Spatial-Temporal DAG Convolutional Networks for End-to-End Joint Effective Connectivity Learning and Resting-State fMRI Classification

Rui Yang[1,2], **Wenrui Dai**[1], **Huajun She**[1], **Yiping P. Du**[1], **Dapeng Wu**[2], **Hongkai Xiong**[1]
[1]Shanghai Jiao Tong University, [2]City University of Hong Kong
{rui_yang, daiwenrui, xionghongkai}@sjtu.edu.cn, dapengwu@cityu.edu.hk

Abstract

Building comprehensive brain connectomes has proved of fundamental importance in resting-state fMRI (rs-fMRI) analysis. Based on the foundation of brain network, spatial-temporal-based graph convolutional networks have dramatically improved the performance of deep learning methods in rs-fMRI time series classification. However, existing works either pre-define the brain network as the correlation matrix derived from the raw time series or jointly learn the connectome and model parameters without any topology constraint. These methods could suffer from degraded classification performance caused by the deviation from the intrinsic brain connectivity and lack biological interpretability of demonstrating the causal structure (i.e., effective connectivity) among brain regions. Moreover, most existing methods for effective connectivity learning are unaware of the downstream classification task and cannot sufficiently exploit useful rs-fMRI label information. To address these issues in an end-to-end manner, we model the brain network as a directed acyclic graph (DAG) to discover direct causal connections between brain regions and propose Spatial-Temporal DAG Convolutional Network (ST-DAGCN) to jointly infer effective connectivity and classify rs-fMRI time series by learning brain representations based on nonlinear structural equation model. The optimization problem is formulated into a continuous program and solved with score-based learning method via gradient descent. We evaluate ST-DAGCN on two public rs-fMRI databases. Experiments show that ST-DAGCN outperforms existing models by evident margins in rs-fMRI classification and simultaneously learns meaningful edges of effective connectivity that help understand brain activity patterns and pathological mechanisms in brain disease.

1 Introduction

Resting-state functional magnetic resonance imaging (rs-fMRI), which detects the spontaneous fluctuations of Blood-Oxygen-Level-Dependent (BOLD) signals in the brain when a specific task is not being performed, has been extensively used to analyze the neural activity patterns of human brain [4]. Over the past few decades, considerable effort has been devoted to estimating the intrinsic connectivity among brain regions of interest (ROIs) from spatial-temporal trajectories in rs-fMRI, which facilitates unveiling the operational principle of human brain architecture [12, 48]. It has now been generally accepted that brain network analysis with the subject at rest plays a crucial part in studying brain development and diagnosis of brain diseases, such as attention deficit hyperactivity disorder [28, 44], Autism spectrum disorder [24, 36] and major depressive disorder [60, 10].

Based on the constructed brain connectome, graph-based deep learning methods have emerged as effective tools to improve the classification accuracy of rs-fMRI time series, and are promising

Temporal Graph Learning Workshop @ NeurIPS 2023, New Orleans.

for disclosing the inherent working mechanism of human brain activity. For example, learning accurate predictors of a particular phenotype of the subjects (e.g., gender) has gained increasing attention in numerous rs-fMRI studies [14, 3, 2, 9, 32, 31]. The core idea behind these studies is the spatial-temporal-based graph convolution, in which the structural information and temporal dynamics of brain graph signals are combined to learn expressive representation for the downstream gender prediction task. Despite achieving remarkable success, several problems still impede the effectiveness of spatial-temporal graph modeling in rs-fMRI analysis, as illustrated below.

Firstly, the brain graph in most existing spatial-temporal-based GCNs is determined based on the correlation (e.g., Pearson's correlation) between each pair of brain ROIs' BOLD signals [14, 3, 2, 23]. Despite being widely adopted in modeling functional connectivity, the symmetric correlation matrix only reflects the simple linear relationships between brain ROIs and cannot explicitly reveal whether the connection is direct or indirect, which leads to a large gap between the estimated connectome and the real complex brain connectivity. Based on the previous studies claiming that graph convolutional networks (GCNs) strongly rely on the extracted graph topology for network analysis and their predictions are vulnerable to the perturbations of graph edges [6], we argue that the simple correlation matrix is inadequate for brain graph modeling and it is indispensable to learn a more complex and meaningful graph structure for more accurate rs-fMRI time series classification.

Secondly, albeit several recent studies [60, 9, 31, 57] were dedicated to addressing the oversimplification problem of the correlation matrix by adaptively learning the graph adjacency matrices in spatial-temporal-based GCNs, they do not consider any explicit topology constraint. In essence, the learned graph structures in [60, 9, 31, 57] and the correlation matrix are all complete graphs, thus lacking biological interpretability of reflecting the effective connectivity of human brain. Contrary to these functional connectivity networks that exhibit the simple fully-connected topology, brain effective connectivity is a causal graph comprising a smaller set of directed edges, in which edge weights characterize the direct causal influence that each neural area exerts over others [27]. Inferring effective connectivity can be regarded as causal discovery from neuroimaging data such as rs-fMRI, which is one of the main end goals in brain network analysis and harder to solve than the simple correlation matrix [48, 12, 27]. Moreover, effective connectivity networks of different population groups with and without certain disease commonly demonstrate distinct patterns and are important for assessing normal brain functions. A myriad of methods have centered on learning effective connectivity and provide promising perspectives for identifying the biomarkers related to certain brain disorder, e.g., Alzheimer's disease [37, 39, 38, 26, 66, 65].

Most existing approaches only use the raw time series to infer the effective connectivity while additional rs-fMRI label information is not fully exploited. The small sample size, high noise level and low temporal resolution characteristics of fMRI time series data pose great challenges to the effectiveness and robustness of these methods [27]. Recently, many studies [14, 3, 2, 9, 32, 31] have demonstrated that the downstream gender prediction task can offer valuable supervised information from accurate rs-fMRI sex labels for training a GCN predictor to effectively discover and analyze human brain connectome, and more accurate and meaningful brain graph modeling could also benefit the performance of the GCN predictor. Based on this motivation, we aim to take advantage of both sides and propose a Spatial-Temporal DAG Convolutional Network (ST-DAGCN) to jointly learn the effective connectivity and enhance time series classification in rs-fMRI analysis.

To be concrete, the adjacency matrix of the directed connectome is treated as tunable parameters and is constrained to ensure the acyclicity, which is primarily dealt with in a broad range of effective connectivity modeling approaches [22, 20, 37, 38, 27]. On such basis, we leverage the nonlinear structural equation model (SEM) [40, 43] to design the spatial-temporal DAG convolution layer, which consists of a DAG convolution that captures the direct causal dependencies among different brain ROIs and an extra one-dimensional (1D) convolution that is cascaded to exploit temporal dynamics for each ROI's time series. Inspired by existing works on causal discovery [64, 61, 43], we formulate a continuous program for causal structure search and jointly optimize the DAG of effective connectivity and model parameters with a score-based learning method via gradient descent. The task-specific score combines the cross entropy between ST-DAGCN's predictions and rs-fMRI labels and an extra graph sparsity regularization. To demonstrate the advantages over existing spatial-temporal-based GCNs, we validate ST-DAGCN in the same gender classification task. ST-DAGCN is evaluated on two public rs-fMRI databases, i.e., disease-irrelevant Human Connectome Project (HCP) and disease-relevant AD (patients with Alzheimer's disease) and CN (control subjects) groups of Alzheimer's Disease Neuroimaging Initiative (ADNI). Experiments show that ST-DAGCN is

effective in learning brain connectivity that boosts time series classification accuracy. Moreover, the extracted causal dependencies among brain ROIs indeed reveal some macroscopic structures as well as some local connectivity that echo previous neuroscience research for disclosing brain activity patterns and biomarkers of brain disease pathology.

2 Related Work

Spatial-Temporal-Based GCNs for Gender Prediction. Predicting a particular phenotype from a subject's rs-fMRI data has gained increasing interest in the neuroscience research community. A number of methods have been developed to solve the gender classification task based on the brain connectivity derived from rs-fMRI signals. Among them, graph-based deep learning has become the mainstream method that applies spatial-temporal-based GCNs to the constructed brain network and rs-fMRI time series to learn effective brain graph representation [14, 3, 2, 9, 32, 31]. Two concurrent pioneering works [14, 3, 2] firstly use the idea of Spatial Temporal Graph Convolutional Networks (ST-GCN) [58] for rs-fMRI analysis, where the brain connectivity is built with the correlation matrix of brain ROIs' raw time series. To further improve the predictor's classification performance, STAGIN [32] builds dynamic correlation matrices with sliding window approach and enhances the network's expressive power via spatial-temporal attention. DAST-GCN [9] adaptively learns the graph adjacency matrix in each spatial-temporal graph convolution layer and uses dilated 1D convolution for temporal feature extraction. Since the learned adjacency matrices of brain graphs are different and essentially complete graphs at each layer, DAST-GCN does not have good interpretability in brain network analysis. FBNetGen [31] employs a time series encoder followed by a graph generator to transform the time series into brain connectivity. For different subjects, it produces different connectome output, while our method learns a shared connectivity across a group of subjects, which enables us to easily identify the alteration in brain connectivity by comparing different populations (e.g., AD versus CN) [37, 38, 19, 8]. Moreover, FBNetGen does not consider topology constraint and also learns fully-connected brain graphs like DAST-GCN, which impedes the network interpretability.

Effective Connectivity Learning. Inferring effective connectivity from neuroimaging data is an ultimate goal of brain network analysis, as it offers a fresh perspective for disclosing the direct causation between brain regions. Most existing approaches can be divided into model-based and data-driven methods. Model-based methods require prior knowledge and depend on presupposed models to fit the neuroimaging data. Typical approaches include structural equation model (SEM) [47], dynamic causal modeling (DCM) [13], and multivariate autoregressive model (MVAR) [15]. However, these methods are limited by model capacity and their performance is largely affected by hyperparameters. By comparison, data-driven methods directly infer effective connectivity from data without any prior knowledge or assumption. Among them, DAG is one common type of causal structure that is primarily dealt with [20, 27], and score-based learning, which uses specific score functions to measure the quality of candidate causal graphs and search for the one with the optimal score, has been widely used in DAG-based effective connectivity learning [37, 38]. Recently, some deep learning approaches have been developed. nCREANN [50] extends MVAR to the nonlinear case via artificial neural network. EC-GAN [39] and EC-RGAN [26] use GANs [17] to learn effective connectivity. CVAEEC [65] and STGCMEC [66] utilize variational autoencoder [34] and GCNs, respectively, to model the generative procedure of rs-fMRI time series. TBDS [62] optimizes the GCN predictor and brain connectivity in an alternating manner, in which the downstream classification loss serves as the regularization term for maximum likelihood estimation of the raw time series in DAG-based effective connectivity learning. Therefore, TBDS does not directly learn task-specific connectivity in an end-to-end manner. Moreover, due to the introduction of the regularization term, the objective function may not satisfy the convergence condition of its employed optimization algorithm [42], causing the learned effective connectivity practically far from the DAG.

3 Method

3.1 Preliminaries

Causal Discovery. In this paper, we adopt the score-based learning of DAG for inferring effective connectivity, which assigns a score to each candidate causal graph based on some pre-defined score function and outputs the DAG with the optimal score. Formally, we denote $\mathcal{G} = (\mathcal{V}, \mathbf{A})$ as the causal graph, where $\mathcal{V} = \{v_i\}_{i=1}^{N}$ is the node set and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the weighted adjacency matrix. Each

non-zero entry $\mathbf{A}[j,i]$ in $\mathbf{A}$ describes the strength of the direct causal influence $v_j \to v_i$. Given the pre-defined score function $F : \mathbb{R}^{N\times N} \to \mathbb{R}$, the causal search problem is defined as

$$\min_{\mathbf{A}} \; F(\mathbf{A}) \quad \text{s.t. } \mathcal{G} \in \text{DAGs}, \tag{1}$$

which is a combinatorial optimization problem that is NP-hard in general and needs specialized search algorithm to solve over the combinatorial space of DAGs. To make the optimization procedure easier, the widely used strategy is to transform it into a continuous program [64, 61, 43]:

$$\min_{\mathbf{A}} \; F(\mathbf{A}) \quad \text{s.t. } h(\mathbf{A}) = 0, \tag{2}$$

where $h : \mathbb{R}^{N\times N} \to \mathbb{R}$ is a smooth function for ensuring the acyclicity constraint with $h(\mathbf{A}) = 0$. By replacing the combinatorial constraint $\mathcal{G} \in$ DAGs with continuous constraint $h(\mathbf{A}) = 0$, the causal discovery problem has been transformed to a continuous optimization problem that can be efficiently solved via standard numerical algorithms such as augmented Lagrangian method.

Structural Equation Model (SEM). SEM is a popular model for describing the data generation procedure with causal relationships and is widely used in the estimation of effective connectivity [47, 48, 12, 39, 26, 27, 65]. Suppose that, for the network system of N nodes, each node v_i is attached with a vector-valued signal or feature $\mathbf{s}_i \in \mathbb{R}^d$. The linear SEM takes the following form:

$$\mathbf{s}_i = \sum_{j\in PA(i)} \mathbf{A}[j,i]\,\mathbf{s}_j + \epsilon_i, \tag{3}$$

where $PA(i)$ denotes the parents of node v_i induced from the causal DAG, i.e., signals are propagated from parent nodes to their children for aggregation, and $\epsilon_i \in \mathbb{R}^d (i = 1, \ldots, N)$ are additive random noise (e.g., Gaussian noise) that are mutually independent between different nodes.

3.2 Problem Set-Up

Given a subject's rs-fMRI data and an atlas, the brain is first parcellated into N ROIs to extract their average BOLD time series $\mathbf{X} \in \mathbb{R}^{N\times T\times 1}$, where T is the number of time points and each brain ROI's BOLD signal is standardized across time. Our goal is to learn a causal structure, i.e., brain effective connectivity, in gender classification for rs-fMRI. To be specific, we represent the brain network with a DAG $\mathcal{G} = (\mathcal{V}, \mathbf{A})$, and here $\mathcal{V} = \{v_i\}_{i=1}^N$ is the set of brain ROIs and each v_i is associated with a temporal BOLD signal $\mathbf{X}[i,:,:]$. Following [14, 37, 38], we assume the same brain (effective) connectivity is shared by all the subjects in the same group (e.g., all the subjects with AD). Based on the extracted time series $\mathbf{X}$ and the brain effective connectivity $\mathcal{G}$, ST-DAGCN is performed for gender prediction of the corresponding subject. Considering the fact that a better graph topology commonly leads to higher classification performance of GCNs [6, 63, 7, 59], we formulate the optimization problem as below, based on the causal discovery problem in (2), to jointly learn the causal structure $\mathbf{A}$ and the model parameters θ of ST-DAGCN.

$$\begin{aligned} \min_{\mathbf{A},\theta} \quad & F(\mathbf{A},\theta) \equiv \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}\left(y'_m, y_m; \mathbf{A}, \theta\right) + \lambda \|\mathbf{A}\|_1 \\ \text{s.t.} \quad & h(\mathbf{A}) \equiv \operatorname{tr}\left[(\mathbf{I} + \alpha \mathbf{A} \odot \mathbf{A})^N\right] - N = 0. \end{aligned} \tag{4}$$

Here, $F(\mathbf{A},\theta)$ is the task-specific score function, in which $\mathcal{L}\left(y'_m, y_m; \mathbf{A}, \theta\right)$ is the loss (i.e., cross entropy) between the ST-DAGCN's prediction y' and the target label y. M is the number of training subjects. $\|\cdot\|_1$ denotes the L_1 norm for sparsity penalty on the causal graph. λ is the balancing hyperparameter for sparsity-based regularization. $\operatorname{tr}\left[(\mathbf{I} + \alpha \mathbf{A} \odot \mathbf{A})^N\right] - N = 0$ is the necessary and sufficient condition to ensure $\mathcal{G} \in$ DAGs [61], in which $\odot$ denotes the Hadamard product of two matrices and α is a hyperparameter. Following [61], we define $\alpha = 1/N$.

3.3 Spatial-Temporal DAG Convolutional Network

The core of ST-DAGCN is the spatial-temporal DAG Convolution (ST-DAGC) layer for nonlinear structural equation modeling in rs-fMRI analysis. ST-DAGC consists of DAG convolution at the spatial level and 1D convolution at the temporal level such that the causal dependencies among brain ROIs and temporal dynamics of each ROI's time series are combined to learn effective representation for the downstream gender prediction task. In this section, we first introduce ST-DAGC and the network architecture for spatial-temporal graph modeling in rs-fMRI. Then we present the optimization procedure of ST-DAGCN to obtain the effective connectivity in gender prediction.

Spatial-Level Convolution. At the spatial level, the linear SEM described by Eq. (3) is widely adopted in effective connectivity learning to model the causal dependencies between brain ROIs [39, 26, 65], but cannot capture the nonlinear relationships in feature generation. The linear SEM can be rewritten in the matrix form as $f(\mathbf{S}, \mathbf{A}) = \mathbf{A}^\top \mathbf{S} + \mathbf{E}$, where $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_N]^\top \in \mathbb{R}^{N \times d}$ is the feature matrix of N ROIs and $\mathbf{E} = [\epsilon_1, \ldots, \epsilon_N]^\top \in \mathbb{R}^{N \times d}$ is the noise matrix. To increase the model expressivity, Ng et al. [43] handle a more complex causal graph model that extends the linear SEM to the nonlinear case:

$$f(\mathbf{S}, \mathbf{A}) = g_2\left(\mathbf{A}^\top g_1\left(\mathbf{S}\right)\right), \tag{5}$$

where $g_2(\cdot)$ and $g_1(\cdot)$ are two neural networks for capturing nonlinear causal relationships. Considering that our model consists of multiple spatial-level DAG convolutions interleaved with nonlinear temporal-level convolutions, we use a linear transformation for $g_1(\cdot)$ and a nonlinear function for $g_2(\cdot)$ in Eq. (5) for one DAG convolution operation. Denote with $\mathbf{H}^i \in \mathbb{R}^{N \times T \times f}$ and $\mathbf{H}^s \in \mathbb{R}^{N \times T \times f_s}$ the input and output features of DAG convolution, where f and f_s are the numbers of input and output feature channels, and $\mathbf{H}^i = \mathbf{X}$ in the first ST-DAGC layer. For $t = 1, \ldots, T$, we have

$$\begin{aligned} \mathbf{H}^s[:, t, :] &= f\left(\mathbf{H}^i[:, t, :], \mathbf{A}\right) \\ &= \sigma\left(\mathrm{BatchNorm}\left(\mathbf{A}^\top\left(\mathbf{H}^i[:, t, :]\mathbf{W}^s + \mathbf{B}\right)\right)\right), \end{aligned} \tag{6}$$

where $\mathbf{H}^i[:, t, :] \in \mathbb{R}^{N \times f}$ and $\mathbf{H}^s[:, t, :] \in \mathbb{R}^{N \times f_s}$ are the t-th slice of features before and after DAG convolution, $\mathbf{W}^s \in \mathbb{R}^{f \times f_s}$ is the learnable weights matrix, $\mathbf{B} \in \mathbb{R}^{N \times f_s}$ is the learnable bias matrix, $\mathrm{BatchNorm}\left(\cdot\right)$ is the batch normalization operation [25] along the feature channel, and $\sigma\left(\cdot\right)$ is the nonlinear activation function. Under this feature generation rule in which each brain ROI's features are generated only dependent on causally relevant brain ROIs at each layer, DAG convolutions capture the nonlinear causal dependencies among brain ROIs by propagating nonlinear node features along the directed edges of effective connectivity for aggregation.

Temporal-Level Convolution. 1D convolution has proved to be an effective tool for analyzing the temporal dynamics of rs-fMRI data, as the BOLD signal itself is typically modeled as the 1D convolution of a stimulus function and the hemodynamic response function. A variety of sophisticated 1D convolution operations have been employed for rs-fMRI time series analysis, such as dilated convolution [9] and causal convolution [2]. In this work, we use the standard 1D convolution at the temporal level attributed to its efficiency and empirical remarkable performance. Specifically, subsequent to DAG convolution, the 1D convolution is cascaded to capture the temporal dynamics for each individual brain ROI. Denote $\mathbf{H}^t \in \mathbb{R}^{N \times T \times f_t}$ as the output features of 1D convolution and f_t as the number of output feature channels. For $n = 1, \ldots, N$:

$$\mathbf{H}^t[n, :, :] = \sigma\left(\mathrm{BatchNorm}\left(\mathrm{Conv1D}\left(\mathbf{H}^s[n, :, :], \mathbf{W}^t\right)\right)\right), \tag{7}$$

where $\mathbf{H}^s[n, :, :] \in \mathbb{R}^{T \times f_s}$ and $\mathbf{H}^t[n, :, :] \in \mathbb{R}^{T \times f_t}$ are the n-th node features before and after 1D convolution, and $\mathbf{W}^t$ is the learnable 1D convolution filter. By analogy with the DAG convolution that aggregates information from the causally dependent spatial neighborhood for each brain ROI, 1D convolution actually also performs feature aggregation for each time point by considering the temporal neighborhood determined by the 1D convolution kernel.

The above two-level convolution is the procedure for obtaining the output features $\mathbf{H}^i \rightarrow \mathbf{H}^s \rightarrow \mathbf{H}^t$ in an ST-DAGC layer, which serves as the building block for spatial-temporal nonlinear structural equation modeling in rs-fMRI analysis.

Network Architecture of ST-DAGCN. Our designed ST-DAGCN architecture consists of three ST-DAGC layers to learn hierarchical graph representations with feature dimension set to 64, i.e., $f = f_s = f_t = 64$ except for $f = 1$ in the first ST-DAGC layer. A global average pooling layer that computes the mean value within each feature channel is then performed on the resulting tensor to get a 64-dimensional feature vector for each input rs-fMRI sequence. Finally, a fully-connected layer with a single output neuron activated by the sigmoid function is cascaded to generate the classification probability of gender prediction for the input subject's rs-fMRI data.

Optimization. Inspired by score-based DAG learning approaches [64, 61, 43], we use the augmented Lagrangian method to solve the causal structure learning problem in (4). To be more specific, the augmented Lagrangian is defined as

$$L_c\left(\mathbf{A}, \theta, \eta\right) = F\left(\mathbf{A}, \theta\right) + \eta h(\mathbf{A}) + \frac{c}{2}|h(\mathbf{A})|^2, \tag{8}$$

where θ is the set of all the learnable parameters of ST-DAGCN, η is the Lagrange multiplier, and $c > 0$ is the penalty parameter. At the k-th iteration of optimization, three steps are performed to alternately update $\{\mathbf{A}^{(k)}, \theta^{(k)}\}$, $\eta^{(k)}$, and $c^{(k)}$:

$$\{\mathbf{A}^{(k)}, \theta^{(k)}\} = \arg\min_{\mathbf{A},\theta} L_{c^{(k)}}\left(\mathbf{A}, \theta, \eta^{(k)}\right), \tag{9}$$

$$\eta^{(k+1)} = \eta^{(k)} + c^{(k)} h(\mathbf{A}^{(k)}), \tag{10}$$

$$c^{(k+1)} = \begin{cases} \beta c^{(k)}, & \text{if } |h(\mathbf{A}^{(k)}| > \gamma |h(\mathbf{A}^{(k-1)}|, \\ c^{(k)}, & \text{otherwise}, \end{cases} \tag{11}$$

where $\beta > 1$ and $\gamma < 1$ are hyperparameters. The subproblem (9) can be solved via gradient descent. Specifically, we use Adam optimizer [33] implemented by PyTorch. To avoid over-fitting in solving (9), dropout [49] is employed after each ST-DAGC layer, and we perform L_2 regularization on all the weights matrices. The optimization procedure guarantees that $\mathbf{A}, \theta$, and η can converge to local optimum, and c progressively increases such that $h(\mathbf{A})$ vanishes to make $\mathbf{A}$ approximate a DAG. Thus, the final output DAG is the causal graph that achieves the optimal score in helping ST-DAGCN classify rs-fMRI time series. We provide detailed implementations in the Appendix.

Post-Processing. Since $h(\mathbf{A})$ cannot be strictly zero at the end of optimization, post-processing is necessary to produce the exact DAG that reveals the potential causal dependencies among different brain ROIs. To be concrete, we first threshold the learned brain connectivity by $\mathbf{A} \odot 1[|\mathbf{A}| > \varepsilon]$ to remove probably falsely discovered connections, where $1[\cdot]$ is the element-wise Iverson bracket and $\varepsilon > 0$ is the threshold value. If the thresholded adjacency matrix still contains cycles, we use depth-first search (DFS) to find the cycle and remove the edge with the smallest absolute value in the cycle, and repeat this procedure until no cycles can be found in the final processed graph $\mathbf{A}_{\text{DAG}}$. The resulting residual caused by $\mathbf{A} - \mathbf{A}_{\text{DAG}}$ can be viewed as the noise matrix part in SEM.

4 Experiments and Analysis

4.1 Datasets and Baselines

In this study, we evaluate all the methods on two publicly available rs-fMRI databases, i.e., Human Connectome Project (HCP) [51] and Alzheimer's Disease Neuroimaging Initiative (adni.loni.usc.edu). The evaluation metrics and scheme are provided in the Appendix.

HCP. For a fair comparison with the previous spatial-temporal graph modeling approach like ST-GCN [14], we use its public pre-processed data that consists of 1089 rs-fMRI scans (496 males and 593 females) with repetition time TR $= 0.72$s and the total number of time points $T_{\text{total}} = 1000$. The brain is parcellated into $N = 22$ ROIs according to [16], and the average BOLD signal in each ROI is standardized to z-scores.

ADNI. We select 132 subjects that are divided into two groups: CN (control subjects) and AD (patients with Alzheimer's disease). The CN group contains 32 females and 40 males, and the AD group consists of 31 females and 29 males. The brain is parcellated into $N = 39$ ROIs via the MSDL atlas [52]. The rs-fMRI data is acquired with the same scanner of TR $= 3$s. Each brain ROI's BOLD signal is also standardized to z-scores and $T_{\text{total}} = 130$.

Baselines. We compare with various spatial-temporal modeling approaches for rs-fMRI: (1) The LSTM model [21] trained on the BOLD time series. (2) ST-GCN [14, 3, 2] that uses 1D convolution at the temporal level and graph convolution at the spatial level based on the pre-defined Pearson's correlation matrix. The specific implementation of convolution is selected from [14, 3, 2] and tuned for each dataset to achieve the best accuracy. (3) DAST-GCN [9] that adaptively learns an adjacency matrix in each spatial graph convolution layer and uses dilated 1D convolution at the temporal level. (4) STAGIN [32] that calculates dynamic Pearson's correlation and attentive graph read-out at the spatial level and employs Transformer-based attention [53] at the temporal level. The original STAGIN model uses the Pearson's correlation matrices in which only the top 30-percentile values are kept. We also consider the variant using the full correlation (FC) matrices. (5) FBNetGen [31] in which a time series encoder is employed to learn temporal features and a graph generator is followed to build the brain network based on the encoded features. The correlation matrix and the learned brain network will serve as the input node features and adjacency matrix for a GCN predictor.

Table 1: rs-fMRI time series classification results on the HCP dataset.

Method	ACC (%)	SEN (%)	SPE (%)	AUC (%)
LSTM	71.53 ± 1.80	60.88 ± 5.75	80.44 ± 3.14	74.98 ± 2.51
ST-GCN	79.52 ± 2.33	76.20 ± 7.28	82.31 ± 6.70	85.99 ± 1.70
DAST-GCN	78.24 ± 2.99	71.97 ± 6.39	83.47 ± 4.05	82.43 ± 3.85
STAGIN (Original)	78.14 ± 1.95	69.74 ± 8.45	85.13 ± 6.57	84.29 ± 2.85
STAGIN (FC)	76.58 ± 2.12	72.17 ± 7.72	80.27 ± 2.81	81.17 ± 3.02
FBNetGen	66.67 ± 0.83	57.85 ± 8.73	74.04 ± 8.09	67.16 ± 1.16
ST-DAGCN	85.03 ± 2.25	84.87 ± 2.16	85.16 ± 3.64	89.90 ± 2.33

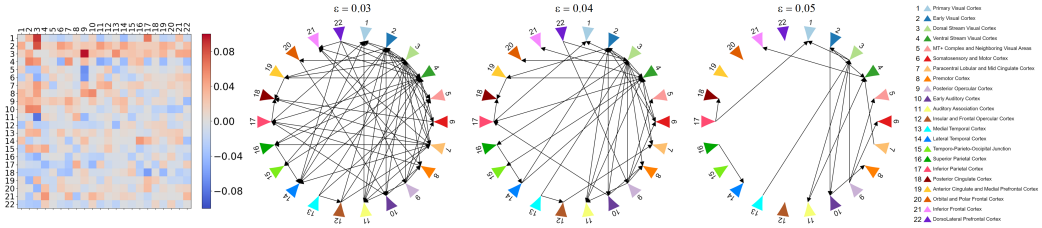

Figure 1: Learned effective connectivity on the HCP dataset, which consists of 22 brain ROIs [16] pre-processed by [14]. The left-most figure is the averaged $\mathbf{A}^\top$, and the right three figures are the produced DAGs with different threshold values.

4.2 Results on Disease-Irrelevant HCP Dataset

Gender Prediction. Table 1 reports the classification results on the HCP dataset. Overall, we make the following observations. (1) Our ST-DAGCN achieves the best classification performance on the HCP database. (2) With the DAG topology constraint based on nonlinear SEM, ST-DAGCN significantly outperforms existing graph-based spatial-temporal modeling approaches. Noticeably, ST-DAGCN achieves about 5% gain in mean ACC and 4% gain in mean AUC compared with the runner-up ST-GCN without topology constraint. (3) Both the correlation matrix and previous adaptive graph learning manners cannot guarantee that the spatial-temporal-based GCNs reach their greatest potential in rs-fMRI gender prediction. (4) In experiments, we find that the complex and large-scale STAGIN models tend to suffer from over-fitting and perform badly on HCP pre-processed by [14]. In contrast, our model demonstrates better generalization ability in terms of four evaluation metrics.

Effective Connectivity Analysis. Here, we visualize and analyze the final learned effective connectivity. We optimize the causal structure learning problem with all the subjects' data for 10 trials and then average the learned $\mathbf{A}$ for producing the DAG with post-processing. Figure 1 visualizes the learned effective connectivity. We find that the visual regions (nodes 1-5) have more neural connections and this result is consistent with previous studies [55, 18] claiming that the visual cortex is highly correlated with rs-fMRI information pattern. The dorsal stream visual cortex (node 3) involves the most direct links with other neural areas, which is aligned with its functionality that it includes higher visual regions responsible for visual stimuli perception and plan of visually guided actions [16]. Moreover, our discoveries are also in agreement with previous research [14, 56] showing that visual regions are crucial in demonstrating gender-related differences in rs-fMRI. These findings physiologically verify that the produced effective connectivity not only substantially improves the rs-fMRI classification performance, but also learns some meaningful macroscopic structures in reflecting the causal dependencies among brain ROIs to help understand normal brain functions.

4.3 Results on Disease-Relevant ADNI Dataset

Existing works separately infer the effective connectivity of AD and CN groups for the downstream analysis [26, 38, 65]. Following this paradigm, we apply ST-DAGCN to the AD and CN groups of ADNI database, respectively, to learn the effective connectivity and perform gender classification.

Gender Prediction. We provide the gender classification results in Table 2. As can be seen, our proposed ST-DAGCN consistently performs best in terms of the ACC and AUC metrics on both AD and CN groups. We can also observe that ST-DAGCN outperforms existing graph modeling approaches without topology constraint by large margins. Averagely, ST-DAGCN achieves over 3.0% and 5.5% accuracy improvement against previous state-of-the-art methods on the AD and CN groups respectively. This result also verifies that more effective modeling of brain connectivity is of great benefit to the spatial-temporal-based GCNs in rs-fMRI time series classification.

Table 2: rs-fMRI time series classification results on the AD and CN groups of the ADNI dataset.

Method	AD				CN			
	ACC (%)	SEN (%)	SPE (%)	AUC (%)	ACC (%)	SEN (%)	SPE (%)	AUC (%)
LSTM	81.67 $\pm$ 8.16	76.00 $\pm$ 7.72	87.14 $\pm$ 12.38	79.33 $\pm$ 12.85	75.14 $\pm$ 10.9	77.50 $\pm$ 21.51	70.95 $\pm$ 17.13	71.96 $\pm$ 15.16
ST-GCN	81.67 $\pm$ 3.33	83.33 $\pm$ 14.91	80.95 $\pm$ 11.57	73.17 $\pm$ 5.48	79.14 $\pm$ 6.22	82.50 $\pm$ 12.75	75.24 $\pm$ 15.47	78.10 $\pm$ 11.94
DAST-GCN	80.00 $\pm$ 8.50	76.00 $\pm$ 7.72	83.81 $\pm$ 10.58	78.22 $\pm$ 9.34	77.62 $\pm$ 5.75	85.00 $\pm$ 12.25	68.10 $\pm$ 11.43	75.83 $\pm$ 9.98
STAGIN (Original)	80.00 $\pm$ 4.08	80.00 $\pm$ 16.33	80.95 $\pm$ 18.87	79.32 $\pm$ 4.22	76.29 $\pm$ 5.81	75.00 $\pm$ 17.68	77.14 $\pm$ 17.21	72.02 $\pm$ 6.54
STAGIN (FC)	81.67 $\pm$ 3.33	83.33 $\pm$ 10.54	80.95 $\pm$ 11.57	79.35 $\pm$ 7.11	80.48 $\pm$ 3.14	87.50 $\pm$ 11.18	70.95 $\pm$ 19.37	75.24 $\pm$ 9.32
FBNetGEN	70.00 $\pm$ 4.08	70.00 $\pm$ 22.11	71.43 $\pm$ 14.75	66.81 $\pm$ 6.03	73.71 $\pm$ 4.52	75.00 $\pm$ 13.69	72.38 $\pm$ 14.57	75.33 $\pm$ 4.55
ST-DAGCN	85.00 $\pm$ 6.24	89.33 $\pm$ 8.79	80.48 $\pm$ 16.45	79.87 $\pm$ 9.88	86.10 $\pm$ 4.54	92.50 $\pm$ 6.12	77.62 $\pm$ 13.85	83.21 $\pm$ 7.53

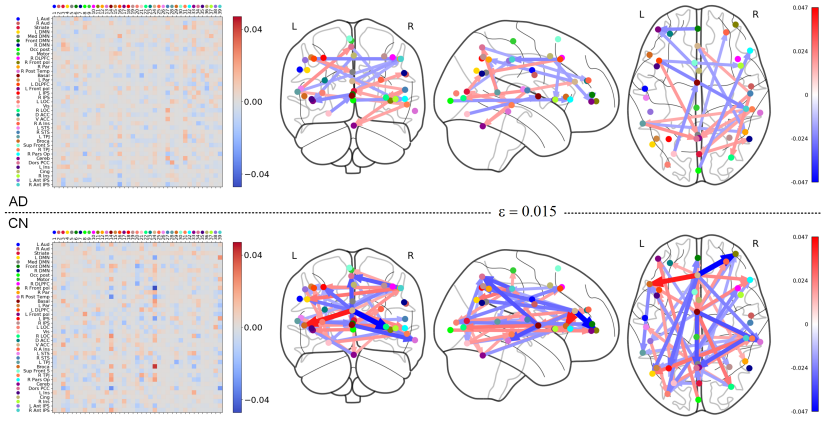

Figure 2: Estimated effective connectivity on the ADNI dataset. The brain is parcellated into 39 ROIs according to the MSDL atlas [52]. The left two matrices are the averaged $\mathbf{A}^\top$, and the right three columns are estimated brain connectomes.

Group Analysis of Estimated Effective Connectivity. For each group, we also estimate the brain network with all the subject's data for 10 trials and average the learned $\mathbf{A}$ for producing the DAG with post-processing. The estimated effective connectivity networks are shown in Figure 2. For a better view of brain connectome, we set the threshold ε to 0.015. In the Appendix, we also provide additional quantitative comparison results. Overall, we make several remarkable observations by comparing the two groups. (1) Compared with the CN group's effective connectivity, it is obvious that the number and strength of edges in AD patients decrease evidently. This observation is supported by a wide range of existing literature [26, 65, 38]. (2) According to [54], the degeneration of basal ganglia (node 14) is considered to be a contributor to the cognitive decline in AD patients. We also identify this brain ROI as the most important biomarker in Alzheimer's disease, as shown in Table 3. Moreover, as illustrated in the 14th column of the adjacency matrices in Figure 2, the basal ROI of CN indeed demonstrates much more connections with other brain regions than that of AD. (3) Many previous works [29, 1] have found degenerative changes and reduced functional activity of the anterior cingulate cortex (ACC) in AD. We also identify ventral ACC (node 24) as the second most salient biomarker and find that the ventral ACC of AD has fewer and weaker neural links with other ROIs than that of CN, as shown in the 24th column of the adjacency matrices in Figure 2. Due to the page limit, we provide some other findings in the Appendix.

The change of effective connectivity is believed to be highly related to brain disease [22, 26]. Our observations verified by existing neuropathological studies indicate that ST-DAGCN provides a promising perspective for identifying the biomarkers associated with Alzheimer's disease, which is worth being further studied to understand the underlying mechanisms in the future.

5 Conclusion

In this paper, we propose the ST-DAGCN model to jointly infer the brain effective connectivity and classify rs-fMRI time series. In ST-DAGCN, DAG convolution and 1D convolution are interleaved for spatial-temporal nonlinear structural equation modeling. A score-based optimization problem is formulated for DAG-based effective connectivity learning and solved by the augmented Lagrangian method. We demonstrate that ST-DAGCN achieves improved gender classification performance against existing approaches on two public rs-fMRI datasets, and yields more interpretability in understanding brain activity patterns and pathology of brain disease such as Alzheimer's disease.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 62250055, Grant 62371288, Grant 62320106003, Grant 61932022, Grant 61971285, Grant 62120106007, and in part by the Program of Shanghai Science and Technology Innovation Project under Grant 20511100100.

References

[1] M. Amanzio, D. M. Torta, K. Sacco, F. Cauda, F. D'Agata, S. Duca, D. Leotta, S. Palermo, and G. C. Geminiani. Unawareness of deficits in Alzheimer's disease: role of the cingulate cortex. *Brain*, 134(4):1061–1076, 2011.

[2] T. Azevedo, A. Campbell, R. Romero-Garcia, L. Passamonti, R. A.I. Bethlehem, P. Liò, and N. Toschi. A deep graph neural network architecture for modelling spatio-temporal dynamics in resting-state functional MRI data. *Med. Image Anal.*, 79:102471, 2022.

[3] T. Azevedo, L. Passamonti, P. Liò, and N. Toschi. A deep spatiotemporal graph learning architecture for brain connectivity analysis. In *2020 42nd Ann. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, pages 1120–1123, 2020.

[4] B. B. Biswal. Resting state fMRI: a personal history. *NeuroImage*, 62(2):938–944, 2012.

[5] R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter. The brain's default network: anatomy, function, and relevance to disease. *Ann. N. Y. Acad. Sci.*, 1124(1):1–38, 2008.

[6] L. Chen, J. Li, Q. Peng, Y. Liu, Z. Zheng, and C. Yang. Understanding structural vulnerability in graph convolutional networks. In *Proc. 30th Int. Joint Conf. Artif. Intell.*, pages 2249–2255, 2021.

[7] Y. Chen, L. Wu, and M. Zaki. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. *Adv. Neural Inf. Process. Syst. 33*, pages 19314–19326, 2020.

[8] O. Demirbilek, I. Rekik, and Alzheimer's Disease Neuroimaging Initiative. Predicting the evolution trajectory of population-driven connectional brain templates using recurrent multigraph neural networks. *Med. Image Anal.*, 83:102649, 2023.

[9] A. El-Gazzar, R. M. Thomas, and G. V. Wingen. Dynamic adaptive spatio-temporal graph convolution for fMRI modelling. In *Int. Workshop Mach. Learn. Clin. Neuroimaging*, pages 125–134, 2021.

[10] Y. Fang, M. Wang, G. G. Potter, and M. Liu. Unsupervised cross-domain functional mri adaptation for automated major depressive disorder identification. *Med. Image Anal.*, 84:102707, 2023.

[11] E. Finger, J. Zhang, B. Dickerson, Y. Bureau, M. Masellis, and Alzheimer's Disease Neuroimaging Initiative. Disinhibition in Alzheimer's disease is associated with reduced right frontal pole cortical thickness. *J. Alzheimer's Dis.*, 60(3):1161–1170, 2017.

[12] K. J. Friston. Functional and effective connectivity: a review. *Brain Connect.*, 1(1):13–36, 2011.

[13] K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302, 2003.

[14] S. Gadgil, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, E. Adeli, and K. M. Pohl. Spatio-temporal graph convolution for resting-state fMRI analysis. In *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, pages 528–538, 2020.

[15] M. Gilson, A. Tauste Campo, X. Chen, A. Thiele, and G. Deco. Nonparametric test for connectivity detection in multivariate autoregressive networks and application to multiunit activity data. *Netw. Neurosci.*, 1(4):357–380, 2017.

[16] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Adv. Neural Inf. Process. Syst. 27*, pages 2672–2680, 2014.

[18] R. Guidotti, C. Del Gratta, A. Baldassarre, G. L. Romani, and M. Corbetta. Visual learning induces changes in resting-state fMRI multivariate pattern of information. *J. Neurosci.*, 35(27):9786–9798, 2015.

[19] M. B. Gürbüz and I. Rekik. MGN-Net: A multi-view graph normalizer for integrating heterogeneous biological network populations. *Med. Image Anal.*, 71:102059, 2021.

[20] T. Henry and K. Gates. Causal search procedures for fMRI: review and suggestions. *Behaviormetrika*, 44(1):193–225, 2017.

[21] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[22] S. Huang, J. Li, J. Ye, A. Fleisher, K. Chen, T. Wu, and E. Reiman. Brain effective connectivity modeling for alzheimer's disease by sparse gaussian bayesian network. In *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pages 931–939, 2011.

[23] S. Huang, J. Xia, L. Xu, and A. Qiu. Spatio-temporal directed acyclic graph learning with attention mechanisms on brain functional time series and connectivity. *Med. Image Anal.*, 77:102370, 2022.

[24] Y. Huang and A. C. S. Chung. Edge-variational graph convolutional networks for uncertainty-aware disease prediction. In *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, pages 562–572, 2020.

[25] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. 32nd Int. Conf. Mach. Learn.*, pages 448–456, 2015.

[26] J. Ji, J. Liu, L. Han, and F. Wang. Estimating effective connectivity by recurrent generative adversarial networks. *IEEE Trans. Med. Imaging*, 40(12):3326–3336, 2021.

[27] J. Ji, A. Zou, J. Liu, C. Yang, X. Zhang, and Y. Song. A survey on brain effective connectivity network learning. *IEEE Trans. Neural Netw. Learn. Syst.*, 2021.

[28] B. Jie, C. Wee, D. Shen, and D. Zhang. Hyper-connectivity of functional networks for brain disease diagnosis. *Med. Image Anal.*, 32:84–100, 2016.

[29] B. F. Jones, J. Barnes, H. B. Uylings, N. C. Fox, C. Frost, M. P. Witter, and P. Scheltens. Differential regional atrophy of the cingulate gyrus in Alzheimer disease: a volumetric MRI study. *Cereb. Cortex*, 16(12):1701–1708, 2006.

[30] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon Kohl, Andrew Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[31] X. Kan, H. Cui, J. Lukemire, Y. Guo, and C. Yang. FBNETGEN: Task-aware GNN-based fMRI analysis via functional brain network generation. In *Med. Imaging Deep Learn.*, 2022.

[32] B.-H. Kim, J. C. Ye, and J.-J. Kim. Learning dynamic graph representation of brain connectome with spatio-temporal attention. In *Adv. Neural Inf. Process. Syst. 34*, pages 4314–4327, 2021.

[33] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd Int. Conf. Learn. Rep.*, 2015.

[34] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *2nd Int. Conf. Learn. Rep.*, 2014.

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

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

[37] J. Liu, J. Ji, X. Jia, and A. Zhang. Learning brain effective connectivity network structure using ant colony optimization combining with voxel activation information. *IEEE J. Biomed. Health Inform.*, 24(7):2028–2040, 2019.

[38] J. Liu, J. Ji, G. Xun, and A. Zhang. Inferring effective connectivity networks from fMRI time series with a temporal entropy-score. *IEEE Trans. Neural Netw. Learn. Syst.*, 2021.

[39] J. Liu, J.and Ji, G. Xun, L. Yao, M. Huai, and A. Zhang. EC-GAN: inferring brain effective connectivity via generative adversarial networks. In *Proc. 34th AAAI Conf. Artif. Intell.*, pages 4852–4859, 2020.

[40] A. R. Mclntosh and F. Gonzalez-Lima. Structural equation modeling and its application to network analysis in functional brain imaging. *Hum. Brain Mapp.*, 2(1-2):2–22, 1994.

[41] F. G. Metzger, B. Schopp, F. B. Haeussinger, K. Dehnen, M. Synofzik, A. J. Fallgatter, and A.-C. Ehlis. Brain activation in frontotemporal and Alzheimer's dementia: a functional near-infrared spectroscopy study. *Alzheimer's Res. Ther.*, 8(56), 2016.

[42] I. Ng, A. Ghassami, and K. Zhang. On the role of sparsity and dag constraints for learning linear dags. In *Adv. Neural Inf. Process. Syst. 33*, pages 17943–17954, 2020.

[43] I. Ng, S. Zhu, Z. Chen, and Z. Fang. A graph autoencoder approach to causal structure learning. In *NeurIPS 2019 Workshop "Do the right thing": machine learning and causal inference for improved decision making*, 2019.

[44] Z. Rakhimberdina and T. Murata. Linear graph convolutional model for diagnosing brain disorders. In *Int. Conf. Complex Netw. Their Appl.*, pages 815–826, 2019.

[45] P. Reiner, E. Jouvent, E. Duchesnay, R. Cuingnet, J.-F. Mangin, H. Chabriat, and Alzheimer's Disease Neuroimaging Initiative. Sulcal span in Alzheimer's disease, amnestic mild cognitive impairment, and healthy controls. *J. Alzheimer's Dis.*, 29(3):605–613, 2012.

[46] E. Salmon, P. Ruby, D. Perani, E. Kalbe, S. Laureys, S. Adam, and F. Collette. Two aspects of impaired consciousness in Alzheimer's disease. *Prog. Brain Res.*, 150:287–298, 2005.

[47] R. Schlösser, T. Gesierich, B. Kaufmann, G. Vucurevic, S. Hunsche, J. Gawehn, and P. Stoeter. Altered effective connectivity during working memory performance in schizophrenia: a study with fMRI and structural equation modeling. *NeuroImage*, 19(3):751–763, 2003.

[48] S. M. Smith, D. Vidaurre, C. F. Beckmann, M. F. Glasser, M. Jenkinson, K. L. Miller, T. E. Nichols, E. C. Robinson, G. Salimi-Khorshidi, M. W. Woolrich, D. M. Barch, K. Uğurbil, and D. C. Van Essen. Functional connectomics from resting-state fMRI. *Trends Cogn. Sci.*, 17(12):666–682, 2013.

[49] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014.

[50] N. Talebi, A. M. Nasrabadi, I. Mohammad-Rezazadeh, and R. Coben. nCREANN: Nonlinear causal relationship estimation by artificial neural network; applied for autism connectivity study. *IEEE Trans. Med. Imaging*, 38(12):2883–2890, 2019.

[51] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, and Wu-Minn HCP Consortium. The WU-Minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

[52] G. Varoquaux, A. Gramfort, F. Pedregosa, V. Michel, and B. Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Inf. Process. Med. Imaging*, pages 562–573, 2011.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Adv. Neural Inf. Process. Syst. 30*, pages 6000–6010, 2017.

[54] K. S. Vitanova, K. M. Stringer, D. P. Benitez, J. Brenton, and D. M. Cummings. Dementia associated with disorders of the basal ganglia. *J. Neurosci. Res.*, 97(12):1728–1741, 2019.

[55] K. Wang, T. Jiang, C. Yu, L. Tian, J. Li, Y. Liu, Y. Zhou, L. Xu, M. Song, and K. Li. Spontaneous activity associated with primary visual cortex: a resting-state FMRI study. *Cereb. Cortex*, 18(3):697–704, 2008.

[56] A. M. Wink. Eigenvector centrality dynamics from resting-state fMRI: gender and age differences in healthy subjects. *Front. Neurosci.*, 13(648), 2019.

[57] J. Yan, Y. Chen, Z. Xiao, S. Zhang, M. Jiang, T. Wang, T. Zhang, J. Lv, B. Becker, R. Zhang, D. Zhu, J. Han, D. Yao, K. M. Kendrick, T. Liu, and X. Jiang. Modeling spatio-temporal patterns of holistic functional brain networks via multi-head guided attention graph neural networks (Multi-Head GAGNNs). *Med. Image Anal.*, 80:102518, 2022.

[58] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proc. 32nd AAAI Conf. Artif. Intell.*, pages 7444–7452, 2018.

[59] R. Yang, W. Dai, C. Li, J. Zou, and H. Xiong. Tackling over-smoothing in graph convolutional networks with em-based joint topology optimization and node classification. *IEEE Trans. Signal Inf. Process. Netw.*, 9:123–139, 2023.

[60] D. Yao, J. Sui, E. Yang, P.-T. Yap, D. Shen, and M. Liu. Temporal-adaptive graph convolutional network for automated identification of major depressive disorder using resting-state fMRI. In *Int. Workshop Mach. Learn. Med. Imaging*, pages 1–10, 2020.

[61] Y. Yu, J. Chen, T. Gao, and M. Yu. DAG-GNN: DAG structure learning with graph neural networks. In *Proc. 36th Int. Conf. Mach. Learn.*, pages 7154–7163, 2019.

[62] Y. Yu, X. Kan, H. Cui, R. Xu, Y. Zheng, X. Song, Y. Zhu, K. Zhang, R. Nabi, Y. Guo, C. Zhang, and C. Yang. Learning task-aware effective brain connectivity for fmri analysis with graph neural networks. *arXiv preprint arXiv:2211.00261*, 2022.

[63] C. Zheng, B. Zong, W. Cheng, D. Song, J. Ni, W. Yu, H. Chen, and W. Wang. Robust graph representation learning via neural sparsification. In *Proc. 37th Int. Conf. Mach. Learn.*, pages 11458–11468, 13–18 Jul 2020.

[64] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Adv. Neural Inf. Process. Syst. 31*, pages 9472–9483, 2018.

[65] A. Zou and J. Ji. Learning brain effective connectivity networks via controllable variational autoencoder. In *2021 IEEE Int. Conf. Bioinform. Biomed. (BIBM)*, pages 284–287, 2021.

[66] A. Zou, J. Ji, M. Lei, J. Liu, and Y. Song. Exploring brain effective connectivity networks through spatiotemporal graph convolutional models. *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

Appendix

A Implementation Details

Following ST-GCN [14], we train ST-DAGCN on the sub-sequences sampled from rs-fMRI time series. Each sub-sequence starts at a random time frame within its corresponding rs-fMRI times series. In the test procedure, the final gender classification for a target subject is determined by the voting (i.e., the average sigmoid values) of its S randomly sampled sub-sequences' predictions. For the HCP dataset, we follow [14] to set the length of sub-sequence T' to 128 and the number of voters S to 64. For ADNI, we set $T' = 100$ and $S = 64$ that empirically achieve good performance.

The other hyperparameter settings are empirically tuned to achieve the best performance and are summarized as follows. For the augmented Lagrangian, we use $\beta = 10$ and $\gamma = 1/4$, same as [61]. The Adam optimizer with a learning rate of 0.001 and batch size of 64 is adopted to solve the subproblem (9). The balancing hyperparameter λ for sparsity-based regularization is 0.001. The dropout rate is set to 0.5 and the weight decay of L_2 regularization is set to 0.001.

B Evaluation Metrics and Scheme

Considering that gender prediction is a binary classification task, we adopt four common metrics to evaluate the prediction performance, i.e., accuracy (ACC), sensitivity (SEN), specificity (SPE), and the area under the receiver operating characteristic curve (AUC). To ensure the credibility of the classification results, 5-fold cross validation is implemented 5 times for each study and we report the mean and standard deviation of these evaluation metrics.

For effective connectivity learning, since there is no ground truth on the employed real-world rs-fMRI databases, we just provide qualitative analysis to verify the learned brain structures based on previous neuroscience research, similar to [14, 39, 26, 37, 38, 8]. Moreover, following [26, 38, 8], we perform group analysis on the ADNI database by estimating the effective connectivity of the CN and AD groups, respectively, and comparing their differences. We also verify some of the important differences according to the existing literature on studying Alzheimer's disease.

C Ablation study of Effective Connectivity

As our defined score function is based on the cross entropy between the model's predictions and the target labels, we provide the corresponding learning curves in Figure 3 to verify that our causal discovery procedure indeed outputs a brain effective connectivity that benefits rs-fMRI time series classification. By solving (9), the cross entropy curves of ST-DAGCN drop much more quickly to smaller values than the counterpart ST-GCN model that uses the correlation matrix as brain connectivity. This result further demonstrates that a more meaningful brain connectivity modeling can lead to better rs-fMRI classification performance.

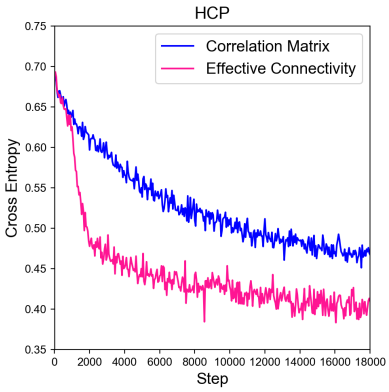

Figure 3: Cross entropy curves of ST-DAGCN in solving (9) and its counterpart model with brain connectivity set as the correlation matrix.

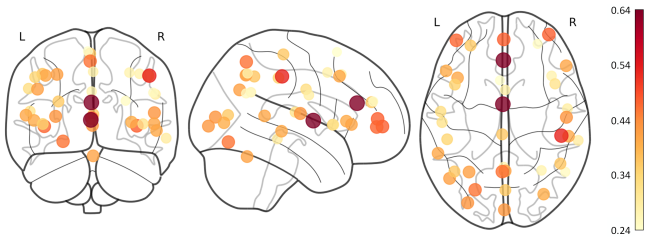

Figure 4: Node difference values by comparing AD versus CN.

Table 3: Top-10 salient brain ROIs in comparing AD versus CN.

ROI Index	ROI Name
14	Basal (basal ganglia)
24	V ACC (ventral anterior cingulate cortex)
39	R Ant IPS (right anterior intraparietal sulcus)
11	R Front pol (right frontal pole)
6	Front DMN (front default mode network)
17	L Front pol (left frontal pole)
34	Dors PCC (dorsal posterior cingulate cortex)
21	Vis (visual cortex)
27	R STS (right superior temporal sulcus)
20	L LOC (left lateral occipital cortex)

D Supplementary Results for Group Analysis on ADNI

To quantify the difference between the two estimated effective connectivity networks, we calculate the difference at both the node level and edge level. Specifically, given the brain graph adjacency matrices of AD and CN groups $\mathbf{A}_{\text{AD}}$ and $\mathbf{A}_{\text{CN}}$, the differences of every node and every edge are:

$$\begin{aligned}\text{Difference of node } v_i = &\sum\nolimits_{j=1}^{N} \left|\mathbf{A}_{\text{AD}}[i,j] - \mathbf{A}_{\text{CN}}[i,j]\right| \\ &+ \sum\nolimits_{j=1}^{N} \left|\mathbf{A}_{\text{AD}}[j,i] - \mathbf{A}_{\text{CN}}[j,i]\right|, \qquad (12)\end{aligned}$$

$$\text{Difference of edge } v_i \to v_j = \left|\mathbf{A}_{\text{AD}}[i,j] - \mathbf{A}_{\text{CN}}[i,j]\right|. \qquad (13)$$

We visualize the node difference values computed by Eq. (12) in Figure 4 and list the top-10 salient brain ROIs in Table 3. We name the selected ROIs as the regional biomarkers in identifying Alzheimer's disease. The top-5 and top-10 discriminative brain graph edges are plotted in Figure 5 based on Eq. (13), which are regarded as connectional biomarkers that differentiate between the AD and CN groups.

In addition to the biological discoveries illustrated in the main text, we have the following observations from group analysis: (1) We find from Figure 5 that most discriminative connections for differentiating AD from CN involve ventral ACC, which also verifies that ACC is an important biomarker in identifying AD. (2) Besides Basal and ACC, some other identified biomarkers in Table 3 have also proved to relate to the differences between AD and CN. For instance, Reiner et al. [45] found remarkable differences of cortical morphology in the IPS (node 39) between AD and CN. Finger et al. [11] demonstrated that the reduced cortical thickness in the R Front pol (node 11) is associated with the presence of disinhibition in Alzheimer's disease. Reduced metabolism in the PCC (node 34) and DMN (node 6) has been identified as a sign of Alzheimer's disease [5, 35]. (3) Aligned with previous study [41], our estimated effective connectivity networks illustrate that the activation pattern of Broca's area (node 29) in AD group is weaker, such as ventral ACC $\to$ Broca, which is inferred as the most discriminative connection with the largest edge difference value, as shown in Figure 5. (4) It is claimed that the metabolism in the temporoparietal junctions (TPJ, nodes 28 and 31) closely correlates with AD patients' cognitive deficits [46]. We find that the neural connections associated with TPJ also demonstrate some remarkable differences between AD and CN groups. For example, lesions in TPJ is demonstrated to relate to AD patients' impaired ability in visual

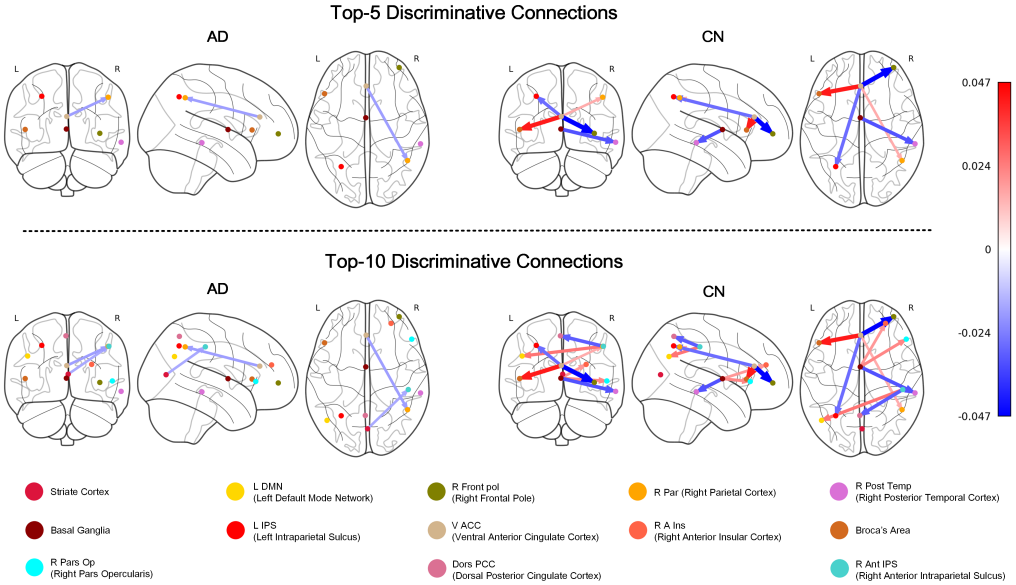

Figure 5: Discriminative connections by comparing AD versus CN.

perspective taking [46]. In Figure 2, we find strong connection from visual region to the right TPJ (i.e., Vis $\rightarrow$ R TPJ) in CN group while the corresponding link in AD group does not exist.

E Potential Future Work

ST-DAGCN outputs a single effective connectivity for a group of subjects based on the optimization problem in (4) and ST-DAGCN is separately run for AD and CN respectively for group-level inference. In fact, existing works also separately infer the effective connectivity of AD and CN groups but treat male and female without distinction [26, 38, 65]. We expect our paper can be further extended for jointly classifying Alzheimer's disease and inferring two effective connectivity networks for the AD and CN groups. A potential method is to additionally learn two bias networks for the two groups, respectively. We can first learn a shared effective connectivity network across the two groups and then separately add the two bias networks to the shared network to obtain the effective connectivity of each group. The two effective connectivity networks are fed back to the model to further train the network to improve the rs-fMRI time series classification performance. The model can be iteratively refined using the recycling strategy [30] that recursively computes the outputs and feeds the outputs into the model. We leave this work for future exploration.