STH-BRAIN: A HIERARCHICAL SPATIOTEMPORAL BRAIN TIME-SERIES FOUNDATION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Resting-state functional magnetic resonance imaging (fMRI) captures spontaneous fluctuations in blood-oxygen-level-dependent signals across the brain. Differences in temporal sampling complicate consistent modeling across scans, while spatial aggregation can obscure the temporal context of individual brain regions. We propose STH-Brain, a brain time-series foundation model combining physical spatiotemporal encoding with hierarchical representation learning. Physical-Time Rotary Encoding maps elapsed seconds to rotary phases, while Spatial-Aware Encoding incorporates atlas coordinates into spatial attention. The hierarchy models temporal context within each region of interest (ROI) and retains regional states alongside network representations. During self-supervised pretraining, dual-level latent prediction supervises both levels, with visibility-based routing selecting same-ROI or network memory for regional prediction according to the available context. Across seven disorder-classification tasks, STH-Brain achieves the best mean performance among the evaluated baselines, improving accuracy by an average of 5.91 percentage points over the strongest baseline for each task. It also achieves the best mean results for all three age-prediction metrics across seven disease groups and outperforms the compared models on the evaluated external-cohort and zero-shot transfer tasks. Interpretability analyses reveal shared and distinct regional patterns across disorders. Code is available at `https://anonymous.4open.science/r/STH-Brain/`.

1 INTRODUCTION

Functional connectivity (FC) is widely used to study brain organization from resting-state functional magnetic resonance imaging (fMRI) Li et al. (2021); Kan et al. (2022). By summarizing associations between blood-oxygen-level-dependent (BOLD) signals from regions of interest (ROIs) over an entire scan, it provides a compact representation for graph-based diagnosis and large-scale pretraining Yang et al. (2024); Wei et al. (2026). Brain activity, however, evolves over time, and a single whole-scan FC matrix does not describe the temporal progression of regional activity and interregional relationships. Dynamic modeling can characterize these changes, providing a more temporally detailed account of brain activity than whole-scan summaries (Figure 1).

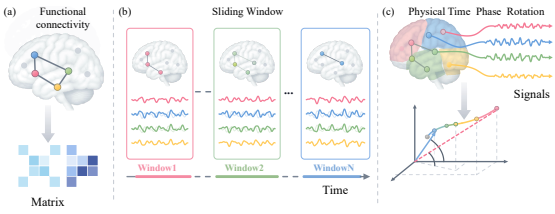


Figure 1: Representations of brain activity: (a) functional connectivity (FC) computed over an entire scan; (b) dynamic FC estimated using sliding windows; (c) physical time mapped to rotary phases.

Dynamic approaches capture temporal information through windowed connectivity or direct signal modeling. MDGL Ma et al. (2023) and MSSTAN Kong et al. (2024) model sequences of sliding-window functional connectomes, while BrainNPT Hu et al. (2024) uses window-derived connectivity snapshots for pretraining. Windowed correlations summarize associations within each window without retaining the ordering of its observations. Window duration determines the observations included in each correlation estimate, while stride determines the spacing between successive estimates Zhao et al. (2024). Direct spatiotemporal models retain the signal sequence: STAR-

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.

Former Dong et al. (2025) uses variable-window and cross-window attention, FROG Zhang et al. (2025) combines temporal and graph convolutions, and DBSTP Li et al. (2026b) exchanges information between spatial and temporal streams before constructing attention-based connectivity. When temporal positions and scales are specified only in frames, their physical-time meaning depends on the sampling interval. Identical frame indices and spans can therefore correspond to different elapsed times and durations across scans. For spatial modeling, connectivity values alone do not specify the anatomical coordinates of individual ROIs.

Pretraining further shapes how regional temporal information is learned and organized. BrainLM learns from masked reconstruction of regional signals Ortega Caro et al. (2024), while NeuroSTORM and MAE-ST reconstruct four-dimensional fMRI Wang et al. (2026a); Gao et al. (2025). Brain Harmony and FlexiBrain address acquisition heterogeneity through repetition-time-adaptive tokenization and physical-unit patches, respectively Dong et al. (2026); Wang et al. (2026b). Beyond reconstruction, TAGSC learns spatiotemporal brain-graph representations through spatiotemporal augmentation and contrastive learning Li et al. (2026c), while Brain-JEPA predicts contextualized regional patches Dong et al. (2024). Brain-Semantoks aggregates regional signals into network–time tokens for masked prediction and global alignment across temporal views Gijsen et al. (2026). Reconstruction, contrastive learning, and latent prediction do not by themselves ensure that ROI-specific temporal states remain separately accessible. When regional states are aggregated into shared network tokens without being retained separately, the resulting representations no longer provide direct access to individual ROI states.

STH-Brain addresses these issues by combining physical spatiotemporal encoding with an ROI-preserving hierarchy. Physical-Time Rotary Encoding (PTRE) maps elapsed time in seconds to rotary phases, while Spatial-Aware Encoding incorporates anatomical coordinates into spatial attention. Hierarchical Spatiotemporal Integration models temporal context within each ROI before aggregating regional information into network tokens. The resulting hierarchy retains regional states alongside representations that integrate context across networks and time. During pretraining, latent prediction supervises both levels, and visibility-based routing selects same-ROI or network memory for ROI prediction according to context availability. Our main contributions are as follows.

- We propose physical spatiotemporal encoding to address temporal-position ambiguity across sampling intervals and the lack of explicit anatomical references in attention. Physical-Time Rotary Encoding maps elapsed seconds to rotary phases, while Spatial-Aware Encoding incorporates ROI coordinates into spatial attention.
- We propose an ROI-preserving hierarchy to prevent the loss of direct access to ROI-specific temporal states during network aggregation. It models temporal context through same-ROI memory and retains regional states alongside network representations, with dual-level latent prediction and visibility-based routing for pretraining.
- We present STH-Brain, a hierarchical spatiotemporal pretraining framework that achieves the best mean performance among the evaluated baselines on seven disorder-classification tasks, age prediction, external-cohort evaluation, and zero-shot transfer. Regional analyses identify shared and distinct patterns across disorders, supporting the interpretability of model predictions.

2 Related Work

2.1 Dynamic Brain Network Modeling

Dynamic brain modeling characterizes changes in regional activity and interregional interactions. Community-aware and hierarchical graph methods organize regional interactions into subnetworks Bannadabhavi et al. (2023); Xue et al. (2025); Ma et al. (2026), while spatiotemporal hypergraphs represent higher-order relationships across regions and time Liu et al. (2023); Li et al. (2026a). Window-based dynamic graph models combine graph readout and Transformer attention to learn from sequences of functional connectomes Kim et al. (2021). Direct signal models extract spatiotemporal features through four-dimensional convolution and attention Jiang et al. (2023); Kim et al. (2023), or learn connectivity from temporally encoded regional signals Thapaliya et al. (2025). Joint approaches also model multiscale regional time series and derived functional connectivity through graph convolution and spatial attention Huang et al. (2022). However, frame indices

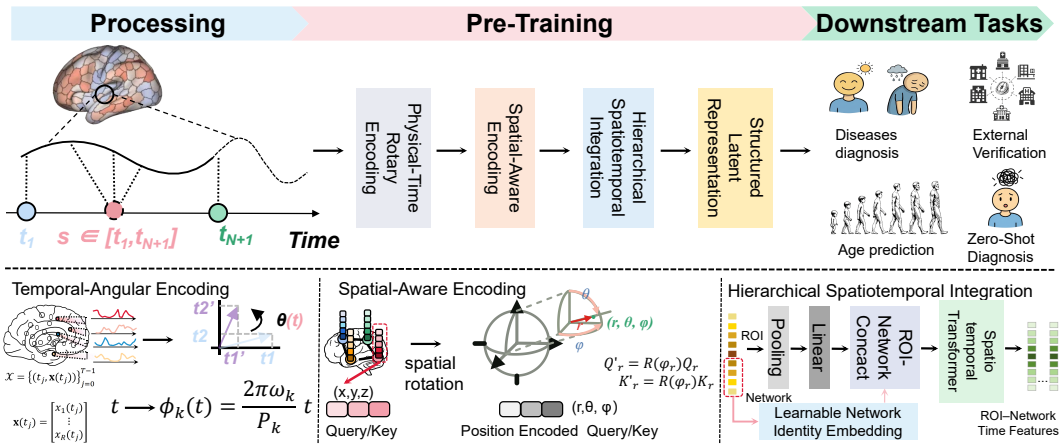


Figure 2: Overall workflow of STH-Brain.

alone do not specify elapsed time across sampling intervals, while connectivity values alone do not specify anatomical coordinates. STH-Brain encodes elapsed seconds as rotary phases and incorporates atlas coordinates into spatial attention, providing explicit temporal and anatomical references for modeling brain dynamics.

2.2 Pretraining for Brain Representation Learning

Pretraining methods differ in their input representations and learning objectives. Connectivity-based methods learn interregional associations from graph summaries Wei et al. (2026); Chen et al. (2025a). Masked autoencoding has also been applied to dynamic functional-connectivity matrices before downstream classification Ma et al. (2024). Regional-signal and volumetric encoders learn from time series and four-dimensional fMRI Ortega Caro et al. (2024); Wang et al. (2026a), while repetition-time-adaptive tokenization and patches defined in physical units address acquisition heterogeneity Dong et al. (2026); Wang et al. (2026b). Generative approaches use flow matching Tew et al. (2026) or conditional diffusion Xia et al. (2026a;b). Latent approaches predict contextualized regional patches Dong et al. (2024) or combine masked prediction of network–time tokens with alignment across temporal views Gijsen et al. (2026). However, reconstruction, alignment, and latent prediction do not by themselves ensure that representations of ROI-specific temporal context remain separately accessible. STH-Brain explicitly models this context through same-ROI temporal memory and retains regional states alongside network representations, with predictive supervision at both levels.

3 Methods

3.1 Overview

STH-Brain combines Physical-Time Rotary Encoding and Spatial-Aware Encoding to incorporate elapsed time and anatomical coordinates into attention. Hierarchical Spatiotemporal Integration models temporal context within each ROI before network aggregation, retaining regional states alongside contextualized network representations. Visibility Routing selects regional or network memory for masked ROI prediction according to the availability of same-ROI temporal context. Dual-level latent prediction supervises both regional and network representations during self-supervised pretraining.

Input Representation and Notation: For scan i, let $X_i \in \mathbb{R}^{T_i \times R}$ contain T_i native observations from R atlas-defined ROIs. Each ROI r has coordinate $c_r \in \mathbb{R}^3$ and fixed network membership $n(r) \in \{1, \ldots, N\}$, with $N = 8$. The scan comprises B_i non-overlapping blocks of duration Δ_{blk}, indexed by b, with centers τ_{ib} measured in seconds from the scan origin. The indicator $v_{ibr} \in \{0, 1\}$ combines block support and ROI validity. Metadata P_i collects timestamps, sampling intervals,

block durations, atlas identities, coordinates, and validity masks; d and D denote the ROI and network feature dimensions, respectively. Appendix A details support and normalization.

3.2 PHYSICAL-TIME ROTARY ENCODING

Local Temporal Encoder: Each ROI is standardized using native observations visible in the current scan view, with frozen training-only statistics used when support is insufficient or its scale is degenerate. Let Δ_q denote the physical query spacing. Supported interpolation onto this grid is followed by additive within-block time features, multiscale convolutions, local attention, and temporal pooling, producing one feature $h^{\mathrm{loc}}_{ibr} \in \mathbb{R}^d$ per ROI and block (Appendix A).

Rotary Phase Definition: To compare blocks in physical units, let t denote elapsed time in seconds, k index temporal channel pairs, $P_k > 0$ be a fixed base period measured in seconds, a_k a learned calibration parameter, and $\kappa > 0$ a fixed bound on log-frequency adjustment. PTRE defines the frequency multiplier ω_k and rotary phase $\phi_k(t)$ as follows:

$$\omega_k = \exp(\kappa \tanh a_k), \qquad \phi_k(t) = 2\pi t \omega_k / P_k. \tag{1}$$

Cross-block ROI and network attention rotate query/key pairs using these phases at $t = \tau_{ib}$. Phase differences depend on elapsed time differences in seconds, while the bounded multiplier ω_k calibrates each temporal scale.

3.3 SPATIAL-AWARE ENCODING

Spherical Coordinate Encoding: Let $\widehat{\rho}_r$, ϑ_r, and ψ_r denote atlas-normalized radius, polar angle, and azimuth derived from c_r. For spatial channel pair h, let $\iota_h \in \{1, 2, 3\}$ select a coordinate component and a^S_h be its fixed harmonic multiplier. Define the coordinate map Γ and channel-pair rotation phase $\varphi_{r,h}$ as follows:

$$\Gamma(c_r) = (\pi\widehat{\rho}_r, \vartheta_r, \psi_r), \qquad \varphi_{r,h} = a^S_h [\Gamma(c_r)]_{\iota_h}. \tag{2}$$

Within-Network Spatial Attention: Let $\varphi_r = (\varphi_{r,h})_h$ collect ROI r's spatial phases and $\mathcal{R}(\varphi_r)$ be their block-diagonal planar rotation matrix. Let d_{att} be the attention-head width. For valid ROIs r, r' in the same network and block, let $q_r, k_{r'} \in \mathbb{R}^{d_{\mathrm{att}}}$ be query and key projections of the local features, suppressing scan and block indices. Each head uses its corresponding phase pairs. With softmax taken over valid member ROIs r' of that network, define the pairwise weight $A_{rr'}$ as follows:

$$A_{rr'} = \mathrm{softmax}_{r'}\left(q_r^\top \mathcal{R}(\varphi_r)^\top \mathcal{R}(\varphi_{r'}) k_{r'} / \sqrt{d_{\mathrm{att}}} \right). \tag{3}$$

Geometry modulates the scores, and projected value aggregation is added residually to the local features, producing $H_i \in \mathbb{R}^{B_i \times R \times d}$ while retaining ROI identity. Network attention uses atlas-derived network geometry statistics (Appendix B).

3.4 HIERARCHICAL SPATIOTEMPORAL INTEGRATION

Regional Temporal and Network Encoders: Let T_R be a shared Transformer acting separately on each ROI's visible blocks. Its output $U_i \in \mathbb{R}^{B_i \times R \times d}$ retains regional trajectories. For network n in block b, define its valid member set $\mathcal{R}_{ibn} = \{r : n(r) = n,\ v_{ibr} = 1\}$. Let $\mathcal{S}_n$ aggregate these regional states by their mean, channel-wise standard deviation, and learned attention summary, followed by projection and network-identity coupling. Let $z_{ibn} \in \mathbb{R}^D$ be the resulting token, T_N the network Transformer, and $V_i \in \mathbb{R}^{B_i \times N \times D}$ its contextualized network output. Using a colon to index the block axis, the hierarchy on valid positions of the current view is:

$$U_{i,:,r} = T_R(H_{i,:,r}; P_i), \qquad z_{ibn} = \mathcal{S}_n(\{U_{ibr} : r \in \mathcal{R}_{ibn}\}), \qquad V_i = T_N(\{z_{ibn}\}; P_i). \tag{4}$$

The shared T_R attends along each ROI's visible valid blocks, and T_N attends across visible valid network–time cells. The retained U_i supplies regional prediction and downstream readout; V_i integrates network summaries across time. Detailed aggregation and dimensions are given in Appendix C.

3.5 Self-Supervised Learning via Dual-Level Prediction

Masking and Teacher Targets: With $|\cdot|$ denoting set cardinality, let $\mathcal{V}_i = \{(b,n) : |\mathcal{R}_{ibn}| > 0\}$ be the valid network–time cells. Temporal and network stripes select target cells $\mathcal{M}_i \subset \mathcal{V}_i$, with disjoint neighboring exclusion cells $\mathcal{E}_i \subset \mathcal{V}_i$ and visible context $\mathcal{C}_i = \mathcal{V}_i \setminus (\mathcal{M}_i \cup \mathcal{E}_i)$. Masks retain nonempty context and target sets, and ROI membership maps these cells to regional positions. Let θ, ξ be the parameters of student encoder F_θ and its exponential-moving-average (EMA) teacher F_ξ, respectively; each encoder includes view-specific signal normalization and the full hierarchy. Superscripts R, N label regional and network outputs U, V, and $\ell \in \{R, N\}$ indexes these levels. Write F_ξ^ℓ for the corresponding teacher output. Let LN denote non-affine layer normalization over feature channels and sg the stop-gradient operator, which preserves its input in the forward pass and blocks gradient propagation through it in the backward pass. Let 0 denote raw zero-valued signals with the same shape as X_i. The teacher target representation Y_i^ℓ is:

$$Y_i^\ell = \mathrm{sg}\big[\mathrm{LN}\big(F_\xi^\ell(X_i; P_i) - F_\xi^\ell(0; P_i)\big)\big]\,. \tag{5}$$

Both teacher passes use complete valid support and shared metadata. Subtraction precedes normalization, and entries in $\mathcal{M}_i$ supply the targets. EMA updates the teacher from the student.

Visibility Routing: Visibility routing selects the ROI predictor's memory according to the availability of visible states from the target ROI. Target queries are constructed from identity and physical metadata. U_i, V_i denote student features encoded from $\mathcal{C}_i$. For target ROI r, define its visible regional memory $\mathcal{U}_{ir}^{\mathcal{C}} = \{U_{ibr} : (b, n(r)) \in \mathcal{C}_i,\ v_{ibr} = 1\}$ and the scan's visible network memory $\mathcal{V}_i^{\mathcal{C}} = \{V_{ibn} : (b,n) \in \mathcal{C}_i\}$. The network predictor uses $\mathcal{V}_i^{\mathcal{C}}$, while the regional predictor selects its memory $\mathcal{K}_{ir}$ as:

$$\mathcal{K}_{ir} = \begin{cases} \mathcal{U}_{ir}^{\mathcal{C}}, & |\mathcal{U}_{ir}^{\mathcal{C}}| > 0, \\ \mathcal{V}_i^{\mathcal{C}}, & |\mathcal{U}_{ir}^{\mathcal{C}}| = 0. \end{cases} \tag{6}$$

The ROI predictor uses the selected memory source at every layer; the network predictor uses visible network memory. Separate projections map memory to predictor width. The first cross-attention layer initializes each target state from visible content, and subsequent residual updates refine it (Appendix D).

Prediction Losses: Let $\widehat{Y}_{ibr}^R \in \mathbb{R}^d$ and $\widehat{Y}_{ibn}^N \in \mathbb{R}^D$ denote predictions, and let k index feature channels. For scalar error e, let $\ell_\beta(e)$ be SmoothL1 with transition parameter $\beta > 0$; use $\beta_R, \beta_N > 0$ for the two levels (Appendix D.4). Let $\mathcal{S}$ be the global batch's subject set, s its subject index, $S = |\mathcal{S}|$, and $\mathcal{I}_s$ the accepted scans of subject s. For a scan quantity f_i, define $\langle f_i \rangle_{\mathrm{subj}} = S^{-1} \sum_{s \in \mathcal{S}} |\mathcal{I}_s|^{-1} \sum_{i \in \mathcal{I}_s} f_i$. The regional and network losses $\mathcal{L}_R, \mathcal{L}_N$ are:

$$\mathcal{L}_R = \left\langle \frac{1}{|\mathcal{M}_i|} \sum_{(b,n)\in\mathcal{M}_i} \frac{1}{|\mathcal{R}_{ibn}| d} \sum_{r \in \mathcal{R}_{ibn}} \sum_{k=1}^{d} \ell_{\beta_R}(\widehat{Y}_{ibrk}^R - Y_{ibrk}^R) \right\rangle_{\mathrm{subj}}, \tag{7}$$

$$\mathcal{L}_N = \left\langle \frac{1}{|\mathcal{M}_i|} \sum_{(b,n)\in\mathcal{M}_i} \frac{1}{D} \sum_{k=1}^{D} \ell_{\beta_N}(\widehat{Y}_{ibnk}^N - Y_{ibnk}^N) \right\rangle_{\mathrm{subj}}. \tag{8}$$

Target members $\mathcal{R}_{ibn}$ use complete valid support. The reduction weights subjects equally and target cells equally within each scan.

Subject-Diversity Losses: For branch ℓ, define feature widths $q_R = d$, $q_N = D$, and let j index its anatomical identities. Complete-view features are averaged over valid blocks, normalized per identity, and averaged over scans within each subject to form $z_{sj}^\ell \in \mathbb{R}^{q_\ell}$. Let $\mathcal{J}_\ell$ contain identities observed in at least two subjects and v_{jk}^ℓ be their channel-wise sample variances across observed subjects. Define $[x]_+ = \max(x, 0)$. The variance loss $\mathcal{L}_{\mathrm{var}}$ is:

$$\mathcal{L}_{\mathrm{var}} = \frac{1}{2} \sum_{\ell \in \{R,N\}} \frac{\sum_{j \in \mathcal{J}_\ell} \sum_{k=1}^{q_\ell} \left[0.5 - \sqrt{v_{jk}^\ell + 10^{-4}}\right]_+}{\max(1, |\mathcal{J}_\ell|) q_\ell}. \tag{9}$$

Let $F_\ell = |\mathcal{J}_\ell| q_\ell$ be the supported feature count and $E_\ell \in \mathbb{R}^{S \times F_\ell}$ the matrix of centered subject–identity residuals defined in Appendix D.5. Set $G_\ell = E_\ell E_\ell^\top$, $K_\ell = G_\ell / \max(\mathrm{tr}\, G_\ell, 10^{-12})$, and

$c_\ell = \min(S-1, F_\ell)$. Let I_S be the identity matrix and $\alpha = 1$ the spectral scale. The log-determinant (LogDet) loss $\mathcal{L}_{\text{logdet}}$ is:

$$\mathcal{L}_{\text{logdet}} = \frac{1}{2} \sum_{\substack{\ell \in \{R,N\} \\ c_\ell > 0}} \left[\log(1+\alpha) - \frac{1}{c_\ell} \log\det(I_S + \alpha c_\ell K_\ell) \right]_+ . \tag{10}$$

Variance encourages cross-subject variation, while LogDet promotes its dispersion across representation directions.

Training Objective: Let $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ weight the four losses, respectively; λ_4 follows the appendix's warm-up schedule. The total pretraining loss $\mathcal{L}_{\text{pre}}$ is:

$$\mathcal{L}_{\text{pre}} = \lambda_1 \mathcal{L}_R + \lambda_2 \mathcal{L}_N + \lambda_3 \mathcal{L}_{\text{var}} + \lambda_4 \mathcal{L}_{\text{logdet}}. \tag{11}$$

Prediction losses train the student and predictors, regularization trains the shared encoder, and EMA updates the teacher. The downstream task loss $\mathcal{L}_{\text{task}}$ is class-weighted cross-entropy or SmoothL1 on standardized ages (Appendix F.2).

4 Experiments

4.1 Datasets and Preprocessing

Datasets. We include 12,774 unique subjects (14,766 scans) from 9 datasets: ABIDE-I, ABIDE-II Di Martino et al. (2017), ADHD Bellec et al. (2017), ADNI Jack Jr et al. (2008), PPMI Marek et al. (2011), SRPBS Tanaka et al. (2021), COBRE Calhoun et al. (2012), REST-MDDI Yan et al. (2019), and REST-MDDII Chen et al. (2025b). Before partitioning the remaining REST-MDDII subjects, separate schizophrenia (SCZ)/healthy controls (HC) and bipolar disorder (BP)/HC cohorts are reserved for external validation and zero-shot transfer, respectively. Dataset characteristics, matching procedures, and task-specific training and evaluation partitions are detailed in Appendix G.1.

Preprocessing. The fMRI data were preprocessed using DPABI (`http://www.rfmri.org/dpabi`). After removing the first 10 volumes, preprocessing included slice-timing correction, six-parameter rigid-body realignment, nuisance-covariate regression, T1 coregistration, tissue segmentation, and DARTEL-based MNI-152 normalization. Linear trends were regressed out, and a 0.01–0.1 Hz band-pass filter was applied. The brain is divided into 200 regions using the Craddock 200-region atlas (CC200) Craddock et al. (2012). Further details are provided in Appendix G.2.

4.2 Experimental Setup

Baseline Comparisons. The protocol specifies 13 baselines: 9 models in the pretraining comparison group: BrainMass Yang et al. (2024), Brain-JEPA Dong et al. (2024), Brain-Semantoks Gijsen et al. (2026), FROG Zhang et al. (2025), BCNSL Chen et al. (2025a), FMM-TC Fan et al. (2025), PHGNN Liu & Rossi (2025), DeepASD Chen et al. (2024), and xAI-ASD Vidya et al. (2025); and four classification frameworks, BrainHGT Ma et al. (2026), BrainIB Zheng et al. (2024), HiGCN Huang et al. (2024), and ALTER Yu et al. (2024). All models shared the same environment, data, and splits, with baselines using their published hyperparameters.

Evaluation Metrics. Classification performance is evaluated using accuracy (ACC), area under the receiver operating characteristic curve (AUC), and F1 score, all reported as percentages. Age prediction is evaluated using mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation R between predicted and chronological ages. MAE and RMSE are reported in years, whereas R is dimensionless. For each task and metric, the top two models are compared using a two-sided paired t-test across 10 matched seeds. Statistical significance is defined as $p < 0.05$. Further details are provided in Appendix G.4.

Implementation Details. Experiments use PyTorch 2.9.1 and AdamW with weight decay 0.05. Pretraining uses distributed data parallelism on 8 NVIDIA RTX PRO 6000D GPUs for up to 500 epochs, with 16 samples per GPU (128 globally), encoder/predictor learning rates of $1.5 \times 10^{-5}/3 \times 10^{-5}$, and a 10-epoch warm-up followed by cosine decay. Fine-tuning uses 1 GPU for up to 200 epochs, an effective batch size of 16, backbone/head learning rates of $10^{-5}/3 \times 10^{-4}$, and validation-based early stopping with a patience of 10 epochs. Further settings appear in Appendix G.3.

Table 1: Classification results for AD, PD, ASD, and MDD (%, mean ± SD across 10 seeds). **Bold** and underlined values denote the best and second-best results, respectively. * denotes statistical significance between the best and second-best models (paired t-test, $p < 0.05$).

Method	Pretraining	AD			PD			ASD			MDD		
		ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑
BrainHGT	—	67.13±2.82	73.74±3.31	67.26±2.83	67.73±2.56	73.98±3.58	69.61±2.68	66.18±2.68	72.89±3.38	65.53±2.69	66.55±3.32	73.41±3.61	70.22±3.14
BrainIB	—	66.21±3.62	72.79±3.75	66.86±3.55	66.69±2.58	73.24±3.66	64.82±2.71	65.69±2.79	72.49±3.48	65.42±2.77	65.85±2.71	72.51±2.55	69.75±2.57
HiGCN	—	65.29±3.16	71.81±2.96	64.56±3.26	65.65±2.96	72.15±3.27	67.57±2.88	65.19±3.06	72.05±3.62	64.77±3.04	65.14±2.79	71.88±2.82	69.19±2.62
ALTER	—	64.37±3.49	71.07±3.59	63.73±3.57	64.62±2.68	71.05±2.72	63.77±2.73	64.71±3.07	71.69±3.11	63.45±3.12	64.44±3.57	71.13±3.67	68.77±3.34
BrainMass	✓	72.64±3.06	79.29±2.83	73.19±3.03	77.08±2.78	83.49±3.18	77.48±2.76	66.68±3.42	73.58±3.05	65.46±3.52	72.91±3.72	79.84±3.39	76.61±3.36
Brain-JEPA	✓	74.49±3.17	81.13±3.39	74.32±3.19	76.04±2.67	82.66±3.38	75.58±2.89	70.14±3.09	77.02±2.72	69.69±3.09	72.19±3.53	79.36±3.59	75.91±3.22
Brain-Semantoks	✓	73.57±3.54	80.01±2.98	73.47±3.54	75.12±2.53	81.59±3.39	74.21±2.75	69.65±3.53	76.43±3.43	67.91±3.68	71.49±3.06	78.68±2.84	75.59±2.76
FROG	✓	75.41±3.24	82.09±2.97	75.65±3.22	73.96±2.99	80.31±3.55	75.33±2.83	69.15±2.63	76.06±3.28	68.23±2.63	70.78±3.07	77.78±3.46	74.34±2.84
BCNSL	✓	71.72±2.95	78.46±3.52	71.51±2.98	72.92±3.09	79.42±2.94	72.31±3.19	68.66±2.59	75.48±3.01	67.32±2.66	70.08±2.85	77.04±2.85	73.66±2.65
FMM-TC	✓	70.82±2.59	77.27±3.48	70.38±2.65	71.88±2.55	78.45±2.59	73.37±2.32	68.16±3.49	74.85±2.93	66.96±3.56	69.37±2.92	76.47±2.81	73.12±2.71
PHGNN	✓	69.88±2.71	76.63±2.91	70.02±2.72	70.85±2.57	77.34±3.14	72.56±2.34	67.67±3.26	74.51±3.57	65.93±3.37	68.67±3.62	75.67±2.65	72.45±3.37
DeepASD	✓	68.97±3.42	75.57±3.27	68.77±3.45	69.81±2.75	76.39±2.61	71.52±2.62	67.17±3.23	73.88±3.17	66.49±3.24	67.96±2.61	74.89±3.68	72.32±2.39
xAI-ASD	✓	68.05±3.13	74.83±2.86	68.06±3.17	68.77±2.52	75.27±3.05	68.23±2.79	70.64±3.51	77.42±2.81	69.09±3.65	67.26±3.26	74.06±3.47	71.56±2.99
STH-Brain	✓	**81.24±2.34***	**87.75±2.15***	**81.12±2.36***	**83.23±2.42***	**89.69±2.24***	**83.42±2.28***	**76.43±2.29***	**83.18±2.13***	**75.57±2.34***	**78.38±2.35***	**85.49±2.23***	**81.03±2.15***

Downstream Tasks. Internal evaluation comprised age prediction and seven binary classification tasks, each distinguishing a disorder from HC: Alzheimer's disease (AD), major depressive disorder (MDD), autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD), mild cognitive impairment (MCI), Parkinson's disease (PD), and prodromal Parkinson's disease (PROPD). Internal tasks used fixed subject-level training, validation, and test partitions in a 7:1:2 ratio. Downstream models were initialized with the student encoder from the checkpoint achieving the lowest epoch-averaged total pretraining loss on the training data. The encoder was then jointly fine-tuned with a newly initialized task-specific multilayer perceptron (MLP) for classification or age prediction. External generalization was assessed by fine-tuning models on SCZ/HC source data and evaluating them on an independent SCZ/HC target cohort. Zero-shot transfer was assessed by directly applying models fine-tuned on MDD/HC source data to BP/HC target cohort without target adaptation. Both target cohorts were excluded from pretraining and supervised model development. Each downstream experiment was repeated with 10 random seeds while keeping the data partitions fixed, and performance is reported as mean ± standard deviation (SD) across runs. Further details are provided in Appendices F and G.1.

4.3 Evaluation on Classification Performance

STH-Brain achieved the highest mean ACC, AUC, and F1 across all seven classification tasks. Table 1 presents the results for AD, PD, ASD, and MDD; results for ADHD, MCI, and PROPD are provided in Appendix H.1 (Table 8). The second-best models in ACC were FROG, BrainMass, xAI-ASD, and BrainMass for AD, PD, ASD, and MDD, respectively. STH-Brain surpassed these models by 5.83, 6.15, 5.79, and 5.47 percentage points in ACC, respectively. Across all seven tasks, the average ACC improvement over the corresponding second-best model was 5.91 percentage points. Relative to the second-best result for each task and metric, improvements ranged from 5.65 to 6.37 percentage points for AUC and from 4.42 to 7.12 percentage points for F1.

4.4 Evaluation on Regression Performance

We evaluated age prediction across 7 diseases using MAE, RMSE, and R. STH-Brain achieved the best mean performance in all 21 group–metric combinations. Table 2 presents the results for AD, PD, ASD, and MDD. Compared with the lowest-MAE baseline in each of these four groups, STH-Brain reduced MAE by 0.157, 0.125, 0.087, and 0.125 years, respectively. Across all 7 groups, RMSE reductions relative to the lowest-RMSE baseline in each group ranged from 0.060 to 0.280 years, and R was higher than the strongest baseline in every group. These results indicate more accurate age estimates and stronger linear associations between predicted and chronological ages. Results for ADHD, MCI, and PROPD are reported in Appendix H.2 (Table 9).

4.5 External Generalization and Zero-Shot Transfer

Generalization Performance on External Dataset. Models fine-tuned on SCZ/HC training subjects and selected using the corresponding source-validation data were evaluated directly on the held-out SCZ/HC cohort (Table 3). STH-Brain achieved the highest mean ACC, AUC, and F1. It

Table 2: Age-prediction results for AD, PD, ASD, and MDD (mean ± SD across 10 seeds). **Bold** and underlined values denote the best and second-best results, respectively. * denotes statistical significance between the best and second-best models (paired t-test, $p < 0.05$).

Method	Pretraining	AD			PD			ASD			MDD		
		MAE↓	RMSE↓	Pearson R ↑	MAE↓	RMSE↓	Pearson R ↑	MAE↓	RMSE↓	Pearson R ↑	MAE↓	RMSE↓	Pearson R ↑
BrainHGT	—	3.765±0.408	4.812±0.700	0.876±0.020	3.268±0.558	4.351±0.875	0.935±0.030	2.216±0.116	2.878±0.106	0.948±0.004	3.313±0.146	4.308±0.184	0.956±0.003
BrainIB	—	3.822±0.509	4.948±0.504	0.887±0.038	3.345±0.645	4.406±1.155	0.930±0.035	2.249±0.115	2.898±0.147	0.950±0.005	3.360±0.154	4.367±0.180	0.955±0.003
HiGCN	—	3.906±0.386	4.923±0.599	0.862±0.037	3.408±0.570	4.192±0.783	0.944±0.019	2.308±0.101	3.010±0.145	0.944±0.004	3.447±0.073	4.469±0.090	0.954±0.002
ALTER	—	4.032±0.561	5.120±0.826	0.857±0.043	3.508±0.679	4.503±0.706	0.936±0.020	2.366±0.134	3.028±0.170	0.945±0.006	3.518±0.227	4.573±0.332	0.951±0.006
BrainMass	✓	2.829±0.371	3.554±0.478	0.929±0.027	2.425±0.464	3.023±0.695	0.971±0.013	1.662±0.095	2.167±0.123	0.971±0.005	2.443±0.133	3.158±0.151	0.976±0.002
Brain-JEPA	✓	2.850±0.343	3.790±0.594	0.911±0.033	2.481±0.447	3.129±0.598	0.968±0.013	1.674±0.082	2.186±0.140	0.970±0.003	2.503±0.114	3.263±0.178	0.974±0.003
Brain-Semantoks	✓	2.821±0.333	3.555±0.300	0.926±0.009	2.457±0.431	3.309±0.728	0.972±0.017	1.660±0.068	2.102±0.103	0.972±0.003	2.479±0.106	3.195±0.138	0.975±0.002
FROG	✓	2.796±0.318	3.497±0.405	0.931±0.017	2.476±0.450	3.080±0.537	0.971±0.009	1.672±0.097	2.158±0.125	0.971±0.004	2.497±0.116	3.226±0.145	0.975±0.003
BCNSL	✓	3.347±0.451	4.458±0.678	0.896±0.033	2.915±0.537	3.627±0.833	0.969±0.015	1.965±0.074	2.553±0.101	0.959±0.003	2.928±0.136	3.791±0.190	0.965±0.004
FMM-TC	✓	3.428±0.423	4.330±0.557	0.895±0.028	3.011±0.571	3.916±0.786	0.946±0.022	2.027±0.130	2.616±0.177	0.957±0.006	3.029±0.155	3.948±0.187	0.963±0.003
PHGNN	✓	3.493±0.499	4.450±0.499	0.891±0.031	3.049±0.602	3.920±0.485	0.960±0.017	2.045±0.138	2.636±0.206	0.957±0.007	3.071±0.200	3.949±0.265	0.962±0.005
DeepASD	✓	3.597±0.465	4.511±0.535	0.896±0.039	3.148±0.586	3.918±0.589	0.953±0.010	2.116±0.114	2.734±0.168	0.956±0.006	3.175±0.186	4.155±0.291	0.959±0.006
xAI-ASD	✓	2.826±0.373	3.657±0.467	0.920±0.012	2.469±0.472	3.081±0.619	0.971±0.015	1.638±0.091	2.110±0.111	0.970±0.003	2.492±0.139	3.251±0.179	0.974±0.003
STH-Brain	✓	**2.639±0.320***	**3.326±0.378***	**0.941±0.018***	**2.300±0.420***	**2.963±0.415**	**0.975±0.007**	**1.551±0.079***	**1.947±0.084***	**0.976±0.002***	**2.318±0.109***	**2.906±0.155***	**0.979±0.002***

Table 3: External SCZ/HC classification results (%, mean ± SD across 10 seeds). **Bold** and underlined values denote the best and second-best results, respectively.

Method	ACC	AUC	F1
Brain-Semantoks	66.26±1.89	73.82±1.92	66.59±1.85
Brain-JEPA	67.92±4.80	74.86±3.24	67.89±4.79
BrainMass	69.84±3.64	74.42±4.43	70.25±3.57
STH-Brain	**71.15±2.10**	**77.99±1.40**	**71.25±2.10**

Table 4: Zero-shot MDD/HC-to-BP/HC transfer on REST-MDDII (%, mean ± SD across 10 seeds). **Bold** and underlined values denote the best and second-best results, respectively.

Method	ACC	AUC	F1
Brain-Semantoks	61.67±3.02	67.91±3.05	60.71±3.01
Brain-JEPA	62.83±1.39	69.52±2.19	62.19±1.41
BrainMass	64.93±2.69	69.72±3.59	65.36±2.56
STH-Brain	**67.09±2.24**	**73.85±2.76**	**66.29±2.19**

exceeded BrainMass, the second-best model for ACC and F1, by 1.31 and 1.00 percentage points, respectively, and surpassed Brain-JEPA, the second-best model for AUC, by 3.13 percentage points. These results support the generalizability of STH-Brain to the evaluated external SCZ/HC cohort.

Zero-Shot Transfer Performance. Models fine-tuned on MDD/HC training subjects and selected using the corresponding source-validation data were applied directly to the held-out REST-MDDII BP/HC cohort, retaining the source decision rule without additional training or threshold adjustment (Table 4). STH-Brain achieved the highest mean ACC, AUC, and F1. It exceeded BrainMass, the second-best model for ACC, AUC, and F1, by 2.16, 4.13, and 0.93 percentage points, respectively. These results support the transferability of MDD/HC-trained models to BP/HC discrimination without using BP/HC target labels for model development.

4.6 Ablation Experiment

Ablation settings. We compared the full model with 3 variants: (1) **w/o Physical-Time Rotary Encoding**, removing Physical-Time Rotary Encoding; (2) **w/o Spatial-Aware Encoding**, removing spatial geometry encoding; and (3) **w/o Visibility Routing**, replacing the exclusive visibility-based memory selection with joint attention to the target ROI's visible temporal states and all visible network states from the same scan, while retaining the ROI predictor, prediction targets, and pretraining loss terms. Classification and age prediction were evaluated using the same fixed subject-level partitions and 10 random seeds.

Ablation results. Table 5 reports classification ablations for AD, PD, ASD, and MDD; classification results for ADHD, MCI, and PROPD are provided in Appendix H.3.1 (Table 10). Age-prediction ablation results for all seven disease groups are reported in Appendix H.3.2 (Table 11). Across all seven disease groups, the full model achieved the best mean performance in all 21 group–metric combinations for classification and all 21 combinations for age prediction. The *w/o Physical-Time Rotary Encoding* variant showed the largest decrease in classification ACC (4.99–5.27 percentage points), followed by the *w/o Spatial-Aware Encoding* (3.75–4.03 percentage points) and *w/o Visibility Routing* (2.62–2.83 percentage points) variants. For age prediction, these ablations increased MAE by 0.936–2.364, 0.660–1.638, and 0.372–0.987 years, respectively. These comparisons support the contribution of all three components, with physical-time rotary encoding showing the largest observed effect on classification ACC and age-prediction MAE.

Table 5: Classification ablation results for AD, PD, ASD, and MDD (%, mean ± SD across 10 seeds). **Bold** and underlined values denote the best and second-best results, respectively.

Model	AD			PD			ASD			MDD		
	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑
w/o Physical-Time Rotary Encoding	76.21±2.86	82.56±3.34	75.94±2.89	78.15±2.67	84.38±3.11	78.48±2.74	71.44±2.94	78.11±2.99	70.54±2.98	73.21±3.02	80.35±3.54	76.96±2.72
w/o Spatial-Aware Encoding	77.38±2.59	83.72±2.87	77.28±2.58	79.42±2.75	85.71±3.34	79.93±2.76	72.68±3.39	79.56±3.62	72.14±3.42	74.59±3.12	81.74±3.52	78.15±2.81
w/o Visibility Routing	78.57±3.27	85.09±2.57	78.75±3.24	80.46±2.82	87.01±2.75	80.92±2.77	73.81±3.57	80.41±3.28	72.88±3.64	75.75±2.69	83.02±2.63	78.69±2.47
STH-Brain (Full)	**81.24±2.34**	**87.75±2.15**	**81.12±2.36**	**83.23±2.42**	**89.69±2.24**	**83.42±2.28**	**76.43±2.29**	**83.18±2.13**	**75.57±2.34**	**78.38±2.35**	**85.49±2.23**	**81.03±2.15**

4.7 Hyperparameter Analysis

Influence of Atlas Choice. To assess how brain parcellation affects classification, we compared AAL116, Dosenbach160, CC200, and Power264, covering different regional definitions and ROI counts. CC200 achieved the highest mean ACC across all seven tasks (Figure 3), exceeding Power264, Dosenbach160, and AAL116 by 1.23–1.86, 3.28–4.57, and 5.12–6.71 percentage points, respectively. Additional plots of AUC and F1 are provided in Appendix H.4.

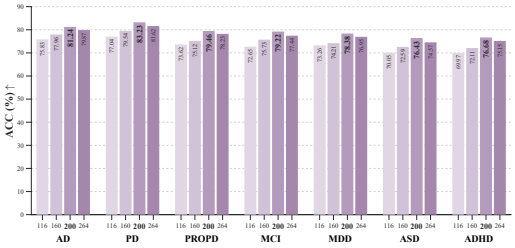


Figure 3: ACC (%) across 7 tasks.

Loss-Weight Coefficients. We evaluated classification sensitivity to the four loss-weight coefficients, λ_1, λ_2, λ_3, and λ_4, using 14 candidate values: 0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.75, 1.00, and 10.00. The respective mean-ACC profiles reached their maxima at $\lambda_1 = 0.75$, $\lambda_2 = 0.05$, $\lambda_3 = 0.50$, and $\lambda_4 = 0.20$. These results illustrate how classification performance varies with each loss-weight coefficient.

4.8 Interpretability Analysis

CC200-based interpretability analysis identified the left dorsolateral prefrontal cortex and midline precuneus across all seven disease groups, alongside distinct regional profiles (Figure 4). AD highlighted bilateral hippocampal, posterior cingulate, and angular regions, whereas PD highlighted striatal, thalamic, and motor regions. ASD involved temporal association, fusiform, cuneus, and frontal opercular regions; MDD involved anterior cingulate, prefrontal, hippocampal, and nucleus accumbens regions. Results for PROPD, ADHD, and MCI are provided in Appendix H.6. These patterns suggest shared involvement of cognitive-control and default-mode regions alongside disease-associated anatomical differences.

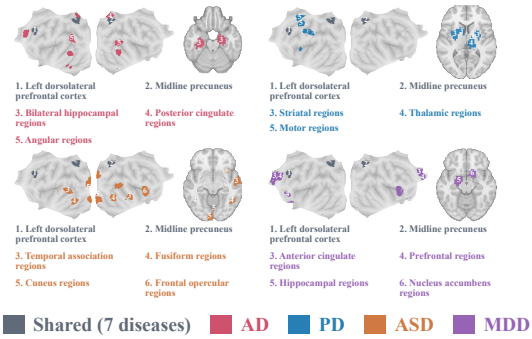


Figure 4: CC200-based interpretability maps for AD, PD, ASD, and MDD.

5 Conclusion

We presented STH-Brain, a brain time-series foundation model that encodes physical time and anatomical space while retaining ROI-specific temporal states alongside network representations. Dual-level latent prediction supervises both representations, with visibility-based routing selecting contextual memory for regional prediction. STH-Brain achieved the best mean performance among the evaluated baselines across seven disorder-classification tasks and age prediction, with further improvements on the evaluated external SCZ/HC cohort and zero-shot MDD/HC-to-BP/HC transfer task. The hierarchy relies on fixed ROI-to-network assignments. Regional analyses identified shared and distinct patterns across disorders, providing anatomical context for interpreting model predictions.

AI USE STATEMENT

Generative AI tools assisted with background research, language polishing, software implementation, and manuscript editing. The authors retained responsibility for the scientific framing, experimental design and controls, and all final scientific and editorial decisions. Portions of the code for data analysis, experimental evaluation, and visualization were developed with AI assistance. The authors reviewed all content and the code used in the reported experiments, and take full responsibility for the manuscript and accompanying research artifacts.

ETHICS STATEMENT

This study used existing, publicly available neuroimaging datasets distributed under their respective access and data-use policies. All analyses were conducted on de-identified data, with no new participant recruitment or data collection. We complied with the applicable data-use agreements and did not attempt to identify individual participants. Differences in demographic composition and acquisition settings across datasets may affect model performance and limit generalizability to underrepresented populations. STH-Brain is intended for research purposes; its predictions should not be interpreted as clinical diagnoses or used for individual clinical decisions without further validation.

REPRODUCIBILITY STATEMENT

The model architecture and learning objectives are described in the Methods section. The Experiments section specifies the datasets, evaluation protocol, and performance metrics. The source code is available via the link provided in the abstract. Additional details on preprocessing, data partitioning, hyperparameter settings, implementation, and metric computation are provided in the supplementary materials to facilitate reproduction of the experiments.

REFERENCES

Anushree Bannadabhavi, Soojin Lee, Wenlong Deng, Rex Ying, and Xiaoxiao Li. Community-aware transformer for autism prediction in fmri connectome. In *International conference on medical image computing and computer-assisted intervention*, pp. 287–297. Springer, 2023.

Pierre Bellec, Carlton Chu, Francois Chouinard-Decorte, Yassine Benhajali, Daniel S Margulies, and R Cameron Craddock. The neuro bureau adhd-200 preprocessed repository. *Neuroimage*, 144:275–286, 2017.

Vince D Calhoun, Jing Sui, Kent Kiehl, Jessica Turner, Elena Allen, and Godfrey Pearlson. Exploring the psychosis functional connectome: aberrant intrinsic networks in schizophrenia and bipolar disorder. *Frontiers in psychiatry*, 2:75, 2012.

Dongdong Chen, Linlin Yao, Mengjun Liu, Zhenrong Shen, Yuqi Hu, Zhiyun Song, Shengyu Lu, Qian Wang, Dinggang Shen, and Lichi Zhang. Brain connectivity network structure learning for brain disorder diagnosis. *arXiv preprint arXiv:2509.16735*, 2025a.

Wanyi Chen, Jianjun Yang, Zhongquan Sun, Xiang Zhang, Guangyu Tao, Yuan Ding, Jingjun Gu, Jiajun Bu, and Haishuai Wang. Deepasd: a deep adversarial-regularized graph learning method for asd diagnosis with multimodal data. *Translational Psychiatry*, 14(1):375, 2024.

Xiao Chen, Bin Lu, Yu-Wei Wang, Xue-Ying Li, Zi-Han Wang, Hui-Xian Li, Yi-Fan Liao, Daniel M Blumberger, Francisco Xavier Castellanos, Eduardo A Garza-Villarreal, et al. Subgenual anterior cingulate cortex functional connectivity abnormalities in depression: insights from brain imaging big data and precision-guided personalized intervention via transcranial magnetic stimulation. *Science Bulletin*, 70(16):2676–2690, 2025b.

R Cameron Craddock, G Andrew James, Paul E Holtzheimer III, Xiaoping P Hu, and Helen S Mayberg. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human brain mapping*, 33(8):1914–1928, 2012.

Adriana Di Martino, David O'connor, Bosi Chen, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Joshua H Balsters, Leslie Baxter, Anita Beggiato, Sylvie Bernaerts, et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. *Scientific data*, 4(1): 170010, 2017.

Wenhao Dong, Yueyang Li, Weiming Zeng, Lei Chen, Hongjie Yan, Wai Ting Siok, and Nizhuan Wang. Starformer: A novel spatio-temporal aggregation reorganization transformer of fmri for brain disorder diagnosis. *Neural Networks*, pp. 107927, 2025.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan T Nguyen, Joanna S Chong, Fang Ji, Nathanael R Tong, Christopher L Chen, and Juan H Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024.

Zijian Dong, Ruilin Li, Joanna Chong, Niousha Dehestani, Yinghui Teng, Yi Lin, Zhizhou Li, Yichi Zhang, Yapei Xie, Leon Ooi, et al. Brain harmony: a multimodal foundation model unifying morphology and function into 1d tokens. *Advances in Neural Information Processing Systems*, 38:122100–122127, 2026.

Wenrui Fan, LM Riza Rizky, Jiayang Zhang, Chen Chen, Haiping Lu, Kevin Teh, Dinesh Selvarajah, and Shuo Zhou. Foundation-model-boosted multimodal learning for fmri-based neuropathic pain drug response prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 238–248. Springer, 2025.

Jie Gao, Bao Ge, Ning Qiang, and Shijie Zhao. 3d masked autoencoder with spatiotemporal transformer for modeling of 4d fmri data. *Medical Image Analysis*, pp. 103861, 2025.

Sam Gijsen, Marc-Andre Schulz, and Kerstin Ritter. Brain-semantoks: Learning semantic tokens of brain dynamics with a self-distilled foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 13186–13208, 2026.

Jinlong Hu, Yangmin Huang, Nan Wang, and Shoubin Dong. Brainnpt: Pre-training transformer networks for brain network classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:2727–2736, 2024.

Shih-Gu Huang, Jing Xia, Liyuan Xu, and Anqi Qiu. Spatio-temporal directed acyclic graph learning with attention mechanisms on brain functional time series and connectivity. *Medical Image Analysis*, 77:102370, 2022.

Yiming Huang, Yujie Zeng, Qiang Wu, and Linyuan Lü. Higher-order graph convolutional network with flower-petals laplacians on simplicial complexes, 2024. URL https://arxiv.org/abs/2309.12971.

Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

Xi Jiang, Jiadong Yan, Yu Zhao, Mingxin Jiang, Yuzhong Chen, Jingchao Zhou, Zhenxiang Xiao, Zifan Wang, Rong Zhang, Benjamin Becker, et al. Characterizing functional brain networks via spatio-temporal attention 4d convolutional neural networks (sta-4dcnns). *Neural Networks*, 158: 99–110, 2023.

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, 2022.

Byung-Hoon Kim, Jong Chul Ye, and Jae-Jin Kim. Learning dynamic graph representation of brain connectome with spatio-temporal attention. *Advances in Neural Information Processing Systems*, 34:4314–4327, 2021.

Peter Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, and Taesup Moon. Swift: Swin 4d fmri transformer. *Advances in Neural Information Processing Systems*, 36:42015–42037, 2023.

Youyong Kong, Xiaotong Zhang, Wenhan Wang, Yue Zhou, Yueying Li, and Yonggui Yuan. Multiscale spatial-temporal attention networks for functional connectome classification. *IEEE Transactions on Medical Imaging*, 44(1):475–488, 2024.

Chaojun Li, Peiliang Gong, Shengrong Li, Chunwei Tian, Yinbo Yu, Ran Wang, Daoqiang Zhang, and Qi Zhu. Spatio-temporal hypergraph attention networks for brain disease analysis. *IEEE Transactions on Image Processing*, 2026a.

Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical image analysis*, 74:102233, 2021.

Zihan Li, Guangping Zeng, Zhenqiao Liu, Han Wen, Yizhi Cui, and Qingchuan Zhang. Dual-branch spatiotemporal transformer pretraining model for asd recognition using functional connectivity. *Biomedical Signal Processing and Control*, 121:110301, 2026b.

Ziyu Li, Zhiyuan Zhu, Qing Li, and Xia Wu. Graph pre-trained framework with spatio-temporal importance masking and fine-grained optimizing for neural decoding. *Pattern Recognition*, 170: 112006, 2026c.

Chenyu Liu and Luca Rossi. Phgnn: a novel prompted hypergraph neural network to diagnose alzheimer's disease. *arXiv preprint arXiv:2503.14577*, 2025.

Jingyu Liu, Weigang Cui, Yipeng Chen, Yulan Ma, Qunxi Dong, Ran Cai, Yang Li, and Bin Hu. Deep fusion of multi-template using spatio-temporal weighted multi-hypergraph convolutional networks for brain disease analysis. *IEEE Transactions on Medical Imaging*, 43(2):860–873, 2023.

Hao Ma, Yongkang Xu, and Lixia Tian. Rs-mae: region-state masked autoencoder for neuropsychiatric disorder classifications based on resting-state fmri. *IEEE transactions on neural networks and learning systems*, 36(6):10707–10720, 2024.

Jiajun Ma, Yongchao Zhang, Chao Zhang, Zhao Lv, and Shengbing Pei. Brainhgt: A hierarchical graph transformer for interpretable brain network analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17617–17625, 2026.

Yunling Ma, Qianqian Wang, Liang Cao, Long Li, Chaojun Zhang, Lishan Qiao, and Mingxia Liu. Multi-scale dynamic graph learning for brain disorder detection with functional mri. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3501–3512, 2023.

Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011.

Josue Ortega Caro, Antonio Henrique de Oliveira Fonseca, Syed Rizvi, Matteo Rosati, Christopher Averill, James Cross, Prateek Mittal, Emanuele Zappala, Rahul Dhodapkar, Chadi Abdallah, et al. Brainlm: A foundation model for brain activity recordings. In *International Conference on Learning Representations*, volume 2024, pp. 565–576, 2024.

Saori C Tanaka, Ayumu Yamashita, Noriaki Yahata, Takashi Itahashi, Giuseppe Lisi, Takashi Yamada, Naho Ichikawa, Masahiro Takamura, Yujiro Yoshihara, Akira Kunimatsu, et al. A multi-site, multi-disorder resting-state magnetic resonance image database. *Scientific data*, 8(1):227, 2021.

Hwa Hui Tew, Junn Yong Loo, Leong Yu, Julia Lau, Ding Fan, Hernando Ombao, Raphaël Phan, Chee Tan, and Chee-Ming Ting. Functional mri time series generation via wavelet-based image transform and spectral flow matching for brain disorder identification. In *International Conference on Learning Representations*, volume 2026, pp. 106373–106393, 2026.

Bishal Thapaliya, Robyn Miller, Jiayu Chen, Yu Ping Wang, Esra Akbas, Ram Sapkota, Bhaskar Ray, Pranav Suresh, Santosh Ghimire, Vince D Calhoun, et al. Dsam: A deep learning framework for analyzing temporal and spatial dynamics in brain networks. *Medical image analysis*, 101: 103462, 2025.

Suryansh Vidya, Kush Gupta, Amir Aly, Andy Wills, Emmanuel Ifeachor, and Rohit Shankar. Identification of critical brain regions for autism diagnosis from fmri data using explainable ai: an observational analysis of the abide dataset. *EClinicalMedicine*, 88, 2025.

Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Chang-bae Bang, Lin Zhao, Wanyi Fu, Jinglei Lv, Jorge Sepulcre, Carl Yang, et al. Towards a general-purpose foundation model for functional mri analysis. *Nature Biomedical Engineering*, pp. 1–12, 2026a.

Mo Wang, Wenhao Ye, Junfeng Xia, Minghao Xu, Hongkai Wen, and Quanying Liu. Flexibrain: Resolution-agnostic voxel-level encoding for native fmri. In *European Conference on Computer Vision*, pp. 432–449. Springer, 2026b.

Xinxu Wei, Kanhao Zhao, Yong Jiao, Lifang He, and Yu Zhang. A brain graph foundation model: Pre-training and prompt-tuning across broad atlases and disorders. In *International Conference on Learning Representations*, volume 2026, pp. 156800–156829, 2026.

Junfeng Xia, Wenhao Ye, Xuanye Pan, Xinke Shen, Mo Wang, and Quanying Liu. Brain-dit: A universal multi-state fmri foundation model with metadata-conditioned pretraining. *arXiv preprint arXiv:2604.12683*, 2026a.

Junfeng Xia, Wenhao Ye, Junxiang Zhang, Xuanye Pan, Mo Wang, and Quanying Liu. Brainworld: A structural-prior-conditioned generative model for whole-brain 4d fmri dynamics. *arXiv preprint arXiv:2606.17742*, 2026b.

Rundong Xue, Hao Hu, Zeyu Zhang, Xiangmin Han, Juan Wang, Yue Gao, and Shaoyi Du. Dhgformer: 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.

Chao-Gan Yan, Xiao Chen, Le Li, Francisco Xavier Castellanos, Tong-Jian Bai, Qi-Jing Bo, Jun Cao, Guan-Mao Chen, Ning-Xuan Chen, Wei Chen, et al. Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. *Proceedings of the National Academy of Sciences*, 116(18):9078–9083, 2019.

Yanwu Yang, Chenfei Ye, Guinan Su, Ziyao Zhang, Zhikai Chang, Hairui Chen, Piu Chan, Yue Yu, and Ting Ma. Brainmass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. *IEEE transactions on medical imaging*, 43(11):4004–4016, 2024.

Shuo Yu, Shan Jin, Ming Li, Tabinda Sarwar, and Feng Xia. Long-range brain graph transformer. *Advances in Neural Information Processing Systems*, 37:24472–24495, 2024.

Shuoyan Zhang, Qingmin Wang, Min Wei, Jiayi Zhong, Ying Zhang, Ziyan Song, Chenyang Li, Xiaochen Zhang, Ying Han, Yunxia Li, et al. Frog: A fine-grained spatiotemporal graph neural network with self-supervised guidance for early diagnosis of alzheimer's disease. *IEEE Journal of Biomedical and Health Informatics*, 29(8):5900–5911, 2025.

Yunxi Zhao, Dong Nie, Geng Chen, Xia Wu, Daoqiang Zhang, and Xuyun Wen. Tardrl: Task-aware reconstruction for dynamic representation learning of fmri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 700–710. Springer, 2024.

Kaizhong Zheng, Shujian Yu, Baojuan Li, Robert Jenssen, and Badong Chen. Brainib: Interpretable brain network-based psychiatric diagnosis with graph information bottleneck. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):13066–13079, 2024.

A Physical Queries and Local Temporal Encoding

A.1 Notation, Identity, and Validity

Scan i contains observations $x_{ijr} = [X_i]_{jr}$ at native times t_{ij}, where $j \in \{1, \dots, T_i\}$ indexes native samples and r indexes ROIs. The CC200 atlas supplies $R = 200$ coordinates c_r and membership $n(r) \in \{1, \dots, N\}$ with $N = 8$ fixed networks. The native support mask is shared across ROIs within each scan–block, while ROI quality validity is fixed per scan. Block b has center τ_{ib} in a non-overlapping physical partition. The joint indicator v_{ibr} combines block support and ROI validity, $\mathcal{R}_{ibn} = \{r : n(r) = n, v_{ibr} = 1\}$ contains usable network members, and $w_{ibn} = \mathbf{1}\{|\mathcal{R}_{ibn}| > 0\}$ determines network validity. These indicators describe valid support; membership in $\mathcal{C}_i$ additionally determines visibility in the masked-context pass. The valid network–time domain is $\mathcal{V}_i = \{(b, n) : w_{ibn} = 1\}$. Position metadata P_i contains physical time, block duration, sampling interval, coordinates, identity, and support. Atlas coordinates provide the shared anatomical reference across subjects.

Packed implementations carry explicit ROI/network identifiers, block identifiers, timestamps, and validity alongside content. Explicit identity indices preserve anatomical ordering throughout packing and readout. The dense shapes below describe the representation layout; computation operates on packed valid rows.

Representation Shapes: X_i has shape $T_i \times R$. Spatially contextualized features H_i and ROI temporal outputs U_i retain shape $B_i \times R \times d$; network outputs V_i have shape $B_i \times N \times D$. The ROI temporal Transformer internally uses width $d_T = 256$ and projects back to $d = 64$. ROI targets have width d and network targets width D, and the downstream readout has width $Rd + ND + R + N$.

A.2 Local Standardization and Supported Queries

For a selected network–time view $\mathcal{A}_i \subseteq \mathcal{V}_i$, let $J_{ir}^{\mathcal{A}}$ contain the indices of finite native observations of ROI r in selected cells $(b, n(r))$, after applying native and ROI validity masks. Write $M_{ir}^{\mathcal{A}} = |J_{ir}^{\mathcal{A}}|$. For positive support, the visible-scan location and population scale are

$$\mu_{ir}^{\mathcal{A}} = \frac{1}{M_{ir}^{\mathcal{A}}} \sum_{j \in J_{ir}^{\mathcal{A}}} x_{ijr}, \qquad \sigma_{ir}^{\mathcal{A}} = \sqrt{\frac{1}{M_{ir}^{\mathcal{A}}} \sum_{j \in J_{ir}^{\mathcal{A}}} (x_{ijr} - \mu_{ir}^{\mathcal{A}})^2}. \tag{A.1}$$

These statistics pool visible native observations across blocks of one scan and ROI. Let $\mu_r^{\mathrm{tr}} \in \mathbb{R}$ and $\sigma_r^{\mathrm{tr}} > 0$ be frozen ROI location and scale statistics fitted on training subjects only. Use the visible estimates when $M_{ir}^{\mathcal{A}} \geq 8$, both estimates are finite, and $\sigma_{ir}^{\mathcal{A}} > 10^{-6}$; otherwise use the frozen pair. Denote the chosen pair for view $\mathcal{A}_i$ by $\mu_{ir}^{*,\mathcal{A}}, \sigma_{ir}^{*,\mathcal{A}}$ and standardize $\widetilde{x}_{ijr}^{\mathcal{A}} = (x_{ijr} - \mu_{ir}^{*,\mathcal{A}})/\sigma_{ir}^{*,\mathcal{A}}$. Token validity is determined by observed support. Context uses $\mathcal{A}_i = \mathcal{C}_i$; complete student and teacher passes use $\mathcal{A}_i = \mathcal{V}_i$ independently. Context estimates use exclusively the native observations selected by the visible cells. This operation removes scan-view location and scale when using visible estimates, while preserving relative differences between its blocks.

For the native times $\mathcal{T}$ of one block, let $t^-(u), t^+(u)$ bracket an interior query u, and let $g_{\max} = 6.5$ s. The admissible time domain is

$$\mathcal{D}(\mathcal{T}) = \mathcal{T} \cup \{u : t^-(u) < u < t^+(u),\ t^+(u) - t^-(u) \leq g_{\max}\}. \tag{A.2}$$

Let $\mathcal{Q}$ be the nominal 1-s query grid inside the block and define the supported queries as $\mathcal{U} = \mathcal{Q} \cap \mathcal{D}(\mathcal{T})$. Queries lie within native support; exact native observations remain valid beside a larger gap. Suppressing scan and view indices, write $\widetilde{x}_r(t_{ij}) = \widetilde{x}_{ijr}^{\mathcal{A}}$ at valid native times. For an admissible interior query, let a_u be its interpolation fraction and $\widetilde{x}_r(u)$ its interpolated value. With stabilizer $\epsilon_t = 10^{-6}$ s, define

$$a_u = \frac{u - t^-(u)}{\max(t^+(u) - t^-(u), \epsilon_t)}, \qquad \widetilde{x}_r(u) = (1 - a_u)\widetilde{x}_r(t^-(u)) + a_u \widetilde{x}_r(t^+(u)). \tag{A.3}$$

Exact native queries copy the observed value. Linear interpolation reproduces affine signals when the time floor is inactive. Reference settings use block duration $\Delta_{\mathrm{blk}} = 8$ s, query spacing $\Delta_q = 1$ s,

and a minimum block-coverage threshold of 0.75. Unsupported blocks are removed before masking and an empty scan is rejected. The minimum of eight normalization samples is evaluated across the visible scan view for each ROI. The loader uses the preprocessed sequences with their registered temporal support.

A.3 Local Multiscale Computation

For each query u, let W_x, b_x project its scalar value, e_{block} be a learned block-type vector, and $p_{\text{loc}}(u)$ a fixed additive sinusoidal basis evaluated at the query offset from the block start in seconds, all producing width d. The query embedding e_u is

$$e_u = W_x \widetilde{x}_r(u) + b_x + e_{\text{block}} + p_{\text{loc}}(u). \tag{A.4}$$

The basis interleaves sine and cosine features at logarithmically spaced frequencies to encode within-block temporal offsets. Cross-block ROI attention and global network–time attention use PTRE to rotate temporal query/key pairs according to block-center timestamps in seconds, as defined in Equation 1. Let $e = (e_u)_{u \in \mathcal{Q}}$ be the sequence on the nominal query grid, with unsupported entries masked to zero. For kernel width $k \in \{1, 3, 5\}$, let DWConv_k denote a depthwise convolution, b_k its shared learned mixing logit, and GELU the Gaussian error linear unit. The mixture weights α_k and combined features e^{conv} are:

$$e^{\text{conv}} = \sum_{k \in \{1,3,5\}} \alpha_k \,\text{GELU}(\text{DWConv}_k(e) + e) \tag{A.5}$$

$$\alpha_k = \exp(b_k) / \sum_{k' \in \{1,3,5\}} \exp(b_{k'}). \tag{A.6}$$

The learned logits b_k are shared across samples. Bias-free convolutions use centered padding; invalid positions are zeroed before and after convolution. Let T_{loc} be the local temporal Transformer and $w_{\text{pool}} \in \mathbb{R}^d$ a learned pooling vector. Its masked outputs y_u, pooling weights α_u^{pool}, and block summary are

$$y = T_{\text{loc}}(e^{\text{conv}}) \tag{A.7}$$

$$\alpha_u^{\text{pool}} = \text{softmax}_{u \in \mathcal{U}}(w_{\text{pool}}^\top y_u / \sqrt{d}) \tag{A.8}$$

$$h^{\text{loc}} = \sum_{u \in \mathcal{U}} \alpha_u^{\text{pool}} y_u. \tag{A.9}$$

The default width is 64 with four heads and feed-forward network (FFN) width 256. Invalid outputs are zeroed before pooling. The kernels cover 1, 3, and 5 query samples at a spacing of 1 s. Pooling summarizes the supported temporal queries separately for each ROI, preserving the ROI axis.

B Physical Rotary Coordinates and Their Properties

B.1 ROI Spherical Coordinates

Write $c_r = (c_{rx}, c_{ry}, c_{rz})$, let $\text{clip}(x, -1, 1)$ clamp x to $[-1, 1]$, and let atan2 denote the quadrant-aware azimuth function. Using the atlas origin, define $\rho_r = \max(\|c_r\|_2, 10^{-6})$, $\widehat{\rho}_r = \rho_r / \max_{r'} \rho_{r'}$, $\vartheta_r = \arccos(\text{clip}(c_{rz}/\rho_r, -1, 1))$, and $\psi_r = \text{atan2}(c_{ry}, c_{rx})$. The eight pairs in each width-16 spatial head receive

$$\varphi_r = (\pi\widehat{\rho}_r, \vartheta_r, \psi_r, 2\pi\widehat{\rho}_r, 2\vartheta_r, 2\psi_r, 4\pi\widehat{\rho}_r, 4\vartheta_r). \tag{B.1}$$

A single within-network spatial-attention module is shared across scans, blocks, and networks. All four width-16 heads use these fixed harmonics. Pre-normalized query/key projections are rotated; values remain content vectors. Projected multi-head outputs update h^{loc} by residual addition, with no additional spatial FFN. The resulting features form H_i. Although the parameters are shared, the attention weights depend on the signal content of each scan and block. The atlas origin, axes, and angular conventions define a consistent spatial reference across scans. The spatial module operates on ROI summaries of 8-s blocks. Its generally asymmetric attention weights describe learned feature interactions. Within the fixed atlas coordinate system, spatial rotary encoding modulates attention through phase differences derived from spherical coordinates, rather than explicit pairwise Euclidean distances. The resulting weights describe learned feature interactions among ROI block summaries within each network.

B.2 Network Position Basis

For axis $a \in \{x, y, z\}$, coordinates are normalized as $\widehat{c}_{ra} = c_{ra} / \max(\max_{r'} |c_{r'a}|, 10^{-6})$. Fixed atlas membership defines nine network spatial statistics: mean normalized radius; sine/cosine of circular mean polar angle; sine/cosine of circular mean azimuth; mean normalized axial x; polar concentration; mean normalized axial z; and azimuth concentration. Circular means are calculated from mean sine/cosine components and mapped back to unit-circle pairs. Near-zero concentration uses a deterministic zero-angle fallback. Separate concentration features encode the directional dispersion of member ROIs.

The network position basis has 51 components: 16 temporal, eight duration, eight sampling-interval, eight network-identity, nine spatial, and two mask-type components. Default target queries disable mask-type conditioning. The geometry-basis contribution to width-256 tokens uses fixed scale 0.01. Each width-32 global head has eight temporal pairs. Half the heads additionally use five spatial pairs; other channels are unrotated. Fixed base periods span multiple physical timescales, with bounded learnable multipliers calibrating the corresponding frequencies. The log-frequency bound in Equation 1 is $\kappa = \log 2$, giving learned multipliers in $[1/2, 2]$. ROI cross-block attention applies PTRE using physical block-center timestamps without mixing ROIs.

B.3 Relative-Phase Identity and Scope

For real angles χ, χ', define the planar rotation $\mathsf{R}(\chi)$ and its relative-phase identity:

$$\mathsf{R}(\chi) = \begin{pmatrix} \cos\chi & -\sin\chi \\ \sin\chi & \cos\chi \end{pmatrix}, \qquad \mathsf{R}(\chi)^\top \mathsf{R}(\chi') = \mathsf{R}(\chi' - \chi). \tag{B.2}$$

For vectors $q, k \in \mathbb{R}^2$, rotations preserve norms and $[\mathsf{R}(\chi)q]^\top [\mathsf{R}(\chi')k] = q^\top \mathsf{R}(\chi' - \chi)k$. For network tokens indexed by u, v, let $q_u, k_v \in \mathbb{R}^{d_{\text{att}}}$ be their query and key vectors, with head width d_{att} as in Equation 3. Let t_u, s_u be token u's time and spatial descriptors, R_T, R_S the corresponding block rotations, and I the identity on unrotated channels. Write $R_u = \operatorname{diag}(R_T(t_u), R_S(s_u), I)$. Superscripts $T, S, 0$ identify the temporal, spatial, and unrotated channel subspaces, respectively; ℓ_{uv} denotes the pre-softmax query–key logit. Then

$$\sqrt{d_{\text{att}}}\,\ell_{uv} = (q_u^T)^\top R_T(t_u)^\top R_T(t_v) k_v^T + (q_u^S)^\top R_S(s_u)^\top R_S(s_v) k_v^S + (q_u^0)^\top k_v^0. \tag{B.3}$$

Separate channel subspaces represent temporal and spatial phase differences. The unrotated subspace supports direct content matching, while additive position features retain absolute coordinate information. Their combined attention score integrates relative phases in physical time and the chosen anatomical coordinates with signal content. The spherical parameterization is tied to the fixed atlas origin and axes.

C ROI Temporal Encoding and Network Aggregation

C.1 Regional Trajectories and Network Summaries

The spatial tokenizer supplies H. Each scan–ROI sequence is packed in chronological order and processed independently by a shared temporal Transformer. An input projection maps $64 \to 256$, eight layers use 16 heads and FFN width 1024, and a normalized output projection maps $256 \to 64$ to produce U. Every width-16 head uses eight channel pairs for PTRE. Invalid keys and outputs are masked. Each temporal sequence contains the visible blocks of one ROI from one scan; its inputs carry the spatial context established by within-network attention.

Network summaries are constructed from U. Let $K_{ibn} = |\mathcal{R}_{ibn}|$ count valid member ROIs and $w_{\text{ROI}} \in \mathbb{R}^d$ be a learned pooling vector. A visible cell has $K_{ibn} > 0$. Let $\odot 2$ denote channel-wise squaring; square root and maximum also act channel-wise. For each visible cell, define channel-wise

mean μ^U_{ibn}, scale σ^U_{ibn}, attention weights π_{ibr}, and pooled vector η^U_{ibn}:

$$\mu^U_{ibn} = K^{-1}_{ibn} \sum_{r \in \mathcal{R}_{ibn}} U_{ibr}, \tag{C.1}$$

$$\sigma^U_{ibn} = \sqrt{\max\left(K^{-1}_{ibn} \sum_{r \in \mathcal{R}_{ibn}} (U_{ibr} - \mu^U_{ibn})^{\odot 2}, 10^{-12} \right)}, \tag{C.2}$$

$$\pi_{ibr} = \mathrm{softmax}_{r \in \mathcal{R}_{ibn}}(w^\top_{\mathrm{ROI}} U_{ibr}/\sqrt{d}), \qquad \eta^U_{ibn} = \sum_{r \in \mathcal{R}_{ibn}} \pi_{ibr} U_{ibr}. \tag{C.3}$$

The statistics are computed in FP32. A learned affine projection $W_{\mathrm{loc}} \in \mathbb{R}^{D \times 3d}$ with bias $b_{\mathrm{loc}} \in \mathbb{R}^D$ produces $a_{ibn} = W_{\mathrm{loc}}[\mu^U_{ibn}; \sigma^U_{ibn}; \eta^U_{ibn}] + b_{\mathrm{loc}}$, where brackets denote concatenation. The pooling weights π_{ibr} assign one contribution weight to each member ROI when constructing a network token; they are distinct from the pairwise spatial-attention weights $A_{rr'}$.

For a scaled network embedding $e_n \in \mathbb{R}^D$, bias-free learned maps $W_s, W_e, W_o \in \mathbb{R}^{D \times D}$, the sigmoid linear unit SiLU, element-wise product $\odot$, and learned scalar g_{id}, define the interaction map $\Phi : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^D$ on inputs $a, e \in \mathbb{R}^D$ and coupling strength γ as follows:

$$\Phi(a, e) = W_o[\mathrm{SiLU}(W_s\,\mathrm{LN}(a)) \odot \mathrm{SiLU}(W_e\,\mathrm{LN}(e))], \tag{C.4}$$

$$\gamma = 0.25 + 0.75\,\mathrm{sigmoid}(g_{\mathrm{id}}), \qquad z_{ibn} = a_{ibn} + e_n + \gamma\Phi(a_{ibn}, e_n). \tag{C.5}$$

Initialization $g_{\mathrm{id}} = \log(1/2)$ gives $\gamma = 0.5$; network embeddings have initialization standard deviation 0.02 and scale 0.1. Layer normalization in this coupling is non-affine. The network Transformer adds its scaled geometry basis and attends noncausally over all visible network–time cells in the same scan, producing V. This forward hierarchy is $H \to U \to z \to V$. The two retained output levels supply regional and network prediction targets and the joint downstream readout. The network output V_i is retained alongside U_i and is not fused back into the ROI encoder states during the forward pass. Network fallback supplies context to the ROI predictor during pretraining; downstream readout uses the two encoder branches.

C.2 Visibility-Based Context Retrieval

Cross-level retrieval occurs in the ROI predictor when same-ROI observations are unavailable. For target ROI r of scan i, define the direct regional and network memory sets

$$\mathcal{U}^{\mathcal{C}}_{ir} = \{U_{ibr} : (b, n(r)) \in \mathcal{C}_i,\ v_{ibr} = 1\}, \qquad \mathcal{V}^{\mathcal{C}}_i = \{V_{ibn} : (b, n) \in \mathcal{C}_i\}. \tag{C.6}$$

Equation 6 selects the direct memory $\mathcal{K}_{ir}$ from these context-only sets. Here $\mathcal{V}_i$ is the set of valid cell indices, whereas $\mathcal{V}^{\mathcal{C}}_i$ contains the encoded network features at context cells. The selected regional or network features are mapped to the predictor width by their respective learned projections. When regional memory exists, network keys and values are masked out at every predictor layer. The route is a deterministic function of visibility. Whole-network masking retains the regional targets and activates network memory from the same scan. Consequently, the local route has zero derivative with respect to direct network memory, while the fallback route can train the network path. The regional memory combines ROI-specific temporal evolution with the within-network spatial context encoded earlier.

D Structured Masking and Dual-Level Prediction

D.1 Mask Construction and Information Boundaries

Let $\mathcal{B}$ select temporal stripes and $\mathcal{N}$ select network stripes. With $\dot\cup$ denoting disjoint union, the target and partition rules are

$$\mathcal{M}_i = \{(b, n) \in \mathcal{V}_i : b \in \mathcal{B} \vee n \in \mathcal{N}\}, \qquad \mathcal{V}_i = \mathcal{C}_i \,\dot\cup\, \mathcal{M}_i \,\dot\cup\, \mathcal{E}_i. \tag{D.1}$$

On a complete grid with B blocks, target coverage is $|\mathcal{B}|/B + |\mathcal{N}|/N - |\mathcal{B}||\mathcal{N}|/(BN)$. Defaults specify 15–30% target coverage, an 8–16-s exclusion neighborhood with at least one adjacent block, and at least 25% visible context. Pretraining requires at least two valid networks and 16 supported

network–time cells. Mask feasibility is checked against each scan's valid support, and only partitions satisfying the coverage constraints with nonempty context and target sets are accepted. Fixed network membership maps target cells to ROI positions.

Context-only support applies before normalization and throughout local convolution, spatial attention, ROI temporal attention, network aggregation, and both predictors. Predictor memory is constructed exclusively from the context pass; the complete-view pass supplies subject-diversity statistics. With preprocessing completed and the validity masks, timestamps, and partitions held fixed, changes restricted to target or excluded amplitudes leave context features and predictions unchanged.

D.2 Student-Teacher Framework and Signal-Residual Targets

The teacher is initialized as a copy of the full student encoder, excluding the prediction heads, and remains in evaluation mode. After each optimizer update, its parameters are updated as $\xi \leftarrow m\xi + (1-m)\theta$, where the EMA momentum m follows a cosine schedule; non-parameter buffers are copied from the student. Regional and network targets are constructed from the teacher's ROI temporal outputs and contextualized network outputs, respectively. At each level, Equation 5 subtracts the complete-view raw-zero response from the real-signal response, then applies non-affine layer normalization over the corresponding feature channels. The resulting targets are detached from gradient computation, so prediction losses update the student encoder and prediction heads, while the teacher parameters evolve only through EMA.

The zero pass replaces the observation values in the preprocessed input X_i with zero before view-specific normalization, while retaining native support, ROI masks, positions, and identities. Each pass independently applies the normalization procedure. The constant zero view uses frozen training statistics, giving standardized baseline $-\mu_r^{\mathrm{tr}}/\sigma_r^{\mathrm{tr}}$ at valid observations. Equation 5 therefore compares real activity with a raw-zero baseline after the view-specific normalization and encoding sequence. Shared positional metadata anchors this signal-response contrast at corresponding anatomical and temporal locations.

D.3 Two Prediction Domains

The network predictor constructs target queries from physical and network-identity metadata and attends to projected visible network representations V. Each target independently retrieves network context from the same scan. Temporal and spatial rotary encoding incorporates physical time and anatomical geometry into query–key matching.

For each ROI target, a learned projection of its normalized anatomical coordinates and temporal features is combined with an ROI-identity embedding to construct the target query. Separate projections map regional and network memories into a common feature space and add level identifiers; regional memory also receives positional encoding. The visibility-based routing rule in Appendix C selects the target ROI's visible temporal memory when available and visible network context otherwise. The selected memory source is used consistently across all predictor layers.

Both predictors use width d_p and depth L_p, with separate learned parameters. For the ROI predictor, let $q \in \mathbb{R}^{d_p}$ denote the target query and $\mathcal{K}$ the selected projected memory, including positional and level encodings. At layer k, let CA_k, FFN_k, and LN_k denote cross-attention, the feed-forward map, and affine layer normalization, respectively. Let $a^{(k)}, h^{(k)} \in \mathbb{R}^{d_p}$ denote the intermediate attention state and layer output. Suppressing scan and target indices, the regional predictor updates are

$$\begin{aligned} a^{(1)} &= \mathrm{CA}_1(q, \mathcal{K}), \\ h^{(1)} &= a^{(1)} + \mathrm{FFN}_1\Big(\mathrm{LN}_1(a^{(1)})\Big), \\ a^{(k)} &= h^{(k-1)} + \mathrm{CA}_k(h^{(k-1)}, \mathcal{K}), \\ h^{(k)} &= a^{(k)} + \mathrm{FFN}_k\Big(\mathrm{LN}_k(a^{(k)})\Big), \quad k = 2, \ldots, L_p. \end{aligned} \tag{D.2}$$

In both predictors, the first cross-attention layer initializes the target state from visible memory without a direct query residual; subsequent layers refine it through residual cross-attention and feed-forward updates.

Let $W_{\text{out}} \in \mathbb{R}^{d \times d_p}$ and $b_{\text{out}} \in \mathbb{R}^d$ be the regional output parameters. For the network predictor, let $h_N^{(L_p)} \in \mathbb{R}^{d_p}$ denote its final target state and $g_N : \mathbb{R}^{d_p} \to \mathbb{R}^D$ its learned output map. The regional and network predictions are

$$\widehat{Y}^R = W_{\text{out}} h^{(L_p)} + b_{\text{out}}, \qquad \widehat{Y}^N = g_N(h_N^{(L_p)}). \tag{D.3}$$

D.4 Prediction Losses and Subject-Balanced Reduction

The prediction objectives in Equations 7 and 8 use the scalar SmoothL1 penalty

$$\ell_\beta(e) = \begin{cases} \dfrac{e^2}{2\beta}, & |e| < \beta, \\ |e| - \dfrac{\beta}{2}, & |e| \geq \beta, \end{cases} \tag{D.4}$$

where e and $\beta > 0$ follow the definitions in the main text. The regional and network terms use β_R and β_N, respectively.

For each target network–time cell $(b, n) \in \mathcal{M}_i$, define the regional and network cell losses L_{ibn}^R and L_{ibn}^N as

$$\begin{aligned} L_{ibn}^R &= \frac{1}{|\mathcal{R}_{ibn}| d} \sum_{r \in \mathcal{R}_{ibn}} \sum_{k=1}^{d} \ell_{\beta_R}\left(\widehat{Y}_{ibrk}^R - Y_{ibrk}^R\right), \\ L_{ibn}^N &= \frac{1}{D} \sum_{k=1}^{D} \ell_{\beta_N}\left(\widehat{Y}_{ibnk}^N - Y_{ibnk}^N\right). \end{aligned} \tag{D.5}$$

The predictions, teacher targets, feature dimensions, and channel index follow the main-text notation. Here $\mathcal{R}_{ibn}$ contains the valid member ROIs of the target cell on complete support; membership is determined by validity rather than context visibility. The regional term averages over member ROIs and feature channels, whereas the network term averages over the channels of a single network representation. Consequently, a target cell's contribution does not scale with its number of valid ROIs.

For branch $\ell \in \{R, N\}$, let L_i^ℓ denote the scan-level loss obtained by averaging over target cells:

$$L_i^\ell = \frac{1}{|\mathcal{M}_i|} \sum_{(b,n) \in \mathcal{M}_i} L_{ibn}^\ell. \tag{D.6}$$

Using the global subject set $\mathcal{S}$, its size S, and the accepted-scan sets $\mathcal{I}_s$ defined in the main text, the branch objective is:

$$\mathcal{L}_\ell = \frac{1}{S} \sum_{s \in \mathcal{S}} \frac{1}{|\mathcal{I}_s|} \sum_{i \in \mathcal{I}_s} L_i^\ell. \tag{D.7}$$

This reduction assigns equal weight to subjects, equal weight to accepted scans within each subject, and equal weight to target cells within each scan. It therefore prevents subjects with more scans or scans with more target cells from receiving greater weight solely because of their counts. When one scan is sampled per subject, the subject-level reduction becomes a mean over the sampled scans.

Prediction errors are evaluated only on valid targets in $\mathcal{M}_i$. The target construction in Equation 5 stops gradients through the teacher, so these losses optimize the student encoder and the corresponding prediction heads.

D.5 Per-Identity Complete-View Moments

The student performs an additional forward pass over all valid signals to compute the variance and LogDet regularizers. The masked-context and complete views use normalization statistics computed independently from their respective available observations. Throughout this subsection, U and V refer to the complete-view student outputs.

Using the branch index $\ell \in \{R, N\}$ and feature widths $q_R = d$ and $q_N = D$ from the main text, write

$$h_{ibj}^R = U_{ibj}, \qquad h_{ibj}^N = V_{ibj}, \qquad \nu_{ibj}^R = v_{ibj}, \qquad \nu_{ibj}^N = w_{ibj}. \tag{D.8}$$

Here h^{ℓ}_{ibj} denotes the feature vector and ν^{ℓ}_{ibj} its validity indicator; the anatomical identity index j refers to an ROI in the regional branch and a network in the network branch.

For each scan–identity pair with at least one valid block, define the temporal mean $\bar{h}^{\ell}_{ij}$ and normalized scan representation $\widetilde{h}^{\ell}_{ij}$ as

$$\bar{h}^{\ell}_{ij} = \frac{\sum_b \nu^{\ell}_{ibj} h^{\ell}_{ibj}}{\sum_b \nu^{\ell}_{ibj}}, \qquad \widetilde{h}^{\ell}_{ij} = \mathrm{LN}(\bar{h}^{\ell}_{ij}). \tag{D.9}$$

The non-affine normalization acts across feature channels separately for each scan and identity, after temporal pooling.

The normalized scan representations are gathered across devices through a differentiable operation and then aggregated by subject. Using the accepted-scan set $\mathcal{I}_s$ defined above, let

$$\mathcal{I}^{\ell}_{sj} = \left\{ i \in \mathcal{I}_s : \sum_b \nu^{\ell}_{ibj} > 0 \right\} \tag{D.10}$$

contain subject s's scans with observed identity j. Define its presence indicator $m^{\ell}_{sj} = \mathbf{1}\{|\mathcal{I}^{\ell}_{sj}| > 0\}$ and subject–identity representation $z^{\ell}_{sj} \in \mathbb{R}^{q_\ell}$ as

$$z^{\ell}_{sj} = \begin{cases} \dfrac{1}{|\mathcal{I}^{\ell}_{sj}|} \displaystyle\sum_{i \in \mathcal{I}^{\ell}_{sj}} \widetilde{h}^{\ell}_{ij}, & m^{\ell}_{sj} = 1, \\ 0, & m^{\ell}_{sj} = 0. \end{cases} \tag{D.11}$$

Only scans containing the identity enter its subject-level average. Missing subject–identity pairs are represented by zero vectors and excluded from the statistical estimates through their presence indicators.

Using the global subject set $\mathcal{S}$ from the main text, the observed-subject count and eligible identity set are

$$n_{\ell j} = \sum_{s \in \mathcal{S}} m^{\ell}_{sj}, \qquad \mathcal{J}_\ell = \{ j : n_{\ell j} \geq 2 \}. \tag{D.12}$$

For each $j \in \mathcal{J}_\ell$, define the observed-subject mean $\bar{z}^{\ell}_{j}$, masked centered residual E^{ℓ}_{sj}, and channel-wise sample variance v^{ℓ}_{jk} as

$$\begin{aligned} \bar{z}^{\ell}_{j} &= \frac{1}{n_{\ell j}} \sum_{s \in \mathcal{S}} m^{\ell}_{sj} z^{\ell}_{sj}, \\ E^{\ell}_{sj} &= m^{\ell}_{sj} (z^{\ell}_{sj} - \bar{z}^{\ell}_{j}), \\ v^{\ell}_{jk} &= \frac{1}{n_{\ell j} - 1} \sum_{s \in \mathcal{S}} (E^{\ell}_{sjk})^2, \qquad k = 1, \ldots, q_\ell. \end{aligned} \tag{D.13}$$

Each observed subject contributes once to these moments, irrespective of its number of scans. Missing pairs contribute neither to the observed-subject mean nor to the residual sum of squares.

Using the positive-part operator defined in the main text, the branch-specific variance penalty is

$$\mathcal{L}_{\mathrm{var},\ell} = \frac{\displaystyle\sum_{j \in \mathcal{J}_\ell} \sum_{k=1}^{q_\ell} \left[0.5 - \sqrt{v^{\ell}_{jk} + 10^{-4}} \right]_+}{\max(1, |\mathcal{J}_\ell|)\, q_\ell}. \tag{D.14}$$

The two branches are weighted equally:

$$\mathcal{L}_{\mathrm{var}} = \frac{1}{2} \left(\mathcal{L}_{\mathrm{var},R} + \mathcal{L}_{\mathrm{var},N} \right). \tag{D.15}$$

The penalty encourages channel-wise variation across subjects within each anatomical identity, with sample variances estimated over observed subjects using the denominator $n_{\ell j} - 1$.

D.6 Rank, Normalization, and Spectral Geometry

Concatenate the centered residuals over eligible identities and feature channels, keeping subjects as rows, to obtain $E_\ell \in \mathbb{R}^{S\times F_\ell}$, with $F_\ell = |\mathcal{J}_\ell| q_\ell$. The subject Gram matrix and its normalized form are

$$G_\ell = E_\ell E_\ell^\top, \qquad K_\ell = \frac{G_\ell}{\max(\operatorname{tr} G_\ell, 10^{-12})}. \tag{D.16}$$

This construction evaluates all eligible identities jointly within each branch.

Using the identity matrix I_S, spectral scale $\alpha = 1$, and nominal capacity $c_\ell = \min(S-1, F_\ell)$ from the main text, the branch-specific LogDet penalty for $c_\ell > 0$ is

$$\mathcal{L}_{\mathrm{logdet},\ell} = \left[\log(1+\alpha) - \frac{1}{c_\ell}\log\det\left(I_S + \alpha c_\ell K_\ell\right)\right]_+. \tag{D.17}$$

The combined objective retains equal weighting of the regional and network branches:

$$\mathcal{L}_{\mathrm{logdet}} = \frac{1}{2}\sum_{\substack{\ell\in\{R,N\}\\ c_\ell>0}} \mathcal{L}_{\mathrm{logdet},\ell}. \tag{D.18}$$

Let $\mathbf{1}_S \in \mathbb{R}^S$ denote the all-ones vector. Centering over observed subjects, together with masked residuals for missing identities, gives $\mathbf{1}_S^\top E_\ell = 0$ and therefore $\operatorname{rank}(E_\ell) \le c_\ell$. When the Gram trace exceeds the numerical floor, K_ℓ is positive semidefinite with unit trace. Concavity of $\log(1+x)$ then gives

$$\frac{1}{c_\ell}\log\det\left(I_S + \alpha c_\ell K_\ell\right) \le \log(1+\alpha). \tag{D.19}$$

Equality requires exactly c_ℓ nonzero eigenvalues of K_ℓ, each equal to $1/c_\ell$. The penalty thus favors a more uniform allocation of normalized variation across representation directions.

The attainable rank also depends on feature normalization and identity support. Non-affine layer normalization imposes a zero channel sum within each identity, yielding

$$\operatorname{rank}(E_\ell) \le \min\left\{S-1,\ |\mathcal{J}_\ell|(q_\ell - 1)\right\}. \tag{D.20}$$

Missing-support patterns can further restrict this rank. Accordingly, c_ℓ is a nominal reference rather than a guaranteed attainable rank; when the attainable rank is smaller, the uniform-spectrum equality cannot be reached and the penalty has a positive lower bound.

The variance penalty encourages the magnitude of cross-subject variation within each identity. Above the numerical floor, trace normalization makes LogDet invariant to common rescaling of E_ℓ, so it regularizes the spectral distribution of that variation.

D.7 Spectral Derivatives and Degenerate Cases

For a branch with $c_\ell > 0$, let $e_\ell = \operatorname{tr} G_\ell$ denote its Gram energy and define $A_\ell = I_S + \alpha c_\ell K_\ell$. When $e_\ell > 10^{-12}$ and the positive-part term in Equation D.17 is active, differentiation with fixed support gives

$$\nabla_{E_\ell}\mathcal{L}_{\mathrm{logdet},\ell} = -\frac{2\alpha}{e_\ell}\left[A_\ell^{-1} - \operatorname{tr}(A_\ell^{-1}K_\ell) I_S\right] E_\ell. \tag{D.21}$$

The gradient is orthogonal to the common scaling direction:

$$\operatorname{tr}\left(E_\ell^\top \nabla_{E_\ell}\mathcal{L}_{\mathrm{logdet},\ell}\right) = 0. \tag{D.22}$$

This agrees with the scale invariance induced by trace normalization above the numerical floor.

At exact cross-subject collapse, $E_\ell = 0$. For positive capacity, the clamped objective then equals $\log(1+\alpha)$, but its gradient with respect to E_ℓ vanishes because the Gram matrix depends quadratically on E_ℓ. The stabilized variance penalty likewise has zero first-order gradient at exact collapse. Thus, these regularizers penalize collapsed representations without ruling out collapsed stationary points.

Table 6: Architecture parameters of the temporal hierarchy and dual-level predictors. Internal widths refer to the Transformer or cross-attention states.

Module	Layers	Internal width	Attention heads
ROI temporal encoder	8	256	16
Network encoder	8	256	8
ROI predictor	4	128	4
Network predictor	4	128	4

For monitoring, let p_a denote the positive eigenvalues of G_ℓ normalized by their sum, with a indexing those eigenvalues. For nonzero Gram energy, the entropy effective rank is:

$$\operatorname{erank}(G_\ell) = \exp\left(-\sum_a p_a \log p_a \right). \tag{D.23}$$

It measures the effective number of directions carrying cross-subject variation. For $G_\ell = 0$, the reported effective rank is defined as zero.

D.8 Loss Weights and Gradient Routing

Equation 11 has exactly four terms. The fixed weights are $(\lambda_1, \lambda_2, \lambda_3) = (0.75, 0.05, 0.5)$. Let $u \geq 0$ denote the optimizer-update count used by the schedule and S_{ep} the number of optimizer updates per epoch. Define the ramp duration $U_{\text{ramp}} = 5S_{\text{ep}}$ and the effective coefficient λ_4 in Equation 11 as

$$\lambda_4 = 0.2 \min\left(\frac{u}{U_{\text{ramp}}}, 1 \right). \tag{D.24}$$

Thus λ_4 increases linearly from 0 to 0.2 over the first five epochs and remains at 0.2 thereafter. The learning-rate warm-up lasts ten epochs. Network prediction backpropagates through the network Transformer, aggregation, ROI temporal Transformer, and tokenizer. ROI prediction follows its visibility-selected direct memory; the network fallback also reaches the earlier ROI temporal path. The variance and LogDet losses update the complete student encoder while bypassing predictors. Each predictor receives only its own prediction loss; the teacher receives no gradients. The optimization contract binds these four terms to their respective prediction and subject-diversity branches.

E Implementation and Training Procedure

E.1 Default Architecture and Optimization

Local temporal encoding uses one width-64 Transformer layer, four heads, and FFN width 256. A single shared within-network spatial-attention module uses four width-16 heads and no spatial FFN. The ROI temporal encoder has eight layers, internal width 256, 16 heads, and FFN width 1024; its input/output projections preserve external ROI width 64. The network Transformer has eight layers, width 256, eight heads, and FFN width 1024. Network summaries project $192 \to 256$. ROI and network predictors each use four independently parameterized layers, internal width 128, four heads, and FFN width 512. Their outputs have widths 64 and 256, respectively. Table 6 summarizes the four principal modules.

Encoder and predictor dropout are zero. Transformer normalizations are affine; identity coupling, teacher targets, readout slots, and moment features use non-affine normalization. The task MLP has hidden width 128 and dropout 0.1. BF16 is used for supported model operations, with normalization statistics and diversity calculations in FP32. ROI temporal activation checkpointing is enabled. AdamW uses weight decay 0.05, learning rates 1.5×10^{-5} for the student and 3×10^{-5} for predictors, and separate gradient clipping at norm 3. The configured budget is 500 epochs on eight devices with 16 subjects per device and no gradient accumulation. The ten-epoch learning-rate warm-up is followed by cosine decay to 0.05 times the peak learning rate.

E.2 One Optimization Step

Each optimization step uses 2 student forward passes and 2 teacher forward passes:

1. Sample 1 scan per subject, construct physical support, and partition valid network–time cells into context, targets, and exclusions.
2. Normalize and encode the visible context to obtain $U^{\mathcal{C}}$ and $V^{\mathcal{C}}$, where the superscript marks the context-only regional and network outputs.
3. Encode complete real and raw-zero signals with the teacher under identical metadata and without gradients. Subtract and normalize the outputs to construct regional and network targets.
4. Predict network targets from visible network memory and ROI targets from visibility-routed memory. Compute both SmoothL1 prediction losses using the subject-balanced reduction.
5. Encode the complete student view, pool and normalize features by identity, gather representations across devices, and aggregate scans by subject. Compute the variance and whole-branch LogDet losses.
6. Combine the 4 loss terms using their coefficients and the update-based LogDet ramp. Backpropagate, clip student and predictor gradients separately, and update the optimizer. Then update the EMA teacher parameters and copy student buffers.

E.3 Factorized Attention Cost

Let B, R, and N denote the numbers of temporal blocks, ROIs, and networks, respectively, and let R_n be the number of ROIs in network n. Let L_R, L_N be the ROI temporal and network encoder depths and d_T, D their internal widths. Dense attention over all ROI–time tokens has an attention interaction cost of $O(B^2R^2d)$ at width d. STH-Brain decomposes these interactions into within-network spatial attention, ROI-specific temporal attention, and network–time attention:

$$\mathcal{C}_{\text{hier}} = O\left(B \sum_{n=1}^{N} R_n^2 d + L_R R B^2 d_T + L_N B^2 N^2 D \right). \tag{E.1}$$

The 3 terms correspond to spatial, regional temporal, and network–time interactions, respectively. This expression characterizes the encoder's attention interactions; the per-block local temporal attention, projections, feed-forward layers, prediction modules, and additional student/teacher passes contribute further computation.

E.4 Coordinate Semantics and Conditional Dependence

For shared physical timestamps u, w and shared rotary parameters, relative time phase is $\phi_k(u) - \phi_k(w) = 2\pi\omega_k(u-w)/P_k$. This expression defines a common physical coordinate scale for cross-block interactions. Native support determines the queried observations, and the atlas origin and axes determine the spatial reference.

Attention expresses content- and position-dependent retrieval, with fixed ROI membership supplying the network prior. Context interventions compare prediction errors under clean and zeroed visible signals while keeping targets and support fixed. This difference quantifies the contribution of visible signal content to prediction. Direct network-memory perturbations assess the regional routing rule, while identity-specific interventions assess the regional information retained after spatial and temporal encoding.

F Downstream Readout and Evaluation Protocol

F.1 Identity-Preserving Readout

For scan i, let v_{ibr} and w_{ibn} be binary validity indicators for ROI r and network n in block b, respectively. Let $\bar{U}_{ir} \in \mathbb{R}^d$ and $\bar{V}_{in} \in \mathbb{R}^D$ denote the complete-view regional and network representations

averaged over valid temporal blocks:

$$\bar{U}_{ir} = \frac{\sum_b v_{ibr} U_{ibr}}{\max(1, \sum_b v_{ibr})}, \qquad \bar{V}_{in} = \frac{\sum_b w_{ibn} V_{ibn}}{\max(1, \sum_b w_{ibn})}. \tag{F.1}$$

Identities with no valid blocks are represented by zero vectors. Binary vectors $m_i^R \in \{0,1\}^R$ and $m_i^N \in \{0,1\}^N$ indicate the presence of each ROI and network, respectively.

Let LN denote non-affine layer normalization over the feature channels of each identity, and let vec flatten the representations in fixed ROI or network order. With regional and network feature widths d and D, let $d_{\text{out}} = Rd + ND + R + N$ be the readout width. To distinguish the readout from the binary validity v_{ibr}, write its vector as $\mathbf{v}_i$. The joint readout is:

$$\mathbf{v}_i = \left[\frac{\text{vec}(\text{LN}(\bar{U}_i))}{\sqrt{Rd}}; \frac{\text{vec}(\text{LN}(\bar{V}_i))}{\sqrt{ND}}; m_i^R; m_i^N \right] \in \mathbb{R}^{d_{\text{out}}}, \qquad d_{\text{out}} = Rd + ND + R + N. \tag{F.2}$$

The fixed scaling factors balance the dimensional contributions of the 2 content branches, while the appended validity indicators identify missing anatomical slots. Temporal pooling preserves a separate readout slot for each ROI and network.

Let $d_{\text{head}} = 128$ be the hidden width. The task-specific MLP consists of layer normalization, a linear projection from d_{out} to d_{head}, GELU, dropout, and an output projection. Binary classification uses 2 output logits, whereas age prediction uses 1 scalar. Both encoder levels and the MLP are jointly fine-tuned; the teacher and pretraining predictors are discarded. Numerical architecture settings are provided in Appendix E.

Supervised Losses and Scaling. Binary classification uses class-weighted cross-entropy, and age prediction uses SmoothL1 loss on standardized ages, as detailed in Appendix F.2. Class weights and age-standardization statistics are estimated from the training subjects. Predicted ages are transformed back to years for evaluation.

F.2 Downstream Tasks and Supervised Objectives

Task Adaptation. We evaluate 7 internal patient-versus-HC classification tasks (AD, MDD, ASD, ADHD, MCI, PD, and PROPD), age prediction, external SCZ/HC generalization, and zero-shot MDD/HC-to-BP/HC transfer. For each supervised task, the pretrained student encoder provides regional and network representations, which are temporally pooled, normalized, and concatenated in fixed anatomical order with validity indicators to form $\mathbf{v}_i$ (Appendix F). A task-specific MLP produces 2 logits for classification or 1 scalar for age prediction. The encoder and MLP are jointly fine-tuned for each task, and the teacher and pretraining predictors are discarded. The age model is trained on eligible training subjects pooled across the 7 internal data sources, with model selection performed on the corresponding validation subjects. Age-prediction performance is reported separately for the 7 diagnosis-specific test groups.

Binary Classification Loss. Let g_{cls} denote the 2-logit classification head and $\mathcal{B}_{\text{sup}}$ the nonempty collection of subject samples accumulated for 1 optimizer step. Here i indexes a subject sample represented by its registered scan. Let $n_c > 0$ be the number of unique training subjects with class label $c \in \{0,1\}$, where each subject contributes once to the class counts. Define the fixed class weights w_c as:

$$w_c = \min\left(\frac{n_0 + n_1}{2n_c}, 5 \right), \qquad c \in \{0,1\}. \tag{F.3}$$

These class weights remain fixed throughout the corresponding supervised fit. For label $y_i \in \{0,1\}$, with patients assigned to class 1, define the predicted patient probability as $p_i = [\text{softmax}(g_{\text{cls}}(\mathbf{v}_i))]_1$. The classification objective is

$$\mathcal{L}_{\text{cls}} = -\frac{1}{|\mathcal{B}_{\text{sup}}|} \sum_{i \in \mathcal{B}_{\text{sup}}} \left[w_1 y_i \log p_i + w_0 (1 - y_i) \log(1 - p_i) \right]. \tag{F.4}$$

Weighted losses are summed across microbatches and normalized by the actual sample count $|\mathcal{B}_{\text{sup}}|$ in the accumulation window, including a partial final window. The age objective uses the same sample-average reduction.

Age Prediction Loss. Let a_i denote the observed age in years and g_{age} the scalar age-prediction head. Let $\epsilon > 0$ be the numerical floor for the age scale. For the set $\mathcal{S}_{\text{tr}}$ of eligible training subjects, let $\bar{a}_s$ be the mean of the available valid ages for subject s. The normalization statistics give each training subject equal weight:

$$\begin{aligned} \mu_{\text{tr}} &= \frac{1}{|\mathcal{S}_{\text{tr}}|} \sum_{s \in \mathcal{S}_{\text{tr}}} \bar{a}_s, \\ \sigma_{\text{tr}} &= \max\left\{ \sqrt{\frac{1}{|\mathcal{S}_{\text{tr}}|} \sum_{s \in \mathcal{S}_{\text{tr}}} (\bar{a}_s - \mu_{\text{tr}})^2}, \epsilon \right\}. \end{aligned} \tag{F.5}$$

These statistics remain fixed for validation and testing. Define the standardized target $a_i^z = (a_i - \mu_{\text{tr}})/\sigma_{\text{tr}}$, the prediction $\widehat{a}_i^z = g_{\text{age}}(\mathbf{v}_i)$, and the residual $e_i = \widehat{a}_i^z - a_i^z$. For transition parameter $\beta > 0$, SmoothL1 is

$$\ell_\beta(e) = \begin{cases} \dfrac{e^2}{2\beta}, & |e| < \beta, \\ |e| - \dfrac{\beta}{2}, & |e| \geq \beta. \end{cases} \tag{F.6}$$

We use $\beta = 1$. The age-prediction objective is

$$\mathcal{L}_{\text{reg}} = \frac{1}{|\mathcal{B}_{\text{sup}}|} \sum_{i \in \mathcal{B}_{\text{sup}}} \ell_\beta(e_i). \tag{F.7}$$

Predictions are transformed back to years using the training statistics:

$$\widehat{a}_i = \mu_{\text{tr}} + \sigma_{\text{tr}} \widehat{a}_i^z. \tag{F.8}$$

MAE and RMSE are computed in years, and Pearson correlation R is computed between predicted and observed ages.

Downstream Objective. Each downstream task uses its corresponding supervised objective:

$$\mathcal{L}_{\text{task}} = \begin{cases} \mathcal{L}_{\text{cls}}, & \text{binary classification}, \\ \mathcal{L}_{\text{reg}}, & \text{age prediction}. \end{cases} \tag{F.9}$$

Downstream fine-tuning optimizes $\mathcal{L}_{\text{task}}$ using separate task-specific fits. For external generalization, the SCZ/HC model is selected using its source-validation data and evaluated on the external SCZ/HC cohort. For zero-shot transfer, the MDD/HC model is selected using its source-validation data and applied directly to the BP/HC cohort. Both evaluations retain the source-selected checkpoint and decision rule, without target-based training, checkpoint selection, or threshold adjustment.

G Experimental Protocol and Evaluation Details

G.1 Data Eligibility, Matching, and Partitions

Inventory and Source-Specific Roles: Section 4 describes the partitioning procedure for each data source. The mutually exclusive groups comprise 9,040 self-supervised pretraining subjects, 905 validation subjects, 1,803 internal test subjects, 620 external SCZ/HC evaluation subjects, and 406 zero-shot BP/HC evaluation subjects, totaling 12,774 unique subjects. SRPBS and COBRE have no separate self-supervised learning validation partition, as all included subjects contribute to self-supervised pretraining. Table 7 summarizes participant counts, age, gender, and diagnostic groups by dataset and experimental use.

Matched Tasks and Longitudinal Subjects. For ADNI, separate AD/HC and MCI/HC cohorts are constructed by matching patients to healthy controls on age and sex. For PPMI, healthy controls serve as matching anchors for selecting age- and sex-matched patients, yielding separate PD/HC and PROPD/HC cohorts. Each matched cohort is divided into training, validation, and test sets at a nominal subject-level ratio of 7:1:2, with matched pairs assigned to the same partition. For subjects with longitudinal scans, all visits remain in the same partition, irrespective of which visit is selected for a particular task. Healthy controls shared between the AD/HC and MCI/HC tasks, or between

Table 7: Dataset characteristics by experimental use. Train/test includes the self-supervised pretraining, validation, and internal test inventory; SRPBS and COBRE also provide supervised source cohorts. Diagnostic counts use one representative visit per subject. Task-specific cohorts are defined separately through visit selection and matching, and subjects may contribute to multiple tasks. The three REST-MDDII rows contain disjoint subjects; external-generalization and zero-shot cohorts are excluded from development. Age is reported in years as mean $\pm$ sample standard deviation from registered subject ages. Gender is reported as male/female (M/F). HC denotes healthy controls.

Type	Dataset	Participants	Age (years) Mean $\pm$ SD	Gender (M/F)	Diagnostic groups (n)
Train/test	ABIDE-I	1,075	16.85 ± 7.97	918/157	ASD (517), HC (558)
	ABIDE-II	1,090	14.45 ± 9.01	847/243	ASD (518), HC (572)
	ADHD	863	12.08 ± 3.30	528/335	ADHD (341), HC (522)
	ADNI	1,095	72.79 ± 8.02	499/596	AD (133), MCI (361), HC (601)
	PPMI	1,307	65.65 ± 8.19	760/547	PD (483), PROPD (755), HC (69)
	SRPBS	1,318	38.46 ± 13.74	756/562	ASD (125), MDD (255) SCZ (147), HC (791)
	COBRE	145	36.96 ± 12.83	108/37	SCZ (71), HC (74)
	REST-MDDI	2,377	36.24 ± 15.07	924/1453	MDD (1276), HC (1101)
	REST-MDDII	2,478	33.72 ± 13.61	921/1557	MDD (1650), HC (828)
External generalization	REST-MDDII	620	26.21 ± 8.20	324/296	SCZ (310), HC (310)
Zero-shot	REST-MDDII	406	29.70 ± 10.47	172/234	BP (203), HC (203)
Total	—	12,774	36.31 ± 21.31	6757/6017	—

the PD/HC and PROPD/HC tasks, retain consistent partition assignments across the corresponding tasks. These constraints prevent the same subject from appearing in different partitions through repeated visits or reuse across tasks. Self-supervised pretraining includes subjects assigned to the matched training cohorts and the remaining unmatched subjects; subjects assigned to the matched validation or test cohorts are excluded from pretraining. Supervised training, validation, and testing use only the corresponding partitions of the matched cohorts.

REST-MDDII Target Cohorts. The SCZ/HC and BP/HC target cohorts are disjoint, including their controls, and are held out from self-supervised pretraining and supervised development. The SCZ/HC cohort contains 620 subjects (310 SCZ/310 HC; 324 male/296 female), and the BP/HC cohort contains 406 subjects (203 BP/203 HC; 172 male/234 female). Controls are matched by age and sex, prioritizing same-site matches with cross-site supplementation. The remaining 2,478 subjects comprise 1,650 MDD patients and 828 HC and are divided into subject-level training, validation, and test sets at a nominal ratio of 7:1:2.

Supervised Sources and Transfer. For external SCZ/HC evaluation, the model is fine-tuned on the SRPBS/COBRE SCZ/HC training cohort and selected using the corresponding source-validation cohort. The external target cohort remains excluded from both pretraining and supervised development. For zero-shot BP/HC transfer, a separate model is fine-tuned on MDD/HC source data and selected using its corresponding source-validation cohort. The selected MDD/HC model is applied directly to the held-out BP/HC cohort, retaining its source decision rule. Both evaluations use fixed source-selected checkpoints and decision rules, without target-based training, checkpoint selection, or threshold adjustment.

ROI-to-Network Assignment. CC200 regions were organized into eight predefined groups comprising the visual (VIS), somatomotor (SMN), dorsal attention (DAN), ventral attention (VAN), limbic (LIM), frontoparietal (FPN), and default mode (DMN) networks, together with a supplementary cerebellar category.

G.2 Preprocessing Details

The fMRI data were preprocessed using Data Processing & Analysis for Brain Imaging (DPABI; `http://www.rfmri.org/dpabi`). The first 10 volumes of each scan were removed. Preprocessing included slice-timing correction, six-parameter rigid-body realignment, nuisance regression

using white matter and cerebrospinal fluid signals, coregistration with T1-weighted images, and tissue segmentation. Spatial normalization to the Montreal Neurological Institute 152 (MNI-152) template was performed using Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra (DARTEL). Linear trends were regressed out, and a 0.01–0.1 Hz band-pass filter was applied. The brain was parcellated into 200 regions using the Craddock 200-region atlas (CC200) Craddock et al. (2012).

G.3 Architecture, Optimization, and Model Selection

Implementation and Architecture. Experiments use PyTorch 2.9.1. Pretraining runs on 8 NVIDIA RTX PRO 6000D GPUs using distributed data parallelism and BF16 precision, while downstream fine-tuning runs on 1 GPU. Inputs are divided into non-overlapping 8-s blocks, with a 1-s physical query grid and a maximum interpolation gap of 6.5 s. The local temporal Transformer has 1 layer and 4 heads, with an ROI output width of 64. The cross-block ROI Transformer has 8 layers, an internal width of 256, and 16 heads. The network Transformer has 8 layers, width 256, and 8 heads. Both predictors have 4 layers, an internal width of 128, and 4 heads. The downstream MLP comprises layer normalization, a linear projection to width 128, GELU, dropout of 0.1, and a task-specific output projection. Further module specifications are provided in Appendix E.

Optimization. Pretraining and fine-tuning use AdamW with weight decay of 0.05 and a gradient-norm limit of 3. Pretraining runs for up to 500 epochs, with a 10-epoch learning-rate warm-up followed by cosine decay to 0.05 times the peak learning rate. After the LogDet ramp, the loss coefficients in Equation 11 are $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.75, 0.05, 0.5, 0.2)$. The effective coefficient λ_4 follows Equation D.24, increasing linearly from 0 to 0.2 over the first 5 epochs and remaining at 0.2 thereafter. Fine-tuning runs for up to 200 epochs, with validation-based early stopping using a patience of 10 epochs. Batch sizes, gradient accumulation, and learning rates are specified in Section 4.

Model Selection and Repeated Runs. Downstream models are initialized with the student encoder from the checkpoint achieving the lowest epoch-averaged total pretraining loss on the training set. Classification checkpoints are selected by the highest validation accuracy (ACC), using a patient-probability threshold of 0.5. Age-prediction checkpoints are selected by the lowest validation MAE. Each downstream task is repeated with 10 random seeds under fixed subject-level partitions, and results are reported as mean $\pm$ standard deviation across runs. Class weights and fitted feature- and age-scaling statistics are estimated from the corresponding training data. External and zero-shot evaluations use source-selected checkpoints and retain the source decision rule.

G.4 Classification Metrics

Let TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively, with patients treated as the positive class. Let $n = \mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}$ be the number of evaluated subjects. Accuracy (ACC) and the F1 score are defined as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{n}, \qquad \mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}. \tag{G.1}$$

The area under the receiver operating characteristic curve (AUC) is computed from continuous patient probabilities across decision thresholds. ACC and F1 use a fixed patient-probability threshold of 0.5. All 3 metrics are reported as percentages, with higher values indicating better performance. Metrics are computed separately for each random seed before summarizing across runs.

G.5 Age Prediction Metrics

Let a_i and $\widehat{a}_i$ denote the true and predicted ages, respectively, of subject i, and let n denote the number of evaluated subjects in the reported group. Predictions are converted from standardized age back to years using the training-set mean and standard deviation before evaluation. Mean absolute error (MAE) and root mean squared error (RMSE) are:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\widehat{a}_i - a_i|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{a}_i - a_i)^2}. \tag{G.2}$$

Table 8: Classification results for ADHD, MCI, and PROPD. Results are reported as mean ± standard deviation (%) across 10 random seeds. **Bold** and underlined values denote the best and second-best results, respectively. * denotes statistical significance between the best and second-best models (paired t-test, $p < 0.05$).

Method	Pretraining	ADHD			MCI			PROPD		
		ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑
BrainHGT	—	66.02±2.82	72.47±3.61	61.35±2.95	66.53±2.72	73.42±2.97	66.19±2.74	66.77±3.07	73.66±2.55	64.92±3.37
BrainIB	—	65.46±3.26	71.99±3.56	60.06±3.48	65.82±2.66	72.61±2.87	66.37±2.63	66.08±2.64	72.86±2.97	64.28±2.84
HiGCN	—	65.03±2.72	71.13±2.92	59.09±2.92	65.11±3.58	71.83±3.28	65.55±3.53	65.19±2.96	71.99±2.72	64.35±3.00
ALTER	—	64.53±3.04	70.92±2.89	59.26±3.21	64.37±2.61	71.13±3.14	65.48±2.58	64.42±2.75	71.17±3.57	62.42±3.16
BrainMass	✓	69.97±3.12	76.82±2.88	64.78±3.39	71.52±3.69	78.57±3.72	71.66±3.69	73.04±2.59	80.09±2.61	71.57±2.83
Brain-JEPA	✓	70.47±2.61	77.09±3.24	65.01±2.87	72.23±3.53	79.15±3.24	72.22±3.54	73.81±2.72	80.98±3.37	73.11±2.76
Brain-Semantoks	✓	69.48±3.35	76.03±2.79	63.38±3.71	72.95±2.98	79.82±3.08	73.56±2.93	72.23±3.12	79.49±2.78	73.68±2.98
FROG	✓	68.98±3.25	75.78±3.58	63.32±3.51	70.81±3.71	77.84±3.18	70.84±3.71	71.54±2.51	78.75±3.56	72.37±2.61
BCNSL	✓	68.49±2.91	75.11±3.21	63.15±3.14	70.09±3.08	76.85±3.25	70.52±3.06	70.65±2.93	77.84±3.26	68.67±3.23
FMM-TC	✓	67.99±3.13	74.71±3.06	62.81±3.32	69.38±3.06	76.47±2.99	69.24±3.11	69.88±2.84	76.91±3.15	70.33±2.89
PHGNN	✓	67.51±3.19	74.08±3.27	62.18±3.42	68.66±2.84	75.73±3.31	69.53±2.78	69.12±2.86	76.12±3.18	68.59±2.99
DeepASD	✓	67.01±2.99	73.59±3.35	60.53±3.26	67.94±3.13	74.77±3.58	68.09±3.14	68.31±2.79	75.46±3.39	67.87±2.93
xAI-ASD	✓	66.52±3.46	72.76±3.38	60.72±3.73	67.24±3.47	74.11±2.95	66.93±3.52	67.58±2.91	74.33±2.98	67.14±3.13
STH-Brain	✓	**76.68±2.18***	**83.46±2.28***	**72.13±2.44***	**79.22±2.26***	**86.14±2.21***	**80.03±2.19***	**79.46±2.46***	**86.74±2.38***	**78.89±2.38***

Let $\bar{a} = n^{-1}\sum_{i=1}^{n} a_i$ and $\bar{\widehat{a}} = n^{-1}\sum_{i=1}^{n} \widehat{a}_i$ denote the mean true and predicted ages. Pearson correlation R is:

$$R = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(\widehat{a}_i - \bar{\widehat{a}})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{n}(\widehat{a}_i - \bar{\widehat{a}})^2}}. \tag{G.3}$$

MAE and RMSE are measured in years, with lower values indicating smaller prediction errors. Pearson R is dimensionless, with higher values indicating stronger positive association between predicted and true ages. It is undefined when either true or predicted ages are constant. Results are computed separately within each reported disease group and random seed. For each task or disease group, metrics are computed separately for each of the 10 random seeds and reported as mean ± sample standard deviation across runs.

H Additional Results

H.1 Classification Performance

Table 8 reports classification results for ADHD, MCI, and PROPD. STH-Brain achieved the highest mean ACC, AUC, and F1 on all three tasks. The second-best models in ACC were Brain-JEPA, Brain-Semantoks, and Brain-JEPA, respectively; STH-Brain improved ACC by 6.21, 6.27, and 5.65 percentage points, respectively.

H.2 Age Prediction

Table 9 presents age-prediction results for ADHD, MCI, and PROPD. STH-Brain achieved the lowest mean MAE and RMSE and the highest mean Pearson R in all three groups. Compared with the lowest-MAE baseline in each group (Brain-JEPA for ADHD, Brain-Semantoks for MCI, and Brain-JEPA for PROPD), STH-Brain reduced MAE by 0.071, 0.150, and 0.125 years, respectively. Relative to the strongest baseline for each metric, RMSE decreased by 0.140, 0.280, and 0.156 years, while Pearson R increased by 0.009, 0.005, and 0.001, respectively. These results demonstrate consistently lower age-prediction errors and higher correlations with chronological age across ADHD, MCI, and PROPD.

H.3 Additional Ablation Studies

Ablation settings. We compared the full model with three ablated variants: (1) **w/o Physical-Time Rotary Encoding**, removing Physical-Time Rotary Encoding; (2) **w/o Spatial-Aware Encoding**, removing spatial geometry encoding; (3) **w/o Visibility Routing**, replacing the exclusive visibility-based memory selection with joint attention to the target ROI's visible temporal states and all visible

Table 9: Age-prediction results for ADHD, MCI, and PROPD. Results are reported as mean ± standard deviation across 10 random seeds. MAE and RMSE are measured in years; Pearson R is dimensionless. ↓ and ↑ indicate that lower and higher values are better, respectively. **Bold** and underlined values denote the best and second-best results, respectively. * denotes statistical significance between the best and second-best models (paired t-test, $p < 0.05$).

Method	Pretraining	ADHD			MCI			PROPD		
		MAE↓	RMSE↓	Pearson R ↑	MAE↓	RMSE↓	Pearson R ↑	MAE↓	RMSE↓	Pearson R ↑
BrainHGT	—	1.578±0.218	2.043±0.287	0.817±0.035	3.976±0.355	4.988±0.433	0.847±0.023	3.048±0.194	3.888±0.372	0.963±0.013
BrainIB	—	1.602±0.212	2.025±0.284	0.823±0.040	4.058±0.359	5.211±0.426	0.845±0.010	3.076±0.178	3.765±0.458	0.955±0.011
HiGCN	—	1.644±0.193	2.085±0.321	0.819±0.062	4.127±0.324	5.385±0.649	0.830±0.046	3.134±0.144	4.077±0.412	0.948±0.015
ALTER	—	1.688±0.255	2.201±0.318	0.785±0.073	4.217±0.446	5.395±0.461	0.842±0.045	3.231±0.267	3.896±0.382	0.961±0.011
BrainMass	✓	1.186±0.177	1.519±0.218	0.889±0.034	2.979±0.302	3.844±0.370	0.905±0.015	2.273±0.151	2.919±0.347	0.975±0.008
Brain-JEPA	✓	1.176±0.163	1.499±0.215	0.894±0.032	3.006±0.284	3.900±0.459	0.906±0.023	2.250±0.143	2.837±0.273	0.976±0.005
Brain-Semantoks	✓	1.188±0.157	1.545±0.191	0.896±0.014	2.933±0.257	3.720±0.320	0.914±0.009	2.280±0.153	2.839±0.359	0.976±0.007
FROG	✓	1.191±0.167	1.527±0.199	0.895±0.024	3.000±0.306	3.858±0.363	0.905±0.015	2.290±0.141	2.887±0.171	0.977±0.005
BCNSL	✓	1.408±0.187	1.764±0.237	0.872±0.027	3.540±0.274	4.594±0.344	0.879±0.016	2.717±0.185	3.544±0.307	0.973±0.003
FMM-TC	✓	1.447±0.211	1.896±0.298	0.844±0.046	3.639±0.393	4.758±0.447	0.878±0.016	2.787±0.174	3.587±0.431	0.964±0.015
PHGNN	✓	1.460±0.240	1.890±0.299	0.840±0.033	3.671±0.395	4.744±0.553	0.870±0.028	2.809±0.200	3.470±0.367	0.967±0.008
DeepASD	✓	1.512±0.225	1.961±0.297	0.830±0.038	3.809±0.404	4.853±0.503	0.851±0.030	2.892±0.222	3.548±0.452	0.962±0.008
xAI-ASD	✓	1.187±0.176	1.495±0.228	0.892±0.038	2.988±0.296	3.881±0.470	0.904±0.032	2.265±0.145	2.929±0.335	0.976±0.006
STH-Brain	✓	**1.105±0.155***	**1.355±0.167***	**0.905±0.022***	**2.783±0.265***	**3.440±0.262***	**0.919±0.010***	**2.125±0.136***	**2.681±0.172***	**0.978±0.003**

Table 10: Classification ablation results for ADHD, MCI, and PROPD. Results are reported as mean ± standard deviation (%) across 10 random seeds. ↑ indicates that higher values are better. **Bold** and underlined values denote the best and second-best means among the evaluated variants within each group and metric, respectively.

Model	ADHD			MCI			PROPD		
	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑
w/o Physical-Time Rotary Encoding	71.65±3.12	78.62±3.32	67.12±3.34	74.02±3.19	80.98±2.97	74.05±3.22	74.19±2.69	81.34±2.65	73.61±2.88
w/o Spatial-Aware Encoding	72.66±3.21	79.42±2.95	68.27±3.47	75.19±3.17	82.08±3.39	75.86±3.08	75.54±2.87	82.63±2.92	75.18±3.02
w/o Visibility Routing	73.85±3.08	80.75±3.39	69.09±3.41	76.58±2.62	83.28±3.01	77.22±2.54	76.73±2.59	84.01±2.56	77.09±2.51
STH-Brain (Full)	**76.68±2.18**	**83.46±2.28**	**72.13±2.44**	**79.22±2.26**	**86.14±2.21**	**80.03±2.19**	**79.46±2.46**	**86.74±2.38**	**78.89±2.38**

network states from the same scan, while retaining the ROI predictor, prediction targets, and pretraining loss terms. Age prediction was evaluated using the same fixed subject-level partitions and 10 random seeds.

H.3.1 Classification Ablation

Table 10 reports classification ablations for ADHD, MCI, and PROPD using the same variants, fixed subject-level partitions, and 10 random seeds as in the main text. The full model achieved the highest mean ACC, AUC, and F1 in all three groups, yielding the best mean performance in all 9 group–metric combinations. Relative to the full model, ACC decreased by 5.03–5.27 percentage points for *w/o Physical-Time Rotary Encoding*, 3.92–4.03 for *w/o Spatial-Aware Encoding*, and 2.64–2.83 for *w/o Visibility Routing*. These results support the contribution of the evaluated components to classification across the three additional disease groups.

H.3.2 Age-Prediction Ablation

Table 11 reports age-prediction ablations across all seven disease groups using the same variants, fixed subject-level partitions, and 10 random seeds as in the main text. The full model achieved the lowest mean MAE and RMSE and the highest mean Pearson R in every group, yielding the best mean performance in all 21 group–metric combinations. The MAE increases were 0.936–2.364 years for *w/o Physical-Time Rotary Encoding*, 0.660–1.638 for *w/o Spatial-Aware Encoding*, and 0.372–0.987 for *w/o Visibility Routing*. These results support the contribution of all three components to age prediction, with physical-time rotary encoding showing the largest observed effect on MAE.

H.4 Influence of Atlas Choice

The atlas comparison showed a consistent ranking across all 7 classification tasks: CC200 achieved the highest mean AUC and F1, followed by Power264, Dosenbach160, and AAL116 (Figures 9

Table 11: Age-prediction ablation results across 7 disease groups. Results are reported as mean ± standard deviation across 10 random seeds. MAE and RMSE are in years; Pearson R is dimensionless. ↓ and ↑ indicate that lower and higher values are better, respectively. **Bold** and underlined values denote the best and second-best means among the evaluated variants within each group and metric, respectively.

Model	AD			PD			ASD			MDD		
	MAE↓	RMSE↓	Pearson R ↑	MAE↓	RMSE↓	Pearson R ↑	MAE↓	RMSE↓	Pearson R ↑	MAE↓	RMSE↓	Pearson R ↑
w/o Physical-Time Rotary Encoding	4.924±0.633	6.498±0.707	0.815±0.042	4.267±0.773	5.678±1.248	0.906±0.031	2.874±0.136	3.647±0.169	0.923±0.006	4.306±0.180	5.529±0.177	0.931±0.005
w/o Spatial-Aware Encoding	4.191±0.479	5.387±0.633	0.847±0.034	3.692±0.662	4.555±0.824	0.934±0.030	2.482±0.117	3.219±0.183	0.938±0.007	3.704±0.164	4.769±0.220	0.947±0.005
w/o Visibility Routing	3.588±0.443	4.508±0.626	0.877±0.036	3.089±0.587	4.005±1.012	0.947±0.025	2.127±0.109	2.724±0.145	0.955±0.006	3.134±0.129	4.024±0.181	0.961±0.003
STH-Brain (Full)	**2.639±0.320**	**3.326±0.378**	**0.941±0.018**	**2.300±0.420**	**2.963±0.415**	**0.975±0.007**	**1.551±0.079**	**1.947±0.084**	**0.976±0.002**	**2.318±0.109**	**2.906±0.155**	**0.979±0.002**

Model	ADHD			MCI			PROPD		
	MAE↓	RMSE↓	Pearson R ↑	MAE↓	RMSE↓	Pearson R ↑	MAE↓	RMSE↓	Pearson R ↑
w/o Physical-Time Rotary Encoding	2.041±0.288	2.611±0.335	0.747±0.039	5.147±0.476	6.629±0.738	0.790±0.020	3.930±0.261	5.186±0.609	0.931±0.013
w/o Spatial-Aware Encoding	1.765±0.240	2.270±0.323	0.784±0.053	4.421±0.405	5.628±0.496	0.825±0.036	3.368±0.240	4.173±0.397	0.952±0.012
w/o Visibility Routing	1.477±0.208	1.906±0.292	0.817±0.053	3.770±0.358	4.880±0.412	0.870±0.009	2.886±0.193	3.454±0.315	0.965±0.012
STH-Brain (Full)	**1.105±0.155**	**1.355±0.167**	**0.905±0.022**	**2.783±0.265**	**3.440±0.262**	**0.919±0.010**	**2.125±0.136**	**2.681±0.172**	**0.978±0.003**

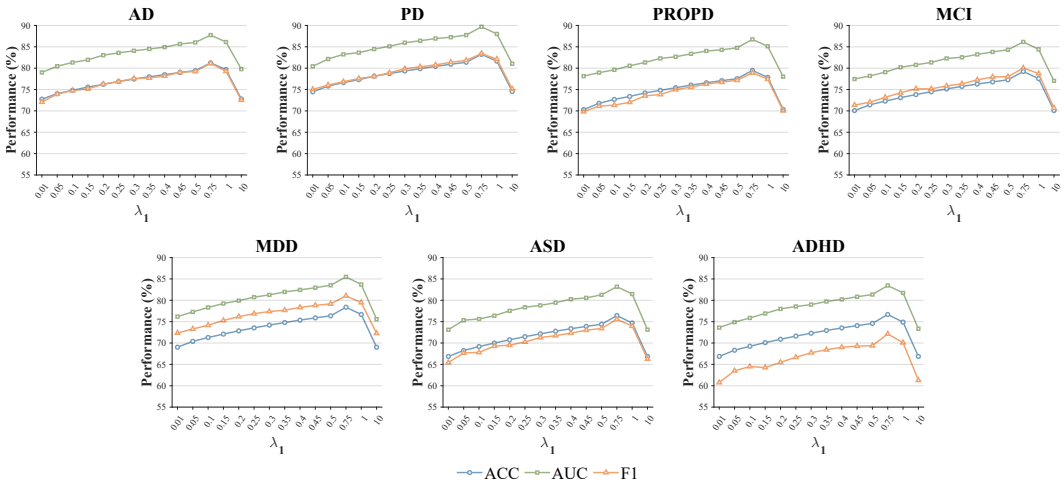


Figure 5: ROI-prediction coefficient sweep across 7 diseases. ACC, AUC, and F1 are reported as %. The horizontal axis varies λ_1 in Equation 11.

and 10). Compared with Power264, the second-best atlas, CC200 improved AUC and F1 by 1.21–1.99 and 1.02–2.26 percentage points, respectively. The corresponding improvements over Dosenbach160 were 3.37–4.68 and 3.48–5.40 percentage points, and those over AAL116 were 5.03–6.81 and 4.85–7.16 percentage points. These results complement the ACC comparison in the main text and show consistent advantages of CC200 across the evaluated classification metrics.

H.5 Loss-Coefficient Sensitivity

To assess how the weighting of prediction and regularization terms affects classification, we evaluated each of the four loss coefficients over 0.01–10.00 using ACC, AUC, and F1 (Figures 5–8). The coefficients λ_1, λ_2, λ_3, and λ_4 weight ROI prediction, network prediction, variance regularization, and LogDet regularization, respectively, in Equation 11.

Across the evaluated range, classification performance generally improved as each coefficient increased toward its best-performing value and declined with further increases. The mean-ACC profiles reached their maxima at $\lambda_1 = 0.75$, $\lambda_2 = 0.05$, $\lambda_3 = 0.50$, and $\lambda_4 = 0.20$. These results indicate that both insufficient and excessive weighting of individual loss terms can impair classification, highlighting the importance of balancing prediction and regularization objectives.

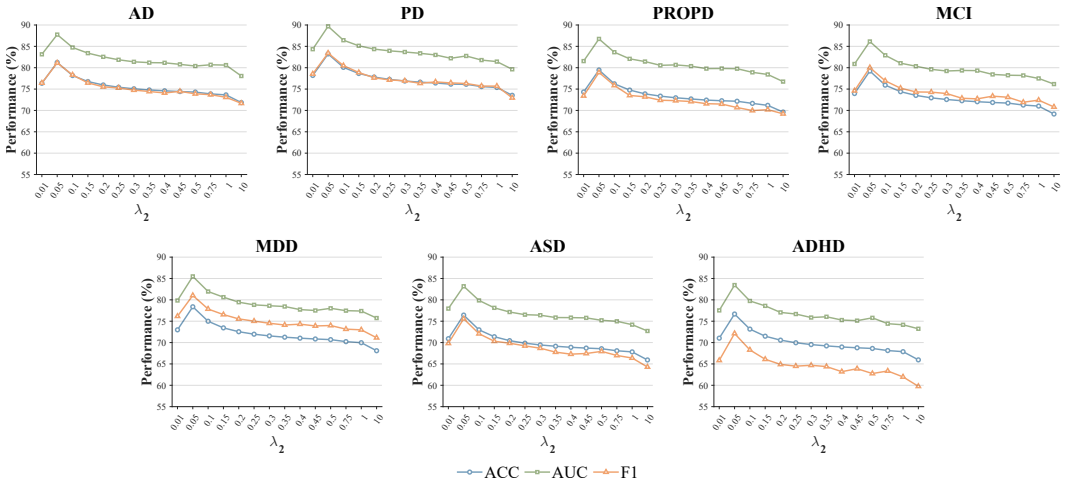


Figure 6: Network-prediction coefficient sweep across 7 diseases. ACC, AUC, and F1 are reported as %. The horizontal axis varies λ_2 in Equation 11.

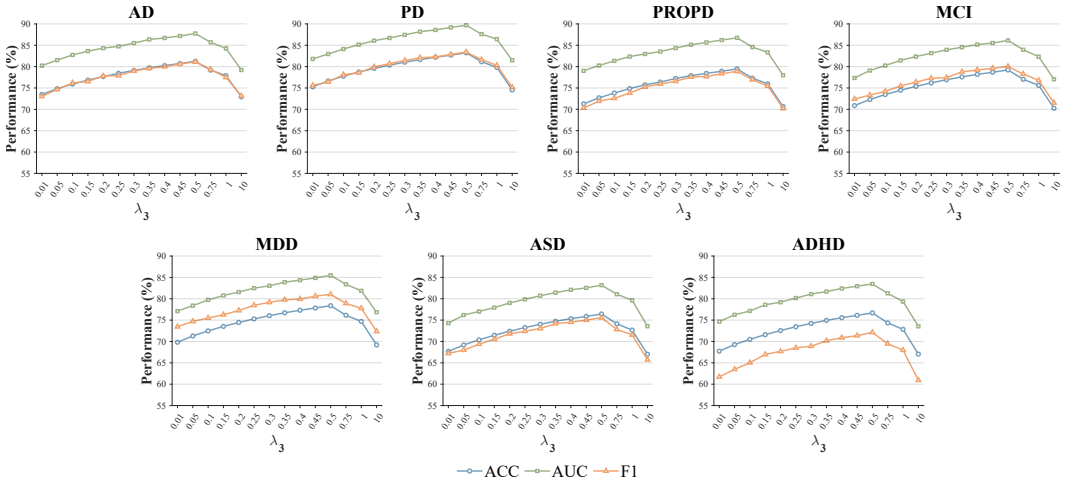


Figure 7: Variance-regularization coefficient sweep across 7 diseases. ACC, AUC, and F1 are reported as %. The horizontal axis varies λ_3 in Equation 11.

H.6 Interpretability Analysis

To complement the main-text analysis, this appendix presents CC200-based regional profiles for all seven disease groups, highlighting shared regions and differences in their anatomical distributions.

The CC200-based anatomical profiles for PROPD (Figure 11), ADHD (Figure 12), and MCI (Figure 13) included the left dorsolateral prefrontal cortex and midline precuneus, two regions shared across all seven disease profiles.

Beyond these shared regions, PROPD included the left putamen, right putamen, right caudate nucleus, left thalamus, right supplementary motor area, right orbitofrontal cortex, left cuneus, and right middle occipital gyrus. The striatal, thalamic, and supplementary motor regions are broadly consistent with motor-related circuitry, while the orbitofrontal and occipital regions are associated with valuation and visual processing, respectively.

ADHD included the right inferior frontal gyrus, right middle frontal gyrus, left middle frontal gyrus, right medial superior frontal gyrus, left caudate nucleus, right caudate nucleus, left supramarginal gyrus, and right medial prefrontal cortex. This distribution encompasses frontal and striatal regions

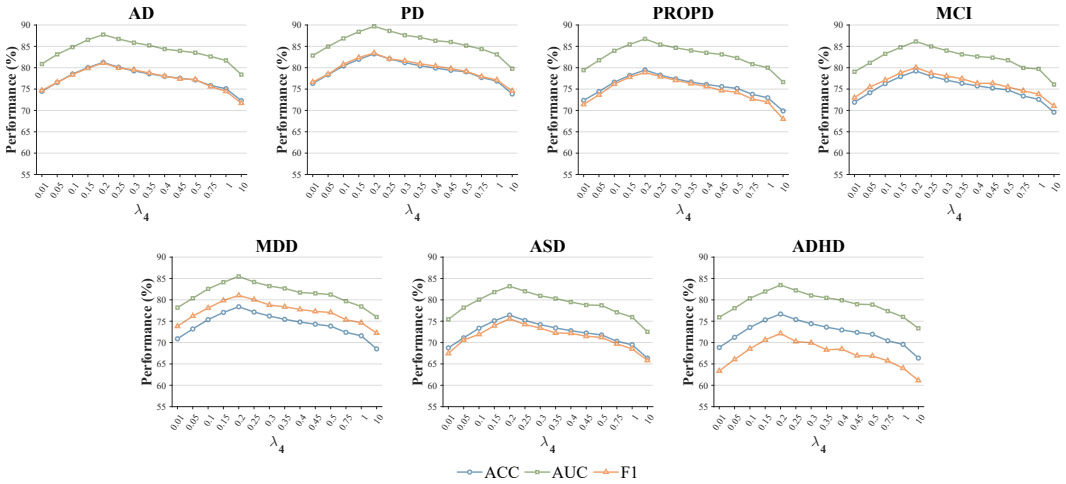


Figure 8: LogDet-regularization coefficient sweep across 7 diseases. ACC, AUC, and F1 are reported as %. The horizontal axis varies λ_4 in Equation 11.

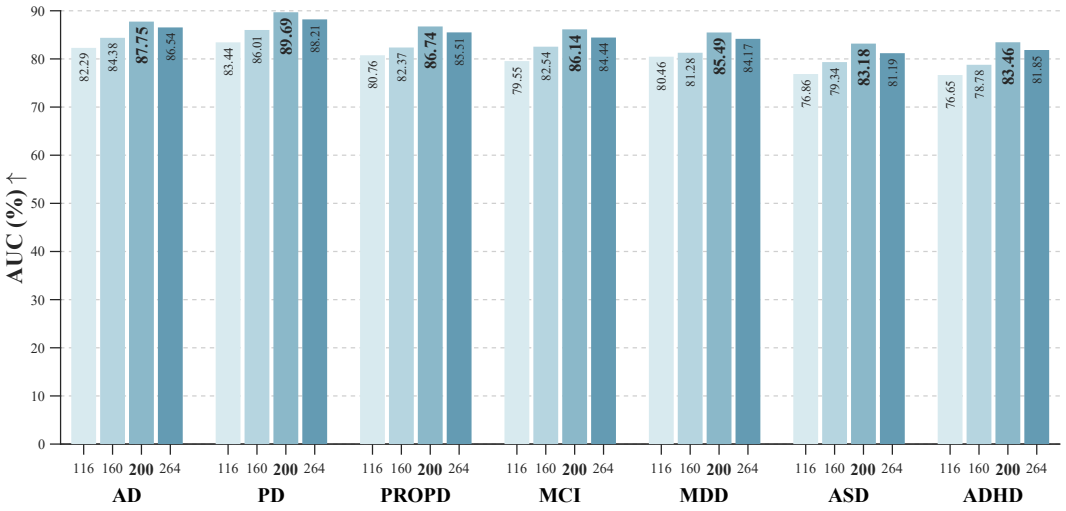


Figure 9: Influence of atlas choice on AUC (%) across 7 binary classification tasks. Within each task, bars are ordered by increasing ROI count: AAL116, Dosenbach160, CC200, and Power264. Bold values indicate the highest mean AUC within each task.

associated with executive control, together with an inferior parietal region associated with attentional processing.

MCI included the left hippocampus, right hippocampus, right posterior cingulate cortex, left angular gyrus, right angular gyrus, right medial prefrontal cortex, right dorsolateral prefrontal cortex, and right precuneus. These regions encompass memory-related hippocampal anatomy, default-mode-associated posterior cingulate, angular, medial prefrontal, and precuneus regions, and a dorsolateral prefrontal cognitive-control component.

Across the seven disease profiles, AD (Figure 14) and MCI included regions associated with memory and default-mode functions, whereas PD (Figure 15) and PROPD included striatal and motor-related regions, with additional occipital and orbitofrontal regions in PROPD. The MDD (Figure 17), ASD (Figure 16), and ADHD profiles included regions associated with affective and reward processing, temporal and visual-association functions, and executive and attentional control, respectively.

Together, these profiles highlight shared prefrontal and precuneus regions alongside differences in the anatomical distribution of model-highlighted regions. These observations provide anatomical

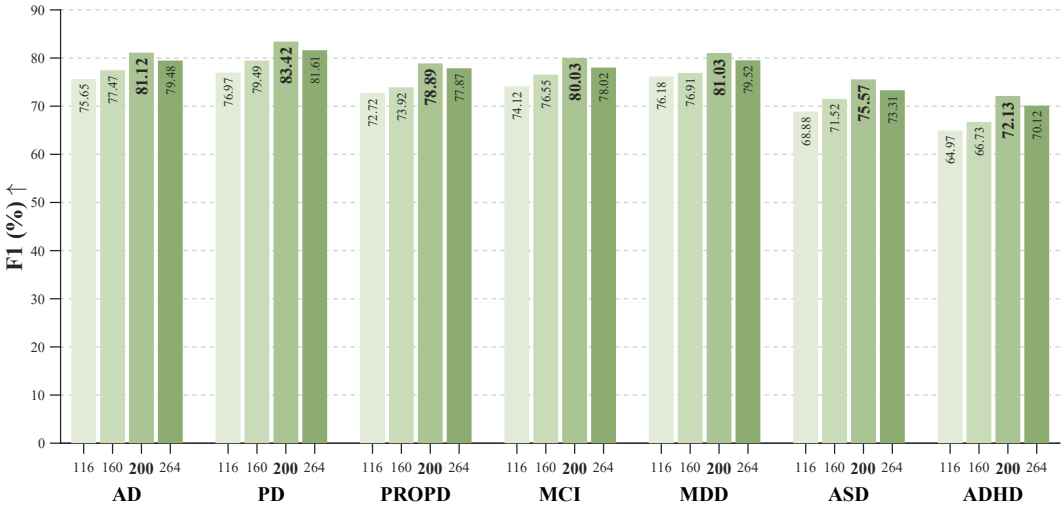


Figure 10: Influence of atlas choice on F1 (%) across 7 binary classification tasks. Within each task, bars are ordered by increasing ROI count: AAL116, Dosenbach160, CC200, and Power264. Bold values indicate the highest mean F1 within each task.

context for interpreting cross-disease commonalities and disease-associated patterns in the model's regional importance profiles.

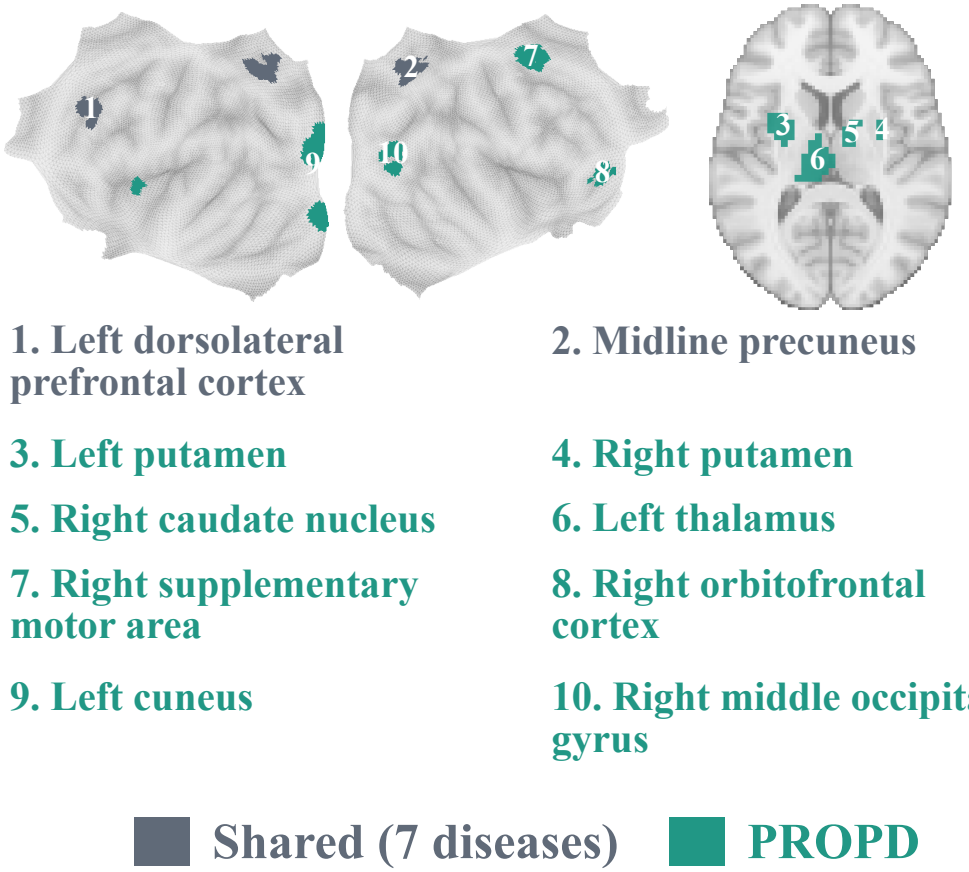


Figure 11: Anatomical visualization of ten CC200 regions for PROPD. Gray denotes regions shared across all seven disease sets; other colors indicate the remaining regions in each displayed set.

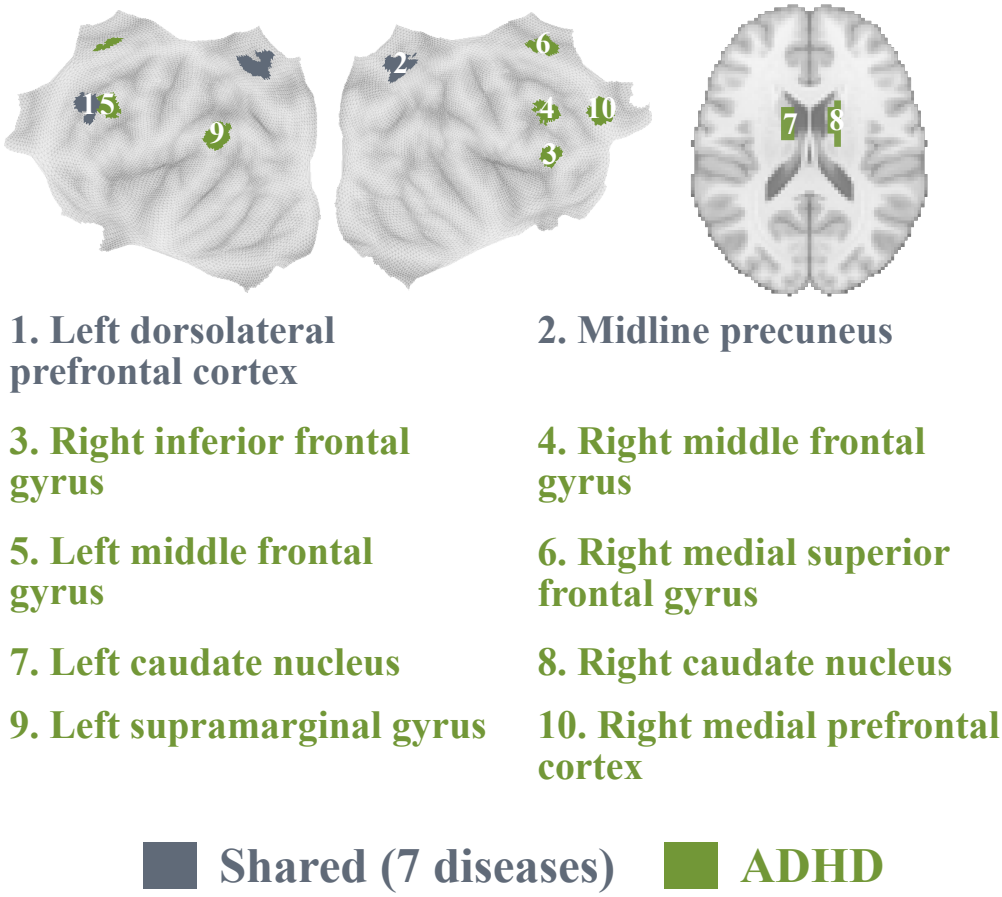


Figure 12: Anatomical visualization of ten CC200 regions for ADHD. Gray denotes regions shared across all seven disease sets; other colors indicate the remaining regions in each displayed set.



1. Left dorsolateral prefrontal cortex

2. Midline precuneus

3. Left hippocampus

4. Right hippocampus

5. Right posterior cingulate cortex

6. Left angular gyrus

7. Right angular gyrus

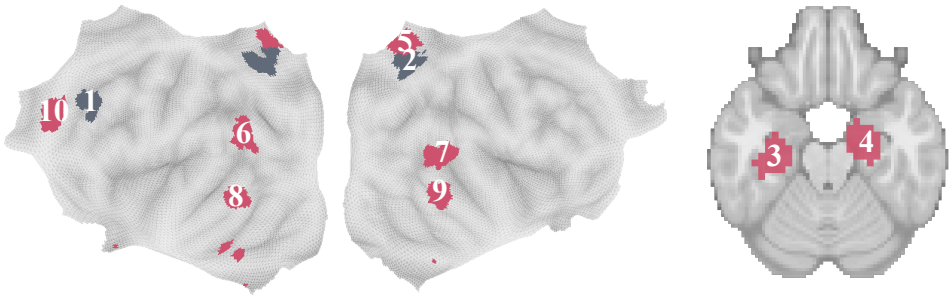
8. Right medial prefrontal cortex

9. Right dorsolateral prefrontal cortex

10. Right precuneus

Shared (7 diseases) MCI

Figure 13: Anatomical visualization of ten CC200 regions for MCI. Gray denotes regions shared across all seven disease sets; other colors indicate the remaining regions in each displayed set.



1. Left dorsolateral prefrontal cortex
2. Midline precuneus
3. Left hippocampus
4. Right hippocampus
5. Right posterior cingulate cortex
6. Left angular gyrus
7. Right angular gyrus
8. Left inferior temporal gyrus
9. Right middle temporal gyrus
10. Left medial prefrontal cortex

Shared (7 diseases) AD

Figure 14: Anatomical visualization of ten CC200 regions for AD. Gray denotes regions shared across all seven disease sets; other colors indicate the remaining regions in each displayed set.

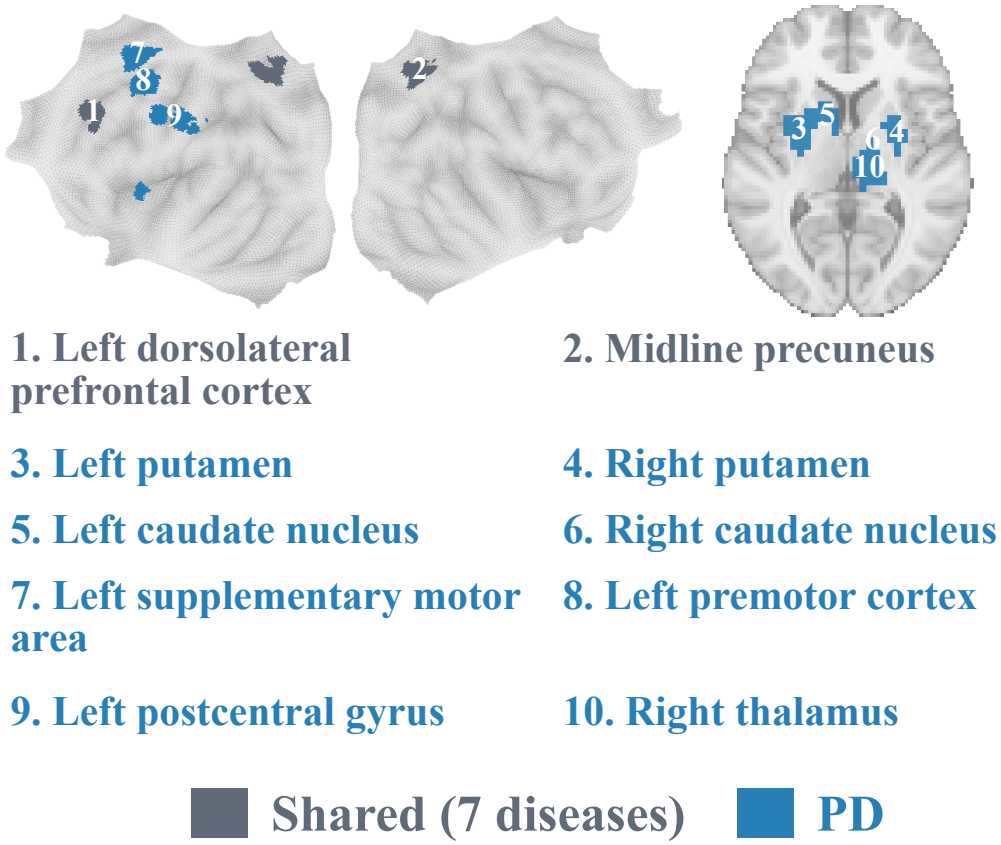


Figure 15: Anatomical visualization of ten CC200 regions for PD. Gray denotes regions shared across all seven disease sets; other colors indicate the remaining regions in each displayed set.

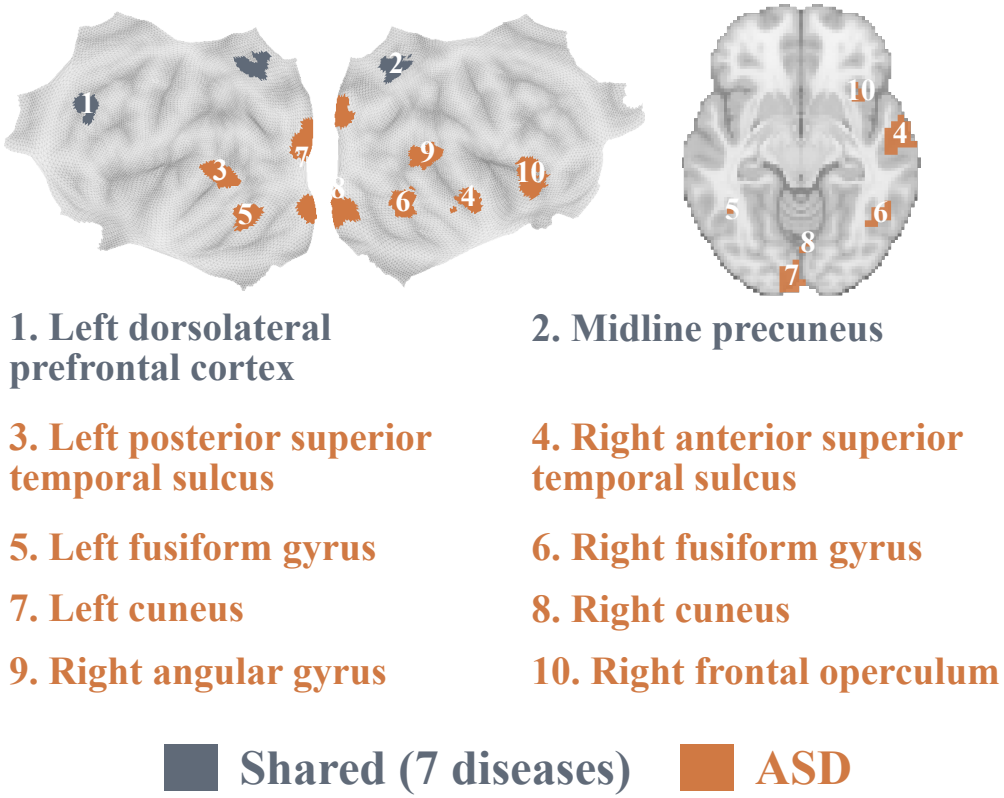
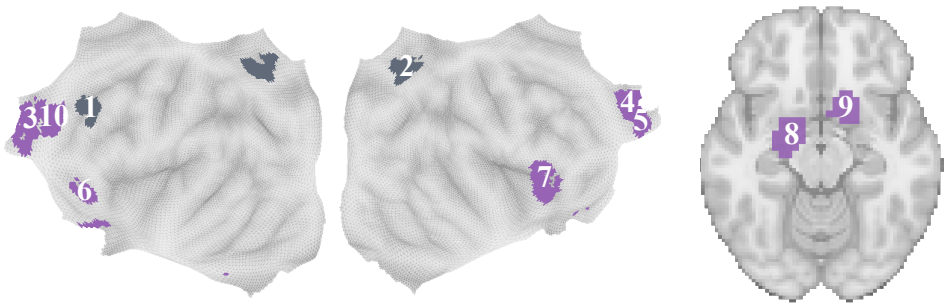


Figure 16: Anatomical visualization of ten CC200 regions for ASD. Gray denotes regions shared across all seven disease sets; other colors indicate the remaining regions in each displayed set.



1. Left dorsolateral prefrontal cortex
2. Midline precuneus
3. Left anterior cingulate cortex
4. Right anterior cingulate cortex
5. Right orbitomedial prefrontal cortex
6. Left orbitofrontal cortex
7. Right frontal operculum
8. Left hippocampus
9. Right nucleus accumbens
10. Left medial prefrontal cortex

Shared (7 diseases) MDD

Figure 17: Anatomical visualization of ten CC200 regions for MDD. Gray denotes regions shared across all seven disease sets; other colors indicate the remaining regions in each displayed set.