CBT-GCL: LEVERAGING CONNECTIONAL BRAIN TEMPLATES FOR ENHANCED ASD DIAGNOSIS VIA GRAPH CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder with a rising global prevalence, making its early and precise diagnosis crucial. Although deep learning has shown great potential in the diagnosis of ASD with Functional Connectivity (FC) data, it still suffers from challenges, including substantial noise in raw FC data, redundant connections, and high inter-individual heterogeneity. In facing these challenges, the Connectional Brain Template (CBT), which represents a group of brain networks, can effectively mitigate them, thereby providing prior guidance for diagnosis. To this end, we propose a CBT-driven graph contrastive learning model, called CBT-GCL, which utilizes CBT as prior knowledge to assist ASD diagnosis. Specifically, our model concurrently trains two encoders to extract multi-level features from both the CBT and individual-specific FC data, subsequently fusing and enhancing these features. A perceptual loss function aligns hierarchical features between the two groups, while a contrastive loss function pulls features of positive samples closer and pushes features of negative samples apart. This effectively extracts critical connections from FC data, thereby enhancing ASD diagnostic performance. Experiments on the Autism Brain Imaging Data Exchange (ABIDE) I and II datasets demonstrate that CBT-GCL outperforms state-of-the-art models in multiple evaluation metrics. Ablation studies further confirm the effectiveness of each component of CBT-GCL.

1 INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disorder that typically emerges in early childhood and is characterized by impaired social communication and interaction, together with restricted or repetitive behaviors Edition et al. (2013). Its prevalence has increased in recent years and is estimated at approximately 1% worldwide Zeidan et al. (2022). Current clinical management mainly relies on behavioral intervention, speech therapy, and educational support, but no curative treatment is available. Moreover, substantial individual heterogeneity limits intervention efficacy and imposes long-term financial and psychological burdens on affected families Lord et al. (2018). Early, objective, and accurate diagnosis is therefore critical for timely intervention and improved outcomes.

Resting-state functional Magnetic Resonance Imaging (rs-fMRI) non-invasively captures spontaneous brain activity through blood oxygen level-dependent signals Biswal et al. (1995), providing an important route toward objective ASD diagnosis. After parcellating the brain into regions of interest, correlations between their time series form Functional Connectivity (FC) matrices that quantify intrinsic brain network organization Fox & Raichle (2007). Widespread FC abnormalities have been observed in ASD; consequently, FC helps reveal potential pathophysiological mechanisms and has become an important imaging biomarker for computer-aided diagnosis Menon (2011).

Machine learning has advanced automated rs-fMRI analysis Vieira et al. (2017), but conventional methods commonly flatten FC matrices into vectors, thereby obscuring graph topology and higher-order, non-local interactions among brain regions Bullmore & Sporns (2009). Graph Neural Networks (GNNs) are better suited to this modality because they propagate information and learn representations directly on brain graphs. Recent ASD studies distinguish connection importance with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

attention Dong & Sun (2024), improve representation robustness through contrastive learning Zhang et al. (2023), and model higher-order relations using hypergraphs Ji et al. (2022). Nevertheless, as shown in Fig. 1, raw FCs contain substantial noise and many redundant connections weakly associated with disease. Together with pronounced inter-subject heterogeneity, these factors continue to hinder stable discriminative representation learning Murphy & Fox (2017).

A Connectional Brain Template (CBT) is a statistically representative connectivity matrix learned from a population and captures stable patterns shared across its brain networks Bessadok et al. (2022). Used as a population prior, it provides a reference for identifying abnormal connections in an individual and may mitigate noise and heterogeneity. A previous approach registers individual FCs to a group CBT using a Generative Adversarial Network (GAN) and then feeds the registered representations to a GNN classifier Gürler et al. (2023). Although demonstrating the value of CBT for ASD diagnosis, the approach suffers from unstable and difficult-to-converge GAN training Goodfellow et al. (2020). Its registration and classification stages are also separated, potentially losing information and precluding optimization toward the diagnostic objective (Fig. 1). This leaves open how population priors can suppress noisy connections while preserving disease-specific individual variation, and how CBT information can be injected at multiple feature levels without forcing every subject toward the same connectivity pattern.

To address these limitations, we propose CBT-driven Graph Contrastive Learning (CBT-GCL), which uses population priors to guide individual FC representation learning (Fig. 1). DGN Gurbuz & Rekik (2020) first generates category-specific CBTs from the training data, and each individual FC is preliminarily fused with templates. Two encoders then extract multi-level graph features from the individual FC and CBT. Semantic Feature Alignment (SFA) aligns the two feature streams at multiple semantic levels, encouraging the individual encoder to suppress noisy and redundant connections. Contrastive Feature Enhancement (CFE) further fuses their highest-level features, treats original and enhanced representations as positive pairs, and uses representations from different classes as negatives. Perceptual Johnson et al. (2016), contrastive, and Binary Cross-Entropy losses jointly optimize feature alignment, representation enhancement, and diagnostic classification.

Unlike staged CBT registration, CBT-GCL does not transform individual networks toward a fixed template before classification. Instead, CBT participates in representation learning through feature-level alignment and enhancement, allowing population regularities to guide the encoder while retaining sample-specific information. With the precomputed CBTs fixed, the entire pipeline is jointly optimized for diagnosis. The main techcontributions of this work are summarized as follows:

- We introduce an CBT-guided ASD diagnostic paradigm that directly integrates population-level connectivity priors into individual representation learning, avoiding information loss and suboptimal optimization in staged pipelines.
- We propose SFA to align individual FC and category-template representations across multiple graph levels, transferring stable population semantics to the individual encoder while preserving subject-specific information.
- We develop a CBT-conditioned contrastive feature enhancement mechanism that fuses individual FC and CBT features into a learnable subject-specific augmented view. A contrastive learning approach is then applied between the original representation and the augmented representation.

We conducted extensive experiments on the publicly available ABIDE I and II datasets. CBT-GCL achieves superior performance across accuracy, sensitivity, specificity, and AUC. Ablation studies and parameter sensitivity analyses further verify the contributions of SFA, CFE, and the CBT prior, demonstrating the effectiveness and robustness of the proposed framework.

2 Related Work

ASD Diagnosis. Clinical ASD diagnosis relies primarily on behavioral assessment, whose subjectivity, sensitivity to comorbidities, and limited ability to identify patients before symptoms fully manifest have motivated objective neuroimaging biomarkers Lord et al. (2000); Loth et al. (2017). Machine learning has shifted ASD research toward individualized prediction, yet heterogeneity, cross-site variation, and limited samples still restrict generalization. Several studies therefore in-

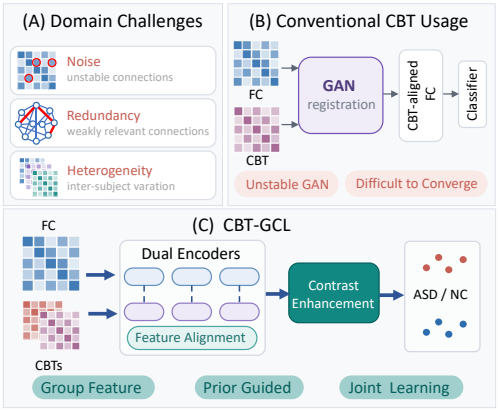

Figure 1: (A) The original FC face challenges, including heavy noise in fMRI data, excessive redundant connections, and strong individual heterogeneity. (B) Existing CBT-based methods suffer from unstable GAN training and may dilute individual-specific diagnostic information. (C) CBT-GCL leverages population-level prior information derived from CBT and performs joint optimization of multi-level features to effectively address these issues while improving ASD diagnostic accuracy.

troduce standardized connectivity templates derived from larger populations to constrain relevant nodes, edges, or subnetworks and improve the biological consistency and reproducibility of learned features Chaari et al. (2023); Gürler et al. (2023); Finn et al. (2015). Existing template-guided approaches typically use population template as a registration or normalization target Gürler et al. (2023), confining template guidance largely to preprocessing and potentially attenuating diagnostically relevant individual variation. In contrast, CBT-GCL construct CBTs and focuses on using them as fixed priors for downstream representation learning, preserving subject-specific information while exploiting stable population patterns.

Connectional Brain Template. CBT integrates population connectomes into a representative atlas of shared connectivity patterns. Early averaging or diffusion-based approaches often neglect brain topology and nonlinear relations among connectomes Zuo et al. (2012). MGINet generates and fuses subgroup templates through hierarchical clustering, but its multi-stage pipeline may accumulate errors Demir et al. (2020). DGN instead uses Edge-Conditioned Convolution and minimizes the Frobenius distance between the template and training connectomes, enabling end-to-end multi-view CBT learning Gurbuz & Rekik (2020). A useful CBT should be representative and topologically consistent. As a population reference, it can reveal disease-specific abnormalities by comparison with individual connectomes Finn et al. (2015) and generate synthetic connectomes for data augmentation through local perturbations Demirbilek et al. (2023).

Contrastive Learning. Contrastive learning builds discriminative representations by modeling sample similarity and dissimilarity: semantically consistent augmented views are treated as positives and pulled together, while negatives are pushed apart Chen et al. (2020). Its effectiveness depends on whether pair construction preserves task semantics; inappropriate brain graph augmentations may alter clinically meaningful connections. For brain network classification, HSGPL combines hierarchical signed graph representations with a contrastive objective Tang et al. (2022), while ContrastPool uses contrastive dual attention and differentiable pooling to capture discriminative fMRI substructures Xu et al. (2024). However, these methods mainly construct relations between individual samples or augmented views, without fully exploiting the stable population patterns encoded by CBT.

3 METHODOLOGY

The proposed CBT-GCL is illustrated in Fig. 2. For the original fMRI data, preprocessing is first performed to obtain standardized graphs. We employ DGN for Category-specific CBT Construction. After obtaining the FC and CBT samples, we conduct preliminary preprocessing, followed by Dual Feature Extraction and Semantic Feature Alignment. For each training sample, we train two encoders to extract multi-scale features from the sample itself and the corresponding category's

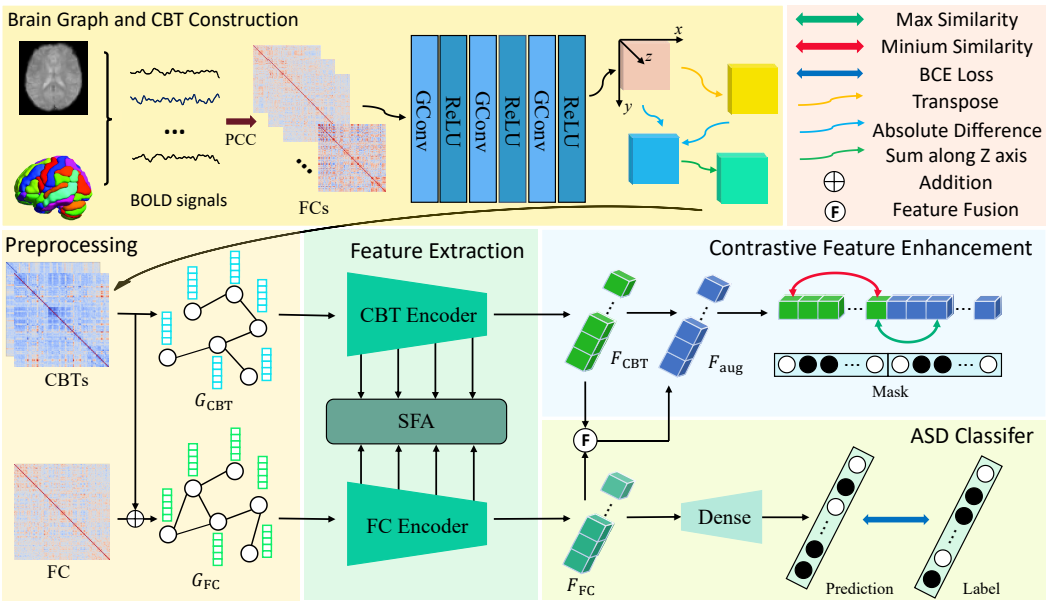

Figure 2: An overview of proposed CBT-GCL. For all raw data, it is first converted into a graph and then construct Category-specific CBTs based on the training set. For each training sample, first fuse it with the both category-specific CBTs preliminarily, then train two encoders to extract multi-scale features from them, respectively. Next, we introduce CFE module to enhance feature representation. Finally, the diagnostic results are obtained using a fully connected layer. During the testing phase, only the FC encoder is used to extract features for subsequent diagnosis.

CBT, respectively. Next, the SFA module is used to introduce CBT priors into the FC features. Subsequently, the CFE module is utilized to fully integrate CBT information. Finally, we employ a fully connected layer to obtain diagnostic results and optimize model performance using multiple loss functions.

3.1 Brain Graph and CBT Construction

Brain Graph Construction. We construct the brain functional connectivity network as follows. For all original fMRIs, a series of preprocessing steps are first applied to obtain standardized brain images. Subsequently, the entire brain is parcellated into multiple regions of interest (ROIs) using a given brain atlas. All ROIs are regarded as the node set V, while the functional connection between any pair of ROIs is considered an edge e between these two nodes; all edges collectively form the set E. To compute the functional connectivity strength, we first extract the BOLD signals of all voxels within each ROI and compute their average time series. The Pearson correlation coefficient is then used to quantify the connectivity strength between the time series of each pair of ROIs. We use the fractional Amplitude of Low-Frequency Fluctuation (fALFF) as the node feature matrix $X = [x_1, x_2, ..., x_n]$, n is the number of nodes.

Category-Specific CBT Construction. We generate category-based CBT prior information from training set, extracting common connectivity patterns within groups for ASD diagnosis. Specifically, we employ the first geometric deep learning CBT generation architecture, DGN, whose core model utilizes a three-layer edge-conditioned graph convolutional network to dynamically learn the mapping relationship between edge features and node embeddings. The convolutional operation at the l-th layer is defined as:

$$x_i^{(l)} = \frac{1}{|\mathcal{N}(i)|+1}(x_i^{(l-1)} + \sum_{j \in \mathcal{N}(i)} F^{(l)}(e_{i,j}, W^{(l)}) x_j^{(l-1)}, \tag{1}$$

where $F^{(l)}(\cdot)$ denotes the filter generation network parameterized by weights $W^{(l)}$, and $\mathcal{N}(i)$ represents the set of neighboring nodes of node v_i. After passing through multiple layers, the model generates the final node embedding $X^{(l)}$.

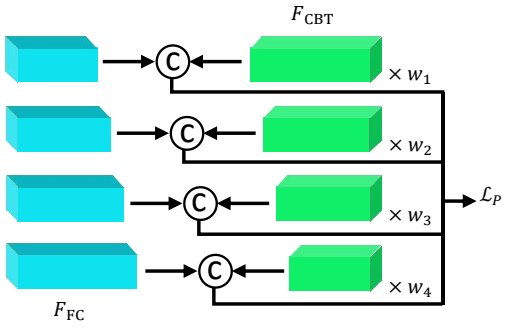

Figure 3: Specific internal structure of the SFA module, "ⓒ" stands for calculating cosine similarity.

During the CBT generation phase, $X^{(l)}$ is replicated t times along the horizontal axis to form the input tensor. After transposing this tensor, the absolute differences are calculated and summed along the embedding dimension, ultimately yielding the CBT C. To optimize the representativeness of CBT, the model employs the Subject Normalization Loss (SNL):

$$\text{SNL}_s = \sum_{v=1}^{n_v} \sum_{i \in \mathcal{S}} \| C_s - T_i^{(v)} \| \cdot \lambda_v, \tag{2}$$

where $\mathcal{S}$ represents a random subset of the training set, T is the individual brain network of the subject and λ_v is used to balance multi-view scale differences. To eliminate input individual biases, the entire training set is fed into the model to generate the CBT set $\mathcal{C}_{\text{train}}$. Finally, by taking the element-wise median in $\mathcal{C}_{\text{train}}$, the most core connections are selected to generate $\mathcal{C} = \{\mathcal{C}_{\text{ASD}}, \mathcal{C}_{\text{NC}}\}$ for training.

For each data split, two DGN models are trained separately using only the ASD and NC subjects in the training partition, producing the training-specific category templates $C_{\text{ASD}}^{\text{tr}}$ and $C_{\text{NC}}^{\text{tr}}$. Subjects in the validation and test partitions are excluded from both DGN optimization and final CBT aggregation. Once constructed, the two CBTs are frozen for the corresponding split. For every subject in the training, validation, and test partitions, both CBTs are added simultaneously to the individual FC without using the subject's category label:

$$E_i' = E_i + C_{\text{ASD}}^{\text{tr}} + C_{\text{NC}}^{\text{tr}}. \tag{3}$$

Thus, CBT construction is strictly nested within the training partition, while the simultaneous use of both fixed templates avoids label-dependent template selection at inference. We then concatenate the current node feature x_i with the fused edge weights $e'_{i,j}$ connecting node i to all other nodes to construct the augmented node feature:

3.2 Dual Feature Extraction and Semantic Feature Alignment

Dual Feature Extraction. For the input data $G = (V, E, X)$, we first generate and load the class-specific CBTs using the two types of data from the training set, and preliminarily fuse them with G to obtain a new edge set E'. Without using their category labels, we simultaneously add both types of CBT and FC together. Since edge features in brain connectivity networks are significantly more important than node features, we further integrate the current node feature x_i with the edge values $e_{i,j}$ of all other nodes to form additional node features, namely:

$$x_i' = \text{Concat}(x_i, e_{i,1}, e_{i,2}, ..., e_{i,n}), \tag{4}$$

where $\text{Concat}(\cdot)$ is concatenation function. The processed node feature X' together with E' forms a new graph $G' = (V, E', X')$.

In the feature extraction stage, we train two encoders, where En_{FC} is used to extract features from individual data, and En_{CBT} is used to extract features from CBT, both sharing the same network architecture. Taking En_{FC} as an example, the model consists of four layers of graph convolution. For the input graph G', after each convolution layer, a node-level feature embedding is generated:

$$h_v^{(k)} = \sigma(\text{BN}(\text{gConv}(X', E'))), \tag{5}$$

where $h_v^{(k)}$ represents the node v at the k-th layer, $\sigma(\cdot)$ is the activation function, $\mathrm{BN}(\cdot)$ is batch normalization, and $\mathrm{gConv}(\cdot)$ is graph convolution. We choose a GIN variant with weighted edge feature enhancement (wGIN) as the convolutional layer of the encoder. wGIN achieves progressive differentiation of graph structures through multi-layer message passing, introducing explicit handling of edge weights to more deeply utilize edge information. Its core principle is:

$$h_v^{(k)} = \mathrm{MLP}^{(k)}\left(h_v^{(k-1)} + \sum_{u \in \mathcal{N}(v)} h_u^{(k-1)} \cdot e_{v,u}\right), \tag{6}$$

where $\mathrm{MLP}^{(k)}(\cdot)$ is a multilayer perceptron, enabling the model to learn complex feature mapping relationships. It means that during aggregation, edge features are used as weights to multiply with neighbor features, enabling the model to adaptively learn the importance distribution of edge weights on neighbor features, which is suitable for brain graph learning tasks that are sensitive to edge information.

After obtaining all the node representations $h^{(k)}$, we use global pooling to derive the graph-level feature embedding $f^{(k)}$.

Semantic Feature Alignment. During the training phase, we obtain two sets of features, F_{FC} and F_{CBT}, using the encoders. To enable the sample encoder to incorporate prior information from the CBT while extracting features, we aim to align the sample features closely with the CBT features. For this purpose, we employ a perceptual loss function to constrain the two feature sets. Unlike Mean Squared Error (MSE), which operates only at the pixel level and fails to capture high-level semantic characteristics, the perceptual loss function measures discrepancies in a high-dimensional feature space, thereby promoting alignment at the semantic level. As shown in Fig. 3, we normalize the two sets of features generated by the encoder, compute the cosine similarity between features at each level, and finally aggregate the similarities across all levels using a weighted sum. The entire calculation process of perceptual loss $\mathcal{L}_{\mathrm{P}}$ can be described as:

$$\mathcal{L}_{\mathrm{P}} = \frac{1}{K}\sum_{i}^{K}\left(1 - w \cdot \mathrm{Sim}(\mathrm{BN}(f_{\mathrm{FC}}^{(\mathrm{i})}), \mathrm{BN}(f_{\mathrm{CBT}}^{(\mathrm{i})}))\right), \tag{7}$$

where, K is the depth of convolution, w is the weights of sum, $\mathrm{Sim}(\cdot)$ is cosine similarity.

3.3 Contrastive Feature Enhancement

To fully exploit the information provided by CBT, we integrate CBT features with sample features to generate enhanced features. Subsequently, via a contrastive loss function, the distance between these enhanced features and other features of the same class (regarded as positive pairs) is minimized, while the distance between them and features of different classes (regarded as negative pairs) is maximized. This process encourages the model to learn more discriminative and globally informative feature representations. Concretely, we first concatenate the top-level components of the two sets of features and then fuse them through a dense layer, forming an enhanced feature representation f_{aug} that incorporates both the original sample information and the global information from CBT:

$$f_{\mathrm{aug}} = \mathrm{Dense}(\mathrm{Concat}(f_{\mathrm{FC}}, f_{\mathrm{CBT}})). \tag{8}$$

Next, the original features and the enhanced features are concatenated along the batch dimension, constructing a feature set with twice the batch size, and the cosine similarity matrix R among all normalized features is computed:

$$R = \mathrm{Sim}(\mathrm{Concat}(f, f_{\mathrm{aug}})). \tag{9}$$

Afterwards, a positive sample mask that identifies pairs of samples from the same class excluding the sample itself is constructed based on the class labels:

$$M_{i,j} = \begin{cases} 1, & \text{if } y_i = y_j \text{ and } i \neq j \\ 0, & \text{otherwise} \end{cases}, \tag{10}$$

where y is the label of the sample. For each sample, its loss value is the ratio of the sum of similarities of all positive sample pairs to the sum of similarities between that sample and all other samples in

Table 1: Statistical distribution of the data used in the experiment.

Dataset	Information	ASD	NC
ABIDE I	Subjects	531	571
	Gender (M/F)	467/64	472/99
	Age (Mean ± std)	17.0±8.4	17.1±7.7
ABIDE II	Subjects	344	402
	Gender (M/F)	291/53	266/136
	Age (Mean ± std)	14.3±10.1	13.9±10.1

the batch (excluding itself), divided by the number of positive samples after taking the negative logarithm.

$$\mathcal{L}_{\mathrm{C}}^{(i)} = -\frac{1}{P_i}\sum_{j=1}^{2B} M_{i,j}\log\left(\frac{\exp(R_{i,j}/\tau)}{\sum_{k\neq i}^{2B}\exp(R_{i,k}/\tau)}\right), P_i = \sum_{j=1}^{2B} M_{i,j}. \tag{11}$$

where B is the batch size, P_i is the number of positive samples for anchor i, and τ is the temperature coefficient. The final contrastive loss is obtained by averaging the loss over all samples in the batch. This process can be formulated as:

$$\mathcal{L}_{\mathrm{C}} = \frac{1}{2B}\sum_{i=1}^{2B}\mathcal{L}_{\mathrm{C}}^{(i)}. \tag{12}$$

3.4 ASD Classifier and Loss Function

To obtain the diagnostic results, we employ a dense layer to transform the FC feature embeddings into probability values. Since some original information may be lost during the process of acquiring multi-level features, we concatenate the highest-level features with the lowest-level features before performing the diagnosis:

$$p = \mathrm{Dense}(\mathrm{Concat}(f^{(1)}, f^{(4)})). \tag{13}$$

BCE Loss is a widely used loss function in binary classification tasks. It operates by measuring the discrepancy between the model's predicted probabilities and the true labels. The value increases rapidly when the true label deviates significantly from the predicted outcome. In this paper, we employ BCE loss to provide fundamental label constraints for the model, encouraging it to generate diagnostic results that closely approximate the ground truth. Its calculation can be expressed as:

$$\mathcal{L}_{\mathrm{BCE}} = -[gt\cdot\log(p) + (1-gt)\cdot\log(1-p)], \tag{14}$$

where gt represents the ground truth label of the sample.

The complete loss function can be expressed as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{BCE}} + \mathcal{L}_{\mathrm{P}} + \mathcal{L}_{\mathrm{C}}. \tag{15}$$

3.5 Implementation Details

The technical implementation of this study is based on the PyTorch 2.5 framework and the PyTorch Geometric (PyG) library for graph neural networks. The encoder consists of 4 graph convolutional layers, with an output feature embedding dimension of 32. The Adam optimizer was employed during training with a fixed learning rate of 0.0001. The batch size was set to 32, and optimization was carried out over 200 epochs.

4 Experiments

4.1 Datasets

We used the public multi-site rs-fMRI datasets ABIDE I and II, selecting 1,102 and 746 samples, respectively; their statistics are reported in Table 1. For ABIDE I, we utilized publicly available CPAC-derived preprocessed data, while ABIDE II was preprocessed using the DPARSF toolbox. Detailed preprocessing parameters are provided in the Supplementary Material. Both datasets were parcellated into 116 ROIs based on the AAL atlas. Samples from all sites were randomly split into training, validation, and test sets at 70%, 20%, and 10%.

Table 2: Quantitative results of ASD diagnosis on the ABIDE I and ABIDE II datasets.

Methods	ABIDE I					ABIDE II				
	Accuracy	AUC	Sensitivity	Specificity	Average	Accuracy	AUC	Sensitivity	Specificity	Average
MLP	0.586	0.565	0.471	0.693	0.578	0.610	0.580	0.528	0.683	0.600
GCN	0.600	0.596	0.736	0.474	0.601	0.623	0.598	0.583	0.659	0.616
GIN	0.600	0.651	0.613	0.588	0.613	0.610	0.639	0.667	0.561	0.619
GATE	0.598	0.597	0.645	0.549	0.597	0.623	0.620	0.658	0.582	0.621
BrainNetCNN	0.613	0.627	0.648	0.565	0.613	0.600	0.642	0.571	0.643	0.614
FBNetGen	0.621	0.703	0.621	0.619	0.641	0.617	0.650	0.619	0.644	0.632
BNT	0.641	0.706	0.680	0.600	0.657	0.643	0.687	0.563	0.711	0.651
BrainGNN	0.672	0.716	0.718	0.618	0.681	0.640	0.680	0.676	0.608	0.651
VanillaTF	0.673	0.723	0.678	0.659	0.683	0.601	0.628	0.555	0.674	0.614
HM-AGPC	0.691	0.730	0.523	**0.816**	0.690	0.693	0.655	0.420	**0.926**	0.673
DHGFormer	0.692	0.702	0.730	0.648	0.693	0.681	0.695	0.670	0.724	0.692
DIBrain	0.693	0.709	0.602	0.788	0.698	0.687	0.708	0.638	0.728	0.690
CBT-GCL	**0.718**	**0.739**	**0.745**	0.693	**0.723**	**0.714**	**0.711**	**0.722**	0.707	**0.713**

Table 3: The ablation experiment results investigating the impact of different modules on model performance, where "Baseline" represents not using any CBT prior guidance and only optimizing diagnostic performance with the BCE loss function.

Baseline	SFA	CFE	Accuracy	AUC	Sensitivity	Specificity	Average
✓			0.654	0.680	0.613	0.693	0.660
✓	✓		0.682	0.698	0.679	0.684	0.686
✓		✓	0.668	0.706	0.613	**0.719**	0.676
✓	✓	✓	**0.718**	**0.739**	**0.74**	0.693	**0.723**

4.2 Competing Methods

We compared CBT-GCL with three baseline groups: a four-layer MLP; standard GCN Kipf (2016) and GIN Xu et al. (2018); and ASD models including BrainGNN Li et al. (2021), GATE Peng et al. (2022), FBNetGen Kan et al. (2022a), BrainNetCNN Kawahara et al. (2017), VanillaTF and BNT Kan et al. (2022b), HM-AGPC Xu et al. (2025), DHGFormer Xue et al. (2025) and DIBrain Hu et al. (2026). MLP and GNN baselines used the same training settings as CBT-GCL, while published settings were retained for the other models. We additionally evaluated ABIDE II using an ABIDE I-derived CBT to assess transfer across populations.

4.3 Experimental Results

Table 2 shows that the proposed method attained the best classification results across most metrics. Compared to the second-best performing method, DHGFormer, CBT-GCL achieved improvements of 2.6% in accuracy, 3.7% in AUC, 1.5% in sensitivity and 4.5% in specificity. These results demonstrate the superiority of our approach over existing state-of-the-art methods and its ability to enhance the accuracy of ASD diagnosis, similar results were also obtained on ABIDE II. As shown in Fig. 5 (a), the datasets exhibit significant class imbalance. This causes many models to easily overfit to a particular category, resulting in a substantial disparity between sensitivity and specificit, whitch is clinically unacceptable. In contrast, CBT-GCL not only outperforms comparative methods in overall performance, but also achieves a more balanced trade-off between sensitivity and specificity.

4.4 Ablation Study

To validate the effectiveness of each component in the proposed model, extensive ablation studies and parameter sensitivity analyses on ABIDE I were conducted.

Key Modules. To evaluate the effectiveness of the SFA and CFE modules, we adopted the baseline model trained solely with the standard BCE loss and compared its performance against variants incorporating different modules. As shown in Table 3, using SFA or CFE alone improved the average score from 0.660 to 0.686 and 0.676, respectively, while combining both modules achieved a score

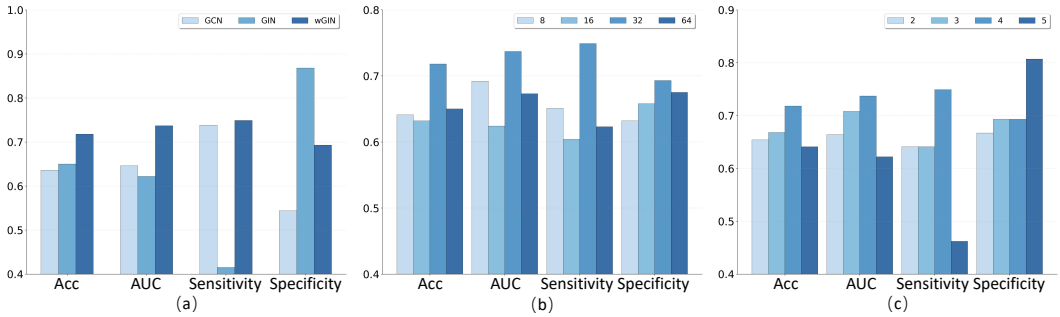

Figure 4: Parameter sensitivity to (a) graph convolution, (b) embedding dimension, and (c) network depth.

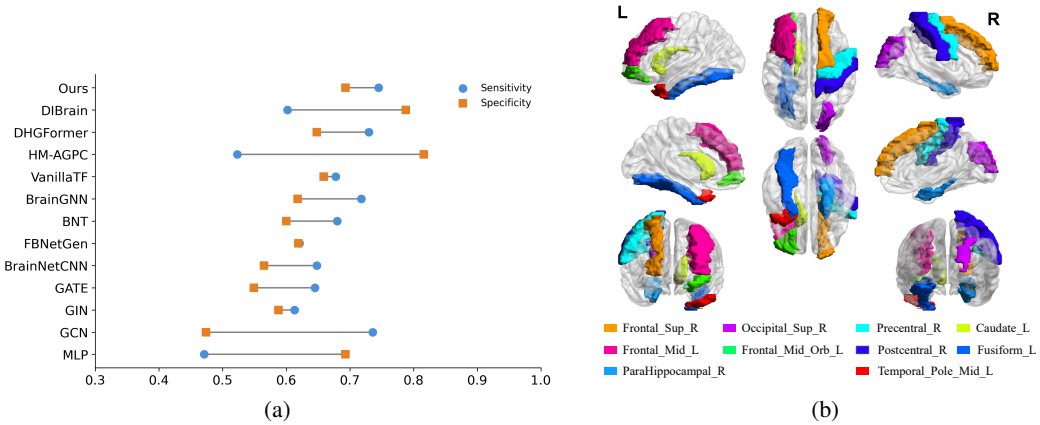

Figure 5: Visualization of CBT-GCL: (a) its ability to overcome data imbalance; (b) the ten brain regions with the greatest impact on ASD diagnosis across two datasets.

of 0.723. These results demonstrate that either the SFA or CFE module was used individually, the model obtained varying degrees of prior information, leading to some improvement in performance. When used together, they also demonstrate complementary gain effects.

Parameter Sensitivity. As shown in Fig. 4, wGIN performs best, supporting explicit use of edge weights for brain graphs. A 32-dimensional embedding provides the best trade-off between representation capacity and redundancy. Performance improves from two to four convolutional layers but drops at five, likely because of over-smoothing.

4.5 Interpretation

To identify key brain regions associated with ASD, we further conducted interpretability analysis on the input brain maps using CBT-GCL. We ranked ROIs by the MSE between the mean final-layer node embeddings of the ASD and NC groups. Fig. 5 (b) visualizes the five highest-ranked ROIs from each dataset. Occipital_Sup_R and Frontal_Sup_R recur across datasets and are consistent with previous ASD findings Zhang et al. (2023). This experiment demonstrates that CBT-GCL can effectively extract key biomarkers for ASD diagnosis from brain maps.

5 Conclusion

This paper presents CBT-GCL, which uses category-specific CBTs, multi-level semantic alignment, and contrastive enhancement to guide individual FC representation learning. Experiments on ABIDE I and II show improved and more balanced ASD classification, while ablations support the complementary contributions of SFA and CFE. Future work will evaluate site-robust generalization and extend CBT-guided learning to additional disorders and imaging modalities.

AI USE STATEMENT

In this work, we used generative AI tools for polishing this paper during the writing process. We have not used generative AI tools for reference search, research idea conceptualization, manuscript drafting, and mathematical proofs, and generating synthetic datasets are not applicable to this work. We have reviewed all AI-assisted work. We manually checked to ensure that the text generated by the LLM contained no exaggerations or fabrications. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alaa Bessadok, Mohamed Ali Mahjoub, and Islem Rekik. Graph neural networks in network neuroscience. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5833–5848, 2022.

Bharat Biswal, F Zerrin Yetkin, Victor M Haughton, and James S Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, 34(4):537–541, 1995.

Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186–198, 2009.

Nada Chaari, Hatice Camgöz Akdağ, and Islem Rekik. Comparative survey of multigraph integration methods for holistic brain connectivity mapping. *Medical Image Analysis*, 85:102741, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Uğur Demir, Mohammed Amine Gharsallaoui, and Islem Rekik. Clustering-based deep brain multigraph integrator network for learning connectional brain templates. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pp. 109–120. Springer, 2020.

Oytun Demirbilek, Islem Rekik, Alzheimer's Disease Neuroimaging Initiative, et al. Predicting the evolution trajectory of population-driven connectional brain templates using recurrent multigraph neural networks. *Medical Image Analysis*, 83:102649, 2023.

Changxu Dong and Dengdi Sun. Brain network classification based on dynamic graph attention information bottleneck. *Computer Methods and Programs in Biomedicine*, 243:107913, 2024.

Fifth Edition et al. Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc*, 21 (21):591–643, 2013.

Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*, 18(11):1664–1671, 2015.

Michael D Fox and Marcus E Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature reviews neuroscience*, 8(9):700–711, 2007.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Mustafa Burak Gurbuz and Islem Rekik. Deep graph normalizer: A geometric deep learning approach for estimating connectional brain templates. In *International conference on medical image computing and computer-assisted intervention*, pp. 155–165. Springer, 2020.

Zeynep Gürler, Mohammed Amine Gharsallaoui, Islem Rekik, Alzheimer's Disease Neuroimaging Initiative, et al. Template-based graph registration network for boosting the diagnosis of brain connectivity disorders. *Computerized Medical Imaging and Graphics*, 103:102140, 2023.

Hao Hu, Rundong Xue, Shaoyi Du, Xiangmin Han, Jingxi Feng, Zeyu Zhang, Wei Zeng, Yue Gao, and Juan Wang. Exploring dynamic interpretable brain networks via hierarchical graph transformer. *Pattern Recognition*, pp. 113371, 2026.

Junzhong Ji, Yating Ren, and Minglong Lei. Fc–hat: Hypergraph attention network for functional brain network classification. *Information Sciences*, 608:1301–1316, 2022.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Xuan Kan, Hejie Cui, Joshua Lukemire, Ying Guo, and Carl Yang. Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation. In *International conference on medical imaging with deep learning*, pp. 618–637. PMLR, 2022a.

Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. Brain network transformer. *Advances in Neural Information Processing Systems*, 35:25586–25599, 2022b.

Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

TN Kipf. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Xiao Li, Jihong Qi, Lei Yi, Mo Xu, Xiaoyu Zhang, Qiang Zhang, and Yige Tang. Hydrochemical characteristics and evolution of geothermal waters in the eastern himalayan syntaxis geothermal field, southern tibet. *Geothermics*, 97:102233, 2021.

Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H Cook Jr, Bennett L Leventhal, Pamela C DiLavore, Andrew Pickles, and Michael Rutter. The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, 30(3):205–223, 2000.

Catherine Lord, Mayada Elsabbagh, Gillian Baird, and Jeremy Veenstra-Vanderweele. Autism spectrum disorder. *The lancet*, 392(10146):508–520, 2018.

Eva Loth, Tony Charman, Luke Mason, Julian Tillmann, Emily JH Jones, Caroline Wooldridge, Jumana Ahmad, Bonnie Auyeung, Claudia Brogna, Sara Ambrosino, et al. The eu-aims longitudinal european autism project (leap): design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders. *Molecular autism*, 8(1):24, 2017.

Vinod Menon. Large-scale brain networks and psychopathology: a unifying triple network model. *Trends in cognitive sciences*, 15(10):483–506, 2011.

Kevin Murphy and Michael D Fox. Towards a consensus regarding global signal regression for resting state functional connectivity mri. *Neuroimage*, 154:169–173, 2017.

Liang Peng, Nan Wang, Jie Xu, Xiaofeng Zhu, and Xiaoxiao Li. Gate: Graph cca for temporal self-supervised learning for label-efficient fmri analysis. *IEEE Transactions on Medical Imaging*, 42(2):391–402, 2022.

Haoteng Tang, Guixiang Ma, Lei Guo, Xiyao Fu, Heng Huang, and Liang Zhan. Contrastive brain network learning via hierarchical signed graph pooling model. *IEEE transactions on neural networks and learning systems*, 35(6):7363–7375, 2022.

Sandra Vieira, Walter HL Pinaya, and Andrea Mechelli. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, 74:58–75, 2017.

Jiakun Xu, Xin Zhang, Tong Xiong, Shengxian Chen, Xiaofen Xing, Jindou Hao, and Xiangmin Xu. Heterogeneous masked attention-guided path convolution for functional brain network analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 383–393. Springer, 2025.

Jiaxing Xu, Qingtian Bian, Xinhang Li, Aihu Zhang, Yiping Ke, Miao Qiao, Wei Zhang, Wei Khang Jeremy Sim, and Balázs Gulyás. Contrastive graph pooling for explainable classification of brain networks. *IEEE Transactions on Medical Imaging*, 43(9):3292–3305, 2024.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Rundong Xue, Hao Hu, Zeyu Zhang, Xiangmin Han, Juan Wang, Yue Gao, and Shaoyi Du. Dhg-former: Dynamic hierarchical graph transformer for disorder brain disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 268–278. Springer, 2025.

Jinan Zeidan, Eric Fombonne, Julie Scorah, Alaa Ibrahim, Maureen S Durkin, Shekhar Saxena, Afiqah Yusuf, Andy Shih, and Mayada Elsabbagh. Global prevalence of autism: A systematic review update. *Autism research*, 15(5):778–790, 2022.

Shengjie Zhang, Xiang Chen, Xin Shen, Bohan Ren, Ziqi Yu, Haibo Yang, Xi Jiang, Dinggang Shen, Yuan Zhou, and Xiao-Yong Zhang. A-gcl: Adversarial graph contrastive learning for fmri analysis to diagnose neurodevelopmental disorders. *Medical Image Analysis*, 90:102932, 2023.

Xi-Nian Zuo, Ross Ehmke, Maarten Mennes, Davide Imperati, F Xavier Castellanos, Olaf Sporns, and Michael P Milham. Network centrality in the human functional connectome. *Cerebral cortex*, 22(8):1862–1875, 2012.

A Resting-State fMRI Preprocessing and FC Construction

For ABIDE I, we used the publicly released CPAC-preprocessed derivatives. The detailed settings below apply exclusively to ABIDE II, which was locally preprocessed using DPARSF V5.5_250415 implemented in DPABI. The first 10 volumes were discarded to allow the MR signal to reach a steady state. Slice-timing correction was not performed. The remaining volumes underwent rigid-body realignment, T1-to-functional coregistration, tissue segmentation with MNI affine regularization, and DARTEL-based normalization. Functional images were normalized to the MNI space with a voxel size of $3\times3\times3\ \mathrm{mm}^3$ and smoothed using a Gaussian kernel with a full width at half maximum of 4 mm.

Nuisance regression was performed after realignment and included a first-order polynomial trend, the Friston 24-parameter head-motion model, the global signal, mean white-matter signal, and mean cerebrospinal-fluid signal. White-matter and cerebrospinal-fluid masks were obtained from the SPM tissue probability maps using a probability threshold of 0.99. The analysis used the global-signal-regressed outputs stored in the `FunImgRglobalCWSF` directory. Motion scrubbing was not performed. Temporal band-pass filtering was applied within 0.01–0.10 Hz after normalization. The ABIDE II preprocessing workflow is summarized in Table 4.

The AAL atlas was used to define 116 regions of interest (ROIs) for both datasets. For ABIDE II, DPARSF extracted the mean BOLD time series within each ROI, whereas the FC matrices were subsequently constructed outside the DPARSF FC module. For subject i, the edge between ROIs p and q was defined as

$$A_i(p,q) = \mathrm{corr}\left(r_{i,p}, r_{i,q}\right), \tag{16}$$

where $r_{i,p}$ is the mean BOLD time series of ROI p. The resulting Pearson correlation coefficients were used directly, without a Fisher z transformation. ROI-level fractional amplitude of low-frequency fluctuations (fALFF) values were used as the initial node features.

Table 4: DPARSF preprocessing and FC construction settings used for ABIDE II.

Component	Setting	Details
Software	DPARSF V5.5_250415	Implemented in DPABI; one functional session.
Acquisition configuration	156 volumes; TR = 2 s	One functional run was used for each subject.
Initial volumes	10 volumes removed	The remaining volumes were processed.
Slice timing	Disabled	No slice-timing correction was applied.
Motion correction	Enabled	Rigid-body realignment followed by Friston 24-parameter regression.
Anatomical processing	Enabled	T1-to-functional coregistration and tissue segmentation with MNI affine regularization.
Spatial normalization	DARTEL	MNI space, $3 \times 3 \times 3\ \mathrm{mm}^3$ voxels.
Nuisance regression	Enabled after realignment	First-order polynomial trend, global, white-matter, cerebrospinal-fluid, and head-motion signals.
Tissue masks	SPM masks	White-matter and cerebrospinal-fluid probability threshold = 0.99; mean-signal regression.
Scrubbing	Disabled	No censoring or spike regressors were used.
Temporal filtering	0.01–0.10 Hz	Performed after spatial normalization.
Spatial smoothing	4-mm FWHM	Gaussian smoothing on normalized functional data.
Brain parcellation	AAL atlas	116 labeled ROIs; mean ROI time series were extracted.
FC construction	Pearson correlation	Pairwise ROI correlations; no Fisher z transformation.
Node feature	ROI-level fALFF	fALFF was calculated in the 0.01–0.10 Hz band.

B CATEGORY-SPECIFIC CBT CONSTRUCTION

The category-specific CBTs were constructed using the original Deep Graph Normalizer (DGN), without modifying its architecture or optimization objective. For every data split, only the training subjects were used. The training FCs were divided into two disjoint category-specific subsets:

$$\mathcal{D}_{\mathrm{tr}}^{s} = \{A_i \mid y_i = s,\ i \in \mathcal{D}_{\mathrm{tr}}\}, \qquad s \in \{\mathrm{ASD}, \mathrm{NC}\}, \tag{17}$$

where s denotes the diagnostic category. Two independent DGN models were trained: one using only $\mathcal{D}_{\mathrm{tr}}^{\mathrm{ASD}}$ and the other using only $\mathcal{D}_{\mathrm{tr}}^{\mathrm{NC}}$. Neither validation nor test subjects were involved in DGN optimization or CBT aggregation.

In the original DGN formulation, subject j is represented by a multi-view tensor $T_j \in \mathbb{R}^{n_r \times n_r \times n_v}$. DGN sets the diagonal connectivity entries to zero, initializes all ROI attributes to one, and applies three edge-conditioned graph-convolution layers. This study used one rs-fMRI FC view, i.e., $T_j^{(1)} = A_j$ and $n_v = 1$. Let $V_{s,j}^{(L)} \in \mathbb{R}^{n_r \times d_L}$ be the final ROI embeddings produced by the category-s DGN. Its subject-conditioned CBT candidate is

$$\widehat{C}_{s,j}(p,q) = \sum_{z=1}^{d_L} \left| V_{s,j}^{(L)}(p,z) - V_{s,j}^{(L)}(q,z) \right|. \tag{18}$$

For an input subject j and a random subset of training-subject indices $\mathcal{S}_{s,j} \subseteq \mathcal{D}_{\mathrm{tr}}^{s}$, the original Subject Normalization Loss (SNL) is

$$\begin{aligned} \mathcal{L}_{\mathrm{SNL}}^{s,j} &= \sum_{v=1}^{n_v} \sum_{i \in \mathcal{S}_{s,j}} \left\| \widehat{C}_{s,j} - T_i^{(v)} \right\|_F \lambda_v, \\ \lambda_v &= \frac{\mu_v^{-1}}{\max_{r \in \{1,\ldots,n_v\}} \mu_r^{-1}}. \end{aligned} \tag{19}$$

Algorithm 1 Category-Specific CBT Construction

Require: Training FCs $\{A_i, y_i\}_{i \in \mathcal{D}_{\text{tr}}}$
1: **for** $s \in \{\text{ASD}, \text{NC}\}$ **do**
2: Form $\mathcal{D}^s_{\text{tr}} = \{A_i : y_i = s\}$.
3: Train an independent DGN g_{θ_s} on $\mathcal{D}^s_{\text{tr}}$ using the original SNL.
4: Generate $\widehat{C}_{s,i} = g_{\theta_s}(A_i)$ for every $A_i \in \mathcal{D}^s_{\text{tr}}$.
5: Aggregate $C_s \leftarrow \text{Median}_i(\widehat{C}_{s,i})$ element-wise.
6: **end for**
7: **return** $\{C_{\text{ASD}}, C_{\text{NC}}\}$

where μ_v is the mean connectivity weight of view v. Hence, $\lambda_1 = 1$ in the single-view setting used here. Each category-specific DGN minimizes

$$\min_{\theta_s} \frac{1}{|\mathcal{D}^s_{\text{tr}}|} \sum_{j \in \mathcal{D}^s_{\text{tr}}} \mathcal{L}^{s,j}_{\text{SNL}}. \tag{20}$$

Resampling $\mathcal{S}_{s,j}$ across iterations regularizes training and encourages every subject-conditioned candidate to represent the corresponding population rather than its input subject alone.

After DGN optimization, all category-s training FCs were passed through g_{θ_s} to obtain a set of candidate CBTs. The final template was computed by element-wise median aggregation:

$$C_s(p, q) = \underset{A_j \in \mathcal{D}^s_{\text{tr}}}{\text{median}} \widehat{C}_{s,j}(p, q). \tag{21}$$

The procedure produces two fixed templates, $\mathcal{C} = \{C_{\text{ASD}}, C_{\text{NC}}\}$.

C CBT–FC Preliminary Fusion and Node-Feature Augmentation

Let $A_i \in \mathbb{R}^{n \times n}$ denote an individual FC matrix, and let C_{ASD} and C_{NC} denote the two category-specific CBTs. Since all three matrices use the same AAL ROI ordering, both templates are injected simultaneously by element-wise addition to construct a single CBT-guided input:

$$\widetilde{A}_i = A_i + C_{\text{ASD}} + C_{\text{NC}}. \tag{22}$$

Thus, the plus sign in the framework denotes direct addition rather than concatenation, candidate-input construction, or a learned transformation. No Fisher z transformation is applied before or after this operation. Because both CBTs are included for every subject, input construction does not require a category label at either training or test time.

The node-feature concatenation reported in Eq. (3) of the main paper is performed *after* the element-wise CBT–FC addition. Specifically, for ROI p, the original node feature $x_{i,p}$ is concatenated with the entire p-th row of the fused connectivity matrix:

$$\widetilde{x}_{i,p} = \text{Concat}\Big(x_{i,p}, \widetilde{A}_i(p, 1), \ldots, \widetilde{A}_i(p, n)\Big). \tag{23}$$

The resulting input graph is $\widetilde{G}_i = (V, \widetilde{A}_i, \widetilde{X}_i)$. Consequently, the same fused connectivity information is available to the encoder both as edge weights during message passing and as explicit connectivity-profile components of the node features.

D Rationale for the Augmented Representation

The element-wise addition in Eq. equation 22 provides a parameter-free structural prior at the input level. CFE serves a distinct purpose at the representation level: it adaptively controls the contribution of the CBT prior after high-level graph features have been extracted. During training, the known label y_i identifies the CBT feature used by the auxiliary CBT branch, and the augmented representation is

$$f_{\text{aug},i} = \phi\Big(f^{(K)}_{\text{FC},i} \,\|\, f^{(K)}_{\text{CBT},y_i}\Big), \tag{24}$$

where $\|$ denotes concatenation and $\phi(\cdot)$ is a dense projection that maps the concatenated vector into the same latent space as $f_{\mathrm{FC},i}^{(K)}$.

Directly using $f_{\mathrm{FC},i}^{(K)}$ and $f_{\mathrm{CBT},y_i}^{(K)}$ as a contrastive positive pair would treat an individual connectome and a population prototype as interchangeable. Such a constraint could over-regularize the individual representation and suppress subject-specific deviations that remain diagnostically relevant. In contrast, $f_{\mathrm{aug},i}$ remains subject-conditioned because it contains both the individual FC representation and the group prior. The learnable projection can adaptively determine the contribution of the CBT rather than imposing a fixed one-to-one match.

Therefore, $(f_{\mathrm{FC},i}^{(K)}, f_{\mathrm{aug},i})$ forms two semantically consistent views of the same subject: an individual representation that serves as the anchor and a CBT-guided representation that serves as its enhanced counterpart. This construction also avoids stochastic edge-dropping augmentations that may remove clinically relevant functional connections. The input-level addition and representation-level augmentation are therefore complementary rather than redundant.

E Training and Inference Procedure

Both training and inference use the single fused graph $\widetilde{G}_i$ defined above. During training, the FC encoder extracts multi-level subject representations from $\widetilde{G}_i$, while the CBT encoder extracts features from the category-specific CBT associated with the known training label. SFA and CFE use this auxiliary CBT branch to regularize and enhance the FC representations. The diagnostic classifier is trained on the low- and high-level FC-encoder features.

At inference, the CBT encoder, SFA, and CFE are not required. The test FC is combined once with both C_{ASD} and C_{NC} using Eq. equation 22; the FC encoder then extracts its features and the trained classifier directly produces the diagnostic probability. Thus, inference uses neither a test label nor category-dependent template selection.

Algorithm 2 Training and Inference of CBT-GCL

Require: FC A_i, node features X_i, templates C_{ASD} and C_{NC}
1: $\widetilde{A}_i \leftarrow A_i + C_{\mathrm{ASD}} + C_{\mathrm{NC}}$.
2: Form $\widetilde{X}_i$ using Eq. equation 23.
3: $F_{\mathrm{FC}} \leftarrow \mathrm{En}_{\mathrm{FC}}(\widetilde{G}_i)$.
4: **if** training **then**
5: $\quad F_{\mathrm{CBT}} \leftarrow \mathrm{En}_{\mathrm{CBT}}(C_{y_i})$.
6: $\quad$ Compute the SFA and CFE training objectives.
7: $\quad$ Optimize both encoders and the diagnostic classifier.
8: **else**
9: $\quad$ Predict directly from the low- and high-level features in F_{FC}.
10: **end if**

F Evaluation Metrics

ASD was treated as the positive class and NC as the negative class. Let TP, TN, FP, and FN denote the numbers of true-positive, true-negative, false-positive, and false-negative predictions, respectively. Accuracy (Acc), sensitivity (Sen), and specificity (Spe) were computed as

$$\begin{aligned} \mathrm{Acc} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \\ \mathrm{Sen} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \qquad \mathrm{Spe} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}. \end{aligned} \tag{25}$$

The area under the receiver operating characteristic curve (AUC) was calculated from the continuous diagnostic probabilities and summarizes discrimination across classification thresholds.

The "Average" column reported in the experiments is the arithmetic mean of the four metrics,

$$\mathrm{Avg} = \frac{\mathrm{Acc} + \mathrm{AUC} + \mathrm{Sen} + \mathrm{Spe}}{4}. \tag{26}$$

G ABIDE I-TO-II CBT TRANSFER

To evaluate cross-dataset transfer of the population connectivity prior, two DGN models were trained only on the ASD and NC subsets of the ABIDE I training set. This produced a fixed template pair $\mathcal{C}^{\mathrm{I}} = \{C^{\mathrm{I}}_{\mathrm{ASD}}, C^{\mathrm{I}}_{\mathrm{NC}}\}$; no ABIDE II subject was involved in its construction.

The templates were frozen and used to guide downstream training on ABIDE II. For every ABIDE II subject i, both were added to the individual FC to construct a single input:

$$\widetilde{A}^{\mathrm{II}\leftarrow\mathrm{I}}_i = A^{\mathrm{II}}_i + C^{\mathrm{I}}_{\mathrm{ASD}} + C^{\mathrm{I}}_{\mathrm{NC}}. \tag{27}$$

This fusion is independent of the ABIDE II subject label. During training, only the auxiliary CBT encoder received the label-matched template $C^{\mathrm{I}}_{y_i}$ for SFA and CFE. At test time, the auxiliary branch was removed and the fused graph was processed only by the FC encoder and classifier. The ABIDE II split and all downstream settings were identical to the native-CBT experiment; only the CBT source was changed.

Table 5: ABIDE II performance using native or transferred CBT pairs. Avg. is defined in Eq. equation 26.

CBT source	Acc.	AUC	Sen.	Spe.	Avg.
ABIDE I	0.701	0.703	0.694	0.707	0.701
ABIDE II	0.714	0.711	0.722	0.707	0.713

Table 5 shows that ABIDE I CBTs achieved 0.701 accuracy and a 0.701 average score on ABIDE II. Relative to native ABIDE II CBTs, accuracy, AUC, sensitivity, and average score decreased by 1.3, 0.8, 2.8, and 1.2 percentage points, respectivelyCompared with using ABIDE II's native CBT, accuracy, AUC, sensitivity, and average score decreased by 1.3, 0.8, 2.8, and 1.2 percentage points, respectively, yet remained higher than those of comparison methods, while specificity stayed unchanged, while specificity was unchanged. This provides preliminary evidence of transferable population patterns.