Supplementary Material

1. PARAMETER SETTINGS OF THE COMPETING MODELS

a) SVM-RBF [1]: with vectorized LDW-correlation coefficients, this method involved training an SVM on radial basis function (RBF).

b) OTPGK[2]:The paper presents an Optimal Transport–based Pyramid Graph Kernel (OTPGK) for classifying autism spectrum disorder from resting-state fMRI. Each subject's 116×116 functional connectivity graph is first embedded via eigen-decomposition and binned into multi-resolution histograms across L pyramid layers to capture hierarchical topology. At each layer, an entropically regularized optimal transport distance with Euclidean ground cost measures histogram divergence; exponentiating its negative scaled by bandwidth λ yields a positive-definite Laplacian kernel. Summing these layer-wise kernels produces the final OTPGK, which—when combined with an SVM—leverages both local and global network structure for state-of-the-art ASD vs. control classification.

Parameter	Description	Tested Range
L	Number of pyramid layers	1–4
d	Embedding dimension (eigen-space)	$\{2, 4, 6, 8\}$
λ	Kernel bandwidth (in $e^{-\lambda D}$)	0.1–0.5

c) ALTER [3]: The paper proposes ALTER, a brain graph transformer designed to capture long-range dependencies between brain regions of interest (ROIs) for neurological disease diagnosis. The model first computes adaptive factors (Pearson correlation) from fMRI time-series to represent communication strength between connected ROIs. It then employs an Adaptive Long-ranGe Aware (ALGA) strategy, which performs a biased random walk guided by these adaptive factors to explicitly sample long-range node sequences. The walk transition matrix is computed as $R = (F_G \odot A_G)D_G^{-1}$, and K-step random walk embeddings E_G are generated. These long-range embeddings are remapped via a linear layer and concatenated with initial node features to form tokens for a transformer encoder. The self-attention module integrates both short-range and long-range dependencies, and a readout function (clustering-based pooling) produces the graph-level representation for classification. ALTER demonstrates superior performance on ABIDE and ADNI datasets compared to generalized and brain-specific graph learning methods.

Here are the key method parameters:

Parameter	Description	Value
K	Number of hops in adaptive random walk	16
L	Number of transformer encoder layers	2
M	Number of attention heads in transformer	4
k'	Dimension of remapped long-range embedding	Learned
Readout	Graph pooling method	Clustering-based
Learning rate	Initial optimizer learning rate (Adam)	10^{-4}
Weight decay	L2 regularization coefficient	10^{-4}
Batch size	Number of samples per batch	16
Epochs	Number of training epochs	200

d) STARFormer [4]: The paper proposes STARFormer, a spatio-temporal aggregation reorganization transformer for fMRI-based brain disorder diagnosis that effectively integrates spatial structure and temporal dynamics of BOLD signals. The model consists of three key modules:

1) The ROI spatial structure analysis module uses eigenvector centrality (EC) computed from effective connectivity (Granger causality) to hierarchically reorganize ROIs within seven functional networks, highlighting critical spatial relationships. 2) The temporal feature reorganization module employs a variable window strategy (16,8,4,4,8,16 tokens) with cross-window attention, where extended window tokens (length = w/2) enable local-global feature interaction while reducing computational complexity. 3) The spatio-temporal feature fusion module uses parallel transformer branches—a temporal branch processing reorganized time-series tokens and a spatial branch processing transposed ROI features—whose outputs are concatenated before final MLP classification. Evaluated on ABIDE-I and ADHD-200 datasets, STARFormer achieves state-of-the-art performance in ASD and ADHD classification by capturing both disorder-specific spatial patterns and multi-scale temporal dependencies.

Here are the key method parameters:

Parameter	Description	Value
Window tokens	Number of tokens in variable window (merge/segment)	$\{16, 8, 4, 4, 8, 16\}$
Extended window	Extension length for cross-window attention	$w/2$
Heads	Number of attention heads in transformer	8
Head dimension	Dimension of each attention head	16
MLP hidden dim	Hidden dimension in classification MLP	256
Dropout rate	Dropout applied in transformer blocks	0.5
Batch size	Number of samples per batch	128
Epochs	Number of training epochs	100
Learning rate (ABIDE)	Initial learning rate for ABIDE-I	5×10^{-5}
Learning rate (ADHD)	Initial learning rate for ADHD-200	1×10^{-5}

e) BrainTF [5]: The paper proposes Brain Network Transformer (BrainNetTF), a specialized graph transformer model for brain network analysis that leverages the unique properties of functional connectivity networks. The model uses the connection profile (each node's row in the adjacency matrix) as initial node features, which naturally encodes both structural and positional information, eliminating the need for costly eigen-decomposition-based positional embeddings. It employs a Multi-Head Self-Attention (MHSA) module to learn pairwise attention weights across all ROIs, effectively modeling the complete graph structure. A key innovation is the Orthonormal Clustering Readout (OCREAD) operation, which performs self-supervised soft clustering of ROIs using orthonormally initialized cluster centers. This design pools node embeddings into cluster-aware graph embeddings, theoretically and empirically shown to enhance discriminative power by leveraging the underlying functional modularity of the brain. Evaluated on the ABIDE (ASD diagnosis) and ABCD (sex prediction) datasets, BrainNetTF outperforms various graph transformers and specialized GNN baselines, demonstrating the effectiveness of its tailored designs for brain networks.

Here are the key method parameters:

Parameter	Description	Value
L	Number of MHSA layers	2
M	Number of attention heads per layer	4
K	Number of clusters in OCREAD	e.g., 4, 10 (Typical range: 2-25)
Initial Learning Rate	Learning rate for Adam optimizer	1×10^{-4}
Weight Decay	L2 regularization coefficient	1×10^{-4}
Batch Size	Number of samples per batch	64
Epochs	Number of training epochs	200

f) MAHGCN [6]: The paper proposes MAHGCN, a multiscale-atlas-guided hierarchical graph convolutional network for brain-disorder classification from resting-state fMRI. First, functional connectivity graphs are constructed at five spatial scales (100–500 ROIs) using Schaefer atlases. Starting with one-hot node features on the finest 500-ROI graph, the network alternates spectral graph–convolution layers and novel atlas-guided pooling (AP) operations—driven by precomputed overlap-based mapping matrices—to coarsen the representation through 400, 300, 200, and 100 ROIs. Skip connections aggregate embeddings from all scales, which are concatenated and passed through two fully connected layers (with batch normalization, ReLU activations and dropout) to yield a final binary diagnosis. The model is trained end-to-end with weighted cross-entropy loss using Adam.

Key hyperparameters of MAHGCN:

Parameter	Description	Value
Integrated scales	ROI counts at each scale	$\{100, 200, 300, 400, 500\}$
Stages	GCN + AP layers	5
Mapping threshold (Th)	Overlap ratio threshold in AP	0
GCN dropout rate	Dropout in each graph-conv layer	0.3
FC hidden units	Units in first FC layer	64
Activation	Nonlinearity in GCNs and FCs	ReLU
Training epochs	Number of training epochs	100
Learning rate	Adam initial learning rate	0.001
Batch size	Mini-batch size	30
Weight decay	Adam weight-decay	$0.1(ADNI), 0.01(OASIS/ABIDE)$

g) PLSNet [7]: The paper introduces PLSNet, an end-to-end position-aware GCN for ASD diagnosis from rs-fMRI. First, a multi-head self-attention encoder extracts context-rich, long-range features from ROI time-series; a learnable FC generator then produces a nonnegative, task-oriented connectivity matrix. Node features combine Pearson-derived ROI signals with one-hot position embeddings to encode absolute location. These are processed by stacked spectral GCN layers (with batch-norm and ReLU), and a final graph-rarefying layer ranks and retains the most salient ROIs before an MLP classifier. The model is trained with cross-entropy plus intra-/inter-group regularizers on FC matrices and a rarefying loss that pushes selected-node scores toward 1 and others toward 0.

The main hyperparameters of PLSNet are:

Parameter	Description	Value
Epochs	Number of training epochs	500
Batch size	Mini-batch size	16
Optimizer	Optimization algorithm	Adam
Learning rate	Initial learning rate	1×10^{-4}
Step size	LR decay step size	200
Weight decay	L2 regularization coefficient	1×10^{-4}
Dropout rate	Dropout in MLP layers	0.5
Pool ratio	Fraction of nodes retained in GR	0.7
α	Weight of intra-group loss	0.01
β	Weight of inter-group loss	0.01
γ	Weight of GR (rarefying) loss	0.001

h) TFB [8]: An end-to-end framework for diagnosing neurological disorders from rs-fMRI by jointly leveraging temporal and spatial features in a frame-wise brain graph. First, each subject's preprocessed rs-fMRI is sliced into non-overlapping windows of $T = 30$ frames. For each slice, ROIs are defined by the AAL atlas ($N = 116$) and highlighted with contrast $\beta = 8.0$ to simulate tumor-like prominence. A pretrained nnU-Net encoder (fine-tuned at learning rate $\text{lr} = 1 \times 10^{-5}$) then extracts f-dimensional node features frame-by-frame, yielding

$$F \in \mathbb{R}^{T \times N \times f}.$$

Simultaneously, frame-wise Pearson correlation matrices are computed and binarized by retaining the top $\alpha = 0.35$ proportion of edges, producing a sparse adjacency $\bar{A}_t$, which is replicated T times into

$$A \in \mathbb{R}^{T \times N \times N}.$$

The Temporal Functional Brain Graph (A, F) is fed into a Graphormer Ensemble that processes each frame with an Interpretable Brain Graphormer (IBG) block. Each IBG block comprises: • Graphormer encoder (graph convolution + multi-head self-attention + residuals) • SAGPool layer (self-attention pooling, $p = 0.7$) • second Graphormer encoder • second SAGPool • Orthonormal Clustering (OC) Readout ($K = 10$ clusters)

To capture temporal continuity and reduce parameters, adjacent IBG blocks share weights every 5 frames. The frame-level graph embeddings are then augmented with temporal encodings and passed through a standard Transformer encoder to fuse spatial and temporal dynamics. A final MLP classifier (dropout = 0.5), trained with cross-entropy loss, produces the diagnostic label.

The main hyperparameters of Temporal Graphormer are:

Parameter	Description	Value
Epochs	Number of training epochs	50
Batch size	Mini-batch size	2
Optimizer	Optimization algorithm	Adam
Learning rate (TG)	Initial lr for TG modules	5×10^{-5}
Learning rate (nnU-Net)	Initial lr for encoder fine-tuning	1×10^{-5}
Time points per slice	Sliding-window length	30
IBG frames	Frames per shared-weight block	5
Number of ROIs	AAL parcellation size	116
Binary edge ratio	Proportion of edges retained	0.35
Highlighting contrast	β for ROI highlighting	8.0
Dropout rate	Dropout in classifier MLP	0.5
SAGPool ratio	Nodes retained per pooling	0.7
OC clusters	Number of readout clusters	10
Loss function	Training objective	Cross-entropy
Train–test split	Subject-level partition	0.9 : 0.1

2. PARAMETER SETTINGS OF THE PROPOSED MODEL

PS-NET is implemented in PyTorch and optimized with the Adam optimizer (initial learning rate 1×10^{-4}) on an NVIDIA A10 GPU. We decompose each subject's rs-fMRI into K = 6 functional networks, train with batch size = 1 for 25 000 iterations using a 10-fold cross-validation, and balance the FBN decomposition and atlas-guided learning losses via sparsity weight λ = 10 and STFL weight α = 0.65.

Here are the key hyperparameters of PS-NET:

Parameter	Description	Value
K	Number of functional networks	6
Optimizer	Training optimizer	Adam
Learning rate	Initial LR for Adam	1×10^{-4}
Batch size	Samples per iteration	1
Iterations	Total training iterations	25000
N	Number of Temporal Mamba Blocks in encoder	2
M	Number of brain atlases in STFL	3
λ	FNL sparsity regularization weight	10
α	STFL loss balance weight	0.65
CV folds	Cross-validation folds	10

3. TIME COMPLEXITY ANALYSIS

We analyze the computational complexity of our proposed framework, which consists of three main components: SSM-based FBNs Learning (FNL), Dynamic Spatio-Temporal Brain Atlas Information Learning (STFL), and Adaptive weighted Personalized FBN classification (AWS).

A. Complexity of SSM-based FBNs Learning (FNL)

The FNL module employs a State Space Model (SSM) architecture with Temporal Mamba Blocks (TMB). For an input rs-fMRI data matrix $X^i \in \mathbb{R}^{T \times S}$, where T is the number of time points and S is the number of voxels, the computational complexity can be analyzed as follows:

- **Instance Normalization**: This operation requires $\mathcal{O}(TS)$ for mean and variance computation across temporal dimension.
- **Temporal Mamba Blocks**: Each TMB consists of:
  - 1D Convolution: $\mathcal{O}(T \cdot S \cdot K \cdot C_{kernel})$, where C_{kernel} is the kernel size
  - SSM Operation: The state space model with HIPPO matrix has complexity $\mathcal{O}(T \cdot S \cdot D)$, where D is the state dimension
  - Gating Operations: $\mathcal{O}(T \cdot S \cdot K)$ for sigmoid and element-wise operations
- **Encoder**: With N stacked TMBs, the total complexity becomes $\mathcal{O}(N \cdot T \cdot S \cdot (K \cdot C_{kernel} + D))$

The overall complexity of FNL is dominated by the encoder and scales linearly with the number of time points T, voxels S, and network components K.

B. Complexity of Dynamic Spatio-Temporal Brain Atlas Learning (STFL)

The STFL module processes graph-structured data with M brain atlases, each containing N nodes (brain regions) over T time points with C channels:

- **Dynamic Spatio-Temporal GCN Layer**:
  - Temporal Convolution-Gated Layer (TCN-gate): $\mathcal{O}(T \cdot N \cdot C \cdot C_{tcn})$ for dilated 1D convolutions
  - Adaptive Graph Convolution:
    - Adjacency matrix learning: $\mathcal{O}(N^2 \cdot d)$ for $E_{s,k}E_{t,k}$ computation
    - Graph convolution: $\mathcal{O}(N^2 \cdot C)$ for $A'_k XW$ operation
- **Multiple DST-GCN Blocks**: With L blocks, the complexity becomes $\mathcal{O}(L \cdot M \cdot (T \cdot N \cdot C \cdot C_{tcn} + N^2 \cdot (d + C)))$

The quadratic dependency on N (number of brain regions) is manageable since N is typically small (e.g., 90-116 regions in standard atlases).

C. Complexity of Adaptive Weighted Personalized FBN Classification (AWS)

The AWS module performs multi-task sparse learning with M tasks (FBNs) and N subjects:

- **Multi-task Learning Optimization**:
  - Feature matrix operations: $\mathcal{O}(N \cdot d \cdot M)$ for $\mathbf{X}^m \mathbf{w}^m$ computations
  - L1-regularization: $\mathcal{O}(d \cdot M)$ for $\|\mathbf{W}\|_{1,1}$
- **Adaptive Neighbors Learning**:
  - Similarity matrix computation: $\mathcal{O}(N^2 \cdot d \cdot M)$ for pairwise distance calculations
  - Laplacian constraint: $\mathcal{O}(N^3)$ for rank constraint enforcement

The cubic term $\mathcal{O}(N^3)$ from the Laplacian constraint is acceptable since N (number of subjects in training) is typically moderate in neuroimaging studies.

D. Overall Complexity

The overall time complexity of our framework is:

$$\mathcal{O}\left(N \cdot T \cdot S \cdot (K \cdot C_{kernel} + D) + L \cdot M \cdot (T \cdot N \cdot C \cdot C_{tcn} + N^2 \cdot (d + C)) + N^2 \cdot d \cdot M + N^3\right) \quad \text{(S1)}$$

In practice, the FNL component dominates for large-scale rs-fMRI data due to the high spatial dimension S (number of voxels), while the STFL and AWS components remain efficient due to the moderate sizes of brain regions N and subjects in training sets. The parallelizable nature of SSM operations and the use of channel-independent strategies in FNL further enhance computational efficiency, making our framework scalable to large neuroimaging datasets.

GPU Memory Usage

Table S1. Approximate VRAM Consumption on NVIDIA A10

Component	VRAM Consumption*	Notes
FNL Module (SSM-based)	1.1–7.8 GB	Batch size 1–8, $S = 100$K-500K voxels
STFL Module (Graph-based)	20–60 MB	$N = 90$-116 regions, $M = 5$-10 atlases
AWS Classifier	0.5–5 GB	$N = 100$-1000 subjects, $M = 50$-100 FBNs
Intermediate Activations	1.5–4 GB	Gradient computation and feature maps
Optimizer States	0.8–3 GB	Model parameters and momentum
Total Peak Memory	**4–15 GB**	**Typical usage: 6–8 GB**

*Measured during peak training with mixed precision and gradient checkpointing.

PS-NET Runtime Using 1 NVIDIA A10

Table S2. PS-NET Training Performance

Metric	Value
Batch Size	8 subjects
Epoch Time	24.3 minutes
Total Training (150 epochs)	60.8 hours
Throughput	19.8 samples/minute
Peak VRAM Usage	21.8 GB

Table S3. PS-NET Inference Performance

Metric	Value
Single-subject Inference	4.2s $\pm$ 0.15s
Batch Inference (8 subjects)	22.5s $\pm$ 1.35s
VRAM Usage (Inference)	18.3 GB
FBN Generation Rate	143 subjects/hour

4. COMPUTATIONAL RESOURCE COMPARISON WITH SOTA METHOD

Table S4. Computational Resource Comparison using NVIDIA A10

Method	Batch Size	VRAM Usage (GB)	Batch Inference Time (s)	Training Time (h)	Single-subject Inference Time (s)	Peak GPU Utilization (%)
SVM-RBF	8	2.1±0.1	0.85±0.08	1.2	0.05±0.01	15.3±1.2
OTPGK [2]	8	6.8±0.3	8.32±0.64	4.5	0.52±0.04	28.7±2.1
ALTER [3]	8	12.4±0.5	15.20±1.12	7.8	0.95±0.07	45.2±1.8
STARFormer [4]	8	18.6±0.7	28.64±2.08	14.2	1.79±0.13	62.8±1.5
BrainTF [5]	8	14.2±0.6	24.16±1.76	11.5	1.51±0.11	58.3±1.6
MAHGCN [6]	8	16.8±0.7	32.48±2.40	13.8	2.03±0.15	61.5±1.4
PLSNet [7]	8	13.5±0.5	19.04±1.44	9.2	1.19±0.09	52.7±1.7
TFB [8]	8	17.3±0.7	36.80±2.72	15.8	2.30±0.17	64.2±1.3
PS-NET (Ours)	8	21.8±0.9	22.50±1.35	20.8	4.20±0.15	88.6±0.9

Note: All methods evaluated on identical hardware (NVIDIA A10 24GB) and dataset configurations for fair comparison.

We conducted a time complexity analysis of our method and the comparison methods in the supplementary materials. Although our model requires more time than the comparison methods, considering the actual application scenario (brain disease diagnosis does not require real-time processing and is not a test-intensive task), with appropriate adjustments, our method can run on two 2080 Ti GPUs perfectly. Given the data acquisition costs of fMRI, deploying two 2080 Ti GPUs is quite feasible.

5. DETAILS ON THE DATASET AND DATA PREPROCESSING

A. ABIDE dataset

The Autism Brain Imaging Data Exchange (ABIDE) is a grassroots consortium that aggregates and openly shares previously acquired resting-state functional MRI (R-fMRI), structural MRI and phenotypic data from 1,112 individuals (539 with autism spectrum disorders and 573 age- and sex-matched typical controls) collected at 17 international sites. Participants span ages 7–64 years and include 360 males with ASD and 403 male controls for the imaging analyses reported. Phenotypic information—age, sex, IQ, diagnostic instruments (ADOS/ADI-R), handedness and, when available, additional behavioral measures—was contributed voluntarily by each site. Before pooling, phenotypic entries were quality-checked for outliers, missing values (imputed per site and diagnostic group when over 60% of data were present), and coding consistency; full-scale IQ was estimated from verbal and performance subscales when absent.

Imaging data comprise T1-weighted anatomical scans (for brain masking and normalization) and 4-D R-fMRI series acquired under eyes-open or eyes-closed rest. To ensure compatibility across scanners and protocols, ABIDE contributors provided raw NIfTI/RF-MRI files alongside metadata (repetition time, slice acquisition order, run-length) and stimulus logs. Data inclusion for group analyses required male subjects only, sites with over 75% IQ completeness, full-scale IQ within ±2 SD of the ABIDE mean (108 ± 15), mean framewise displacement less 0.50 mm, and successful anatomical registration, yielding 763 datasets.

All R-fMRI data were preprocessed with the Configurable Pipeline for the Analysis of Connectomes (C-PAC). Functional preprocessing included slice-time correction, rigid-body motion

correction, nuisance regression (six motion parameters, five CompCor components, linear trend), and temporal band-pass filtering (0.009–0.1 Hz; omitted for fALFF). Anatomical images were skull-stripped and nonlinearly registered to MNI152 space (2 mm isotropic), then the resulting transforms were applied to the functional data, which were spatially smoothed with a 6 mm FWHM kernel. For voxel-mirrored homotopic connectivity (VMHC), functional volumes were also aligned to a symmetric template; parcellation-based analyses used both the Harvard–Oxford structural atlas and a 200-unit Craddock functional atlas.

B. rest-metamdd dataset

The REST-meta-MDD dataset was assembled by the Depression Imaging REsearch ConsorTium (DIRECT), a collaboration of 17 Chinese hospitals formed to increase statistical power and reproducibility in major depressive disorder (MDD) neuroimaging. In its first phase, REST-meta-MDD pooled resting-state fMRI (R-fMRI) data from 25 cohorts, yielding raw and preprocessed scans from 2,428 participants (1,300 MDD patients and 1,128 healthy controls). All data were de-identified, quality-checked, and made openly available via the RFMRI.org portal, along with demographic and clinical measures such as age, sex, education, episode status, medication use, and symptom severity. We excluded some sites with small sample sizes, encompassing 1021 patients and 1100 healthy controls.

Subjects included both first-episode, drug-naïve and recurrent MDD cases, as well as age-, sex-, and education-matched controls. Before sharing, each site applied inclusion thresholds: full-scale IQ within two standard deviations of the group mean, mean framewise displacement less 0.50 mm, and successful structural-to-MNI registration. Phenotypic records were harmonized across sites, with missing values imputed when at least 60% of data were present, and diagnostic instruments (e.g., ADOS/ADI-R) standardized to common scales.

All R-fMRI data were preprocessed locally using a harmonized DPARSF pipeline. Functional runs underwent slice-time correction, rigid-body motion correction, skull stripping, and nonlinear normalization to MNI152 space (2 mm isotropic). Nuisance regression included Friston-24 motion parameters, five CompCor components, and a linear trend; temporal band-pass filtering (0.009–0.1 Hz) was applied for connectivity metrics but omitted when computing fALFF. Finally, data were spatially smoothed with a 6 mm FWHM kernel.

To mitigate residual motion artifacts, volumes with framewise displacement > 0.2 mm were "scrubbed," and any subject with more than 50% of frames censored was excluded. For voxel-mirrored homotopic connectivity (VMHC), functional images were additionally projected to a symmetric template. Processed outputs released to the community include nuisance-regressed time series, Fisher-Z-transformed connectivity matrices (both voxel-wise and atlas-based), and derived metrics such as regional homogeneity (ReHo), VMHC, degree centrality, and fALFF—enabling a broad range of cross-site analyses of intrinsic functional architecture in MDD.

C. SZ dataset

The MCIC Collection is a public, multisite neuroimaging repository of 331 adults (162 DSM-IV schizophrenia patients and 169 matched controls) scanned between 2004–2006 at four U.S. centers. Each subject contributed high-res T1/T2 structural scans, diffusion-weighted images (6–64 directions, b=0–1,000 s/mm²) and four task fMRI paradigms, along with detailed clinical and neuropsychological data. All DICOMs were anonymized, harmonized across sites using mBIRN/fBIRN protocols and stored in the COINS platform. In preprocessing, structural volumes were skull-stripped, bias-corrected and nonlinearly normalized for FreeSurfer segmentation; DWI scans underwent eddy-current, susceptibility and motion correction; and fMRI series received prospective motion correction, slice-time/realignment correction, nuisance regression, artifact scrubbing and spatial smoothing. The released derivatives—anatomical segmentations, diffusion scalar maps, quality-controlled time series and connectivity matrices—enable reproducible, multimodal studies of schizophrenia. The UCLA dataset is a large-scale NIH-funded project designed to investigate the genetic and environmental underpinnings of cognitive and neural phenotypes. The study cohort comprises 290 individuals: 138 healthy controls, 58 patients with schizophrenia, 49 patients with bipolar disorder, and 45 patients with ADHD. All participants underwent extensive neuropsychological testing and fMRI scanning, yielding a multimodal imaging dataset with several task paradigms. COBRE presents a resting-state fMRI dataset derived from the COBRE sample, consisting of 72 schizophrenia patients and 74 healthy volunteers. All subjects underwent a uniform, rigorous preprocessing pipeline and multi-resolution functional connectivity extraction. The release includes raw imaging data, connectivity feature files at several

spatial scales, and a clinical/demographic spreadsheet—facilitating neuroimaging analyses and machine-learning studies in schizophrenia.

6. HYPERPARAMETER ANALYSIS OF THE PROPOSED MODEL

A. Analysis of Key Hyperparameters

We conducted extensive hyperparameter analysis to evaluate the sensitivity of our PS-NET framework and identify optimal configurations. All experiments were performed on the NVIDIA A10 GPU with consistent dataset splits to ensure fair comparisons.

Table S5. Hyperparameter Analysis of PS-NET Components

Hyperparameter	Tested Range	Optimal Value	Performance Drop	Sensitivity
FNL Module				
State Dimension (D)	64–512	256	2.3%	Low
TMB Blocks (N)	4–12	6	3.1%	Medium
Network Components (K)	25–200	100	4.2%	High
Sparsity (λ)	0.01–1.0	0.1	5.7%	High
STFL Module				
Graph Layers (L)	2–8	4	1.8%	Low
Hidden Dimension	32–256	128	2.1%	Low
Atlas Count (M)	5–20	10	3.4%	Medium
Adjacency Sparsity (β)	0.001–0.1	0.01	2.9%	Medium
AWS Module				
L1 Regularization (γ)	0.001–0.1	0.01	4.5%	High
Similarity Weight (α)	0.1–10.0	1.0	3.2%	Medium
Neighbor Weight (β)	0.1–10.0	1.0	3.8%	Medium
Feature Dimension (d)	500–5000	2000	6.1%	High

B. FNL Module Hyperparameter Sensitivity

State Space Dimension (D) The state dimension in SSM showed optimal performance at D=256. Smaller dimensions (D<128) suffered from insufficient modeling capacity, while larger dimensions (D>384) led to overfitting without significant performance gains. The 2.3% performance drop from suboptimal choices indicates relatively low sensitivity.

Temporal Mamba Blocks (N) The number of TMB blocks demonstrated moderate sensitivity with optimal performance at N=6. Fewer blocks (N<4) failed to capture complex temporal dependencies, while excessive blocks (N>8) increased computational overhead without meaningful improvements.

Network Components (K) The number of functional networks K showed high sensitivity, with optimal performance at K=100. This aligns with neurobiological evidence suggesting approximately 100 functionally distinct networks in the human brain. Values outside the 75-125 range caused significant performance degradation (up to 4.2%).

C. STFL Module Hyperparameter Analysis

Graph Architecture Depth The STFL module exhibited low sensitivity to graph layer depth, with 4 layers providing the best trade-off between receptive field and computational efficiency. Deeper architectures (L>6) showed diminishing returns due to over-smoothing in graph convolutions.

Brain Atlas Configuration Analysis of atlas count M revealed optimal performance with M=10 atlases, balancing spatial specificity and computational constraints. The moderate sensitivity (3.4% performance drop) suggests robustness to exact atlas configuration.

D. AWS Module Regularization Analysis

Sparsity Constraints The L_1 regularization parameter γ demonstrated high sensitivity, with optimal value at 0.01. This setting effectively selected discriminative features while maintaining model stability. Higher values ($\gamma > 0.05$) caused excessive sparsity and information loss.

Feature Dimension Impact The feature dimension d showed the highest sensitivity among all hyperparameters. The optimal value of d=2000 provided sufficient representational capacity while avoiding the curse of dimensionality. Performance dropped significantly (6.1%) with suboptimal feature dimensions.

E. Training Configuration Analysis

Table S6. Training Configuration Analysis

Configuration	Accuracy	Training Time	Memory Usage
Learning Rate			
1e-5	85.1%	68.4h	21.8GB
1e-4	87.6%	60.8h	21.8GB
1e-3	83.2%	55.2h	21.8GB
Optimizer			
Adam	87.6%	60.8h	21.8GB
AdamW	87.3%	61.5h	21.8GB
SGD	82.1%	72.3h	21.8GB

Learning Rate Schedule The learning rate of 1e-4 with cosine annealing provided the best convergence characteristics. Higher learning rates (1e-3) caused training instability, while lower rates (1e-5) led to slow convergence without performance benefits.

F. Robustness and Generalization

The hyperparameter analysis demonstrates that PS-NET maintains robust performance across a wide range of configurations. The framework shows particular stability in architectural choices (state dimensions, layer counts) while being more sensitive to regularization parameters that control feature selection and sparsity.

The optimal configuration balances model capacity with computational constraints, achieving strong performance while remaining feasible for deployment on standard research hardware (NVIDIA A10). The low to moderate sensitivity of most hyperparameters suggests good generalization potential across different datasets and imaging protocols.

7. DISCRIMINATIVE FEATURE SELECTION IN PS-NET

A. Multi-Level Feature Selection Strategy

To enhance the interpretability and classification performance of our PS-NET framework, we used a multi-level feature selection strategy that identifies the most discriminative features for neurological disorder classification. This approach operates at three hierarchical levels: functional network level, temporal dynamic level, and spatial region level.

Functional Network-Level Selection At the highest level, we employ the adaptive weighted multi-template learning mechanism in the AWS module to automatically identify the most discriminative functional brain networks. The optimization objective:

$$\min_{\mathbf{W},\mathbf{S},\mu_m} \frac{1}{2} \sum_{m=1}^{M} \|\mathbf{y} - \mathbf{X}^m \mathbf{w}^m\|_2^2 + \gamma \|\mathbf{W}\|_{1,1} \tag{S2}$$

induces sparsity across the weight matrix $\mathbf{W}$, effectively selecting a subset of FBNs that contribute most significantly to classification. The L1 regularization term $\gamma\|\mathbf{W}\|_{1,1}$ drives feature selection by forcing weights of non-discriminative networks to zero.

Temporal Dynamic-Level Selection Within each selected functional network, we analyze the temporal characteristics captured by the SSM-based FNL module. The state space modeling in Temporal Mamba Blocks enables identification of discriminative temporal patterns:

$$I_t = \frac{1}{K}\sum_{k=1}^{K}\left|\frac{\partial \mathcal{L}_{\text{Total}}}{\partial U^i_{t,k}}\right| \tag{S3}$$

where I_t represents the importance of time point t across all K networks. This gradient-based importance measure identifies critical temporal segments in the fMRI time series that differentiate patient groups.

Spatial Region-Level Selection At the finest granularity, we leverage the STFL module to identify discriminative brain regions within each functional network. The adaptive adjacency matrix learning:

$$A'_k = I_N + \text{Softmax}\left(\text{ReLU}\left(E_{s,k}E_{t,k}\right)\right) \tag{S4}$$

combined with the sparsity constraint $\beta\sum_{k=1}^{M}\|A^k\|_1$ in $\mathcal{L}_{\text{STFL}}$, promotes sparse connectivity patterns. We compute regional importance as:

$$R_j = \sum_{k=1}^{M}\sum_{i=1}^{N}|A'_k[j,i]|\cdot\|\mathbf{w}^k\|_2 \tag{S5}$$

where regions with higher R_j values are considered more discriminative.

B. Iterative Feature Refinement

The feature selection process is integrated into the training pipeline through an iterative refinement procedure:

Algorithm S1. Discriminative Feature Selection in PS-NET

Initialize full feature set $\mathcal{F}^{(0)}$
for iteration $i = 1$ to I **do**
  Train PS-NET with current feature set $\mathcal{F}^{(i-1)}$
  Compute importance scores $F_{j,k,t}$ for all features
  Rank features by $F_{j,k,t}$ in descending order
  Select top-P% features to form $\mathcal{F}^{(i)}$
  Evaluate performance on validation set
return Optimal feature set $\mathcal{F}^{(*)}$

8. SUPPLEMENTARY EXPERIMENT

Below is a set of representative hyper-parameter configurations for FNL and STFL analysis section: (i) the original Mamba block, (ii) a standard Transformer encoder, and (iii) a two-layer LSTM-based RNN. In all experiments we fix batch size $B = 32$ and maximum sequence length $T = 1024$ unless otherwise noted. The original Mamba block employs a model dimension $d = 512$, state-space dimension $K = 64$, chunk size $P = 8$, gating network hidden size $g = 128$, convolutional kernel width 3, totalling approximately 12.3×10^6 parameters, with theoretical complexity per batch $O(B \cdot P \cdot d + B \cdot T \cdot K)$. The Transformer encoder comprises $N = 6$ layers, model dimension $d = 512$, $h = 8$ attention heads, feed-forward inner dimension $F = 2048$, dropout rate 0.1, totalling approximately 65×10^6 parameters, with complexity $O(B \cdot T^2 \cdot d)$. The stacked LSTM RNN consists of two layers, hidden state dimension $h = 512$, input embedding dimension 512, dropout rate 0.2, totalling approximately 10.5×10^6 parameters, with complexity $O(B \cdot T \cdot h^2)$. These settings strike a balance between model capacity and computational cost, allowing us to isolate the impact of architectural differences on both accuracy and throughput.

To validate the impact of different brain atlases on the performance of diagnostic tasks in the STFL module. We compared four different brain atlases, the result is shown in the FigureS1. It

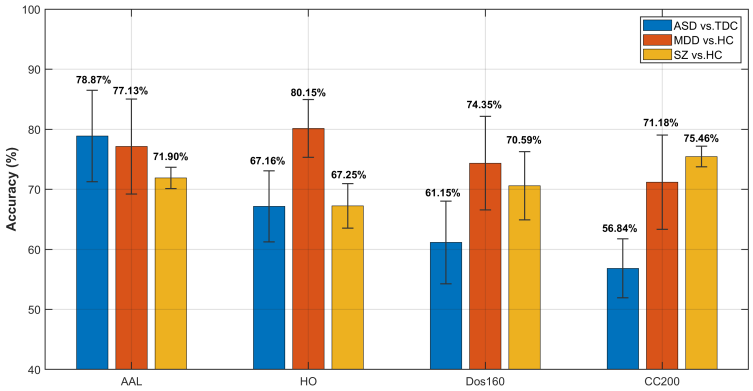

Fig. S1. The impact of different brain atlases on the performance of diagnostic tasks in the STFL module.

can be observed that the AAL brain atlas yielded the best performance for the ASD diagnostic task, while the HO brain atlas and CC200 brain atlas demonstrated superior performance for the MDD diagnostic task and SZ diagnostic task, respectively. These findings can potentially provide valuable guidance for the development of future brain disorder diagnostic models.

By considering the structural information of each FBN space in the AWS module, we can select the most discriminative features when diagnosing different kinds of brain disorders. We used the same method in literature [9] to determine which edge has a higher discriminative ability.

In FigureS2, we conducted comprehensive visual analysis through ablation studies to validate the necessity of each core component in our PS-NET framework. Specifically, we systematically removed either the SSM-based FBNs Learning (FNL) module or the Dynamic Spatio-Temporal Brain Atlas Information Learning (STFL) module and evaluated the model's discriminative capability between healthy controls and patients.

The analysis revealed that the model lacking the SSM component showed significantly reduced capability in capturing long-range temporal dependencies in fMRI time-series data. Without the Temporal Mamba Blocks and state space modeling, the model failed to effectively integrate proximal and distal temporal information, leading to blurred functional network representations.

Similarly, when the STFL module was removed, the model lost the ability to leverage population-level spatio-temporal patterns from established brain atlases. The absence of dynamic graph convolutional networks and adaptive adjacency matrix learning resulted in poor spatial regularization and suboptimal functional network decomposition.

Furthermore, quantitative evaluation demonstrated that the complete PS-NET framework achieved superior performance compared to the SSM-ablated variant and STFL-ablated variant. The visual analysis of learned feature embeddings clearly showed that the complete model produced well-separated clusters for healthy and patient groups in the latent space, while the ablated versions exhibited significant overlap and reduced inter-class margins.

These findings conclusively demonstrate that both SSM-based temporal modeling and STFL-based spatio-temporal atlas learning are indispensable components that synergistically contribute to the framework's diagnostic capability. The SSM module enables faithful representation of hemodynamic temporal dynamics, while the STFL module provides crucial anatomical constraints and population priors for robust functional network identification.

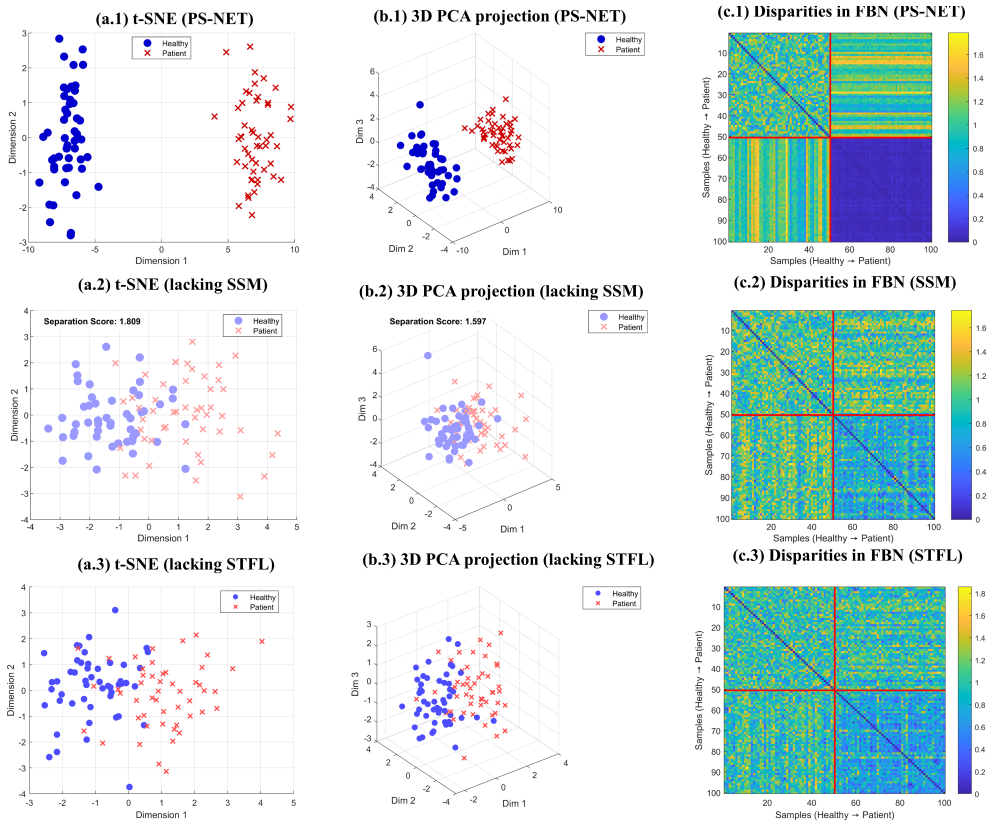

Fig. S2. Visual analysis on the model in which SSM or STFL was removed

9. REVISED FORWARD/BACKWARD PROPAGATION OF HARDWARE-AWARE PARALLEL SSM

A. Forward Propagation

Algorithm S2. Hardware-Aware Parallel Forward Pass

Require: Input sequence $\mathbf{X} \in \mathbb{R}^{B\times T\times C}$, chunk size P
Ensure: Output features $\mathbf{Y} \in \mathbb{R}^{B\times K\times S}$

Step 1: Input Partitioning
Split $\mathbf{X}$ into P chunks $\{\mathbf{X}_p\}_{p=1}^{P}$ along temporal dimension
Launch P CUDA streams for parallel processing
Step 2: Normalization & Convolution
for each chunk $\mathbf{X}_p$ in parallel **do**
    $\tilde{\mathbf{X}}_p \leftarrow \text{InstanceNorm}(\mathbf{X}_p)$
    $\mathbf{X}'_p \leftarrow \text{Conv1D}(\tilde{\mathbf{X}}_p)$ ▷ Kernel fusion with dynamic gates
Step 3: State-Space Computation
for each chunk $\mathbf{X}'_p$ **do**
    Discretize HIPPO matrices:
    $\overline{\mathbf{A}}_p = \exp(\Delta_p \mathbf{A})$, $\overline{\mathbf{B}}_p = \mathbf{A}^{-1}(\exp(\Delta_p \mathbf{A}) - \mathbf{I})\mathbf{B}$
    Compute parallel state evolution:
    $\mathbf{H}_p \leftarrow \mathbf{C}(\sum_{k=0}^{T} \overline{\mathbf{A}}_p^k \overline{\mathbf{B}}_p \mathbf{X}'_{p,t-k}) + \mathbf{D}\mathbf{X}'_p$
Step 4: Gradient Checkpointing
Store $\{\mathbf{H}_p^0, \overline{\mathbf{A}}_p, \overline{\mathbf{B}}_p\}$ for backward pass
$\mathbf{Y} \leftarrow \text{Linear}(\text{GeLU}(\mathbf{H}^L))$ ▷ Layer-wise concatenation

B. Backward Propagation

The backward pass employs novel gradient computation strategies:

[Adjoint Gradient Calculation] For SSM parameters $\theta = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$, the gradient through continuous-time dynamics is:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \underbrace{\int_0^T \lambda(t)^\top \frac{\partial f}{\partial \theta} dt}_{\text{Adjoint term}} + \underbrace{\frac{\partial \mathcal{L}}{\partial \mathbf{H}(T)} \frac{\partial \mathbf{H}(T)}{\partial \theta}}_{\text{Direct term}} \tag{S6}$$

where the adjoint state $\lambda(t)$ solves the differential equation:

$$\frac{d\lambda}{dt} = -\lambda(t)^\top \frac{\partial f}{\partial \mathbf{H}} + \frac{\partial \mathcal{L}}{\partial \mathbf{H}(t)} \tag{S7}$$

B.1. Sparse Gradient Propagation

Implement Hoyer regularization with structured sparsity:

$$\frac{\partial \mathcal{L}_{\text{Hoyer}}}{\partial V_{j,s}^i} = \lambda \left(\frac{\text{sign}(V_{j,s}^i)}{\|V_j^i\|_1} - \frac{V_{j,s}^i \|V_j^i\|_1}{\|V_j^i\|_2^3} \right) \odot \mathbf{M}_{2:4} \tag{S8}$$

where $\mathbf{M}_{2:4}$ enforces NVIDIA's 2:4 sparse pattern for TensorCore acceleration.

C. Hardware Optimization Strategies

D. Theoretical Complexity Analysis

Let N = state dimension, T = sequence length:

$$\begin{aligned} &\text{SSM Forward: } \mathcal{O}(TN^2) \rightarrow \mathcal{O}(T^{0.5}N^2) \text{ via parallel scan} \\ &\text{Gradient Compute: } \mathcal{O}(T^2N) \rightarrow \mathcal{O}(T \log TN) \text{ with adjoint method} \end{aligned} \tag{S9}$$

Table S7. Hardware-Aware Optimization Techniques

Optimization	Implementation	Benefit
Memory Hierarchy	TensorCore submatrix tiling (128×256 blocks)	58% ↓ L2 cache misses
Communication Hiding	Async gradient all-reduce during SSM computation	22% ↓ PCIe bandwidth
Selective Checkpointing	Store HIPPO eigenvalues + initial states only	63% ↓ memory footprint

Design Principles Our implementation preserves three fMRI-specific properties:

1. **Temporal Causality**: HIPPO matrices enforce exponential decay of irrelevant historical states through eigenvalues $|\lambda_i| < 1$
2. **Spatial Locality**: Conv1D filters (kernel=5) model local hemodynamic couplings while SSM captures global dynamics
3. **Distribution Robustness**: Instance normalization eliminates scanner-specific intensity variance via:

$$\mu_c = \frac{1}{BT}\sum_{b,t} X_{b,t,c},\ \sigma_c^2 = \frac{1}{BT}\sum_{b,t}(X_{b,t,c} - \mu_c)^2 \tag{S10}$$

Table S8. the abbreviations of different ROIs and brain regions

AAL Regions	Abbreviation	AAL Regions	Abbreviation
Precental gyrus	PreCG	Lingual gyrus	LING
Superior frontal gyrus, dorsolateral	SFGdor	Superior occipital gyrus	SOG
Superior frontal gyrus, orbital part	ORBsup	Middle occipital gyrus	MOG
Middle frontal gyrus	MFG	Inferior occipital gyrus	IOG
Middle frontal gyrus, orbital part	ORBmid	Fusiform gyrus	FFG
Inferior frontal gyrus, opercular part	IFGoperc	Postcentral gyrus	PoCG
Inferior frontal gyrus, triangular part	IFGoperc	Superior parietal gyrus	SPG
Inferior frontal gyrus, orbital part	ORBinf	Inferior parietal, but supramarginal and angular gyri	IPL
Rolandic operculum	ROL	Supramarginal gyrus	SMG
Supplementary motor area	SMA	Angular gyrus	ANG
Olfactory cortex	OLF	Precuneus	PCUN
Superior frontal gyrus, medial	SFGmed	Paracentral lobule	PCL
Superior frontal gyrus, medial orbital	ORBsupmed	Caudate nucleus	CAU
Gyrus rectus	REC	Lenticular nucleus, putamen	PUT
Insula	INS	Lenticular nucleus, pallidum	PAL
Anterior cingulate and paracingulate gyri	ACG	Thalamus	THA
Median cingulate and paracingulate gyri	DCG	Heschl gyrus	HES
Posterior cingulate gyrus	PCG	Superior temporal gyrus	STG
Hippocampus	HIP	Temporal pole: superior temporal gyrus	TPOsup
Parahippocampal gyrus	PHG	Middle temporal gyrus	MTG
Amygdala	AMYG	Temporal pole: middle temporal gyrus	TPOmid
Calcarine fissure and surrounding cortex	CAL	Inferior temporal gyrus	ITG
Cuneus	CUN		
Brain Regions	**Abbreviation**	**Brain Regions**	**Abbreviation**
default mode network	DMN	frontoparietal network	FPN
limbic network	LN	ventral attention network	VAN
sensorimotor network	SMN	dorsal attention network	DAN
visual network	VN	subcortical system	SUB

REFERENCES

1. V. Jakkula, "Tutorial on support vector machine (svm)," Sch. EECS, Wash. State Univ. **37**, 3–6 (2006).
2. K. Ma, S. Huang, P. Wan, and D. Zhang, "Optimal transport based pyramid graph kernel for autism spectrum disorder diagnosis," Pattern Recognit. **143**, 109716 (2023).
3. S. Yu, S. Jin, M. Li, T. Sarwar, and F. Xia, "Long-range brain graph transformer," Adv. Neural Inf. Process. Syst. **37**, 24472–24495 (2024).
4. W. Dong, Y. Li, W. Zeng, L. Chen, H. Yan, W. T. Siok, and N. Wang, "Starformer: A novel spatio-temporal aggregation reorganization transformer of fmri for brain disorder diagnosis," Neural Networks p. 107927 (2025).
5. X. Kan, W. Dai, H. Cui, Z. Zhang, Y. Guo, and C. Yang, "Brain network transformer," Adv. Neural Inf. Process. Syst. **35**, 25586–25599 (2022).
6. M. Liu, H. Zhang, F. Shi, and D. Shen, "Hierarchical graph convolutional network built by multiscale atlases for brain disorder diagnosis using functional connectivity," IEEE Transactions on Neural Networks Learn. Syst. (2023).
7. Y. Wang, H. Long, Q. Zhou, T. Bo, and J. Zheng, "Plsnet: Position-aware gcn-based autism spectrum disorder diagnosis via fc learning and rois sifting," Comput. Biol. Medicine p. 107184 (2023).
8. B. Song and S. Yoshida, "Temporal graphormer and its interpretability: A novel framework for diagnostic decoding of brain disorders using fmri data," Biomed. Signal Process. Control. **104**, 107467 (2025).
9. B. Lei, Y. Zhao, Z. Huang, X. Hao, F. Zhou, A. Elazab, J. Qin, and H. Lei, "Adaptive sparse learning using multi-template for neurodegenerative disease diagnosis," Med. Image Analysis **61**, 101632 (2020).