challenge (1), current methods typically rely solely on BOLD signals and constructed graph structures for analysis, emphasizing statistical connectivity while failing to integrate the functional

roles of brain regions with disease-related semantics. For example, although the prefrontal cortex and hippocampus are clinically recognized as highly associated with the diagnosis of Alzheimer's disease, early-stage functional impairments in the prefrontal cortex may not manifest as significant changes in cerebral blood flow, resulting in low statistical correlation based on fMRI-derived BOLD signals. For challenge (2), the black-box nature of transformers makes it difficult to align model predictions with biological knowledge or known cognitive functions of brain regions. These limitations hinder the further advancement of transformer-based models in brain network analysis.

To address the aforementioned challenges, we propose a large language model (LLM) driven causal knowledge brain Transformer framework, termed BrainCKT. Specifically, to solve challenge (1), we construct a neuroscience knowledge graph based on existing literature and open-access medical databases to enrich the semantic context of brain network analysis. To solve challenge (2), we use the relations between brain region entities in the knowledge graph to construct a brain causal graph based on the structural equation model (SEM) to guide the masked attention learning process in the Transformer framework, thereby improving the interpretability of the model. To our best knowledge, we are the first to utilize causal knowledge for interpretable research in brain network analysis. It is worth noting that our proposed method is plug-and-play and interpretable, which can be integrated seamlessly into most of existing graph transformer-based brain network analysis model and provide effective evidence for brain network analysis. To demonstrate the effectiveness and robustness of the BrainCKT, we conduct comprehensive experiments on two brain network analysis datasets. In summary, our key contributions are as follows:

- We propose a plug-and-play LLM-driven causal knowledge transformer framework for brain network analysis, providing sufficient semantic information and interpretability for existing transformer-based approaches.
- We propose a causal knowledge-driven brain augmentation mechanism. Specifically, we construct a neuroscience knowledge graph that supplies comprehensive semantic relationships to brain network analysis.
- We integrate BrainCKT into existing Transformer-based brain network models and conduct extensive experiments, which demonstrate the superiority and robustness of our method.

Related Work

Graph Neural Network-based Models Graph neural network (GNN)-based brain network analysis methods aim to leverage the message-passing capabilities of GNNs to learn global representations of brain regions. Within this framework, brain network analysis is formulated as a graph-level classification task. For example, BrainGNN (Li et al. 2021) introduces novel region-of-interest (ROI)-aware graph convolutional layers that achieve interpretable disease prediction. BN-GNN (Zhao et al. 2022) employs a deep reinforcement learning framework to search for the optimal GNN architecture tailored to each individual brain network, enabling personalized brain network analysis. IBGNN (Cui et al. 2022b) presents an interpretable framework for analyzing disease-specific regions of interest and significant connections. Overall, these methods have achieved promising results in brain network analysis tasks. However, the inherent over-smoothing issue in GNNs poses challenges in capturing such long-range dependencies, thereby limiting the further advancement of these methods.

Graph Transformer-based Models Compared to traditional Graph Neural Network (GNN) architectures (Meng et al. 2023), the Graph Transformer represents a more advanced graph learning paradigm with enhanced expressive power. For instance, BrainNETTF (Kan et al. 2022b) utilizes connectivity profiles as node features, providing natural and low-cost positional information for brain graphs. Furthermore, BrainNETTF learns pairwise connection strengths between regions of interest (ROIs) and assigns effective attention weights across individuals. GBT (Peng et al. 2024) presents a simple yet effective geometry-guided brain Transformer that approximates the learning of the most relevant and representative graph representations using an attention-weighted matrix module. ALTER (Yu et al. 2024) proposes an innovative long-range perception strategy capable of explicitly capturing long-distance dependencies between brain ROIs. However, existing transformer-based methods tokenize graphs as input, resulting in the loss of structural information. Moreover, the inherent black-box nature of transformer models limits interpretability, posing a key challenge for brain network analysis.

Interpretable Brain Network Analysis Models The interpretability of predictions from brain network analysis methods is crucial for identifying and localizing biomarkers that cause diseases. Currently, a variety of interpretable brain network analysis methods have been proposed, including BrainGNN (Li et al. 2021), IBGNN (Cui et al. 2022b), and BPI-GNN (Zheng et al. 2024). However, these methods are all designed based on the GNN architecture, and the interpretability of Transformer-based brain network analysis methods has not been fully studied, leaving us a gap to fill. In this paper, we attempt to provide sufficient traceable guidance for the model's predictions by constructing a brain causal graph and a knowledge graph. To the best of our knowledge, we are the first to propose an interpretable Transformer-based brain network analysis method.

Methodology

The proposed BrainCKT consists of 3 main components, including knowledge-driven brain augmentation, SEM-based causal graph construction, and causal-aware transformer. First, we constructed a brain knowledge graph based on open-source databases and LLMs, providing sufficient semantic information for raw brain features. After that, we constructed a simple brain causal graph based on a structural equation model (SEM). Finally, we used the adjacency matrix of the causal graph to guide transformer attention learning, providing sufficient interpretability for the model.

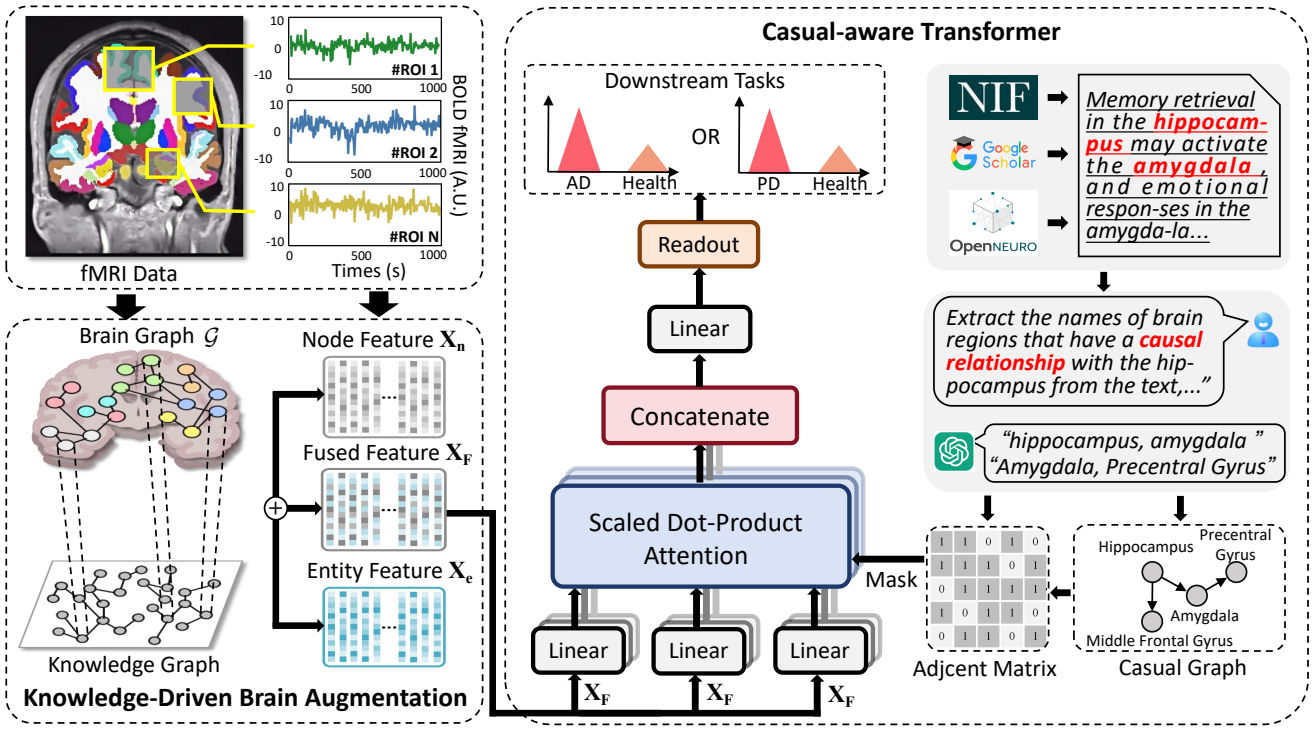


Figure 2: Illustration of the framework of the proposed BrainCKT. Our model consists of three module, *i.e.*, knowledge-driven brain augmentation module, SEM-based causal graph construction module, and causal aware transformer.

Preliminaries

In this paper, we proposed a novel plug-and-play framework for transformer-based brain network analysis task, with the goal of effectively identifying neurological disorders. The problem is formulated as a graph-level classification task. Let $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_i\}_{i=1}^n$ represent a collection of fMRI time-series data, where each $\mathbf{X}_i \in \mathbb{R}^{T \times O}$. Here, T indicates the number ofobservationservations, and O refers to the count of brain voxels. The associated set of ground truth labels is denoted by $\mathbf{Y} = \{y_1, y_2, ..., y_n\}$. For each matrix $\mathbf{X}_i$, a row vector $\mathbf{x}_t \in \mathbb{R}^V$ captures the voxel-level fMRI signals at the specific time point t. To construct the brain graph, a standard brain atlas is employed, resulting in a graph $G = (V, E)$, where the node set $V = \{v_1, v_2, ..., v_N\}$ consists of N brain regions. The edges $E \subseteq V \times V$ reflect inter-regional relationships that may be derived from either anatomical structures or functional connectivity. Each edge is assigned a weight w_{ij}, which quantifies the interaction strength between regions i and j, typically estimated using statistical correlations or structural metrics.

Knowledge-Driven Brain Augmentation

Knowledge-driven brain augmentation module aims to provide features containing rich semantic information for brain region nodes. Specifically, we first construct a brain knowledge graph over LLMs based on open sourced neuralscience database and related papers. Subsequently, the constructed brain knowledge graph is encoded using a pre-trained knowledge graph embedding (KGE) model to obtain knowledge representation. Finally, the features of the brain region are fused with the original physiological signal features.

Brain Knowledge Graph Construction To provide richer semantic information for brain network analysis, we constructed a human brain knowledge graph that integrates entities and relations relevant to brain disease prediction from the fields of neuroscience and clinical medicine. The resulting knowledge graph comprises 237 entities and 23 relations, amounting to a total of 5056 triples. Figure 3 illustrates the construction process of the human brain knowledge graph. Specifically, we first collected textual descriptions of brain regions, neural functions, and brain-related diseases from existing literature databases such as Google Scholar and PubMed, as well as from open-access neuroscience and clinical medical databases, including BrainBase and the Disease Ontology. Subsequently, we employed a large language model (Roumeliotis and Tselikas 2023) to extract triples from these texts. The prompt of triplet extraction is shown in Appendix.

Finally, we remove similar entities by calculating the semantic similarity matrix between entities over a similarity function $s(\cdot)$:

$$\mathbf{S} = \sum_{i=1}^{N} \sum_{j=1}^{N} s(e_i, e_j), i \neq j, \tag{1}$$

where $\mathbf{S}$ denotes the entity semantic similarity matrix, e_i and e_j represent entities in the knowledge graph. Due to

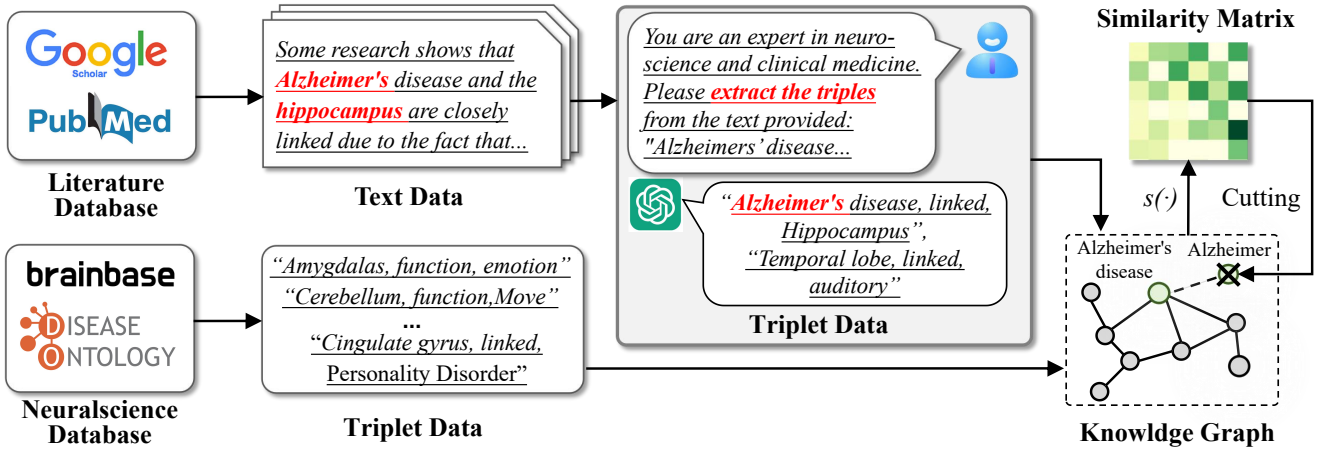


Figure 3: Illustration of the brain knowledge graph construction process.

the differences in data types and features in literature and databases, the definition of relations in the knowledge graph is manual. We define a total of 23 relations, including direct relations, indirect relations, affiliation relations, and synergy relations. These relations mainly describe the association between brain regions and neural functions, and the relationship between neural diseases and brain regions.

Knowledge Encoding and Fusion To further exploit the semantic relations in the knowledge graph, we use a pre-trained knowledge graph embedding (KGE) model to encode entities and relations into a vector space as follows:

$$\mathbf{X_E}, \mathbf{X_R} = \mathrm{KGE}(\mathcal{G}), \tag{2}$$

where $\mathbf{X_E} \in \mathbb{R}^{N \times d}$ and $\mathbf{X_R} \in \mathbb{R}^{m \times d}$ represent the learned embeddings of entities and relations, respectively. Here, $\mathcal{G}$ stands for the constructed knowledge graph, N is the total number of entities involved, and m denotes the number of distinct types of relations.

After that, we retrieve features corresponding to brain region entities from the knowledge graph feature matrix. These features are then fused with the original node features $\mathbf{X_n}$ obtained from the brain network data. The final feature $\mathbf{X_F}$ is obtained through the concatenation operation:

$$\mathbf{X_F} = \mathbf{X_n} \oplus \mathbf{X_e}, \tag{3}$$

where $\mathbf{X_e}$ represents the embedding vector corresponding to a specific brain region entity and $\oplus$ represents the concatenation operator applied along the feature dimension. Note that our approach is robust to the choice of fusion technique. In this paper, we choose the concatenation operation for feature fusion due to its simplicity and effectiveness.

SEM-based Causal Graph Construction

The objective of the causal graph construction module is to generate a sparse directed acyclic graph (DAG) $G = (V, E)$ based on brain region features. In this graph, each node $v_i \in V$ represents a region of interest (ROI), and each directed edge $e_{ij} \in E$ indicates a presumed causal influence from region i to region j. The graph is constructed entirely offline and remains fixed during the model training phase. To discover causal structure, we adopt an offline algorithm based on linear structural equation modeling (SEM), aiming to estimate an interpretable and sparse causal network from observational data.

We begin by evaluating potential causal influence between each pair of brain regions. Specifically, for each candidate edge e_{ij}, a local linear SEM is fitted as

$$X_j = \beta_{ij} X_i + \eta_j, \tag{4}$$

where η_j denotes a Gaussian noise term. Statistical tests (e.g., based on partial correlation analysis) are used to assess the significance of β_{ij}, resulting in a set of directed edges with associated scores or p-values. These form a pool of candidate directed connections across all region pairs.

Next, we select a subset of high-confidence edges from this candidate pool to construct the final graph. Acyclicity principle is strictly enforced throughout the edge selection process. Edges are added incrementally according to their scores, and the process terminates once no further statistically significant edges can be incorporated without violating the DAG constraint. The generated causal graph $\hat{G}_{\text{SEM}}$ is a sparse DAG that captures plausible directional relationships among brain regions.

Finally, given the fixed structure $\hat{G}_{\text{SEM}}$, we perform global SEM parameter estimation via multivariate linear regression for each node. For node j, the model takes the form:

$$X_j = \sum_{i \in \mathrm{Pa}(j)} B_{ij} X_i + \varepsilon_j, \tag{5}$$

where $\mathrm{Pa}(j)$ denotes the set of parent nodes of j, and B_{ij} are the regression coefficients. The full coefficient matrix B is obtained by fitting these models in parallel. To produce the final unweighted graph, we binarize the entries of B using a threshold δ: if $|B_{ij}| < \delta$, we set $A_{ij} = 0$; otherwise, $A_{ij} = 1$. The resulting adjacency matrix A encodes a binary DAG that serves as the final causal graph used in our framework.

Model Type	Methods	Dataset: ABIDE				Dataset: OSF			
		ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE
Learnable Network	BrainGNN	59.42	62.13	47.72	70.71	45.22	51.79	55.10	44.35
	BrainGB	63.64	69.71	64.05	61.31	46.70	49.38	52.71	49.32
	BrainNetCNN	67.80	74.98	64.32	71.04	46.70	49.38	52.71	49.32
	FBNETGNN	68.01	75.67	65.71	62.94	46.70	49.38	52.71	49.32
Graph Transformer	Graphormer	63.57	60.84	78.72	36.72	55.00	58.88	43.33	52.51
	TC-BrainTF	69.70	77.71	69.12	70.15	61.11	62.08	45.34	67.85
	BrainNETTF	71.00	80.20	72.50	69.30	60.00	62.59	66.67	55.56
	BrainNETTF+ours	**75.00**	**81.25**	**76.60**	**73.59**	**62.21**	**68.52**	**44.44**	**83.33**
	Com-BrainTF	72.50	79.60	80.10	65.70	63.33	75.00	71.80	68.13
	Com-BrainTF+ours	**78.00**	**83.11**	**81.81**	**73.33**	**62.79**	**76.55**	**75.32**	**67.78**
	GBT	72.00	81.00	88.00	56.00	60.71	77.80	58.86	73.33
	GBT+ours	**73.16**	**81.47**	**73.47**	**60.78**	**66.67**	**80.00**	**62.50**	**75.00**
	ALTER	82.80	77.00	77.40	76.60	66.67	74.55	81.80	55.15
	ALTER+ours	**83.87**	**79.00**	**78.51**	**79.31**	**73.33**	**81.48**	**62.50**	**60.00**

Table 1: Performance comparison of our proposed model with latest unsupervised, supervised and semi-supervised baseline models on three brain network analysis datasets.

Causal-aware Brain Transformer

We use the constructed causal graph to guide the learning process of the graph transformer. Specifically, we integrate $\mathbf{A} \in \{0,1\}^{N\times N}$ into a graph transformer by using it as a hard mask in the self-attention layers. Let $\mathbf{X} \in \mathbb{R}^{N\times d}$ be the node feature matrix at some layer. For each of H heads we form queries, keys, and values by learned projections:

$$Q^{(h)} = \mathbf{X}\mathbf{W}_{\mathbf{Q}}^{(h)}, \tag{6}$$

$$K^{(h)} = \mathbf{X}\mathbf{W}_{\mathbf{K}}^{(h)}, \tag{7}$$

$$V^{(h)} = \mathbf{X}\mathbf{W}_{\mathbf{V}}^{(h)}, \tag{8}$$

where $\mathbf{W_Q}, \mathbf{W_K}, \mathbf{W_V} \in \mathbb{R}^{d\times d_k}$, $d_k = d/H$. In standard multi-head attention this is followed by scaled dot-product and softmax. Here we modify it by adding the adjacency-based mask $\mathbf{M}$. Formally, we define a mask matrix:

$$\mathbf{M}_{ij} = \begin{cases} 0, & \text{if } \mathbf{A}_{ij} = 1, \\ -\infty, & \text{if } \mathbf{A}_{ij} = 0. \end{cases} \tag{9}$$

Then each head computes attention weights only over neighbors. For head h, compute the unnormalized logits

$$L^{(h)} = \frac{Q^{(h)} {K^{(h)}}^{\top}}{\sqrt{d_k}} + \mathbf{M}, \tag{10}$$

where $\mathbf{M}_{ij}$ is as above. Applying the softmax yields

$$\alpha_{ij}^{(h)} = \frac{\exp\big(L_{ij}^{(h)}\big)}{\sum_{j'} \exp\big(L_{ij'}^{(h)}\big)}. \tag{11}$$

Since $\mathbf{M}_{ij} = -\infty$ for $\mathbf{A}_{ij} = 0$, we have $\alpha_{ij}^{(h)} = 0$ whenever $e_{ij} \notin E$, i.e., each node attends only to its causally connected neighbors. The output of head h is $\mathbf{Z}^{(h)} = \alpha^{(h)} V^{(h)}$, a weighted sum of value vectors over neighbors. Finally, all heads are concatenated and linearly transformed:

$$\mathrm{MHA}(\mathbf{X}) = [\mathbf{Z}^{(1)}; \ldots; \mathbf{Z}^{(H)}]\, \mathbf{W}^{O}. \tag{12}$$

Finally, we use an efficient readout function to derive the whole brain map representation, and also train an additional classifier for downstream tasks:

$$Y = Softmax(MLP(Readout(\mathbf{Z_G}))). \tag{13}$$

It should be noted that our method can be applied to a variety of readout functions. In this paper, we adopt clustering-based pooling as the readout function.

Experiments

In this section, we first introduce the experimental setup, including the dataset, compared baselines, and implementation details. After that, we demonstrate the effectiveness of the proposed BrainCKT through sufficient experiments, including superiority analysis, ablation study, and interpretability analysis.

Experiment Settings

Used Datasets In this paper, we employ two brain network analysis datasets of varying sizes to assess the performance of our proposed model: the Autism Brain Imaging Data Exchange (ABIDE) dataset (Craddock et al. 2013) and the Theory of Mind task dataset available on the Open Science Framework (OSF) (Richardson et al. 2018).

Compared Baselines In this paper, we select two types of models as comparison baselines, including learnable networks and graph transformer models. Specifically, learnable networks include 4 comparison baselines, including BrainGNN (Li et al. 2021), BrainGB (Cui et al. 2022a), BrainNetCNN (Kawahara et al. 2017), and FBNETGNN (Kan et al. 2022a). Graph transformer models include

Models	Components		Dataset: ABIDE				Dataset: OSF			
	w. KG	w. Causal	ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE
ALTER+ours			82.8	77.00	77.4	76.6	66.67	74.55	81.8	55.15
	✓		83.15	77.63	78.36	78.42	71.21	80.84	82.29	53.37
		✓	83.33	77.82	78.05	77.96	72.14	79.36	60.75	58.88
	✓	✓	83.87	79	78.51	79.31	73.33	81.48	62.5	60
BrainNETTF+ours			71	80.2	72.5	69.3	60	62.59	66.67	55.56
	✓		73.52	81.17	75.17	71.25	61.87	66.54	65.32	71.18
		✓	74.15	80.36	74.96	72.33	61.85	65.32	67.87	59.6
	✓	✓	75	81.25	76.6	73.59	62.21	68.52	64.44	73.33

Table 2: Ablation study of two main components of BrainCKT, including knowledge graph enhancement mechanism and the causal-aware attention learning module.

Dataset	Method	Metrics			
		ACC	AUC	SEN	SPE
ABIDE	BrainCKT w. TransE	80.45	75.67	72.81	74.04
	BrainCKT w. CompGCN	**83.87**	**79.00**	**78.51**	**79.31**
	BrainCKT w. RGCN	83.30	77.85	77.21	78.14
OSF	BrainCKT w. TransE	70.81	76.38	59.32	56.15
	BrainCKT w. CompGCN	73.33	**81.48**	**62.50**	60.00
	BrainCKT w. RGCN	**73.91**	81.02	60.50	**61.33**

Table 3: Ablation study on adaptability of the KGE models.

Dataset	Method	Metrics			
		ACC	AUC	SEN	SPE
ABIDE	BrainCKT w. Prompt1	**83.87**	79.00	**78.51**	**79.31**
	BrainCKT w. Prompt2	83.45	77.91	78.32	78.85
	BrainCKT w. Prompt3	**83.87**	**79.41**	75.38	79.01
OSF	BrainCKT w. Prompt1	73.33	**81.48**	62.50	60.00
	BrainCKT w. Prompt2	71.20	80.83	60.00	59.33
	BrainCKT w. Prompt3	**73.67**	**81.48**	**63.17**	**61.30**

Table 4: Ablation study on adaptability of the LLM prompts.

6 baseline models, Graphormer (Ying et al. 2021), TC-BrainTF (Yang et al. 2024), BrainNETTF (Kan et al. 2022b), ComBrainTF (Bannadabhavi et al. 2023), GBT (Peng et al. 2024), and ALTER (Yu et al. 2024). Details of all comparison baseline models are introduced in the Appendix.

Implementation Details All experiments for the proposed model were implemented using the PyTorch framework on a single NVIDIA GeForce RTX 3090Ti GPU. For brain region segmentation, the Craddock 200 atlas (comprising 200 regions) was applied to the ABIDE dataset, while the AAL atlas (116 regions) was used for the OSF datasets. A pre-trained CompGCN served as the knowledge graph encoder. For hyper-parameters in our model, we utilizes 4 attention heads. Across all datasets, we partition the training, validation, and test sets in a 70:10:20 ratio. During training, the Adam optimizer is used in conjunction with the Cosine Learning Rate (CosLR) scheduler, both initialized with a learning rate and weight decay of 1e-4. The training process uses a batch size of 16 over 200 epochs.

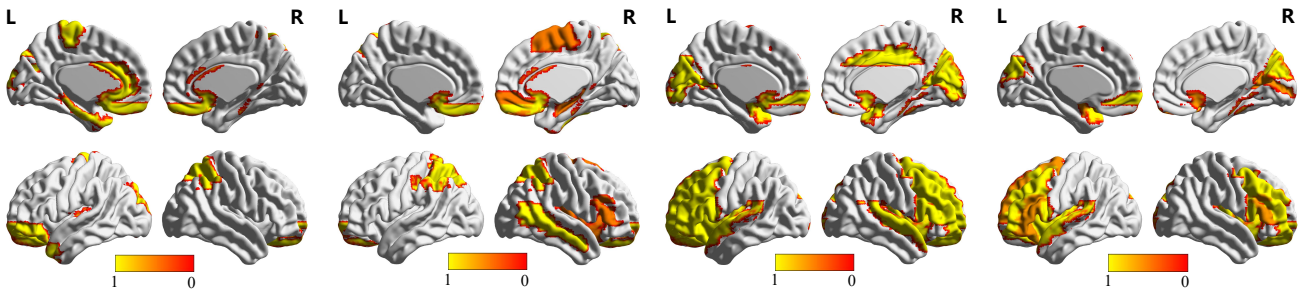
Superiority Analysis

To demonstrate the superiority of BrainCKT, we integrated BrainCKT into several state-of-the-art baselines and conducted extensive comparative experiments on two public brain imaging datasets. We divided the baselines into two categories according to their technical types, including learnable networks and Graph Transformer methods. The results shown in Table 1 show that BrainCKT has achieved satisfactory performance improvements. Specifically, after integrating BrainCKT into BrainNETTF, ComBrainTF, GBT, and ALTER, the ACC on the ABIDE dataset increased by 5.63%, 7.59%, 1.61%, and 1.29%, respectively. On the OSF dataset, the AUC of the model after integrating BrainCT increased by 9.46%, 2.07%, 2.82%, and 9.30%, respectively. Compared with the Learnable Network method, the Graph Transformer-based method achieved better performance. Experimental results show that our proposed BrainCKT is able to facilitate various downstream tasks of brain network analysis.

Ablation Study

To demonstrate the effectiveness of the BrainCKT component, we conducted extensive ablation experiments. Firstly, we integrated BrainCKT into ALTER and BrainNETTF, and explored the impact of the knowledge augmentation module and the causal augmentation module on model performance through four experimental settings. After that, we utilize different KGE models to pretrain the constructed KG. Finally, various LLM prompts are explored to show the robustness of the triplet extraction process.

Effectiveness of Causal and Knowledge Modules We integrated BrainCKT into ALTER and BrainNETTF, and explored the impact of the knowledge augmentation module and the causal augmentation module on model performance through four experimental settings. w.KG and w.causal indicates that the model is augmented by the knowledge graph and the causal graph, respectively. Table 2 shows the effectiveness of the above two modules. The experimental results show that both the knowledge graph and the causal graph bring performance improvement. For example, the knowledge augmentation module improves the AUC of the AL-



(a). OSF, Middle-aged (b).OSF,Old (c).ABIDE, ASD HC (d). ABIDE, ASD Patient

Figure 4: Interpreting the relevant significant ROIs of different groups of individuals in the two datasets. The color bar ranges from 0 to 1.0. Bright yellow indicates a high score and dark red indicates a low score.

TER on the ABIDE and OSF datasets by 0.63% and 8.44%, respectively. The causal augmentation mechanism improves the AUC of the ALTER on the ABIDE and OSF datasets by 0.82% and 6.45%, respectively. For BrainNETTF, the knowledge augmentation module improves the AUC of the two datasets by 1.21% and 6.31%, respectively.

Influences of Different KGE Models In order to explore the impact of different pretrained KGE models on the performance of BrainCKT, we conducted experiments on two datasets using different KGE models based on the ALTER model. Specifically, we selected TransE (Wang et al. 2014), CompGCN (Vashishth et al. 2019) and RGCN (Schlichtkrull et al. 2018) as knowledge graph encoders. The experimental results in Table 3 show that compared with the traditional KGE model TransE, the graph neural network-based methods CompGCN and RGCN achieved better performance on both datasets, demonstrating the importance of structural information for knowledge graph encoding. In addition, CompGCN and RGCN achieved comparable performance, demonstrating the robustness of BrainCKT to the selection of KG encoders.

Influences of Different LLM Prompts In our proposed BrainCKT, the construction process of the knowledge graph relies on a large language model and prompts. Therefore, we designed three different triple extraction prompts to explore their impact on model performance (prompt1-prompt3). Specifically, the three prompts we designed have similar semantics and consistent input and output formats. The difference lies in the question pattern. Due to space limitations, the detailed information of the three prompts is shown in the appendix. The experimental results in Table 4 show that the model performance is the same after using the three prompts to build the knowledge graph. It should be noted that the performance of the method using prompt2 is reduced to a small extent. This is because prompt2 defines stricter rules, which makes the triples higher quality, but also actively sacrifices some recall rate. In general, the experimental results confirm the adaptability of the proposed method to different prompts.

Interpretability Analysis

In this section, we discuss the interpretability of our method through visualization analysis. In BrainCKT, the adjacency matrix of the causal graph is designed to guide the training process of self-attention mechanism. Therefore, we map the attention weight matrix into the brain voxel space and visualize the activation states of brain regions through brain analysis tools. Figure 4 shows the activation states of our method under different labels on two datasets. Specifically, for the ABIDE dataset, we select ASD patients (Figure 4.(d)) and healthy subjects (Figure 4.(c)) as samples, while for the OSF dataset, we select middle-aged (Figure 4.(a)) and elderly (Figure 4.(b)) as control subjects. The visualization results show that BrainCKT can accurately locate key brain regions related to tasks. For example, compared with healthy samples (Figure 4.(c)), the BOLD signals of the prefrontal cortex, amygdala and superior temporal sulcus of ASD patients (Figure 4.(d)) are weakened. These brain regions are highly related to social interaction and cognitive control. The results of the visualization analysis show that our method provides interpretable evidence for the results while improving the performance of brain network analysis.

Conclusion

In this paper, we proposed a novel graph transformer-based brain network analysis model, which is plug-and-play and provides additional semantic knowledge for brain network analysis by constructing a knowledge graph. Furthermore, we construct a brain causal graph and guide the learning of the transformer attention mechanism through its adjacency matrix. Finally, we integrated BrainCKT into four mainstream graph transformer baselines for verification. Experimental results on two brain imaging datasets demonstrated that BrainCKT provides an interpretable solution for various downstream tasks of brain network analysis.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (no. 62325604, 62441618, 62276271, 62376039, 62506371, XJJC2025013).

References

Bannadabhavi, A.; Lee, S.; Deng, W.; Ying, R.; and Li, X. 2023. Community-aware transformer for autism prediction

in fmri connectome. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 287–297. Springer.

Craddock, C.; Benhajali, Y.; Chu, C.; Chouinard, F.; Evans, A.; Jakab, A.; Khundrakpam, B. S.; Lewis, J. D.; Li, Q.; Milham, M.; et al. 2013. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, 7(27): 5.

Cui, H.; Dai, W.; Zhu, Y.; Kan, X.; Gu, A. A. C.; Lukemire, J.; Zhan, L.; He, L.; Guo, Y.; and Yang, C. 2022a. Braingb: a benchmark for brain network analysis with graph neural networks. *IEEE transactions on medical imaging*, 42(2): 493–506.

Cui, H.; Dai, W.; Zhu, Y.; Li, X.; He, L.; and Yang, C. 2022b. Interpretable Graph Neural Networks for Connectome-Based Brain Disorder Analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 375–385. Springer.

Feng, Y.; Liang, W.; Wan, X.; Liu, J.; Liu, S.; Qu, Q.; Guan, R.; Xu, H.; and Liu, X. 2025. Incremental Nyström-based Multiple Kernel Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16613–16621.

Genon, S.; Reid, A.; Langner, R.; Amunts, K.; and Eickhoff, S. B. 2018. How to characterize the function of a brain region. *Trends in cognitive sciences*, 22(4): 350–364.

Guan, R.; Tu, W.; Wang, S.; Liu, J.; Hu, D.; Tang, C.; Feng, Y.; Li, J.; Xiao, B.; and Liu, X. 2025. Structure-Adaptive Multi-View Graph Clustering for Remote Sensing Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16933–16941.

Hu, D.; Liu, S.; Wang, J.; Zhang, J.; Wang, S.; Hu, X.; Zhu, X.; Tang, C.; and Liu, X. 2024. Reliable Attribute-missing Multi-view Clustering with Instance-level and feature-level Cooperative Imputation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1456–1466.

Kan, X.; Cui, H.; Lukemire, J.; Guo, Y.; and Yang, C. 2022a. Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation. In *International conference on medical imaging with deep learning*, 618–637. PMLR.

Kan, X.; Dai, W.; Cui, H.; Zhang, Z.; Guo, Y.; and Yang, C. 2022b. Brain network transformer. *Advances in Neural Information Processing Systems*, 35: 25586–25599.

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

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

Liu, M.; Liu, Y.; Liang, K.; Tu, W.; Wang, S.; Zhou, S.; and Liu, X. 2024. Deep Temporal Graph Clustering. In *The 12th International Conference on Learning Representations*.

Luo, X.; Wu, J.; Yang, J.; Xue, S.; Beheshti, A.; Sheng, Q. Z.; McAlpine, D.; Sowman, P.; Giral, A.; and Yu, P. S. 2024. Graph neural networks for brain graph learning: A survey. *arXiv preprint arXiv:2406.02594*.

Meng, X.; Wei, W.; Liu, Q.; Wu, S.; and Wang, L. 2023. TiBGL: Template-induced Brain Graph Learning for Functional Neuroimaging Analysis. *arXiv preprint arXiv:2309.07947*.

Mohammadi, H.; and Karwowski, W. 2024. Graph neural networks in brain connectivity studies: Methods, challenges, and future directions. *Brain Sciences*, 15(1): 17.

Peng, Z.; He, Z.; Jiang, Y.; Wang, P.; and Yuan, Y. 2024. GBT: Geometric-oriented Brain Transformer for Autism Diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 142–152. Springer.

Richardson, H.; Lisandrelli, G.; Riobueno-Naylor, A.; and Saxe, R. 2018. Development of the social brain from age three to twelve years. *Nature communications*, 9(1): 1027.

Roumeliotis, K. I.; and Tselikas, N. D. 2023. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6): 192.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, 593–607. Springer.

Su, T.; Gong, J.; Tang, G.; Qiu, S.; Chen, P.; Chen, G.; Wang, J.; Huang, L.; and Wang, Y. 2021. Structural and functional brain alterations in anorexia nervosa: A multimodal meta-analysis of neuroimaging studies. *Human brain mapping*, 42(15): 5154–5169.

Vashishth, S.; Sanyal, S.; Nitin, V.; and Talukdar, P. 2019. Composition-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1911.03082*.

Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28.

Yang, Y.; Zhao, B.; Ni, Z.; Zhao, Y.; and Li, X. 2024. Learnable Community-Aware Transformer for Brain Connectome Analysis with Token Clustering. *arXiv preprint arXiv:2403.08203*.

Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; and Liu, T.-Y. 2021. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34: 28877–28888.

Yu, S.; Jin, S.; Li, M.; Sarwar, T.; and Xia, F. 2024. Long-range brain graph transformer. *Advances in Neural Information Processing Systems*, 37: 24472–24495.

Zhao, X.; Wu, J.; Peng, H.; Beheshti, A.; Monaghan, J. J.; McAlpine, D.; Hernandez-Perez, H.; Dras, M.; Dai, Q.; Li, Y.; et al. 2022. Deep reinforcement learning guided graph neural networks for brain network analysis. *Neural Networks*, 154: 56–67.

Zheng, K.; Yu, S.; Chen, L.; Dang, L.; and Chen, B. 2024. BPI-GNN: Interpretable brain network-based psychiatric diagnosis and subtyping. *NeuroImage*, 292: 120594.