CONTEXT-CONDITIONED ANOMALOUS SUBGRAPH REPRESENTATION LEARNING FOR BRAIN DISEASE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing approaches for brain network analysis typically perform whole-graph classification or rely on dynamic functional connectivity sequences, overlooking the fact that healthy brains exhibit a stable contextual regularity where local subnetworks are consistently constrained by their specific functional contexts. We discover that this fundamental organizational principle is largely preserved in neuropsychiatric disorders, with disease-related alterations manifesting primarily as subtle perturbations to its concrete instantiation. Context-conditioned anomalous subgraph representation construction module is proposed to localize abnormal brain subnetwork. Role-aware brain subnetwork guidance encoder is introduced to encode multi-modal deviation directions, while anomaly-guided signed brain graph network is designed to propagate subgraph-level signals across the whole brain. Theoretically, we establish controlled false-positive bounds, abnormal detectability conditions, and region recovery guarantees via total-variation-regularized residual fields, showing that CCA-SRL provably distinguishes normal from abnormal brains when local subnetworks exhibit sufficiently strong conditional-representation shifts. Empirically, our model achieves new state-of-the-art performance on three public datasets—ADHD-200, OpenNeuro-ds002424, and ABIDE-II—outperforming 13 baselines.

1 INTRODUCTION

The human brain operates as a complex network of functionally interconnected regions, where the precise orchestration of neural activity across distributed circuits underpins cognition, behavior, and mental health. Resting-state functional magnetic resonance imaging (rs-fMRI) has emerged as a powerful non-invasive tool for mapping these functional connectivity patterns, enabling the construction of subject-specific brain graphs in which nodes represent anatomical regions of interest (ROIs) and edges encode statistical dependencies between their time series (Biswal et al., 2010; Power et al., 2011). A central challenge in computational neuroimaging is the identification of localized network disruptions associated with neurodevelopmental and psychiatric disorders—a task that holds promise for biomarker discovery, early diagnosis, and targeted intervention planning.

Traditional approaches to brain network analysis have largely focused on global graph-theoretic properties, such as clustering coefficients, characteristic path lengths, and modularity (Rubinov & Sporns, 2010; Bullmore & Sporns, 2009). While these summary statistics have revealed macroscopic differences between patient populations and healthy controls, they inevitably average over regional heterogeneity and may obscure the precise loci of pathology. More recent methods leverage deep learning architectures—including graph neural networks (GNNs) and graph transformers—to perform whole-brain classification (Kipf & Welling, 2017; Veličković et al., 2018; Cui et al., 2023; Li et al., 2021). However, these models are predominantly designed for subject-level diagnostic decisions and offer limited interpretability regarding where and how the brain deviates from normative organization. Furthermore, many existing approaches rely on dynamic functional connectivity sequences, which introduce substantial computational overhead and require carefully curated temporal segmentation (Preti et al., 2017; Lurie et al., 2020).

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.

Local subgraph properties have also been studied, including node-level centrality, community affiliation, and within- versus between-module connection strengths (Guimerà & Nunes Amaral, 2005; Bassett et al., 2011). These analyses have linked specific functional systems—the Default Mode Network (DMN), Dorsal Attention Network (DAN), and Fronto-Parietal Network (FPN)—to attention-deficit/hyperactivity disorder (ADHD) and related conditions (Castellanos & Proal, 2012; Cortese et al., 2012). However, such local metrics typically treat each subgraph in isolation and ignore the functional context in which it is embedded. Healthy brain organization is inherently relational: the internal connectivity of a cortical region is constrained by its interactions with surrounding areas, its boundary connectivity, and its role in broader functional hierarchies (Friston, 2011; Park & Friston, 2013). A subgraph that appears abnormal in isolation may simply reflect normal contextual variability, whereas a genuinely pathological alteration departs from the context-conditional regularity observed in healthy brains.

In healthy brains, the internal connectivity representation of a local subnetwork is predictable from its functional context—surrounding regions, boundary interactions, and control variables such as subgraph size, volume, and network membership. In abnormal brains, local mechanism shifts break this context-conditional regularity. We therefore propose context-conditioned anomalous subgraph representation learning (CCA-SRL), a framework for localizing abnormal subnetworks in static brain functional graphs without relying on dynamic sequences or multi-branch fusion.

The main contributions of this work are as follows:

- We show that functional brain networks exhibit a stable, identity-specific, and continuously graded subnetwork-level contextual constraint: local subnetworks preferentially match their specific functional contexts, with consistency systematically degrading as the context is disrupted. This organizational principle remains largely preserved in neuropsychiatric disorders, where disease-related alterations primarily perturb its concrete instantiation.
- We propose context-conditioned anomalous subgraph representation construction module to localize abnormal brain subnetwork, and role-aware brain subnetwork guidance encoder is introduced to encode multi-modal deviation directions, while anomaly-guided signed brain graph network is designed to propagate subgraph-level signals across the whole brain.
- We establish controlled false-positive bounds, abnormal detectability conditions, and region recovery guarantees via total-variation-regularized residual fields, showing that CCA-SRL provably distinguishes normal from abnormal brains when local subnetworks exhibit sufficiently strong conditional-representation shifts.
- CCA-SRL achieves new state-of-the-art performance on ADHD-200, OpenNeuro-ds002424, and ABIDE-II, outperforming 13 baselines spanning machine learning, graph neural networks, graph transformers, and hypergraph-based methods.

2 Related Work

2.1 Graph Neural Networks for Brain Disease Analysis

Graph neural networks (GNNs) have become a major tool for modeling functional connectomes because they can directly operate on ROI-level brain graphs while integrating regional features and inter-regional connectivity. BrainGNN (Li et al., 2021) introduced ROI-aware graph convolution and ROI-selection pooling to support both disorder classification and biomarker identification from fMRI. BrainGB (Cui et al., 2023) subsequently systematized the design space of GNN-based brain-network analysis, including node features, message passing, attention, and pooling. More recent work has addressed clinically relevant sources of heterogeneity and graph construction: GroupBNA (Peng et al., 2024) improves robustness across subject groups, BrainOOD (Xu et al., 2025) targets multi-site out-of-distribution generalization while extracting interpretable disorder-related subgraphs, and EBIGNN (Mamoon et al., 2026) jointly infers dynamic effective connectivity and graph representations for neurological-disorder recognition. Despite these advances, most GNN-based approaches remain optimized for subject-level discrimination or global graph representation, rather than explicitly testing whether the internal state of a local subnetwork is compatible with its expected functional context.

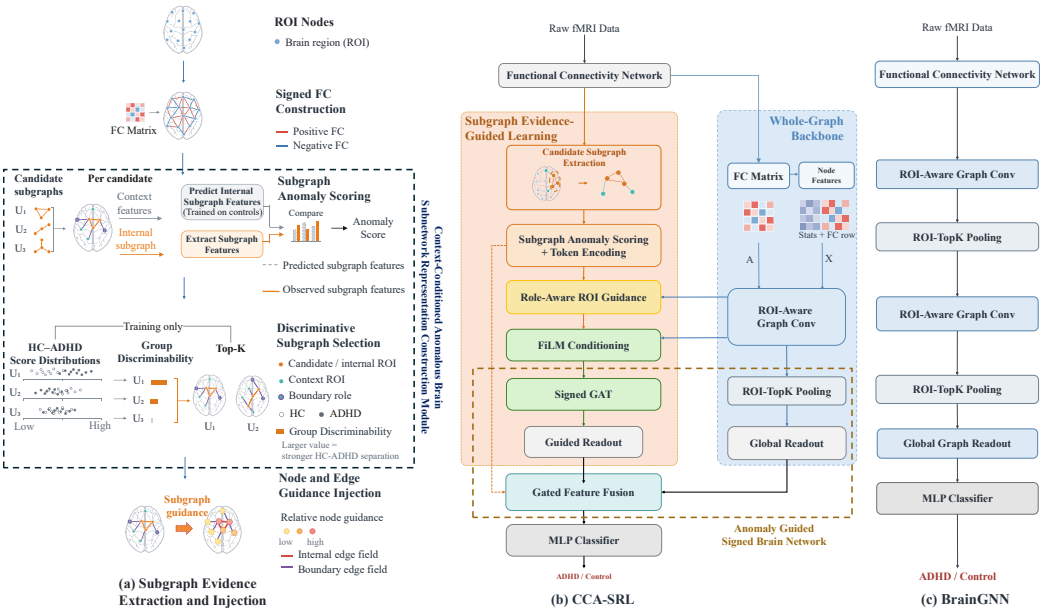


Figure 1: Comparison of the proposed CCA-SRL framework with BrainGNN. (a) Schematic illustration of subgraph evidence extraction and injection, from signed functional connectivity construction and candidate subgraph selection to contextual anomaly scoring and node- and edge-level guidance. (b) Architecture of CCA-SRL, consisting of a whole-graph backbone and a subgraph evidence-guided branch coupled through gated feature fusion. (c) Architecture of BrainGNN.

2.2 Graph Transformers for Brain Disease Analysis

Graph transformers extend brain-network modeling by using self-attention to capture long-range interactions between ROIs. Brain Network Transformer (BNT) (Kan et al., 2022) represents each ROI by its connection profile and combines self-attention with an orthonormal clustering readout for brain-network prediction, including ASD analysis on ABIDE. Multiview Brain Network Transformer (MBT) (Dong et al., 2024) further integrates common and individualized functional-connectivity views for ASD diagnosis, using diffusion-kernel attention to model indirect inter-regional interactions. BioBGT (Peng et al., 2025) incorporates brain-specific structural priors through node-importance encoding and functional-module-aware self-attention, and is evaluated on ADHD-200, ABIDE, and ADNI. These transformer-based models improve global dependency modeling and can highlight disease-relevant regions or connections, but they still formulate diagnosis primarily as graph-level representation learning. In contrast, CCA-SRL focuses on whether a local subnetwork violates the context-conditioned regularity learned from healthy brains and uses this localized inconsistency as explicit evidence for downstream diagnosis.

3 Methods

3.1 Motivation

Existing brain network methods usually treat diagnosis as whole-graph classification or rely on dynamic connectivity sequences. This works reasonably well at the subject level, but it misses a basic property of healthy brains: a local subnetwork is not defined by its internal edges alone. Its connectivity is constrained by the surrounding functional context, boundary interactions, and factors such as subgraph size and network membership. Abnormal organization should therefore be understood not only as a global graph-level shift, but also as a local breakdown of this context-conditioned regularity.

A subgraph can look unusual simply because its context varies. Pathology is different: it breaks the regularity that healthy brains follow. Thus, the key representation-learning question is not how to

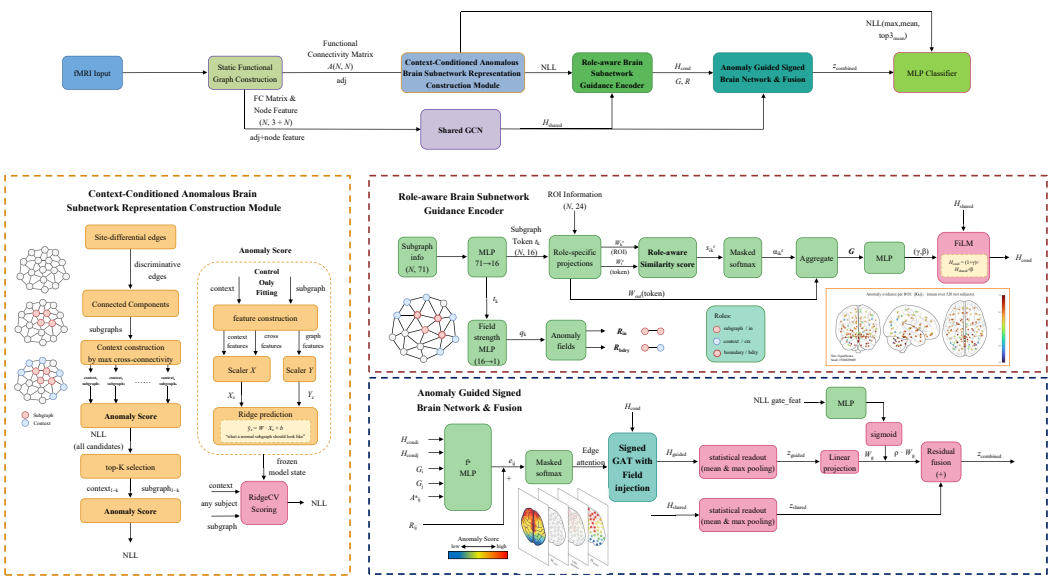


Figure 2: Overview of our proposed model.

build a better global graph embedding, but how to learn local subnetwork representations that are conditioned on context and can expose when that context fails to explain them. Existing GNNs, graph transformers, and local subgraph measures are not built for this—they either optimize graph-level representations or treat subgraphs in isolation.

Our empirical results in Section 5.5 and Appendices C-F support this picture. In healthy subjects, local subnetworks match their own context much better than mismatched ones (Table 5). This preference is identity-specific, not a byproduct of low-order graph statistics (Tables 6-7), and it fades gradually as the context is corrupted (Figure 6). It also survives split-half and site-level checks (Figure 7). In clinical groups, the same organization is largely preserved but systematically shifted (Table 10, Figure 8), and controlled violations leave spatially identifiable subnetwork signatures (Table 12, Figure 9). These findings point to a representation-learning formulation that works on static functional graphs and learns context-conditioned, anomaly-aware representations for localizing abnormal subnetworks.

3.2 Context-Conditioned Anomalous Brain Subnetwork Representation Construction Module

We mine candidate subnetworks and fits all selection and normality models using training subjects only. For each eligible site, disorder–control FC edges are ranked by absolute standardized effect $|d_{ij}|$, connected components from the highest-ranked edges and their cross-site union form the deduplicated candidate pool.

For candidate U_k and subject s, its context is the C outside ROIs with the largest total absolute connectivity to the candidate:

$$q_i^{(s,k)} = \sum_{j \in U_k} |A_{ij}^{(s)}|, \qquad C_k^{(s)} = \mathrm{TopC}_{i \notin U_k}\, q_i^{(s,k)}. \tag{1}$$

Thus, the candidate is fixed within a fold, whereas its functional context is subject-specific.

Healthy contextual consistency is modeled from 14 graph features describing subgraph size, density, signed/absolute FC statistics, edge-sign ratios, nodal strength/degree statistics, clustering, and global efficiency. The same feature family represents the context, while six additional features summarize subgraph–context interactions. A Ridge model fitted only on controls predicts the internal features from context, cross-interaction, normalized-size, and candidate-identity information. After

standardization, anomaly magnitude is measured by the Mahalanobis NLL

$$a_k = \frac{1}{2}\big(\tilde{\mathbf{r}}_k^\top \boldsymbol{\Sigma}^{-1}\tilde{\mathbf{r}}_k + \log\det\boldsymbol{\Sigma} + d_y \log 2\pi\big), \tag{2}$$

where $\tilde{\mathbf{r}}_k$ is the standardized residual and $\boldsymbol{\Sigma}$ is the shrunk covariance of control-training residuals. The prediction and original-space residual are retained to encode deviation direction in later stages.

The scorer is first applied to the full training candidate pool. Candidates are ranked by the absolute standardized separation between disorder and control NLL distributions, the top K are retained, and the normal scorer is re-fitted on this selected set. Final NLLs are standardized using training statistics.

3.3 Role-Aware Brain Subnetwork Guidance Encoder

The selected anomaly evidence is converted into learnable subgraph tokens. Each token combines observed and control-predicted subgraph features, their residual, context and cross-interaction features, standardized anomaly magnitude, and a learnable candidate identity embedding. A two-layer MLP with LayerNorm and ReLU maps this 71-dimensional input to the token space, while a shared GCN produces ROI embeddings $\mathbf{H}_{\text{shared}}$ for both the base and guided pathways.

To map subgraph evidence back to ROIs, each ROI is treated as *internal*, *context*, or *boundary* relative to a candidate. Role-specific attention combines the ROI embedding, subgraph token, anomaly magnitude, and a size correction:

$$s_{ik}^{(r)} = (\mathbf{W}_h^{(r)}\mathbf{h}_i)^\top(\mathbf{W}_t^{(r)}\mathbf{t}_k) + \gamma_r \cdot \text{softplus}(\tilde{a}_k) - \delta_r \cdot \log|U_k| + b_r. \tag{3}$$

Masked softmax over covered candidates yields ROI-level guidance $\mathbf{G}$. A Conditioner MLP then applies a learned FiLM scale and shift to the shared ROI embeddings, allowing anomaly evidence to selectively modulate their features.

To complement this node-wise conditioning, each token also produces a scalar strength q_k that defines internal and boundary anomaly fields:

$$R_{ij}^{\text{in}} = \sum_k q_k \mathbb{I}(i \in U_k)\mathbb{I}(j \in U_k), \tag{4}$$

$$R_{ij}^{\text{bdry}} = \sum_k q_k [\mathbb{I}(i \in U_k)\mathbb{I}(j \in C_k) + \mathbb{I}(j \in U_k)\mathbb{I}(i \in C_k)]. \tag{5}$$

These fields distinguish anomalous connectivity within a candidate from anomalous interactions across its context boundary.

3.4 Anomaly-Guided Signed Brain Graph Network

The conditioned evidence is propagated over positive and negative FC channels separately. Channel-specific attention depends on conditioned node embeddings, ROI guidance, FC strength, and the two anomaly fields:

$$\begin{aligned} e_{ij}^{\pm} = f_{\pm}\big([\mathbf{h}_{\text{cond},i}; \mathbf{h}_{\text{cond},j}; \mathbf{g}_i; \mathbf{g}_j; A_{ij}^{\pm}]\big) \\ + \phi_{\text{in}}^{\pm}(R_{ij}^{\text{in}}) + \phi_{\text{bdry}}^{\pm}(R_{ij}^{\text{bdry}}), \end{aligned} \tag{6}$$

where A^+ and A^- denote positive and negative FC magnitudes. Separate two-layer MLPs parameterize the two channels, and the edge fields modulate attention per head. Positive and negative messages are aggregated independently and combined with the conditioned representation through a residual LayerNorm update to obtain $\mathbf{H}_{\text{guided}}$.

Mean/max pooling of $\mathbf{H}_{\text{shared}}$ and $\mathbf{H}_{\text{guided}}$ gives base and anomaly-guided graph representations. To account for subject-wise variation in evidence reliability, each anomaly score is converted to an empirical upper-tail probability under controls and then to calibrated negative-log evidence. A sigmoid MLP maps six summary statistics—mean, top-three mean, standard deviation and maximum calibrated evidence, residual strength, and support fraction—to the reliability coefficient ρ.

Finally, the guided representation is added to the base representation after projection and reliability weighting:

$$\mathbf{z}_{\text{combined}} = \mathbf{z}_{\text{base}} + \rho\, \mathbf{W}_g(\mathbf{z}_{\text{guided}}). \tag{7}$$

When $\rho \to 0$, the model falls back to the anomaly-independent base pathway. The classifier receives this fused representation together with the mean, top-three mean, and maximum standardized NLL, and is optimized with focal loss.

4 Theoretical Analysis

We analyze the context-conditioned anomaly formulation underlying CCA-SRL, together with concise guarantees on calibration, detectability, and spatial recoverability.

Assumption 1 (Normal conditional consistency)**.** *For every normal subject $s \in \mathcal{S}_0$ and every candidate subgraph $U \in \mathcal{C}$,*

$$z_{\text{in}}^{(s)}(U) = \mu_\star\big(z_{\text{ctx}}^{(s)}(U), c(U)\big) + \varepsilon_U^{(s)}, \tag{8}$$

where $\varepsilon_U^{(s)}$ is zero-mean and $\Sigma_\star^{-1/2}\varepsilon_U^{(s)}$ is sub-Gaussian. In the ideal Gaussian case, $\Sigma_\star^{-1/2}\varepsilon_U^{(s)} \sim \mathcal{N}(0, I_d)$.

Assumption 2 (Local mechanism shift in abnormals)**.** *For an abnormal subject $s \in \mathcal{S}_1$, there exists at least one subgraph $U^\star \in \mathcal{C}$ such that*

$$z_{\text{in}}^{(s)}(U^\star) = \mu_\star\big(z_{\text{ctx}}^{(s)}(U^\star), c(U^\star)\big) + \Delta_{U^\star}^{(s)} + \varepsilon_{U^\star}^{(s)}, \tag{9}$$

with $\Delta_{U^\star}^{(s)} \neq 0$. The whitened shift strength is

$$\lambda_{U^\star}^{(s)} = \left\|\Sigma_\star^{-1/2}\Delta_{U^\star}^{(s)}\right\|_2^2. \tag{10}$$

Assumption 3 (Candidate coverage)**.** *For any true abnormal region $U^\star$, there exists $\widetilde{U} \in \mathcal{C}$ with $\mathrm{IoU}(\widetilde{U}, U^\star) \geq \rho_0 > 0$.*

Theorem 1 (False-positive control)**.** *Let $M = |\mathcal{C}|$. Under Assumption 1, the ideal Gaussian case, and ideal estimation ($\widehat{\mu} = \mu_\star, \widehat{\Sigma} = \Sigma_\star$), for any normal subject,*

$$\mathbb{P}\left(\max_{U \in \mathcal{C}} T_\star^{(s)}(U) \geq \tau_\alpha\right) \leq \alpha, \quad \tau_\alpha = d + 2\sqrt{d\log(M/\alpha)} + 2\log(M/\alpha). \tag{11}$$

Proof. For fixed U, $T_\star^{(s)}(U) = \|\xi_U\|_2^2$ with $\xi_U \sim \mathcal{N}(0, I_d)$. The Laurent–Massart bound gives

$$\mathbb{P}\left(T_\star^{(s)}(U) \geq \tau_\alpha\right) \leq \frac{\alpha}{M}. \tag{12}$$

Taking a union bound over all $U \in \mathcal{C}$ yields the result.

Theorem 2 (Abnormal detectability)**.** *Under Assumptions 1–2 and the ideal Gaussian case, given $\alpha, \beta \in (0, 1)$, let τ_α be as above and*

$$r_\beta = \sqrt{d + 2\sqrt{d\log(1/\beta)} + 2\log(1/\beta)}. \tag{13}$$

If

$$\sqrt{\lambda_{U^\star}^{(s)}} \geq \sqrt{\tau_\alpha} + r_\beta, \tag{14}$$

then

$$\mathbb{P}\Big(T_\star^{(s)}(U^\star) \geq \tau_\alpha\Big) \geq 1 - \beta. \tag{15}$$

Proof. Write $T_\star^{(s)}(U^\star) = \|v + \xi\|_2^2$ with $\|v\|_2^2 = \lambda_{U^\star}^{(s)}$. By the reverse triangle inequality,

$$\|v + \xi\|_2 \geq \|v\|_2 - \|\xi\|_2. \tag{16}$$

The chi-square tail bound gives $\mathbb{P}(\|\xi\|_2 > r_\beta) \leq \beta$. Hence, if $\|v\|_2 \geq \sqrt{\tau_\alpha} + r_\beta$, then $\|v + \xi\|_2 \geq \sqrt{\tau_\alpha}$ with probability at least $1 - \beta$, so $T_\star^{(s)}(U^\star) \geq \tau_\alpha$.

Table 1: Performance comparison (Mean ± Std) and p-values on OpenNeuro-ds002424 and ADHD-200.

Method	OpenNeuro-ds002424				ADHD-200			
	ACC (%)	p-value	AUC (%)	p-value	ACC (%)	p-value	AUC (%)	p-value
Machine Learning Methods								
Naive Bayes	61.01±2.50	3.43×10^{-7}	63.07±2.61	1.10×10^{-7}	57.59±3.90	2.47×10^{-4}	58.19±3.75	1.05×10^{-4}
Decision Tree	60.73±2.59	5.66×10^{-7}	59.83±2.90	3.15×10^{-7}	58.13±2.36	2.68×10^{-6}	56.15±2.22	3.15×10^{-5}
Graph Transformer Methods								
SAN (NeurIPS'21)	57.50±4.66	6.86×10^{-2}	61.20±2.46	4.25×10^{-3}	57.45±2.64	1.10×10^{-2}	47.17±2.96	1.52×10^{-1}
Polynormer (ICLR'24)	57.50±0.99	6.24×10^{-4}	54.74±3.85	9.07×10^{-2}	60.85±1.45	9.92×10^{-4}	47.02±2.41	1.22×10^{-1}
BioBGT (ICLR'25)	57.50±3.74	2.78×10^{-2}	53.17±3.69	1.84×10^{-1}	56.90±4.01	4.12×10^{-2}	55.74±3.85	5.85×10^{-2}
Graph Neural Networks								
BrainGNN (MIA'21)	66.82±3.82	4.68×10^{-3}	65.29±4.07	7.37×10^{-3}	61.30±0.80	1.50×10^{-4}	51.19±1.34	2.22×10^{-1}
BrainGB (TMI'23)	54.55±2.23	2.66×10^{-2}	53.39±2.77	9.19×10^{-2}	60.19±2.13	3.68×10^{-3}	57.51±3.75	4.04×10^{-2}
GroupBNA (ICLR'24)	55.00±1.64	1.32×10^{-2}	49.10±0.72	1.18×10^{-1}	38.10±7.32	6.69×10^{-2}	49.87±0.13	1.82×10^{-1}
BrainOOD (ICLR'25)	68.07±4.64	6.65×10^{-3}	66.42±4.27	6.90×10^{-3}	60.84±2.57	5.30×10^{-3}	55.38±0.89	1.86×10^{-3}
EBIGNN (JBHI'26)	57.50±7.14	1.66×10^{-1}	57.02±7.28	1.93×10^{-1}	61.30±1.09	2.45×10^{-4}	53.62±4.23	1.85×10^{-1}
Hypergraph-based Methods								
HGNN (AAAI'19)	65.65±4.79	2.73×10^{-6}	69.81±4.61	2.65×10^{-7}	51.13±2.59	2.01×10^{-1}	51.12±1.44	3.62×10^{-2}
I^{2}HBN (MICCAI'24)	59.82±1.94	6.41×10^{-8}	64.71±5.51	1.44×10^{-5}	62.23±1.63	2.00×10^{-9}	53.70±3.36	6.91×10^{-3}
T-HyperGNN (TNNLS'25)	64.05±4.22	2.33×10^{-6}	66.43±7.70	8.39×10^{-5}	56.79±7.64	2.04×10^{-2}	52.30±5.67	2.32×10^{-1}
Ours	**70.38±4.76**	5.06×10^{-3}	**74.52±4.51**	2.54×10^{-3}	**64.33±2.14**	1.37×10^{-3}	**63.04±5.40**	2.49×10^{-3}

Theorem 3 (Spatial recoverability from residual fields)**.** *Let r_i be a node-level residual obtained from local ego-subgraphs. Suppose the true abnormal region $U^\star$ satisfies*

$$\mathbb{E}[r_i] \geq \mu_1 \quad (i \in U^\star), \qquad \mathbb{E}[r_i] \leq \mu_0 \quad (i \notin U^\star), \tag{17}$$

with $\mu_1 > \mu_0$, and that $r_i - \mathbb{E}[r_i]$ is sub-Gaussian with parameter σ_r. For auxiliary spatial analysis, consider the TV-regularised objective

$$\widehat{m} = \arg\max_{m \in [0,1]^N} \left\{ \sum_i m_i r_i - \lambda m^\top L m - \eta(\mathbf{1}^\top m - s_0)^2 \right\}. \tag{18}$$

If

$$(\mu_1 - \mu_0)|U^\star| > 2\lambda \,\mathrm{cut}(U^\star, V \setminus U^\star) + C\sigma_r \sqrt{s_{\max} \log N}, \tag{19}$$

then $\widehat{m}$ overlaps with $U^\star$ with high probability.

Proof. For the true indicator $m^\star = \mathbf{1}_{U^\star}$, the expected data term is at least $\mu_1|U^\star|$, while for any alternative region S of similar size, it is at most $\mu_0|S|$. The difference is $(\mu_1 - \mu_0)|U^\star|$. The TV penalty adds $2\lambda \,\mathrm{cut}(U^\star, V \setminus U^\star)$ to the true region's cost. The maximum deviation due to random noise is bounded by $C\sigma_r\sqrt{s_{\max} \log N}$ via sub-Gaussian maximal inequalities. When the stated inequality holds, the true region has higher objective value than comparable non-target regions up to the stated noise bound, yielding high-probability overlap with $U^\star$.

Lemma 1 (Normal calibration)**.** *On normal subjects only, the maximum score over $\mathcal{C}$ stays below τ_α with probability at least $1 - \alpha$.*

Testable implication 1. For normal subgraphs, disrupting the matched context is expected to increase $T(U)$ relative to the true context.

Testable implication 2. Under local mechanism shifts, abnormal subjects are expected to exhibit elevated anomaly scores on affected subnetworks.

5 Experimental Results

5.1 Datasets and Experimental Setting

We evaluate CCA-SRL on three fMRI datasets: ADHD-200 (ADHD-200 Consortium, 2012), OpenNeuro-ds002424 (WMRC) (Booth et al., 2021), and ABIDE-II (Di Martino et al., 2017).

Table 2: Performance comparison and component ablation on ABIDE-II. Results are Mean±Std with one-sample t-test p-values; SUB, ENC, BAS, GUI, and FUS denote the subgraph anomaly module, guidance encoder, base representation, guided signed graph network, and calibrated fusion.

(a) Performance Comparison

Method	ACC	p-value	AUC	p-value
Machine Learning				
Naive Bayes	57.26±3.97	5.06×10^{-2}	58.43±3.74	2.98×10^{-2}
Decision Tree	49.82±0.81	7.33×10^{-1}	49.87±0.60	7.34×10^{-1}
Graph Transformer				
SAN	56.07±3.69	6.52×10^{-2}	56.30±2.33	1.84×10^{-2}
Polynormer	57.47±2.03	7.82×10^{-3}	59.35±2.01	3.99×10^{-3}
BioBGT	56.47±3.30	4.26×10^{-2}	55.39±3.24	6.35×10^{-2}
Graph Neural Networks				
BrainGNN	55.49±1.06	2.92×10^{-3}	54.95±1.77	1.68×10^{-2}
BrainGB	56.03±2.32	2.05×10^{-2}	61.57±0.61	6.20×10^{-5}
GroupBNA	56.62±3.78	3.94×10^{-2}	57.01±3.61	3.02×10^{-2}
BrainOOD	57.73±2.63	1.47×10^{-2}	57.59±2.95	2.10×10^{-2}
EBIGNN	55.92±2.06	1.56×10^{-2}	56.66±1.64	5.90×10^{-3}
Hypergraph-based Methods				
HGNN	57.49±2.86	2.00×10^{-2}	58.04±3.36	2.55×10^{-2}
I^2HBN	55.17±3.97	1.09×10^{-1}	57.54±6.05	1.20×10^{-1}
T-HyperGNN	54.00±4.35	2.09×10^{-1}	52.23±5.32	5.20×10^{-1}
Ours	**63.07±0.73**	7.52×10^{-5}	**64.19±1.18**	2.44×10^{-4}

(b) Component Ablation

Variant	ACC	p-value	AUC	p-value
w/o SUB	55.87±2.63	3.05×10^{-2}	54.87±4.77	1.75×10^{-1}
w/o ENC	57.26±2.53	1.57×10^{-2}	59.20±3.53	2.03×10^{-2}
w/o BAS	55.98±2.43	2.38×10^{-2}	58.87±3.73	2.60×10^{-2}
w/o GUI	54.71±2.90	6.74×10^{-2}	56.21±3.44	5.23×10^{-2}
SUB only	53.78±4.49	2.41×10^{-1}	53.79±3.68	1.72×10^{-1}
SUB+ENC	56.91±4.05	5.98×10^{-2}	57.67±3.11	2.34×10^{-2}
SUB+ENC+BAS	56.57±2.33	1.64×10^{-2}	59.99±1.09	5.36×10^{-4}
w/o FUS	57.49±2.45	1.32×10^{-2}	58.42±2.74	1.29×10^{-2}
Full	**63.07±0.73**	7.52×10^{-5}	**64.19±1.18**	2.44×10^{-4}

Table 3: Component ablation on ADHD-200 and OpenNeuro-ds002424. SUB, ENC, BAS, GUI, and FUS denote the subgraph branch, learned encoder, base readout, guided channel, and gated fusion; p-values are from one-sample t-tests.

SUB	ENC	BAS	GUI	FUS	ADHD-200		OpenNeuro-ds002424	
					ACC (%) / p-value	AUC (%) / p-value	ACC (%) / p-value	AUC (%) / p-value
×	✓	✓	✓	✓	$61.45 \pm 3.48 / 1.07 \times 10^{-2}$	$59.96 \pm 2.73 / 8.01 \times 10^{-3}$	$65.77 \pm 6.07 / 2.05 \times 10^{-2}$	$72.09 \pm 6.45 / 9.58 \times 10^{-3}$
✓	×	✓	✓	✓	$59.75 \pm 2.85 / 9.60 \times 10^{-3}$	$57.22 \pm 3.26 / 3.12 \times 10^{-2}$	$57.31 \pm 2.00 / 7.96 \times 10^{-3}$	$59.09 \pm 4.86 / 4.78 \times 10^{-2}$
✓	✓	×	✓	✓	$61.24 \pm 3.66 / 1.30 \times 10^{-2}$	$59.69 \pm 4.20 / 2.80 \times 10^{-2}$	$63.08 \pm 2.43 / 2.62 \times 10^{-3}$	$63.73 \pm 3.27 / 5.36 \times 10^{-3}$
✓	✓	✓	×	✓	$59.85 \pm 2.07 / 3.75 \times 10^{-3}$	$57.31 \pm 2.42 / 1.36 \times 10^{-2}$	$65.38 \pm 4.39 / 8.95 \times 10^{-3}$	$71.20 \pm 6.13 / 9.31 \times 10^{-3}$
✓	×	×	×	×	$61.34 \pm 0.32 / 9.66 \times 10^{-6}$	$49.80 \pm 3.54 / 9.27 \times 10^{-1}$	$52.12 \pm 1.26 / 6.22 \times 10^{-2}$	$52.73 \pm 3.09 / 2.23 \times 10^{-1}$
✓	✓	×	×	×	$58.89 \pm 0.85 / 3.65 \times 10^{-4}$	$54.99 \pm 4.84 / 1.73 \times 10^{-1}$	$57.31 \pm 5.06 / 8.76 \times 10^{-2}$	$53.06 \pm 4.71 / 3.43 \times 10^{-1}$
✓	✓	✓	×	×	$60.81 \pm 0.39 / 1.97 \times 10^{-5}$	$58.03 \pm 2.47 / 1.11 \times 10^{-2}$	$65.77 \pm 2.00 / 8.48 \times 10^{-4}$	$69.31 \pm 4.16 / 4.02 \times 10^{-3}$
✓	✓	✓	✓	×	$61.45 \pm 0.42 / 2.15 \times 10^{-5}$	$58.59 \pm 1.26 / 1.32 \times 10^{-3}$	$69.81 \pm 2.33 / 6.81 \times 10^{-4}$	$74.18 \pm 3.35 / 1.11 \times 10^{-3}$
✓	✓	✓	✓	✓	$64.33 \pm 2.14 / 1.37 \times 10^{-3}$	$63.04 \pm 5.40 / 2.49 \times 10^{-2}$	$70.38 \pm 4.76 / 5.06 \times 10^{-3}$	$74.52 \pm 4.51 / 2.54 \times 10^{-3}$

ADHD-200 contains 357 ADHD subjects and 582 controls after removing samples with missing labels. From ABIDE-II, we retain 400 ASD subjects and 461 controls after excluding samples with constant ROI parcellation or missing data. OpenNeuro contains 232 ADHD subjects and 314 healthy controls. Data are split into training, validation, and test sets (60/20/20), and performance is assessed with four-fold cross-validation. We report ACC and AUC as mean±std with one-sample t-test p-values. All experiments are run on an NVIDIA RTX 3060 Laptop GPU. Further implementation details are provided in Appendix J.

5.2 Overall Performance

We compare CCA-SRL against 13 baselines from four families: machine-learning methods (Naive Bayes (Mitchell, 1997), Decision Tree (Breiman et al., 1984)), graph transformers (SAN (Kreuzer et al., 2021), Polynormer (Deng et al., 2024), BioBGT (Peng et al., 2025)), graph neural networks (BrainGNN (Li et al., 2021), BrainGB (Cui et al., 2023), GroupBNA (Peng et al., 2024), BrainOOD (Xu et al., 2025), EBIGNN (Mamoon et al., 2026)), and hypergraph-based methods (HGNN (Feng et al., 2019), I^2HBN (Han et al., 2024), T-HyperGNN (Wang et al., 2025)).

Tables 1 and 2 show the overall results on all three datasets. CCA-SRL improves AUC over the best baseline by 4.71 points on OpenNeuro and 5.53 on ADHD-200, and improves ABIDE-II accuracy by 5.34 points, indicating complementary value beyond conventional whole-graph representations.

5.3 Ablation Studies

Ablations in Table 2 and 3 consistently favor the full architecture, though component impact varies by dataset. Removing `ENC` causes large losses on ADHD-200 and the largest OpenNeuro drop, supporting scalar-to-structured ROI guidance; removing `GUI` is most damaging on ABIDE-II, consistent with signed whole-brain propagation. `BAS` contributes across datasets, while removing `SUB` yields the largest ABIDE-II AUC drop, indicating complementary generic and anomaly-guided representations. `FUS` has a smaller effect on OpenNeuro but remains beneficial overall, consistent with reliability-dependent residual correction. Progressive variants recover performance as components are added, with the full model strongest.

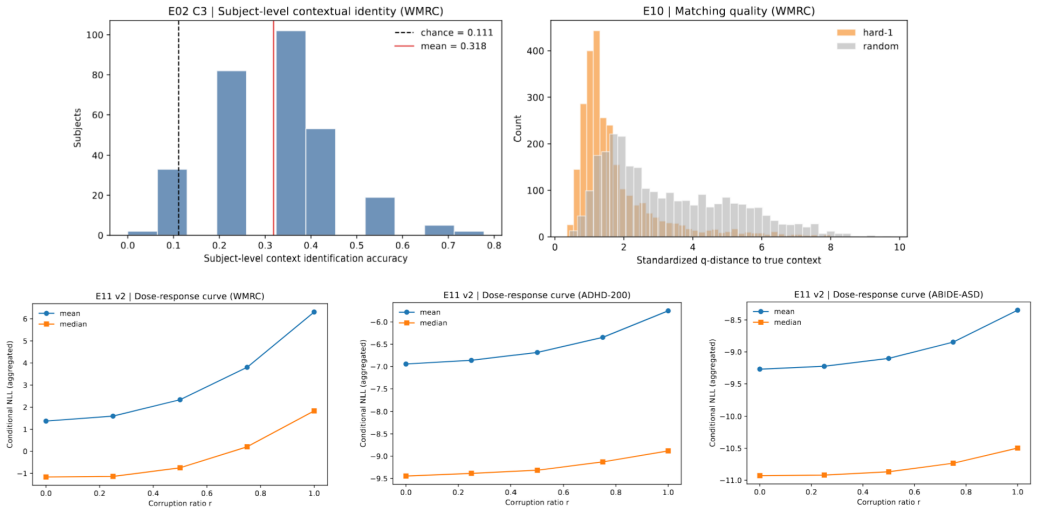


Figure 3: SCC validation. (a) Subject-level context identification. (b) Hard-negative matching after controlling low-order context statistics. (c) Conditional NLL under progressively stronger hard-matched context corruption. Complete statistics are reported in the appendix.

5.4 Analysis of SCC Existence, Identity Specificity, and Robustness

We validate SCC using subject–candidate pairs and conditional NLL, comparing true and wrong contexts under four outer folds. Identity specificity is tested by nine-way context identification (Fig. 3(a)) and hard-negative matching (Fig. 3(b)). Robustness is assessed by context corruption (Fig. 3(c)), leave-one-site-out transfer, and a healthy-control-fitted scorer on patients.

The true-context NLL is consistently lower, with positive fraction ≥ 0.997 and fold-cluster 95% CIs excluding zero (Appendix C.1, Table 5). Identification exceeds chance, and the advantage persists under hard-negative matching (Appendix C.2 and C.3, Tables 6 and 7). Corruption monotonically increases NLL (Fig. 3(c)), the effect is stable across halves, and all 28 held-out sites remain positive (Appendix D.1, D.2, and D.3, Tables 8 and 9). Patients still show the same organization (Appendix E, Tables 10 and 11). Thus, SCC is stable, identity-specific, graded, and largely preserved.

5.5 Analysis of Anomaly Localization, Finite-Context Sufficiency, and Candidate Attribution Control

We test identifiability with edge perturbations against random controls, reporting top-1/top-3 recovery. Finite-context sufficiency uses m_{95}, smallest context size reaching 95% of full-context R^2. Attribution controls use alternative contexts, within-subject swaps, candidate-mining baselines. Controlled perturbations yield top-1 recovery of 0.939–0.996 and top-3 of 0.967–0.996 (Appendix F.1, Table 12, Fig. 9). WMRC is compact, while ADHD-200 and ABIDE-II are more distributed yet retain high $Q(12)$ (Appendix F.2, Table 13, Fig. 10). Effects survive alternative contexts and within-subject swaps, and candidate mining alone is insufficient (Appendix G, Figs. 11, 13, and 15). CCA-SRL thus benefits from combining candidate discovery with context-conditioned anomaly modeling.

6 Conclusion

We identify a stable, identity-specific, graded subnetwork contextual constraint (SCC) in functional brain networks, largely preserved across clinical populations. We therefore propose CCA-SRL, modeling context-conditioned subnetwork inconsistency and integrating anomaly evidence via role-aware guidance, signed graph propagation, and reliability-calibrated fusion. It guarantees false-positive control, abnormal-shift detection, and region recovery, and achieves state-of-the-art performance on ADHD-200, OpenNeuro-ds002424, and ABIDE-II.

AI USE STATEMENT

Generative AI tools were used solely to assist with LaTeX typesetting, the formatting and organization of mathematical expressions, and limited language polishing. The authors reviewed these edits and take full responsibility for the scientific content and final manuscript.

REFERENCES

ADHD-200 Consortium. The ADHD-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, 6:62, 2012.

Danielle S. Bassett, Nicholas F. Wymbs, Mason A. Porter, Peter J. Mucha, Jean M. Carlson, and Scott T. Grafton. Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences*, 108(18):7641–7646, 2011.

Bharat B. Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Stephen M. Smith, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739, 2010.

James R. Booth, Grace E. Cooke, J. Gayda, R. Hammer, M. N. Lytle, M. A. Stein, and M. Tennekoon. Working memory and reward in children with and without attention deficit hyperactivity disorder (ADHD). *OpenNeuro*, 2021.

Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. *Classification and Regression Trees*. Wadsworth, 1984.

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

F. Xavier Castellanos and Erika Proal. Large-scale brain systems in ADHD: Beyond the prefrontal–striatal model. *Trends in Cognitive Sciences*, 16(1):17–26, 2012.

Samuele Cortese, Clare Kelly, Camille Chabernaud, Erika Proal, Adriana Di Martino, Michael P. Milham, and F. Xavier Castellanos. Toward systems neuroscience of ADHD: A meta-analysis of 55 fMRI studies. *American Journal of Psychiatry*, 169(10):1038–1055, 2012.

Hejie Cui, Wei Dai, Yanqiao Zhu, Xuan Kan, Antonio Aodong Chen Gu, Joshua Lukemire, Liang Zhan, Lifang He, Ying Guo, and Carl Yang. BrainGB: A benchmark for brain network analysis with graph neural networks. *IEEE Transactions on Medical Imaging*, 42(2):493–506, 2023.

Chenhui Deng, Zichao Yue, and Zhiru Zhang. Polynormer: Polynomial-expressive graph transformer in linear time. In *The 12th International Conference on Learning Representations*, pp. 18954–18979, 2024.

Adriana Di Martino et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. *Scientific Data*, 4(1):170010, 2017.

Qunxi Dong, Hongxin Cai, Zhigang Li, Jingyu Liu, and Bin Hu. A multiview brain network transformer fusing individualized information for autism spectrum disorder diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 28(8):4854–4865, 2024.

Y. Feng, H. You, et al. Hypergraph neural networks. In *AAAI*, pp. 3558–3565, 2019.

Karl J. Friston. Functional and effective connectivity: A review. *Brain Connectivity*, 1(1):13–36, 2011.

Roger Guimerà and Luís A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

X. Han, R. Xue, S. Du, and Y. Gao. Inter-intra high-order brain network for ASD diagnosis via functional MRIs. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, LNCS 15002, pp. 216–226, 2024.

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

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, pp. 1–14, 2017.

Devin Kreuzer, Dominique Beaini, William Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 21618–21629, 2021.

Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Pamela Ventola, and James S. Duncan. BrainGNN: Interpretable brain graph neural network for fMRI analysis. *Medical Image Analysis*, 74:102233, 2021.

Daniel J. Lurie, Daniel Kessler, Danielle S. Bassett, Richard F. Betzel, Michael Breakspear, Shella Kheilholz, et al. Questions and controversies in the study of time-varying functional connectivity in resting fMRI. *Network Neuroscience*, 4(1):30–69, 2020.

Saqib Mamoon, Zhengwang Xia, Wang Jin, Amani Alfakih, and Jianfeng Lu. Joint dynamic brain network estimation and graph representation learning for the recognition of neurological disorders. *IEEE Journal of Biomedical and Health Informatics*, 30(4):2986–2999, 2026.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

Hae-Jeong Park and Karl Friston. Structural and functional brain networks: From connections to cognition. *Science*, 342(6158):1238411, 2013.

C. Peng, M. Liu, C. Meng, S. Yu, and F. Xia. Adaptive brain network augmentation based on group-aware graph learning. In *Tiny Papers Track at ICLR*, 2024.

C. Peng, Y. Huang, Q. Dong, S. Yu, F. Xia, C. Zhang, and Y. Jin. Biologically plausible brain graph transformer. In *International Conference on Learning Representations (ICLR)*, pp. 75071–75097, 2025.

Jonathan D. Power, Alexander L. Cohen, Steven M. Nelson, Gagan S. Wig, Kelly Anne Barnes, Jessica A. Church, et al. Functional network organization of the human brain. *Neuron*, 72(4): 665–678, 2011.

Maria Giulia Preti, Thomas A. Bolton, and Dimitri Van De Ville. The dynamic functional connectome: State-of-the-art and perspectives. *NeuroImage*, 160:41–54, 2017.

Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, pp. 1–12, 2018.

Fuli Wang, Karelia Pena-Pena, Wei Qian, and Gonzalo R. Arce. T-HyperGNNs: Hypergraph neural networks via tensor representations. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):5044–5058, 2025.

Jiaxing Xu, Yongqiang Chen, Xia Dong, Mengcheng Lan, Tiancheng Huang, Qingtian Bian, James Cheng, and Yiping Ke. BrainOOD: Out-of-distribution generalizable brain network analysis. In *International Conference on Learning Representations*, pp. 6751–6765, 2025.

A Theoretical Analysis

A.1 Notation and Problem Setup

We are given a set of subjects $\mathcal{S}$. Each subject $s \in \mathcal{S}$ is associated with a global static brain functional graph

$$G^{(s)} = (V, E^{(s)}, A^{(s)}, X), \tag{20}$$

where $V = \{1, \dots, N\}$ is a fixed set of ROI nodes, $A^{(s)} \in \mathbb{R}^{N \times N}$ is the static functional connectivity matrix of the subject, and $X \in \mathbb{R}^{N \times d_x}$ contains node attributes such as ROI coordinates, functional network labels, or community labels. All subjects share the same ROI set, so subgraphs across subjects are aligned by ROI indices.

Our focus is not whole-graph classification but rather **local anomalous brain subnetwork detection**. For any ROI subset $U \subseteq V$, denote the induced subgraph by

$$G_U^{(s)} = G^{(s)}[U]. \tag{21}$$

The goal is to identify, for an abnormal subject $G^{(s)}$, a collection of abnormal subgraphs

$$\widehat{\mathcal{U}}^{(s)} = \{\widehat{U}_1^{(s)}, \dots, \widehat{U}_K^{(s)}\}, \tag{22}$$

and to output an anomaly score $T^{(s)}(U)$ for each candidate subgraph U. A higher score indicates that the subgraph is less consistent with the context-conditional regularity of normal brain graphs.

Definition 1 (Subgraph context)**.** *Given a subgraph $U \subseteq V$, define its k-hop external context as*

$$C_k(U) = \{v \in V \setminus U : \mathrm{dist}(v, U) \le k\}, \tag{23}$$

and its boundary edge set as

$$\partial U = \{(i, j) \in E : i \in U,\ j \notin U\}. \tag{24}$$

In the brain graph setting, U represents a candidate anomalous ROI subnetwork, $C_k(U)$ denotes the functional network context in which the subnetwork is embedded, and ∂U captures the interaction boundary between the subnetwork and outside regions.

Definition 2 (Internal representation and context representation)**.** *For subject s and subgraph U, define the internal representation of the subgraph as*

$$z_{\mathrm{in}}^{(s)}(U) = \Phi_{\mathrm{in}}\left(G^{(s)}[U]\right) \in \mathbb{R}^d, \tag{25}$$

and the context representation as

$$z_{\mathrm{ctx}}^{(s)}(U) = \Phi_{\mathrm{ctx}}\left(G^{(s)}[C_k(U)], \partial U\right) \in \mathbb{R}^q. \tag{26}$$

We also introduce control variables

$$c(U) = \left[|U|, \mathrm{vol}(U), |\partial U|, \mathrm{network}(U)\right], \tag{27}$$

which account for subgraph size, volume, boundary size, and functional network membership.

The fundamental idea of CCA-SRL is that local subgraphs in a normal brain are not generated in isolation, but are constrained by their functional context. Hence, for normal brains, there should exist a conditional mapping $\mu_\star$ such that

$$z_{\mathrm{in}}^{(s)}(U) \approx \mu_\star\left(z_{\mathrm{ctx}}^{(s)}(U), c(U)\right). \tag{28}$$

An abnormal brain, on the other hand, exhibits a local mechanism shift on a specific subgraph $U^\star$, so that the internal representation cannot be explained by the normal context.

A.2 Core Assumptions

Our theory rests on two assumptions: local conditional consistency in normal brains and local mechanism shift in abnormal brains. These assumptions provide the fundamental reason why CCA-SRL can discriminate between normal and abnormal subjects.

Assumption 4 (Local conditional consistency in normal brains)**.** *For a normal subject $s \in \mathcal{S}_0$ and any candidate subgraph $U \in \mathcal{C}$,*

$$z_{\mathrm{in}}^{(s)}(U) = \mu_\star\left(z_{\mathrm{ctx}}^{(s)}(U), c(U)\right) + \varepsilon_U^{(s)}, \tag{29}$$

where $\mu_\star$ is the conditional representation law of normal brains, and $\varepsilon_U^{(s)}$ is zero-mean noise. Moreover, the whitened noise

$$\xi_U^{(s)} = \Sigma_\star^{-1/2}\varepsilon_U^{(s)} \tag{30}$$

is sub-Gaussian, i.e., there exists a constant $\sigma > 0$ such that for any unit vector $a \in \mathbb{R}^d$,

$$\mathbb{P}\left(|a^\top \xi_U^{(s)}| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{31}$$

This assumption states that in a normal brain, the internal connectivity pattern of a local subnetwork is not arbitrary; it is consistent with the surrounding functional systems, neighbouring regions, and boundary connections. The control variables $c(U)$ prevent natural variations due to size, hubness, or network assignment from being misinterpreted as anomalies.

Assumption 5 (Local mechanism shift in abnormal brains)**.** *For an abnormal subject $s \in \mathcal{S}_1$, there exists at least one abnormal subgraph $U^\star \in \mathcal{C}$ such that*

$$z_{\mathrm{in}}^{(s)}(U^\star) = \mu_\star\left(z_{\mathrm{ctx}}^{(s)}(U^\star), c(U^\star)\right) + \Delta_{U^\star}^{(s)} + \varepsilon_{U^\star}^{(s)}, \tag{32}$$

where $\Delta_{U^\star}^{(s)} \neq 0$ represents the local mechanism shift. Define the whitened shift strength as

$$\lambda_{U^\star}^{(s)} = \left\|\Sigma_\star^{-1/2}\Delta_{U^\star}^{(s)}\right\|_2^2. \tag{33}$$

If $\lambda_{U^\star}^{(s)}$ is sufficiently large, the abnormal subgraph can be detected by the CCA-SRL score.

This assumption clarifies that abnormal brains need not exhibit a global shift; rather, local subnetwork mechanisms may change. For instance, in ADHD, the attention control system, the default-mode network, or fronto-parietal pathways may show local disruptions while their surrounding context remains largely normal. CCA-SRL is designed precisely to detect such "internal vs. contextual" inconsistencies.

Assumption 6 (Candidate subgraph coverage)**.** *The candidate set $\mathcal{C}$ covers potential abnormal regions. Specifically, for any abnormal subgraph $U^\star$, there exists $\widetilde{U} \in \mathcal{C}$ such that*

$$\mathrm{IoU}(\widetilde{U}, U^\star) = \frac{|\widetilde{U} \cap U^\star|}{|\widetilde{U} \cup U^\star|} \geq \rho_0, \tag{34}$$

where $\rho_0 \in (0, 1]$ is a coverage constant.

The coverage assumption separates the theory into two parts: the anomaly detection theory (which assumes candidate subgraphs are given) and the subgraph extraction theory (which shows how to recover high-anomaly regions from the residual field).

A.3 CCA-SRL Anomaly Score

We estimate the conditional mean function $\widehat{\mu}$ and residual covariance $\widehat{\Sigma}$ from normal subjects. For any subject s and candidate subgraph U, define the conditional residual

$$\widehat{e}_U^{(s)} = z_{\mathrm{in}}^{(s)}(U) - \widehat{\mu}\left(z_{\mathrm{ctx}}^{(s)}(U), c(U)\right). \tag{35}$$

The core anomaly score of CCA-SRL is the Mahalanobis-type statistic

$$T^{(s)}(U) = \left(\widehat{e}_U^{(s)}\right)^\top \widehat{\Sigma}^{-1}\widehat{e}_U^{(s)}. \tag{36}$$

Under a Gaussian conditional model, this score equals the negative log-likelihood (up to constants):

$$-\log p_{\widehat{\theta}}\left(z_{\text{in}}^{(s)}(U) \mid z_{\text{ctx}}^{(s)}(U), c(U)\right) = \frac{1}{2}T^{(s)}(U) + \frac{1}{2}\log|\widehat{\Sigma}| + C. \tag{37}$$

Thus, anomaly detection is equivalent to testing whether the internal representation of a subgraph deviates significantly from its conditional distribution given the context.

A.4 Concentration of Anomaly Scores for Normal Brains

To highlight the main theoretical line, we first consider the ideal case $\widehat{\mu} = \mu_\star$ and $\widehat{\Sigma} = \Sigma_\star$. Estimation errors will be treated in a later proposition.

Lemma 1 (Chi-square distribution for normal subgraph scores)**.** *If Assumption 1 holds and the whitened noise satisfies*

$$\Sigma_\star^{-1/2}\varepsilon_U^{(s)} \sim \mathcal{N}(0, I_d), \tag{38}$$

then for a normal subject $s \in \mathcal{S}_0$ and any candidate subgraph $U \in \mathcal{C}$,

$$T_\star^{(s)}(U) = \left(\varepsilon_U^{(s)}\right)^\top \Sigma_\star^{-1}\varepsilon_U^{(s)} \sim \chi_d^2. \tag{39}$$

Proof. Let $\xi_U^{(s)} = \Sigma_\star^{-1/2}\varepsilon_U^{(s)}$. Then $\xi_U^{(s)} \sim \mathcal{N}(0, I_d)$. Hence

$$T_\star^{(s)}(U) = \left\|\xi_U^{(s)}\right\|_2^2 = \sum_{j=1}^{d}(\xi_{U,j}^{(s)})^2, \tag{40}$$

which by definition of the chi-square distribution yields $T_\star^{(s)}(U) \sim \chi_d^2$. □

Lemma 2 (Concentration bound for normal subgraph scores)**.** *If $T \sim \chi_d^2$, then for any $x > 0$,*

$$\mathbb{P}\left(T \geq d + 2\sqrt{dx} + 2x\right) \leq e^{-x}. \tag{41}$$

Proof. This is a direct form of the Laurent–Massart chi-square concentration inequality. It shows that normal subgraph scores are concentrated around d with deviations controlled by $\sqrt{dx}$ and x. □

Theorem 4 (False-positive control for multiple subgraph scanning)**.** *Let the number of candidate subgraphs be $M = |\mathcal{C}|$. Under the ideal estimation setting, choose the threshold*

$$\tau_\alpha = d + 2\sqrt{d\log(M/\alpha)} + 2\log(M/\alpha), \tag{42}$$

Then for a normal subject $s \in \mathcal{S}_0$,

$$\mathbb{P}\left(\max_{U\in\mathcal{C}} T_\star^{(s)}(U) \geq \tau_\alpha\right) \leq \alpha. \tag{43}$$

Proof. For a fixed U, set $x = \log(M/\alpha)$. From equation 41,

$$\mathbb{P}\left(T_\star^{(s)}(U) \geq \tau_\alpha\right) \leq \frac{\alpha}{M}. \tag{44}$$

Applying the union bound over all candidate subgraphs gives

$$\mathbb{P}\left(\max_{U\in\mathcal{C}} T_\star^{(s)}(U) \geq \tau_\alpha\right) \leq \sum_{U\in\mathcal{C}}\mathbb{P}\left(T_\star^{(s)}(U) \geq \tau_\alpha\right) \leq M \cdot \frac{\alpha}{M} = \alpha. \tag{45}$$

□

Remark 1. *This theorem shows that even when scanning a large number of candidate subgraphs in a normal brain, the maximum CCA-SRL score can be controlled at the family-wise error rate by using the appropriate threshold. This is the fundamental statistical guarantee of the method for anomaly subgraph detection.*

A.5 Detectability in Abnormal Brains

Now consider an abnormal subgraph $U^\star$ in an abnormal subject. By Assumption 2, the whitened conditional residual is

$$\Sigma_\star^{-1/2}\left[z_{\text{in}}^{(s)}(U^\star) - \mu_\star(z_{\text{ctx}}^{(s)}(U^\star), c(U^\star))\right] = v_{U^\star}^{(s)} + \xi_{U^\star}^{(s)}, \tag{46}$$

where

$$v_{U^\star}^{(s)} = \Sigma_\star^{-1/2}\Delta_{U^\star}^{(s)}, \qquad \left\|v_{U^\star}^{(s)}\right\|_2^2 = \lambda_{U^\star}^{(s)}. \tag{47}$$

Thus the anomaly score is

$$T_\star^{(s)}(U^\star) = \left\|v_{U^\star}^{(s)} + \xi_{U^\star}^{(s)}\right\|_2^2. \tag{48}$$

Lemma 3 (Lower bound for anomaly score)**.** *For any $r > 0$, if*

$$\left\|v_{U^\star}^{(s)}\right\|_2 \geq \sqrt{\tau} + r, \tag{49}$$

then

$$\mathbb{P}\left(T_\star^{(s)}(U^\star) \geq \tau\right) \geq 1 - \mathbb{P}\left(\left\|\xi_{U^\star}^{(s)}\right\|_2 > r\right). \tag{50}$$

Proof. By the reverse triangle inequality, if $\left\|\xi_{U^\star}^{(s)}\right\|_2 \leq r$, then

$$\left\|v_{U^\star}^{(s)} + \xi_{U^\star}^{(s)}\right\|_2 \geq \left\|v_{U^\star}^{(s)}\right\|_2 - \left\|\xi_{U^\star}^{(s)}\right\|_2 \geq \sqrt{\tau}. \tag{51}$$

Hence $T_\star^{(s)}(U^\star) \geq \tau$. Taking probabilities gives the result. □

Theorem 5 (Detection condition for abnormal subgraphs)**.** *Let $\xi \sim \mathcal{N}(0, I_d)$. Given a false-positive level α and a missed-detection level β, let τ_α be as in equation 42 and set*

$$r_\beta = \sqrt{d + 2\sqrt{d\log(1/\beta)} + 2\log(1/\beta)}. \tag{52}$$

If the abnormal subgraph $U^\star$ satisfies

$$\sqrt{\lambda_{U^\star}^{(s)}} \geq \sqrt{\tau_\alpha} + r_\beta, \tag{53}$$

then

$$\mathbb{P}\left(T_\star^{(s)}(U^\star) \geq \tau_\alpha\right) \geq 1 - \beta. \tag{54}$$

Proof. From the chi-square concentration bound,

$$\mathbb{P}\left(\|\xi\|_2^2 > d + 2\sqrt{d\log(1/\beta)} + 2\log(1/\beta)\right) \leq \beta, \tag{55}$$

i.e., $\mathbb{P}(\|\xi\|_2 > r_\beta) \leq \beta$. Applying the previous lemma with $\tau = \tau_\alpha$ yields the claim. □

Remark 2. *Condition equation 53 provides the core condition under which CCA-SRL can distinguish abnormal from normal brains: the abnormal brain must exhibit a sufficiently large conditional-representation shift on some local subgraph. This shift is not merely a difference in connection strength, but a systematic deviation of the internal subgraph representation from its contextual expectation.*

A.6 Robustness under Estimation Error

In practice, $\mu_\star$ and $\Sigma_\star$ are unknown and must be estimated from a normal training set. We now analyse the effect of estimation errors on the anomaly score.

Assumption 7 (Conditional mean and covariance estimation errors)**.** *There exist constants $\epsilon_\mu, \epsilon_\Sigma \geq 0$ such that for all $U \in \mathcal{C}$,*

$$\|\widehat{\mu}(z_{\text{ctx}}(U), c(U)) - \mu_\star(z_{\text{ctx}}(U), c(U))\|_{\Sigma_\star^{-1}} \leq \epsilon_\mu, \tag{56}$$

and

$$\left\|\Sigma_\star^{1/2}\widehat{\Sigma}^{-1}\Sigma_\star^{1/2} - I_d\right\|_{\text{op}} \leq \epsilon_\Sigma < 1. \tag{57}$$

Proposition 1 (Perturbation of anomaly score due to estimation error)**.** *Under Assumption 4, for any candidate subgraph U, let $T_\star(U)$ be the ideal score and $\widehat{T}(U)$ the estimated score. Then there exists a constant $C > 0$ such that*

$$\left|\widehat{T}(U) - T_\star(U)\right| \le C\left[\epsilon_\Sigma T_\star(U) + \epsilon_\mu\sqrt{T_\star(U)} + \epsilon_\mu^2\right]. \tag{58}$$

Proof sketch. Let $y = \Sigma_\star^{-1/2} e_U$ be the ideal whitened residual, and let b be the whitened mean estimation error, so $\|b\|_2 \le \epsilon_\mu$. The covariance estimation error can be written as

$$\widehat{T}(U) = (y-b)^\top (I + E_\Sigma)(y-b), \tag{59}$$

with $\|E_\Sigma\|_{\mathrm{op}} \le \epsilon_\Sigma$. Expanding the quadratic form and applying the Cauchy–Schwarz inequality yields

$$|\widehat{T}(U) - \|y\|_2^2| \le C\left[\epsilon_\Sigma \|y\|_2^2 + \|b\|_2 \|y\|_2 + \|b\|_2^2\right], \tag{60}$$

which, since $T_\star(U) = \|y\|_2^2$, gives equation 58. □

Lemma 2 (Detection condition under estimation errors)**.** *If the abnormal shift satisfies*

$$\sqrt{\lambda_{U^\star}} \ge \sqrt{\tau_\alpha} + r_\beta + C'(\epsilon_\mu + \epsilon_\Sigma\sqrt{\tau_\alpha}), \tag{61}$$

then the abnormal subgraph can still be detected with probability at least $1-\beta$ in the presence of estimation errors.

A.7 Recovery of Abnormal Regions via Residual Field

The preceding theory shows that if candidate subgraphs are given, CCA-SRL scores can separate normal from abnormal. We now show how to recover abnormal regions from the representation residual field over the whole graph.

For each ROI node i, define a node-level context residual

$$r_i^{(s)} = \left\| \Sigma_i^{-1/2}\left[z_{\mathrm{in}}^{(s)}(B_i) - \widehat{\mu}_i(z_{\mathrm{ctx}}^{(s)}(B_i), c(B_i))\right]\right\|_2^2, \tag{62}$$

where B_i is a small ego subgraph centred at i (or a functional neighbourhood). This yields a residual field on the graph

$$r^{(s)} : V \to \mathbb{R}_+, \qquad i \mapsto r_i^{(s)}. \tag{63}$$

Abnormal subgraph extraction can be formulated as a graph total-variation regularised region search:

$$\widehat{m} = \arg\max_{m\in[0,1]^N}\left\{\sum_{i=1}^N m_i r_i^{(s)} - \lambda m^\top L^{(s)} m - \eta(\mathbf{1}^\top m - s_0)^2\right\}, \tag{64}$$

where $L^{(s)}$ is the graph Laplacian of the static brain graph, m_i denotes a soft selection weight indicating whether node i belongs to the abnormal subgraph, $m^\top Lm$ penalises unsmooth region boundaries, and s_0 controls the target subgraph size.

Assumption 8 (Signal separation in abnormal regions)**.** *There exists a true abnormal region $U^\star$ such that*

$$\mathbb{E}[r_i] \ge \mu_1, \quad i \in U^\star, \tag{65}$$

and

$$\mathbb{E}[r_i] \le \mu_0, \quad i \notin U^\star, \tag{66}$$

with $\mu_1 > \mu_0$. Furthermore, $r_i - \mathbb{E}[r_i]$ is sub-Gaussian with parameter σ_r.

Theorem 6 (Region recovery condition from residual field)**.** *Let $m^\star = \mathbb{I}_{U^\star}$. Consider all candidate regions of size at most $s_{\max}$. If the signal separation satisfies*

$$(\mu_1 - \mu_0)|U^\star| > 2\lambda\,\mathrm{cut}(U^\star, V\setminus U^\star) + C\sigma_r\sqrt{s_{\max}\log N}, \tag{67}$$

then the optimal solution of equation 64 will, with high probability, select a region that highly overlaps with $U^\star$.

Proof sketch. The objective consists of a data term and a smoothness term. For the true region $U^\star$, the expected data term is at least $\mu_1|U^\star|$; for any normal region S of similar size, the expected data term is at most $\mu_0|S|$. Their expected difference is governed by $(\mu_1 - \mu_0)|U^\star|$. The graph total-variation penalty for the true region is proportional to $\mathrm{cut}(U^\star, V \setminus U^\star)$. The maximum deviation due to random noise can be bounded by $C\sigma_r\sqrt{s_{\max}\log N}$ using sub-Gaussian maximal inequalities. When equation 67 holds, the true abnormal region attains a higher objective value than any normal region, so the optimal solution covers the true abnormal region. □

Remark 3. *This theorem shows that CCA-SRL's subgraph extraction is not merely thresholding high-score nodes, but balances residual intensity with region connectivity. If the abnormal region has high residuals, small boundary, and local coherence, the graph total-variation regularisation helps recover the abnormal brain subnetwork.*

A.8 Theoretical Implications and Summary

Combining the assumptions and theorems above, the detectability of CCA-SRL follows from the following logical chain:

1. **Normal brains satisfy local conditional consistency.** In a normal brain, the internal functional connectivity representation of any local ROI subgraph can be predicted from its external functional context, boundary interactions, and size variables. Thus, normal subgraphs have small conditional residuals and satisfy concentration.
2. **Abnormal brains exhibit local mechanism shifts.** In an abnormal brain, some functional subnetworks change their internal organisation, no longer obeying the normal context-conditional regularity. This change is quantified by the whitened shift λ_U.
3. **Anomaly scores provide statistical separation.** Normal subgraph scores approximately follow a χ^2_d distribution and can be controlled using equation 42; abnormal subgraphs satisfying equation 53 exceed the threshold with high probability.
4. **Abnormal regions can be recovered from the residual field.** If the abnormal shift forms a locally coherent region in ROI space and the signal separation exceeds boundary penalties and random noise, the graph total-variation search recovers the abnormal subgraph.

Therefore, CCA-SRL exploits a local brain-network mechanism: **a local subnetwork in a normal brain should be consistent with its functional context, while pathological subnetworks in abnormal brains break this consistency.**

Testable implications. The theory yields several directly testable implications for empirical validation.

Lemma 3 (Normal calibration)**.** *If trained and tested only on a normal control group, the top-K subgraph anomaly scores should follow a null distribution controlled by equation 42, and the proportion exceeding the threshold should not significantly exceed α.*

Lemma 4 (Context shuffling increases residuals)**.** *If the context $C_k(U)$ of a subgraph U is randomly permuted, destroying the true functional context, then in normal subjects the anomaly score*

$$T(U; \text{shuffled ctx}) \tag{68}$$

should be significantly larger than the score with the true context

$$T(U; \text{true ctx}). \tag{69}$$

This verifies the core premise of Assumption 1 that the internal representation depends on context.

Lemma 5 (Abnormal group top-K scores are elevated)**.** *If abnormal brains satisfy the local mechanism shift assumption, then the distribution of top-K subgraph scores in the abnormal group should be significantly higher than that in the normal group, and the abnormal subgraphs should be enriched in known relevant functional systems.*

Together, these results show that a stable context-conditional law in normal brains, combined with a sufficiently strong local representation shift in abnormal brains, yields both statistical separation and recoverable subnetwork-level anomaly structure.

Table 4: Summary of the SCC validation protocol.

Setting	Protocol
Outer evaluation	4-fold cross-validation
Inference unit	Subject–candidate pair
Candidate/context estimation	Outer-training partition only
Normality model	Outer-training controls only, with cross-fitting
Outer test fold	Scored once after all fitting steps
Primary interval estimate	Fold-cluster bootstrap 95% CI
Multi-seed reporting	Direction consistency and effect ranges
Exact fold-level check	Two-sided 4-fold sign-flip test

B SCC Validation Protocol and Statistical Inference

This section summarizes the evaluation protocol used for the empirical validation of the Subnetwork Contextual Constraint (SCC). The purpose of these analyses is to test whether a candidate subnetwork is more compatible with its own functional context than with appropriately mismatched contexts, and whether this relationship is stable under controlled perturbations and data resampling. The analyses reported in Appendices C–F follow the same leakage-controlled outer cross-validation scheme.

SCC scoring and evaluation units. For a candidate subnetwork U, let Y_U denote its internal state and X_U its functional context. Conditional consistency is quantified by

$$\mathrm{NLL}(Y_U \mid X_U) = -\log p(Y_U \mid X_U), \tag{70}$$

and the preference for the true context over a mismatched context is measured by

$$\Delta_{\mathrm{NLL}} = \mathrm{NLL}(Y_U \mid X_U^{\mathrm{wrong}}) - \mathrm{NLL}(Y_U \mid X_U^{\mathrm{true}}). \tag{71}$$

A positive Δ_{NLL} therefore indicates that the internal state of the subnetwork is more compatible with its true functional context.

The SCC validation experiments use three resting-state fMRI cohorts: WMRC ($n = 298$ controls), ADHD-200 ($n = 582$ controls for SCC scoring), and ABIDE-II ($n = 461$ controls). The basic inference unit is a subject–candidate pair, yielding 2682 units in WMRC, 5093–5238 units in ADHD-200 across split seeds, and 4149 units in ABIDE-II. The archived SCC experiments retain their prespecified candidate settings; in particular, the context-identity experiment evaluates nine candidate alternatives, giving a chance identification rate of $1/9$. These validation settings are distinct from the final classifier hyperparameters reported in Appendix H.

Leakage-controlled cross-validation. All SCC validation analyses use four outer folds. Candidate construction and any statistics required to define candidate contexts are estimated from the outer-training partition only. The normality model is fitted exclusively on training controls, with cross-fitting among those controls so that no subject is scored by a model that used that subject during fitting. The held-out outer fold is scored only after candidate construction, context definition, and normality-model estimation have been completed. Test labels are never used for candidate construction or normality-model fitting.

Statistical inference. Because all subjects in an outer fold share the same fitted model, confidence intervals are computed with a fold-cluster bootstrap in which complete folds are resampled with replacement. For experiments repeated over multiple data-splitting seeds, we report the consistency of the effect direction and the range of effect estimates rather than pooling p-values across seeds. The prespecified seeds are $\{0, 1, 2, 42\}$ for WMRC and $\{0, 1, 42\}$ for ADHD-200 and ABIDE-II when the corresponding experiment is available.

For fold-level directional statistics we additionally use an exact two-sided sign-flip test over the $2^4 = 16$ possible sign assignments of the four outer-fold effects. The minimum attainable two-sided value is therefore $p = 0.125$. We treat this value as a consequence of the four-fold design rather than as conventional significance evidence; substantive conclusions rely on fold-cluster confidence intervals, consistency across folds, and replication across datasets or split seeds.

Table 5: True-context consistency. Δ_{NLL} is the wrong-subject minus true-context conditional NLL gap. Positive values indicate preference for the true functional context.

Dataset	n	Mean Δ_{NLL}	$P(\Delta > 0)$	Cohen's d	Fold-cluster 95% CI	Positive folds
WMRC	298	3.369	0.997	1.022	[1.680, 5.080]	4/4
ADHD-200	582	145.271	1.000	0.784	[113.638, 176.903]	4/4
ABIDE-II	461	5627.303	1.000	0.278	[93.775, 11184.891]	4/4

Table 6: Context-identity analysis. Subject-level identification is evaluated among nine candidate contexts (chance $= 1/9$).

Dataset	Candidate dominance	ID accuracy	Chance	Excess accuracy	Excess 95% CI	Margin 95% CI
WMRC	1.000	0.318	0.111	0.207	[0.185, 0.229]	[-0.310, 3.065]
ADHD-200	1.000	0.755	0.111	0.644	[0.634, 0.654]	[7.452, 13.783]
ABIDE-II	1.000	0.727	0.111	0.616	[0.598, 0.635]	[7.018, 134.056]

C Validation of SCC Existence and Context Identity

The first set of analyses evaluates the two defining properties of SCC: the internal state of a subnetwork should be more compatible with its true functional context than with a mismatched context, and this compatibility should depend on the identity of the context rather than on trivial low-order graph statistics.

C.1 Context Consistency

For each held-out subject–candidate pair, we compare the conditional NLL obtained with the true functional context against a wrong-subject context. All three datasets show a positive wrong-minus-true NLL gap in every outer fold. The positive fraction is at least 0.997, and the fold-cluster 95% confidence interval excludes zero in each dataset (Table 5). Because absolute NLL scales differ substantially across datasets, raw gaps are interpreted only within a dataset; the direction of the gap and standardized effect sizes provide the cross-dataset evidence.

These results establish the basic existence of SCC: a local subnetwork is systematically more consistent with its own functional context than with an unrelated context.

C.2 Context Identity

We next evaluate whether contextual compatibility is identity-specific. Each subject is scored against nine candidate contexts, and the true context is treated as the correct identification target. Candidate-level diagonal dominance is 1.0 in all three datasets, indicating that every candidate is best matched to its own context. At the subject level, identification accuracy substantially exceeds the $1/9$ chance level, with excess accuracies of 0.207 in WMRC, 0.644 in ADHD-200, and 0.616 in ABIDE-II (Table 6). The true-versus-mismatch margin is also positive in ADHD-200 and ABIDE-II; the corresponding WMRC margin is weaker and is not used as primary evidence.

C.3 Hard-Negative Context Matching

To test whether the context-identity effect can be explained by simple graph statistics, each true context is matched to alternative contexts with similar values of eleven low-order features: context size, mean absolute and signed coupling, positive and negative edge fractions, context strength, density, internal absolute FC, edge variance, and the mean and spread of degree. The standardization used for matching is fitted on outer-training controls only. Across all dataset–seed combinations, hard negatives are substantially closer to the true context in this feature space than randomly sampled contexts.

Despite this matching, the true context remains preferred (Table 7). For the nearest hard negative, the median hard-minus-true NLL gap is 0.194–0.239 in WMRC, 2.335–2.524 in ADHD-200, and 2.158–2.348 in ABIDE-II. Averaging the three nearest hard negatives increases the corresponding

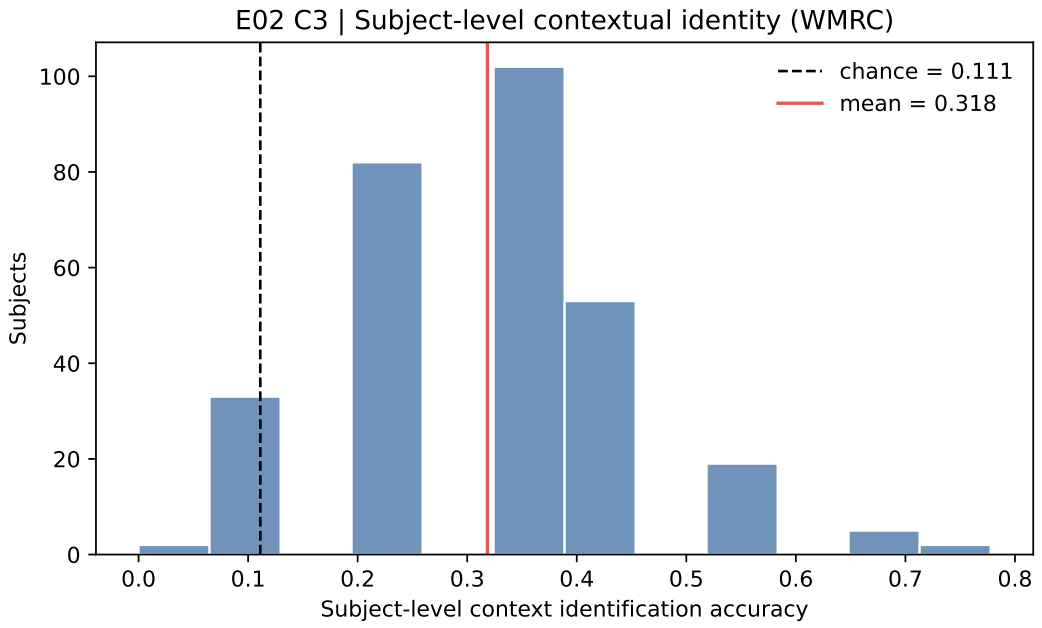


Figure 4: Subject-level context identification in WMRC. Excess accuracy over the $1/9$ chance level is shown with the fold-cluster 95% confidence interval.

Table 7: Hard-negative context matching. ΔNLL is the matched-context minus true-context NLL gap. "Nearest" uses the closest matched context and "Top-3" averages the three closest matched contexts.

Dataset	Seed	Matched set	Median ΔNLL	95% CI	True-context win rate	Cohen's d	Context ID acc.
WMRC	42	Nearest	0.239	[0.221, 0.257]	0.616	0.282	0.406
WMRC	42	Top-3	0.479	[0.369, 0.641]	0.752	0.392	0.406
WMRC	0	Nearest	0.194	[0.161, 0.226]	0.599	0.281	0.394
WMRC	0	Top-3	0.379	[0.303, 0.469]	0.715	0.415	0.394
ADHD-200	42	Nearest	2.524	[2.115, 3.099]	0.872	0.243	0.797
ADHD-200	42	Top-3	5.786	[4.805, 6.657]	0.965	0.327	0.797
ADHD-200	0	Nearest	2.335	[1.924, 2.787]	0.872	0.282	0.799
ADHD-200	0	Top-3	4.890	[4.019, 6.038]	0.956	0.339	0.799
ABIDE-II	42	Nearest	2.348	[1.799, 2.986]	0.859	0.150	0.775
ABIDE-II	42	Top-3	5.479	[3.812, 6.928]	0.952	0.162	0.775
ABIDE-II	0	Nearest	2.158	[1.547, 2.900]	0.835	0.138	0.745
ABIDE-II	0	Top-3	5.608	[3.844, 8.067]	0.941	0.176	0.745

gaps. Context identification under the matched setting also remains above its 0.25 chance level. Thus, SCC cannot be reduced to context size, mean connectivity strength, or the other matched low-order properties.

D Gradedness, Temporal Stability, and Cross-Site Generalization

After establishing context specificity, we evaluate whether SCC behaves as a graded organizational constraint and whether it remains detectable under independent temporal segments and held-out acquisition sites.

D.1 Context-Corruption Dose Response

The true context is progressively corrupted by replacing a fraction $r \in \{0, 0.25, 0.5, 0.75, 1.0\}$ of its nodes. Hard-matched replacement is the primary condition, while random replacement serves as

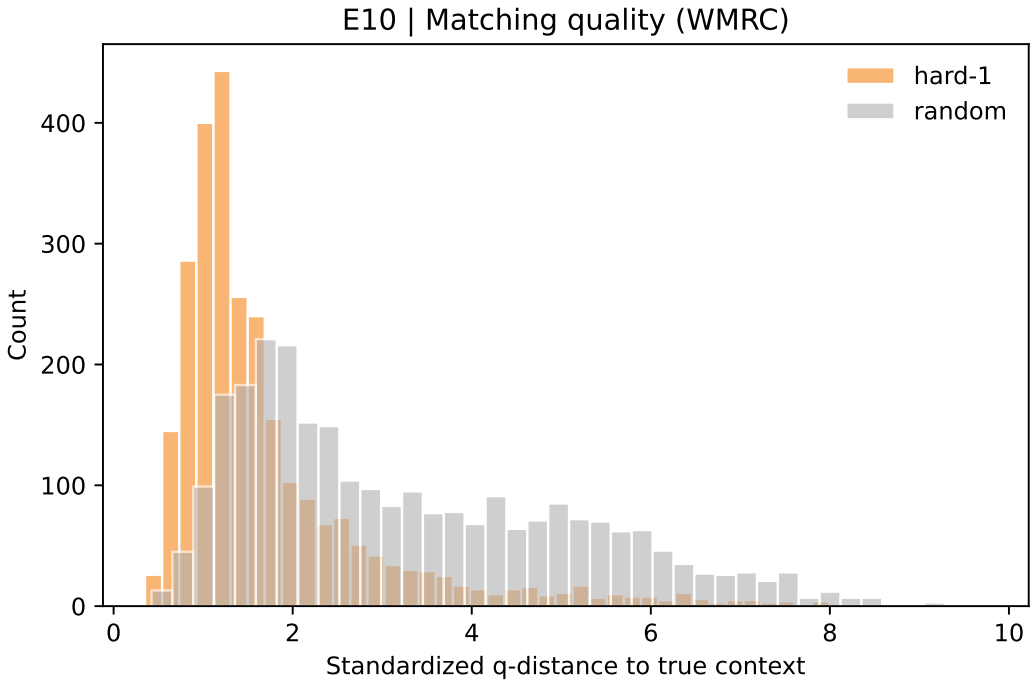


Figure 5: Quality of hard-negative matching in WMRC. Matched alternatives are substantially closer to the true context than random contexts across the eleven low-order features.

Table 8: Context-corruption dose response. The primary condition replaces increasing fractions of the true context with hard-matched alternative nodes.

Dataset	Seed	Replacement	Units	Spearman ρ	Positive-ρ fraction	Monotonic fraction	NLL slope	95% CI
WMRC	42	Hard-matched	2682	1.000	0.937	0.565	4.838	[4.511, 5.166]
WMRC	42	Random	2682	1.000	1.000	0.802	29.453	[27.203, 31.086]
WMRC	0	Hard-matched	2682	1.000	0.918	0.566	4.535	[4.225, 4.785]
ADHD-200	42	Hard-matched	5238	0.900	0.678	0.439	1.160	[0.914, 1.406]
ADHD-200	42	Random	5238	1.000	0.956	0.703	43.302	[30.038, 60.700]
ADHD-200	0	Hard-matched	5093	0.900	0.674	0.408	1.064	[0.836, 1.292]
ABIDE-II	42	Hard-matched	4149	0.700	0.650	0.411	0.889	[0.752, 1.037]
ABIDE-II	42	Random	4149	1.000	0.966	0.817	3177.670	[26.896, 9277.873]
ABIDE-II	0	Hard-matched	4149	0.800	0.642	0.425	0.628	[0.557, 0.728]

a stronger perturbation reference. For each nonzero corruption rate, 50 replacements are generated and aggregated at the subject–candidate level before inference.

A genuine graded constraint predicts increasing conditional NLL as more of the true context is replaced. This pattern is observed in all three datasets (Table 8). Under hard-matched corruption, the median Spearman correlation between corruption rate and NLL is 1.00 in WMRC, 0.90 in ADHD-200, and 0.70–0.80 in ABIDE-II. The NLL slope is positive in every outer fold and its fold-cluster confidence interval excludes zero for all reported hard-matched conditions. The result therefore supports a continuous, rather than purely categorical, dependence on contextual integrity.

D.2 Temporal Stability

Temporal stability is evaluated in ADHD-200, for which aligned raw time series are available. Each subject's time series is divided into two independent halves, and SCC estimated from one half is evaluated on the other. Across 582 subjects, the cross-half wrong-minus-true NLL gap remains strongly positive (median 11.161, 95% CI [9.804, 12.498]), with a positive fraction of 0.971, Cohen's $d = 0.556$, and cross-half $R^2 = 0.968$. The cross-half effect retains 73.1% of the same-half

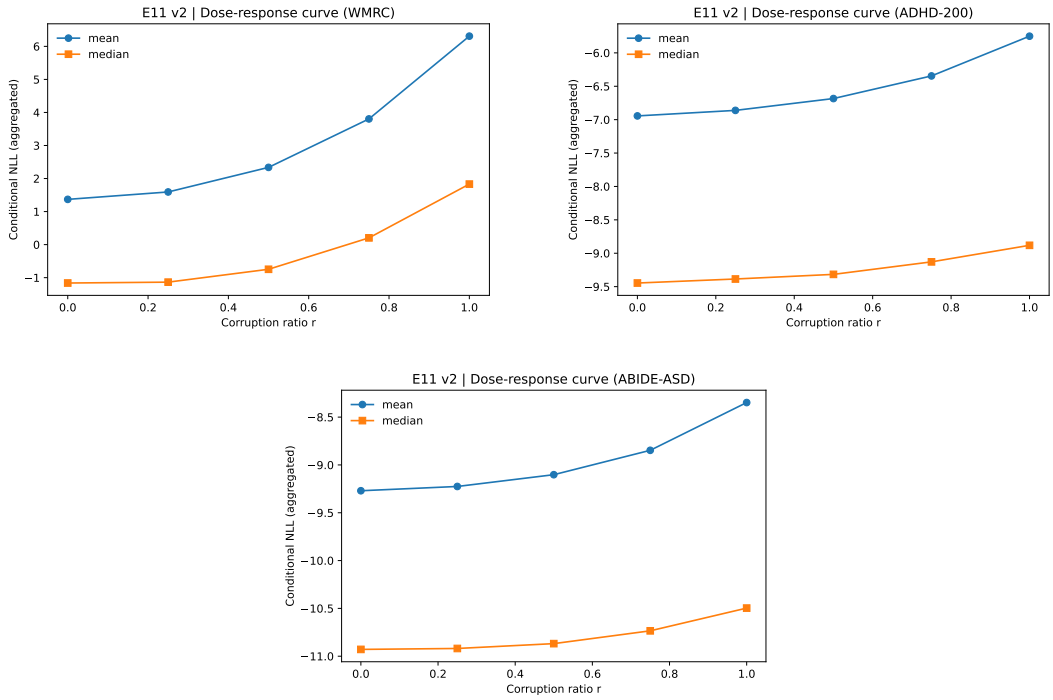


Figure 6: Conditional NLL as a function of context-corruption rate under hard-matched replacement (seed 42). Panels are shown separately because absolute NLL scales are not comparable across datasets.

Table 9: Leave-one-site-out transfer of the true-context advantage.

Dataset	Held-out sites	Positive-site fraction	Meta mean	95% CI
WMRC	8	1.000	6.845	[6.037, 7.652]
ADHD-200	7	1.000	1.135	[0.845, 1.424]
ABIDE-II	13	1.000	1.195	[0.824, 1.565]

effect, showing that the observed context preference is not explained solely by statistical coupling within a single estimated FC matrix.

D.3 Cross-Site Generalization

We further evaluate SCC with leave-one-site-out transfer. For each held-out site, candidate construction, candidate selection, and control normality fitting are performed without access to that site. The wrong-minus-true context gap is positive in all 28 held-out sites across the three datasets. Random-effects meta-analytic means are 6.845 for WMRC, 1.135 for ADHD-200, and 1.195 for ABIDE-II, with confidence intervals above zero (Table 9). Together with the consistent effect direction across the prespecified split seeds, this result indicates that SCC generalizes beyond individual acquisition sites and scanning protocols.

E SCC in Disease Groups

The preceding analyses characterize SCC primarily from healthy controls. We next evaluate whether the same control-defined contextual organization remains detectable in clinical populations. All patient analyses use the scorer fitted to healthy controls, no patient-specific organizational law is estimated.

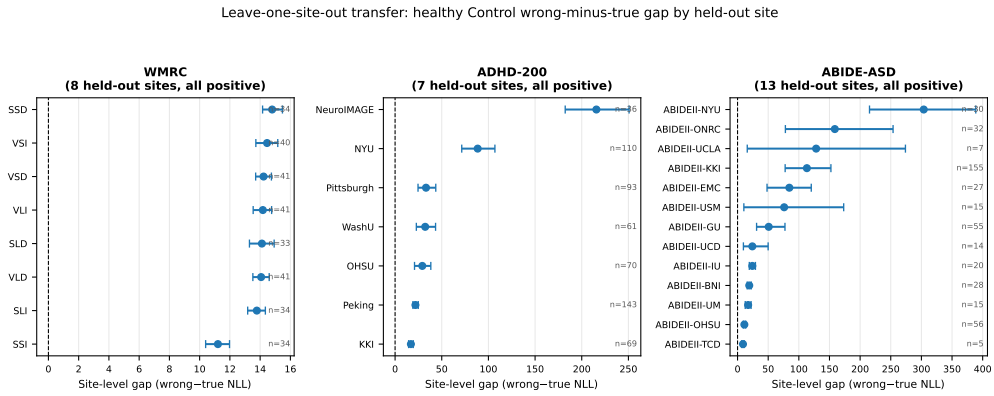


Figure 7: Leave-one-site-out transfer of the healthy-control wrong-minus-true NLL gap. Points and intervals show site-level estimates and 95% bootstrap confidence intervals; all 28 held-out sites show positive effects.

Table 10: Preservation of the true-context advantage in clinical groups under the healthy-control-fitted SCC scorer.

Dataset	Patient n	Positive-gap fraction	Patient/control gap ratio
WMRC	222	1.000	0.927
ADHD-200	357	1.000	1.090
ABIDE-II	400	0.995	1.103

E.1 Preservation of the True-Context Advantage

Patients retain a strong preference for their true functional contexts under the healthy-control-fitted scorer. The fraction of patients with a positive wrong-minus-true NLL gap is at least 0.995 in all three datasets, while the median patient-to-control gap ratios are 0.927, 1.090, and 1.103 for WMRC, ADHD-200, and ABIDE-II, respectively (Table 10). The corresponding distributions substantially overlap between controls and patients (Fig. 8). Thus, the control-defined contextual regularity remains strongly expressed in the clinical groups rather than disappearing at the group level.

E.2 Relative Magnitude of Disease-Related Shifts

To place disease-related deviations on the same conceptual scale as contextual organization strength, we compute η, the ratio between the estimated disease-related shift and the contextual organization strength. Values below one indicate that the disease-related deviation is smaller than the background organization captured by SCC. Median η is below one in every dataset, and the fraction of subjects with $\eta < 1$ ranges from 0.861 to 1.000 (Table 11). These results support the interpretation that clinical alterations perturb the realization of a largely preserved contextual regularity rather than replacing that regularity altogether. This comparison is descriptive under the healthy-fitted scoring framework and is not interpreted as a disease-specific directional law.

F Additional Validation Analyses

This section reports two supplementary analyses that characterize the spatial identifiability of controlled SCC violations and the amount of context required to recover the full-context predictive relationship.

F.1 Localization under Controlled Perturbations

A controlled perturbation is injected into the internal edges of a target subnetwork, with perturbation doses of $\{0.25, 0.5, 1.0, 1.5\}$ and dose-scaled noise. The conditional residual scores are then used

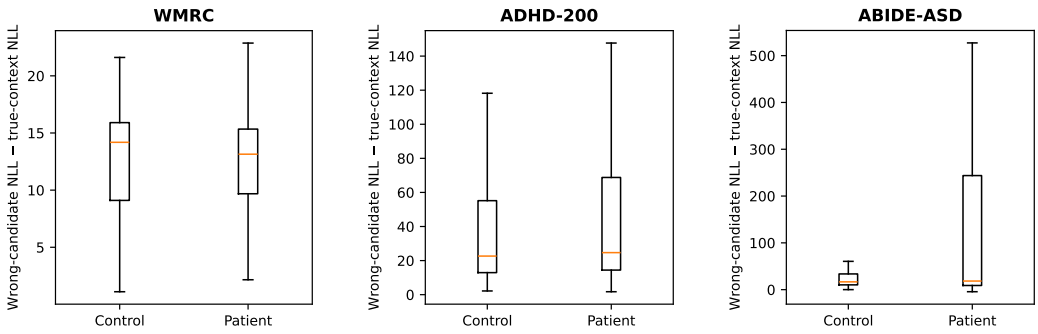


Figure 8: Distributions of contextual organization strength in controls and patients under the healthy-control-fitted scorer (seed 42).

Table 11: Disease-related shift relative to contextual organization strength.

Dataset	Median η	95% CI	Fraction with $\eta < 1$
WMRC	0.155	[0.060, 0.256]	0.861
ADHD-200	0.005	[0.003, 0.008]	1.000
ABIDE-II	0.013	[0.003, 0.017]	1.000

to rank candidate regions. An equal-size perturbation applied to randomly located non-target edges provides the negative control. Across 23,040 trials per dataset, the perturbed target is recovered as the top-ranked region in 0.939–0.996 of trials and within the top three in 0.967–0.996 of trials (Table 12). Random-edge controls produce much lower recovery rates. This demonstrates that controlled violations of SCC leave a spatially identifiable subnetwork-level signature.

The dose dependence of the localization signal is strong in ADHD-200 and ABIDE-II (Spearman $\rho = 1.0$ and monotonic fractions above 0.94) but weak in WMRC (Spearman $\rho = 0.40$, monotonic fraction 0.154). We therefore use this experiment as evidence for spatial identifiability, while avoiding a general claim of monotonic localization strength across all datasets.

F.2 Finite-Context Sufficiency

We quantify how many context ROIs are required to recover the predictive information available from the full context. For each candidate-fold unit, m_{95} is defined as the smallest tested context size whose held-out R^2 reaches at least 95% of the full-context R^2. Tested context sizes are $\{1, 2, 4, 6, 8, 12, 16, \text{all}\}$. The finite fraction is the proportion of units for which this threshold is achieved before the full context is used; $Q(12)$ is the median ratio between R^2 at context size 12 and the full-context R^2 among units with positive full-context R^2.

WMRC exhibits highly compact contextual dependence: the finite fraction is 0.972–1.000 across seeds, and 92–94% of units reach the 95% recovery threshold with a single context ROI. ADHD-200 and ABIDE-II show more distributed dependence. Their finite fractions are 0.618–0.828 and 0.500–0.727, respectively, and a larger fraction of units requires the full context. Nevertheless, $Q(12)$ remains high (0.986–0.992 in ADHD-200 and 0.874–0.991 in ABIDE-II), indicating that a limited context captures most of the predictable information for many candidates even when strict 95% recovery requires more ROIs.

G Additional Controls for Context Specificity and Candidate Attribution

The preceding analyses establish the existence, gradedness, and robustness of the Subnetwork Contextual Constraint (SCC). We further perform targeted controls to examine three alternative explanations: whether SCC depends on the subject-specific adaptive context construction, whether it can be

Table 12: Localization of controlled SCC violations. Target top-1/top-3 denote recovery of the perturbed subnetwork; random top-1/top-3 use equal-size non-target perturbations.

Dataset	Target top-1	Target top-3	Random top-1	Random top-3	Chance	Spearman	Monotonic
WMRC	0.996	0.996	0.021	0.098	0.111	0.40	0.154
ADHD-200	0.939	0.976	0.032	0.187	0.111	1.00	0.946
ABIDE-II	0.940	0.967	0.049	0.267	0.111	1.00	0.943

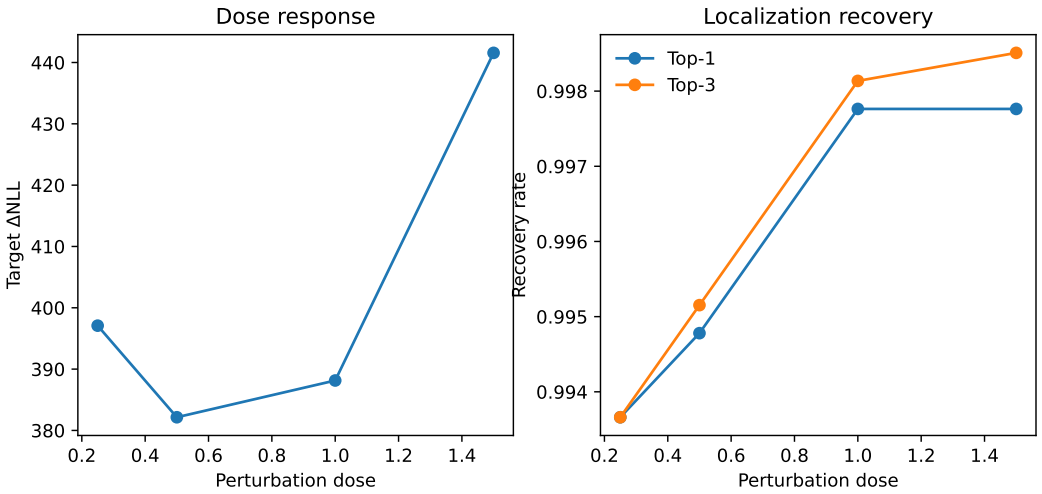


Figure 9: Controlled-perturbation localization in WMRC. Target top-1 recovery is substantially above both the equal-size random-edge control and chance.

explained by a generic subject-level connectome fingerprint, and whether the classification improvement of CCA-SRL can be reduced to disease-informed candidate mining alone. Unless otherwise stated, all analyses follow the same outer-fold protocol and use training data only for candidate construction, context definition, and normality-model fitting.

G.1 Robustness of SCC to Alternative Context Definitions

The original CCA-SRL pipeline defines the context of each candidate subnetwork using the subject-specific Top-C ROIs with the strongest total absolute connectivity to that candidate. A potential concern is therefore that the observed candidate–context preference may be induced by this adaptive construction itself. We test this possibility using alternative context definitions that do not depend on the held-out subject.

First, for each candidate, we construct a *fixed training-control context* from the group-average FC of the outer-training controls. The same context indices are then used for all subjects in the corresponding outer fold. We evaluate both the absolute connectivity of the signed group-average FC and the group-average absolute FC. Second, we construct *network-membership contexts*. For ADHD-200 and ABIDE-II, the network partition is estimated only from the outer-training controls using the group-average absolute FC, whereas OpenNeuro additionally permits an atlas-derived AAL116 network partition independent of FC estimation.

Across all three datasets, the true candidate–context pairing remains preferential under these subject-independent context definitions. For the fixed training-control contexts, candidate-identification accuracy ranges from 0.457 to 0.809, compared with the chance level of $1/K = 0.167$, while retrieval AUC ranges from 0.587 to 0.916. All four outer folds exhibit a positive wrong-minus-true NLL gap for every reported fixed-context condition. Network-membership contexts show the same qualitative pattern: identification accuracy ranges from 0.426 to 0.742 and retrieval AUC from 0.596

Table 13: Finite-context sufficiency across datasets and split seeds.

Dataset	Seed	Finite fraction	Full context required	Units	$Q(12)$
WMRC	42	0.972	1	36	0.998
WMRC	0	0.972	1	36	0.998
WMRC	1	1.000	0	36	0.998
WMRC	2	1.000	0	36	0.998
ADHD-200	42	0.618	13	36	0.986
ADHD-200	0	0.828	5	35	0.992
ADHD-200	1	0.677	10	34	0.988
ABIDE-II	42	0.727	9	36	0.991
ABIDE-II	0	0.500	18	36	0.874
ABIDE-II	1	0.543	16	36	0.978

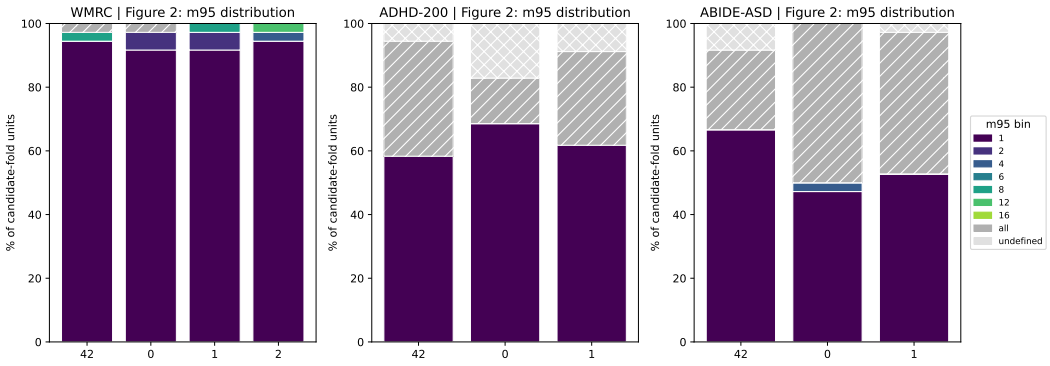


Figure 10: Distribution of the 95%-recovery context size m_{95}. WMRC is concentrated at small contexts, whereas ADHD-200 and ABIDE-II contain a larger full-context component.

to 0.900, with positive NLL gaps in all four folds. The OpenNeuro atlas-based context also preserves the effect (identification accuracy 0.447, retrieval AUC 0.596).

These results indicate that SCC is not specific to the original adaptive Top-C construction and remains detectable when candidate contexts are defined independently of the held-out subject.

We additionally verify that the alternative context controls are not degenerate. Fixed training-control contexts remain candidate-specific rather than collapsing to the same ROI set: pairwise Jaccard overlap between fixed candidate contexts remains moderate across all three datasets. Conversely, the original adaptive Top-C construction exhibits substantial variation across subjects, confirming that it is genuinely subject-specific rather than an approximately fixed group-level template.

G.2 Candidate-Specific Context Correspondence

We next examine whether SCC reflects a specific correspondence between each candidate subnetwork and its associated context, rather than a generic preference for any context derived from the same subject. For each subject, we evaluate every retained candidate U_k against the contexts associated with all retained candidates C_j, producing a $K \times K$ candidate–context score matrix. For each pair, we compute the excess conditional NLL relative to the matched context,

$$\Delta_{kj} = \mathrm{NLL}(U_k \mid C_j) - \mathrm{NLL}(U_k \mid C_k). \tag{72}$$

The diagonal is therefore zero by construction, whereas positive off-diagonal values indicate that the matched context provides a better explanation of the candidate than a mismatched candidate context.

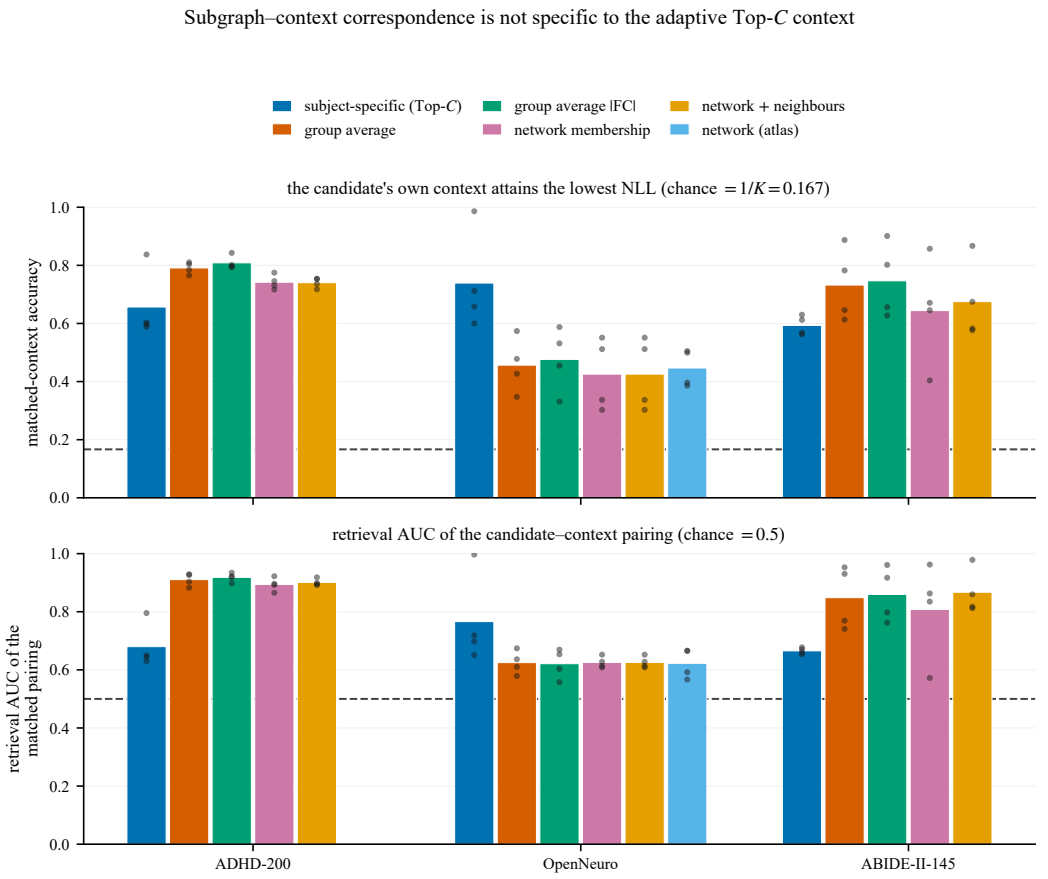


Figure 11: **SCC under alternative context constructions.** Candidate-identification accuracy and candidate–context retrieval AUC are shown for adaptive Top-C, fixed training-control, and network-membership context definitions. Points denote individual outer folds. Dashed horizontal lines indicate chance performance ($1/K = 0.167$ for identification and 0.5 for retrieval AUC). The true candidate–context correspondence remains detectable when context construction is independent of the held-out subject.

The resulting matrices consistently show lower conditional NLL along the matched candidate–context diagonal. Importantly, this structure is observed not only for the original adaptive contexts but also for fixed training-control and network-membership contexts, providing a direct candidate-level view of the contextual correspondence underlying SCC.

A related alternative explanation is that this preference simply reflects a generic subject-specific connectome fingerprint: subnetworks and contexts obtained from the same subject may be more compatible because they share subject-level FC structure. To separate subject specificity from candidate specificity, we therefore perform all candidate–context substitutions strictly within the same subject. The subject identity and global FC matrix remain unchanged, while only the candidate–context assignment is altered.

Within-subject context swaps consistently increase conditional NLL relative to the matched context. The median excess NLL is approximately $+1.9$ for ADHD-200, $+4.2$ for OpenNeuro, and $+1.6$ for ABIDE-II, with a positive effect for the large majority of subject–candidate pairs. Replacing the context with one from another subject produces a substantially larger penalty under the adaptive construction, indicating that context compatibility contains both subject-specific and candidate-specific components.

Because the within-subject control holds subject identity fixed, the preference for the matched candidate context cannot be explained solely by a global subject-level connectome fingerprint.

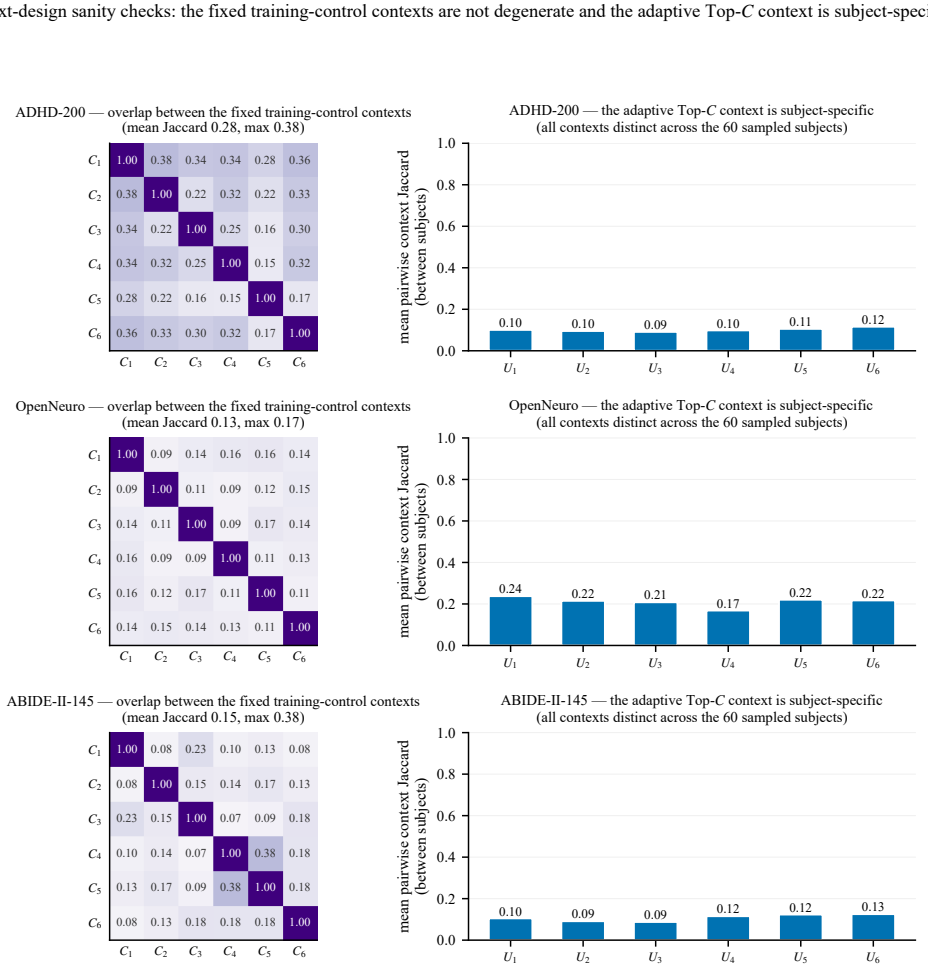


Figure 12: **Design checks for fixed and adaptive context constructions.** Left: pairwise Jaccard overlap between candidate-specific fixed training-control contexts, showing that the fixed context library is non-degenerate. Right: mean pairwise context Jaccard between subjects under the adaptive Top-C construction, confirming that the original context definition is subject-specific.

G.3 Controlling for Disease-Informed Candidate Mining

CCA-SRL uses disease labels during candidate discovery: disorder–control FC differences are used to construct the candidate pool and retain informative candidates within each training fold. We therefore test whether the final classification performance can be explained by this candidate-selection procedure alone.

We hold the candidate-mining pipeline, outer folds, and retained candidates fixed and compare the full model against matched baselines that remove the context-conditioned anomaly mechanism. These controls include a candidate-statistics MLP, a whole-brain backbone concatenated with the same candidate statistics, and simple linear classifiers operating on candidate statistics with or without raw context and cross-interaction features.

Across all three datasets, the candidate-based controls remain below the full CCA-SRL model. For the neural matched baseline combining the whole-brain backbone and candidate statistics, mean AUC decreases from 0.630 to 0.564 on ADHD-200, from 0.745 to 0.614 on OpenNeuro, and from 0.651 to 0.591 on ABIDE-II. Candidate-statistics linear classifiers similarly achieve AUCs of approximately 0.57–0.62, remaining below CCA-SRL on all three datasets.

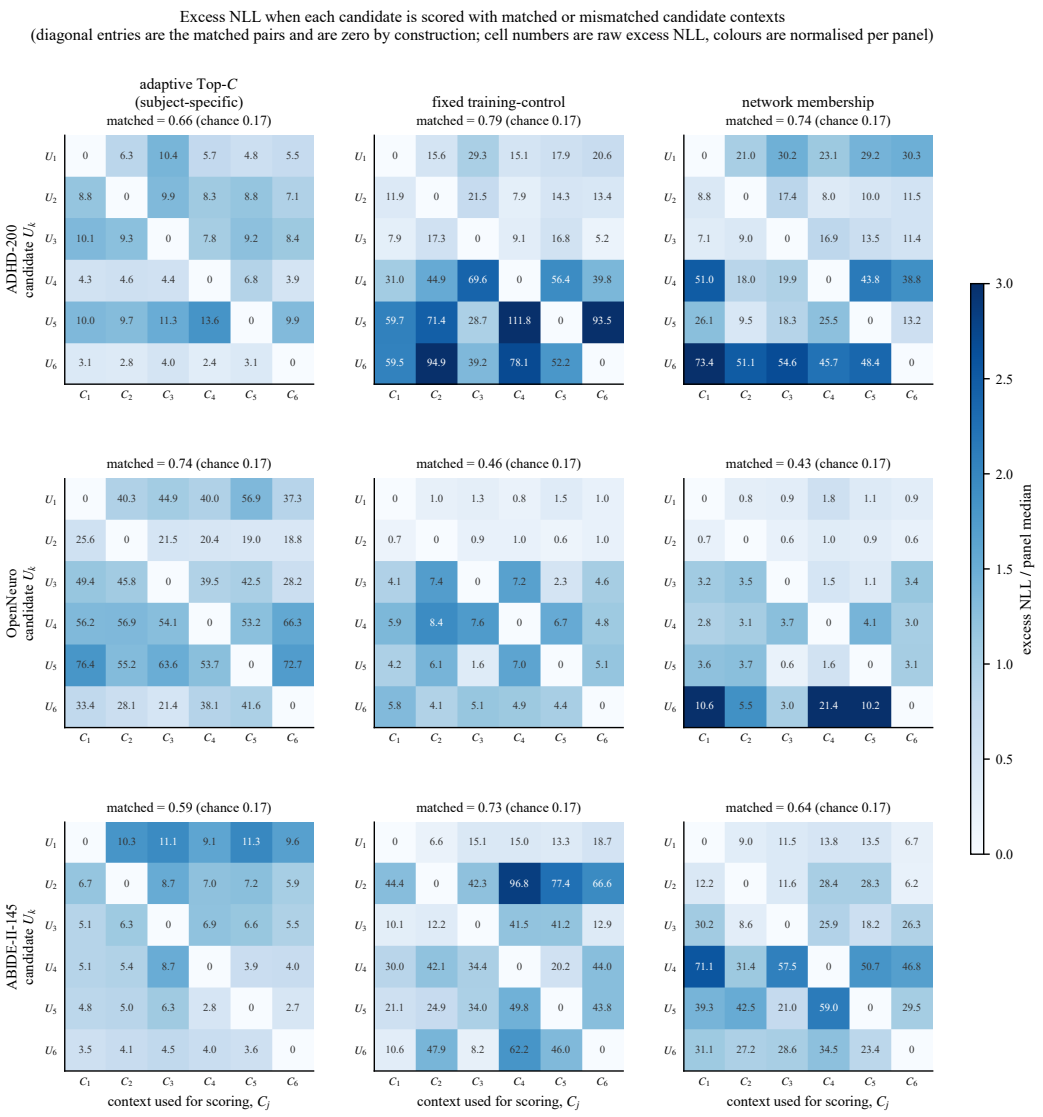


Figure 13: **Candidate–context swap analysis.** Each matrix shows the excess conditional NLL obtained when candidate U_k is evaluated using context C_j, relative to its matched context C_k. Diagonal entries correspond to matched candidate–context pairs and are zero by construction. Cell annotations report raw excess NLL values, whereas colors are normalized within each panel for visual comparability. Positive off-diagonal values indicate preferential compatibility between a candidate and its matched context.

The magnitude of the difference varies across datasets and is largest on OpenNeuro. These results therefore do not imply that disease-informed candidate mining is uninformative; rather, they show that candidate selection alone is insufficient to account for the performance of the full model. The context-conditioned anomaly representation provides additional discriminative information beyond the selected candidate statistics.

H Parameter Sensitivity and Validation

The final CCA-SRL experiments use a prespecified structural configuration of six retained candidates ($K = 6$), a context size of 16 ROIs ($C = 16$), and 120 retained site-level edges. These values were fixed before the final comparative evaluation and were used consistently across all three

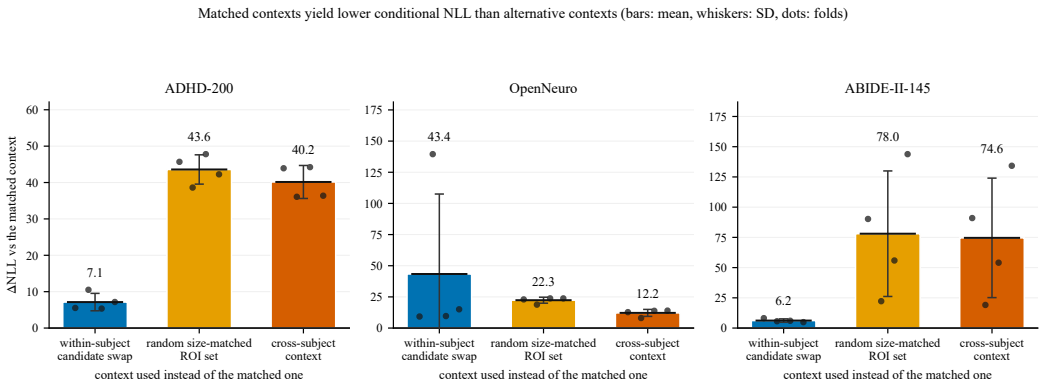


Figure 14: **Matched contexts yield lower conditional NLL than alternative contexts.** Conditional-NLL increases are shown for within-subject candidate–context swaps, random size-matched contexts, and cross-subject contexts. Within-subject swaps alter only the candidate–context assignment while preserving the subject and global FC matrix, thereby isolating candidate-specific contextual compatibility from generic subject-level effects.

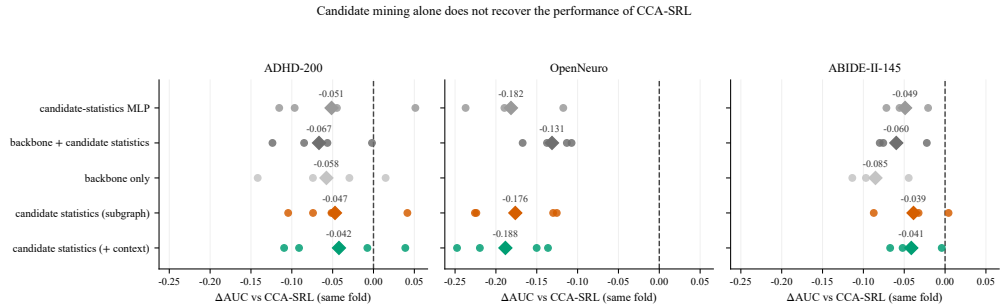


Figure 15: **Candidate mining alone does not recover the performance of CCA-SRL.** Fold-wise AUC differences between matched candidate-based baselines and the full CCA-SRL model are shown for each dataset. Candidate construction and selection are held fixed across compared variants. Negative values indicate lower AUC than CCA-SRL. Diamonds denote the mean paired difference across outer folds.

datasets; the outer-test results reported in the main paper were not used to select these hyperparameters.

We subsequently perform a post-hoc sensitivity analysis around this prespecified configuration to assess whether the reported performance is strongly dependent on these choices. Each parameter is varied independently while the remaining settings are kept at the prespecified defaults.

Table 14: Parameter sensitivity on ADHD-200. The prespecified configuration used in the main experiments is shown in bold.

Parameter	Value	AUC (%)	p-value	ACC (%)	p-value	Balanced ACC (%)	Macro-F1	Score
Candidate count K	3	58.78 ± 3.77	2.74×10^{-2}	60.38 ± 2.01	2.96×10^{-3}	53.61	51.81	59.58
Candidate count K	**6**	63.04 ± 5.40	2.49×10^{-2}	64.33 ± 2.14	1.37×10^{-3}	58.45	57.06	63.69
Candidate count K	9	58.32 ± 2.64	1.21×10^{-2}	60.49 ± 2.54	5.60×10^{-3}	55.22	54.44	59.40
Context size C	8	57.22 ± 3.11	2.77×10^{-2}	61.24 ± 2.51	4.46×10^{-3}	53.55	50.95	59.23
Context size C	12	58.18 ± 3.14	2.03×10^{-2}	60.06 ± 2.46	5.80×10^{-3}	52.83	50.52	59.12
Context size C	**16**	63.04 ± 5.40	2.49×10^{-2}	64.33 ± 2.14	1.37×10^{-3}	58.45	57.06	63.69
Retained site edges	80	57.14 ± 2.39	1.40×10^{-2}	61.34 ± 1.60	1.17×10^{-3}	53.82	51.79	59.24
Retained site edges	**120**	63.04 ± 5.40	2.49×10^{-2}	64.33 ± 2.14	1.37×10^{-3}	58.45	57.06	63.69
Retained site edges	160	59.10 ± 3.61	2.23×10^{-2}	61.56 ± 2.16	2.66×10^{-3}	54.28	50.35	60.33

Table 15: Parameter sensitivity on OpenNeuro-ds002424. The prespecified configuration used in the main experiments is shown in bold.

Parameter	Value	AUC (%)	p-value	ACC (%)	p-value	Balanced ACC (%)	Macro-F1	Score
Candidate count K	3	65.74 ± 4.34	8.17×10^{-3}	62.12 ± 4.66	2.04×10^{-2}	59.94	58.29	63.93
Candidate count K	**6**	74.52 ± 4.51	2.54×10^{-3}	70.38 ± 4.76	5.06×10^{-3}	68.58	68.21	72.45
Candidate count K	9	66.85 ± 4.82	9.02×10^{-3}	63.85 ± 4.71	1.46×10^{-2}	63.35	62.81	65.35
Context size C	8	70.01 ± 6.45	1.26×10^{-2}	66.92 ± 3.81	4.56×10^{-3}	66.06	66.06	68.47
Context size C	12	68.91 ± 5.80	1.10×10^{-2}	65.38 ± 5.36	1.56×10^{-2}	65.04	64.77	67.15
Context size C	**16**	74.52 ± 4.51	2.54×10^{-3}	70.38 ± 4.76	5.06×10^{-3}	68.58	68.21	72.45
Retained site edges	80	65.61 ± 5.46	1.58×10^{-2}	61.35 ± 4.46	2.17×10^{-2}	59.50	58.84	63.48
Retained site edges	**120**	74.52 ± 4.51	2.54×10^{-3}	70.38 ± 4.76	5.06×10^{-3}	68.58	68.21	72.45
Retained site edges	160	71.13 ± 3.29	1.55×10^{-3}	63.65 ± 5.31	2.11×10^{-2}	62.10	61.35	67.39

Table 16: Parameter sensitivity on ABIDE-II. The prespecified configuration used in the main experiments is shown in bold.

Parameter	Value	AUC (%)	p-value	ACC (%)	p-value	Balanced ACC (%)	Macro-F1	Score
Candidate count K	3	57.72 ± 2.69	1.56×10^{-2}	55.41 ± 3.38	6.96×10^{-2}	55.22	53.55	56.56
Candidate count K	**6**	64.19 ± 1.18	2.44×10^{-4}	63.07 ± 0.73	7.52×10^{-5}	62.43	62.21	63.63
Candidate count K	9	58.73 ± 2.67	1.10×10^{-2}	56.92 ± 2.78	2.31×10^{-2}	55.98	53.39	57.82
Context size C	8	60.84 ± 3.26	1.03×10^{-2}	58.42 ± 2.54	1.05×10^{-2}	57.84	54.55	59.63
Context size C	12	62.92 ± 1.76	1.05×10^{-3}	59.69 ± 2.83	9.55×10^{-3}	59.53	57.55	61.31
Context size C	**16**	64.19 ± 1.18	2.44×10^{-4}	63.07 ± 0.73	7.52×10^{-5}	62.43	62.21	63.63
Retained site edges	80	62.58 ± 3.37	7.52×10^{-3}	57.37 ± 3.40	3.29×10^{-2}	56.70	54.34	59.98
Retained site edges	**120**	64.19 ± 1.18	2.44×10^{-4}	63.07 ± 0.73	7.52×10^{-5}	62.43	62.21	63.63
Retained site edges	160	63.43 ± 4.77	1.65×10^{-2}	63.30 ± 3.57	7.58×10^{-3}	62.62	62.38	63.36

Parameter sensitivity varies across datasets. OpenNeuro shows the largest variation with respect to the number of retained candidates and the site-level edge budget: the prespecified setting $K = 6$ yields Scores approximately 6.5–8.5 points higher than the neighboring candidate-count settings, while 120 retained site-level edges yields a Score approximately 5.1–9.0 points higher than the neighboring edge-budget settings. ADHD-200 exhibits more moderate variation, whereas ABIDE-II is comparatively stable with respect to the retained-edge setting, with 120 and 160 edges yielding similar Scores (63.63 vs. 63.36).

Importantly, these comparisons are post-hoc sensitivity analyses around a configuration fixed before the final comparative evaluation. They were not used for hyperparameter selection or model configuration.

I Visualization of Selected Subnetworks

To qualitatively examine the subnetworks selected by CCA-SRL, we visualize the candidate structures identified on ADHD-200 and highlight several representative patterns. These visualizations are intended to illustrate the spatial organization of the selected subnetworks rather than to provide independent confirmatory localization evidence. Throughout this section, the reported anomaly score denotes the context-conditioned subnetwork inconsistency measured by NLL and should not be interpreted as mean functional-connectivity strength. In the network plots, large nodes and edges denote selected candidate structures, whereas the smaller nodes provide the full AAL116/Yeo-informed functional-network context. Edge colors indicate the direction of the group-average FC difference (ADHD $>$ Control or Control $>$ ADHD); individual displayed edges are not interpreted as separately significant after multiple-comparison correction.

Table 17 summarizes three representative examples chosen to illustrate both cross-network and within-network patterns. The DMN–DAN example provides the clearest candidate-level separation, whereas the Visual and DMN cases are retained as complementary exploratory examples.

DMN–DAN pathway. The DMN–DAN pattern provides the strongest representative example among the three cases. For the representative 4-ROI DMN Angular–DAN Parietal candidate, the

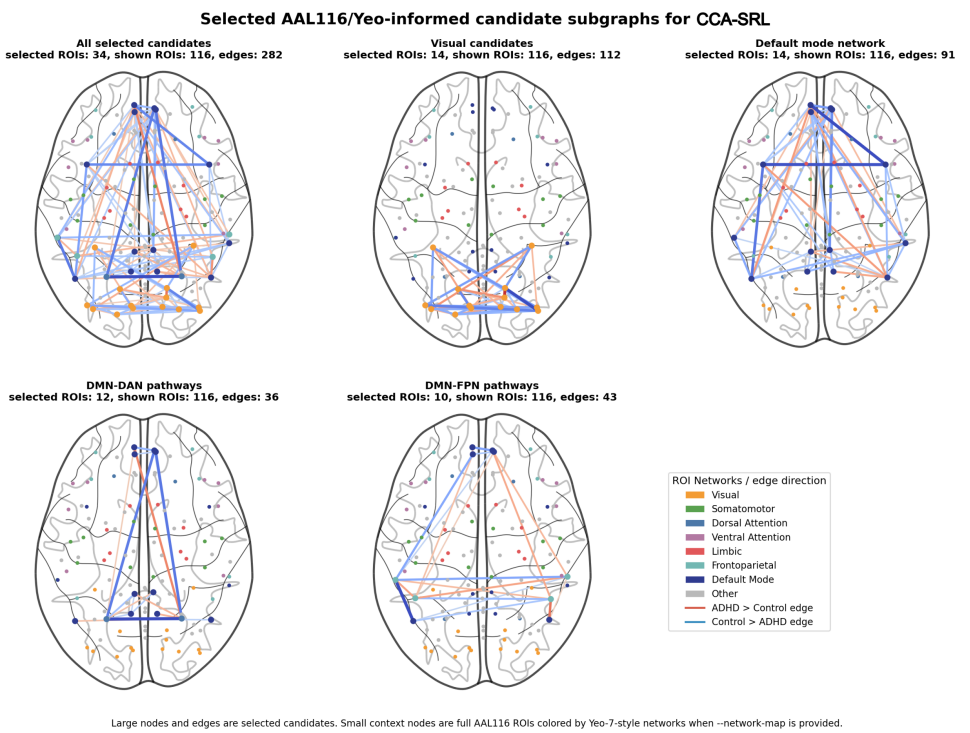


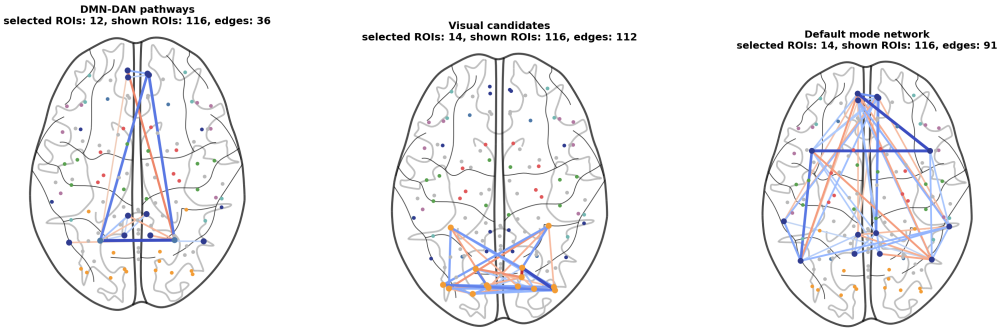
Figure 16: Overview of the CCA-SRL selected subnetworks on ADHD-200. The visualization includes all selected candidates together with functional subsets corresponding to Visual, default-mode-network (DMN), DMN–dorsal-attention-network (DMN–DAN), and DMN–frontoparietal-network (DMN–FPN) patterns. Large nodes and connecting edges indicate selected candidate structures, while smaller nodes show the AAL116/Yeo-informed whole-brain context. Edge color encodes the direction of the group-average FC difference and is used only for visualization.

Table 17: Candidate-level summary for three representative selected subnetworks. HC and ADHD denote mean context-conditioned anomaly scores (NLL), not mean FC. For the DMN–DAN visualization, the displayed 12-ROI panel aggregates multiple candidates; the listed statistics correspond specifically to the representative 4-ROI DMN Angular–DAN Parietal candidate.

Representative pattern	**HC NLL**	**ADHD NLL**	Δ	**Mann–Whitney** p	**Welch** p	**Welch-BH** q
DMN–DAN pathway	6.099	13.248	+7.149	0.0589	0.000948	0.00853
Visual network	2.723	4.072	+1.349	0.0304	0.0434	0.0872
Default Mode Network	1.920	2.912	+0.992	0.0624	0.0485	0.0872

mean anomaly score increases from 6.099 in controls to 13.248 in ADHD ($\Delta = +7.149$). This candidate contains bilateral angular and bilateral superior parietal regions, providing a concrete example of a cross-network contextual inconsistency. The effect is strong under Welch's test and remains below 0.05 after the reported nine-comparison Welch-BH adjustment ($q = 0.00853$), although the Mann–Whitney test does not cross the 0.05 threshold ($p = 0.0589$). Importantly, the 12-ROI visualization in Fig. 17(a) aggregates several DMN–DAN candidates; therefore, the candidate-level statistics above should not be attributed to the complete displayed panel.

Visual network. The Visual example shows a smaller but directionally consistent increase in context-conditioned inconsistency, from 2.723 in controls to 4.072 in ADHD ($\Delta = +1.349$). Both unadjusted Mann–Whitney and Welch tests are below 0.05, whereas the reported Welch-BH value is $q = 0.0872$. We therefore interpret this pattern as exploratory evidence that the selected abnormalities are not restricted to higher-order association systems and may also involve within-network deviations in sensory systems.



(a) DMN–DAN pathways (b) Visual network (c) Default Mode Network

Figure 17: Representative selected-subnetwork visualizations. (a) DMN–DAN pathways: 12 selected ROIs and 36 displayed selected edges; this panel aggregates multiple DMN–DAN candidates. (b) Visual network: 14 selected ROIs and 112 selected edges. (c) Default Mode Network: 14 selected ROIs and 91 selected edges. These examples illustrate both cross-network and within-network organizational patterns captured by the CCA-SRL selection procedure.

Default Mode Network. The DMN example exhibits the same directional trend, with the mean anomaly score increasing from 1.920 in controls to 2.912 in ADHD ($\Delta = +0.992$). Statistical support is weaker (Mann–Whitney $p = 0.0624$; Welch-BH $q = 0.0872$), so this example is used primarily to illustrate a functionally interpretable within-network pattern rather than a robust correction-surviving group effect.

Taken together, the visualizations show that the selected CCA-SRL candidates can span both within-network and cross-network configurations. The DMN–DAN example provides the clearest representative cross-network case, while the Visual and DMN examples illustrate complementary within-network patterns. Because these cases are drawn from the same candidate-selection pipeline used to identify informative subnetworks, the associated group comparisons should be interpreted as descriptive post-selection analyses rather than as independent validation of disease-specific biomarkers. Likewise, the displayed edge directions are visualization aids and do not imply edge-wise statistical significance.

J Implementation Details of CCA-SRL

This section provides implementation details for the final CCA-SRL classifier described in the main Methods. The final classifier and the SCC validation analyses above serve different purposes and therefore use different prespecified candidate settings. The final classification model uses the configuration selected in Appendix H: six retained candidates ($K = 6$), a context size of 16 ROIs ($C = 16$), and 120 retained site-level edges.

Candidate subnetwork mining and selection. Candidate subnetworks are constructed independently within each outer-training fold. Within every eligible acquisition site, edges are ranked by the absolute standardized disorder–control difference $|d_{ij}|$, and the 120 highest-ranked edges are retained under the default configuration. Connected components of these high-discriminability edges form site-level candidate subnetworks; components from the union of retained edges across sites are added and duplicate ROI sets are removed. All candidate construction uses training subjects only.

For each candidate U_k and subject s, the functional context is computed from that subject's FC matrix rather than from a fixed site-level template. Each ROI outside the candidate is ranked by its total absolute connectivity to the candidate,

$$q_i^{(s,k)} = \sum_{j \in U_k} |A_{ij}^{(s)}|, \qquad i \notin U_k, \tag{73}$$

and the $C = 16$ highest-ranked outside ROIs define $C_k^{(s)}$ under the default configuration. Thus, candidate membership is fixed within a fold, whereas the context is subject-specific.

A temporary normal contextual anomaly scorer is fitted to the candidate pool and applied to the training subjects. Candidates are ranked by the absolute standardized separation between disorder and control anomaly scores,

$$D_k = \left| \frac{\mu(a_k^{\mathrm{dis}}) - \mu(a_k^{\mathrm{ctrl}})}{\sqrt{[\mathrm{Var}(a_k^{\mathrm{dis}}) + \mathrm{Var}(a_k^{\mathrm{ctrl}})]/2 + \epsilon}} \right|, \tag{74}$$

and the top K candidates are retained. The temporary scorer is then discarded and the final normal scorer is re-fitted on the selected candidate set before generating anomaly evidence for downstream training and evaluation.

Contextual anomaly scoring. For each selected candidate, 14 internal graph features describe its size, density, signed and absolute FC statistics, edge-sign ratios, nodal strength and degree statistics, clustering, and global efficiency. The same 14-dimensional feature family is extracted from the subject-specific context, together with six subgraph–context interaction features. A ridge-regularized predictor is fitted only on controls using the context features, the six cross-interaction features, relative candidate size, and candidate identity. The ridge penalty is selected from $\{0.01, 0.1, 1, 10, 100\}$ on training controls. Residual covariance is estimated with shrinkage 0.9 and numerical jitter 10^{-6}.

The contextual anomaly score is the Gaussian negative log-likelihood of the standardized residual under the control residual covariance. In addition to the scalar NLL, the predicted internal feature vector and the signed residual vector are retained because the downstream model uses both anomaly magnitude and deviation direction. Raw NLL values are standardized with training statistics before they enter the neural model.

Neural classification architecture. A one-layer shared GCN first produces a whole-brain ROI representation $\mathbf{H}_{\mathrm{shared}}$. For each selected candidate, the model forms a 71-dimensional evidence vector from the observed internal features, control-predicted features, residuals, context features, cross-interaction features, standardized NLL, and an eight-dimensional candidate embedding. A two-layer token MLP maps this vector to the learned subgraph token space.

The role-aware guidance encoder projects candidate tokens to ROIs according to internal, context, and boundary membership. The resulting ROI guidance $\mathbf{G}$ conditions the shared node representation through FiLM, while token-derived internal and boundary anomaly fields modulate signed graph attention. Positive and negative FC channels are propagated separately to obtain the guided whole-brain representation $\mathbf{H}_{\mathrm{guided}}$.

The base pathway pools $\mathbf{H}_{\mathrm{shared}}$, while the guided pathway pools $\mathbf{H}_{\mathrm{guided}}$, using concatenated mean and max statistics. A control-calibrated reliability gate ρ scales the guided residual before fusion,

$$\mathbf{z}_{\mathrm{combined}} = \mathbf{z}_{\mathrm{base}} + \rho\, \mathbf{W}_g(\mathbf{z}_{\mathrm{guided}}). \tag{75}$$

The mean, top-three mean, and maximum standardized NLL are concatenated with the fused graph representation before the final MLP classifier. Focal loss is used for classification.

K Computational Cost and Complexity

This appendix section reports the measured computational cost and asymptotic complexity of CCA-SRL. FLOP accounting is performed for the exact model configurations used on the three benchmarks (ADHD-200, OpenNeuro-ds002424, and ABIDE-II).

K.1 FLOP Accounting

We count multiply–accumulate operations (MACs), with 1 MAC = 2 FLOPs. The reported values are obtained through operator-level profiling of the fixed experimental implementation, accounting for linear transformations, normalization layers, batched matrix multiplications, and tensor-contraction operations. The resulting values therefore correspond to measured operation counts for the reported configurations rather than analytical estimates.

Table 18: Model size and inference cost per subject (single forward pass, batch size 1).

Dataset	N	K	H	Attention heads	Parameters	MACs / subject	MFLOPs / subject
ADHD-200	137	6	32	1	38,488	160,318,822	320.6
OpenNeuro	116	6	24	2	26,658	66,337,552	132.7
ABIDE-II	145	6	24	1	24,424	102,296,790	204.6

Table 19: Per-module MAC attribution for one forward pass of one subject (percent of the subject total).

Module	ADHD-200	OpenNeuro	ABIDE-II
Signed GAT (pairwise attention + fields)	157,714,948 (98.4%)	64,830,080 (97.7%)	100,161,940 (97.9%)
Structural backbone (GCN stem)	1,214,368 (0.8%)	654,240 (1.0%)	1,019,640 (1.0%)
Role-aware subgraph→ROI projection	597,120 (0.4%)	314,496 (0.5%)	382,752 (0.4%)
FiLM conditioning	420,864 (0.3%)	267,264 (0.4%)	334,080 (0.3%)
Edge-level anomaly fields	339,474 (0.2%)	243,840 (0.4%)	379,314 (0.4%)
Subgraph tokenizer	16,704	16,704	7,584
Classifier + fusion + gate	10,576	7,760	7,760
Total	160,314,054	66,334,384	102,293,070

Table 20: Training and inference compute (MACs; multiply by 2 for FLOPs). The reported epoch count is measured on the first outer fold under the standard training protocol; early stopping (patience 25, minimum 30 epochs) determines E.

	ADHD-200	OpenNeuro	ABIDE-II
Mini-batches / epoch ($B = 16$)	25	15	28
Forward MACs / epoch	64.13 G	15.92 G	45.83 G
Forward + backward MACs / epoch ($3\times$)	192.38 G	47.76 G	137.49 G
Epochs completed (fold 1)	84	66	30
Training MACs / fold (measured E)	16.16 T	3.15 T	4.12 T
Training MACs / 4-fold CV	64.6 T	12.6 T	16.5 T
Upper bound ($E = 500$) / 4-fold CV	384.8 T	95.5 T	275.0 T
Inference MACs, whole cohort (one pass)	150.5 G	34.5 G	88.1 G

Table 21: Order-of-magnitude cost of the non-neural stages (MACs). The normality model is fitted twice per fold (pool scorer and refit); even so, every classical stage stays below 0.5% of the per-subject network forward pass (Table 18).

Classical stage	ADHD-200	OpenNeuro	ABIDE-II
Site-wise effect sizes + components (per fold)	5.0×10^7	2.0×10^7	5.1×10^7
Normality-model fit (one ridge/GCV pass)	2.2×10^7	3.2×10^7	6.1×10^7
Normality-model scoring (per subject)	$< 10^4$	$< 10^4$	$< 10^4$
Feature extraction (per subject)	3.2×10^5	3.2×10^5	3.2×10^5

Table 22: Performance and measured computational cost of the reported configurations. Fold times include data loading, candidate mining, both normality-model fitting stages, and the complete training loop, measured on a single NVIDIA GeForce RTX 3060 Laptop GPU (6 GB).

Dataset	AUC	ACC	Balanced ACC	Fold times (s)	CV time (s)	GPU hours
ADHD-200	0.6304	0.6433	0.5845	145 / 103 / 123 / 98	470	0.13
OpenNeuro	0.7452	0.7038	0.6858	116 / 131 / 115 / 89	451	0.13
ABIDE-II	0.6419	0.6307	0.6243	126 / 109 / 100 / 109	444	0.12

Computational characteristics. (i) The model is deliberately small (24–38k parameters) and its inference cost is 0.13–0.32 GFLOPs per subject, so scoring an entire cohort takes well under a minute of GPU time; (ii) almost all compute ($\approx 98\%$) is spent in the dense pairwise attention of the signed graph attention network — exactly where the anomaly signal is propagated — so the quadratic-in-N cost is intrinsic to the propagation mechanism rather than an implementation artefact; (iii) training is dominated by the same term and costs at most a few tens of TFLOPs per four-fold cross-validation with early stopping; (iv) the classical pipeline (candidate mining, normality model, feature extraction) accounts for less than 1% of the total compute and is therefore not a bottleneck.

K.2 Complexity Analysis

Per-subject inference. The tokeniser costs $\mathcal{O}(K(d_x T + T^2))$ with $d_x = 21 + M$; the structural backbone costs $\mathcal{O}(NDH)$ for the feature projection plus $\mathcal{O}(N^2 H)$ for graph propagation; the role-aware projection $\mathcal{O}(NKH)$; the edge-field assembly $\mathcal{O}(N^2 K)$; FiLM $\mathcal{O}(N(TH_{\text{film}} + H_{\text{film}}H))$; the signed graph attention processes *all ROI pairs* in two channels, $\mathcal{O}\big(N^2(4H+1)H_{\text{att}} + N^2 H\big)$; readouts, gating, fusion and the classifier are $\mathcal{O}(NH + H^2)$. Since $H, K \ll N$, the leading term is

$$\mathcal{O}\big(N^2(H^2 + K)\big) \qquad \text{per subject, per forward pass,} \tag{76}$$

i.e. *quadratic in the number of ROIs*, dominated by the dense pairwise attention over ROI pairs (Sec. K.1).

Training. With n_{tr} training subjects and n_{tr}/B balanced mini-batches, one fold costs

$$\mathcal{O}\Big(E \cdot \tfrac{n_{\text{tr}}}{B} \cdot B \cdot N^2(H^2 + K)\Big) = \mathcal{O}\big(E\, n_{\text{tr}} N^2(H^2 + K)\big) \tag{77}$$

multiply–accumulate operations for the forward pass, and roughly twice that for the backward pass.

Classical (non-neural) stages. Site-wise effect sizes cost $\mathcal{O}(N^2 \sum_g n_g)$ plus $\mathcal{O}(N^2 \log N)$ sorting per site; context construction is $\mathcal{O}(KN\bar{n})$ per subject with $\bar{n} = |U_k|$; fitting the normality model costs $\mathcal{O}(|\Lambda|\,(n_{\text{ctl}} M d_x^2 + d_x^3))$ (ridge with GCV over $|\Lambda|$ candidates); scoring costs $\mathcal{O}(K(d_x d_y + d_y^2))$ per subject. These are below 1% of the neural computation (Table 21).

Memory. Activations are dominated by the dense pairwise tensors of the signed graph attention: the attention input is $(B, N, N, 4H+1)$ per channel, i.e. $BN^2(4H+1)$ floating-point values; for $B = 16$, $N = 145$, $H = 24$ this is ≈ 131 MB per channel (≈ 262 MB for both channels), plus the $(B, N, N, 2)$ anomaly fields and their gradients.

L NOTATION SUMMARY

Table 23: Summary of notation used in the algorithms.

Symbol	Description	Symbol	Description
B	Batch size	N	Number of brain ROIs
M	Candidate-pool size before selection	K	Retained candidates per fold
D	Node-feature dimension $(3 + N)$	H	Hidden width of the graph network
T	Subgraph-token dimension	H_{film}	FiLM conditioner width
d_y	Graph-feature dimension (14)	d_x	Regressor input dimension $(21 + M)$
d_c	Candidate-identity embedding size (8)	C	Context size
U_k	ROI set of candidate k	$C_k^{(s)}$	Its context for subject s
$\mathbf{A}$	Signed functional connectivity	$\mathbf{X}$	Node features
$\hat{\mathbf{A}}$	Symmetrically normalised adjacency	$\mathbf{H}_{\text{shared}}$	Shared GCN node representation
$\mathbf{M}$	Role membership masks (in/ctx/bdry)	$\mathbf{G}$	ROI-level anomaly guidance
q_k	Scalar field modulation strength	$\mathbf{R}^{\text{in}}, \mathbf{R}^{\text{bdry}}$	Internal / boundary anomaly fields
$\boldsymbol{\gamma}, \boldsymbol{\beta}$	FiLM gain and shift	$\mathbf{H}_{\text{cond}}$	FiLM-conditioned representation
$\boldsymbol{\alpha}^{\pm}$	Signed attention weights	$\mathbf{H}_{\text{guided}}$	Signed-GAT node representation
a_k	Anomaly score (Mahalanobis NLL)	c_k	Control-calibrated evidence
p_k	Empirical upper-tail probability	$\mathbf{u}$	Gate features (6-dimensional)
$\mathbf{z}_{\text{base}}, \mathbf{z}_{\text{guided}}$	Statistical readouts	$\boldsymbol{\rho}$	Reliability gate coefficient
$\mathbf{z}$	Fused representation	$\hat{s}$	Class logit
D_k	Candidate-level $\lvert$Cohen's $d\rvert$	$\mathbf{r}_k$	Residual vector $y_k - \hat{y}_k$
$\sigma(\cdot)$	Sigmoid function	$\odot$	Element-wise (Hadamard) product
$\Vert$	Concatenation	$\mathbf{I}_N$	Identity matrix

M PSEUDOCODE

The following algorithms specify the complete leakage-controlled CCA-SRL pipeline, from candidate discovery and normality modeling to anomaly-guided graph encoding, evidence fusion, and outer-fold evaluation. All procedures that touch data are executed *inside the outer training fold only*; the outer validation and test folds are scored exactly once by models that never saw them.

M.1 CANDIDATE DISCOVERY AND SELECTION

Algorithm 1: Candidate Discovery, Context Construction, and Selection

// Candidate Mining and Context Construction
// Part 1: Site-Wise Differential Edge Mining
Input: $\mathcal{D}_{\text{tr}} = \{(\mathbf{A}^{(s)}, y^{(s)}, g^{(s)})\}$: outer-training graphs, labels, site tags; $n_{\min}$: minimum group size; E_{top}: edges per site; $[n_{\min}^{\text{nodes}}, n_{\max}^{\text{nodes}}]$: candidate size range
Output: $\mathcal{P}$: candidate pool (list of ROI sets)
// Group training subjects by site
$\mathcal{G} \leftarrow$ group $\mathcal{D}_{\text{tr}}$ by $g^{(s)}$
$\mathcal{E} \leftarrow \emptyset$; $\quad \mathcal{P} \leftarrow [\,]$
// $\mathcal{E}$: edge set (empty); $\mathcal{P}$: candidate list, i.e. a list of ROI sets (empty)
foreach site $g \in \mathcal{G}$ **do**
  $\mathcal{C}_g \leftarrow \{s : g^{(s)} = g,\ y^{(s)} = 0\}$; $\quad \mathcal{A}_g \leftarrow \{s : g^{(s)} = g,\ y^{(s)} = 1\}$
  if $|\mathcal{C}_g| \geq n_{\min}$ **and** $|\mathcal{A}_g| \geq n_{\min}$ **then**
    // Per-edge Cohen's d between the two groups
    $\boldsymbol{\mu}_0 \leftarrow \text{mean}_{s \in \mathcal{C}_g}(\mathbf{A}^{(s)})$
    $\boldsymbol{\mu}_1 \leftarrow \text{mean}_{s \in \mathcal{A}_g}(\mathbf{A}^{(s)})$
    $\mathbf{V}_0 \leftarrow \text{var}_{s \in \mathcal{C}_g}(\mathbf{A}^{(s)})$
    $\mathbf{V}_1 \leftarrow \text{var}_{s \in \mathcal{A}_g}(\mathbf{A}^{(s)})$
    $\mathbf{S} \leftarrow \sqrt{\frac{(n_0-1)\mathbf{V}_0 + (n_1-1)\mathbf{V}_1}{n_0+n_1-2}} + \epsilon$
    $\mathbf{D} \leftarrow (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)/\mathbf{S}$
    $\text{diag}(\mathbf{D}) \leftarrow 0$ *// remove self-loops*

// Top edges by absolute effect size
$E_g \leftarrow \{(i,j) : i < j\}$ sorted by $|\mathbf{D}_{ij}|$, truncated to E_{top}
$\text{score}_i^{\text{node}} \leftarrow \sum_j |\mathbf{D}_{ij}|$ *// node-level effect mass*
// Edges → connected components
foreach component $\mathcal{K}$ of UnionFind(E_g) **do**
  if $|\mathcal{K}| < n_{\min}^{\text{nodes}}$ **then**
    continue
  end if
  if $|\mathcal{K}| > n_{\max}^{\text{nodes}}$ **then**
    $\mathcal{K} \leftarrow$ top-$n_{\max}^{\text{nodes}}$ nodes of $\mathcal{K}$ by $\text{score}^{\text{node}}$
  end if
  $\mathcal{P}$.append($\mathcal{K}$); $\mathcal{E} \leftarrow \mathcal{E} \cup E_g$
end
end if
end
// Cross-site union and deduplication
foreach component $\mathcal{K}$ of UnionFind($\mathcal{E}$) **do**
  if $n_{\min}^{\text{nodes}} \le |\mathcal{K}| \le n_{\max}^{\text{nodes}}$ **then**
    $\mathcal{P}$.append($\mathcal{K}$)
  end if
end
$\mathcal{P} \leftarrow$ deduplicate $\mathcal{P}$ by ROI set
return $\mathcal{P} = \{U_1, \ldots, U_M\}$

// Part 2: Subject-Specific Context Construction
Input: $\mathbf{A}^{(s)} \in \mathbb{R}^{N \times N}$: FC matrix of subject s; U_k: candidate ROI set; C: context size
Output: $C_k^{(s)} \subset \{1, \ldots, N\}$: functional context of U_k for subject s
$\mathcal{O} \leftarrow \{i : i \notin U_k\}$ *// outside-ROI set*
if $|\mathcal{O}| \le C$ **then**
  return $\mathcal{O}$
end if
foreach $i \in \mathcal{O}$ **do**
  $r_i \leftarrow \sum_{j \in U_k} |\mathbf{A}_{ij}^{(s)}|$ *// total coupling to the subgraph*
end
$C_k^{(s)} \leftarrow$ the C indices $i \in \mathcal{O}$ with largest r_i
return $C_k^{(s)}$

// Two-Stage Candidate Selection
Input: $\mathcal{D}_{\text{tr}}$: outer-training graphs; $\mathcal{P}$: candidate pool; K: number of retained candidates
Output: $\mathcal{S} \subset \mathcal{P}$: retained candidates; refitted normality model
// Stage 1: score the whole pool with a temporary model
$\mathcal{M}_{\text{pool}} \leftarrow \textsc{NormalityFit}(\{\mathbf{A}^{(s)} : s \in \mathcal{D}_{\text{tr}}, y^{(s)} = 0\}, \mathcal{P})$
// Alg. 2
foreach $s \in \mathcal{D}_{\text{tr}}$ **do**
  $\{a_k^{(s)}\}_{k=1}^{M} \leftarrow \textsc{NormalityScore}(\mathcal{M}_{\text{pool}}, \mathbf{A}^{(s)}, \mathcal{P})$
  // Alg. 2
end
// Stage 2: rank candidates by cross-group separation of their anomaly scores
for $k \leftarrow 1$ **to** M **do**
  $\mu_k^{\text{ctl}} \leftarrow \text{mean}\{a_k^{(s)} : y^{(s)} = 0\}$
  $\mu_k^{\text{pat}} \leftarrow \text{mean}\{a_k^{(s)} : y^{(s)} = 1\}$
  $\sigma^2_{\text{ctl},k} \leftarrow \text{var}\{a_k^{(s)} : y^{(s)} = 0\}$
  $\sigma^2_{\text{pat},k} \leftarrow \text{var}\{a_k^{(s)} : y^{(s)} = 1\}$
  $D_k \leftarrow \left| (\mu_k^{\text{pat}} - \mu_k^{\text{ctl}}) \Big/ \sqrt{(\sigma^2_{\text{pat},k} + \sigma^2_{\text{ctl},k})/2 + \epsilon} \right|$
end for
$\mathcal{S} \leftarrow$ the K candidates with largest D_k
// Discard the temporary model and refit on the retained candidates only
$\mathcal{M}_{\text{final}} \leftarrow \textsc{NormalityFit}(\{\mathbf{A}^{(s)} : y^{(s)} = 0\}, \mathcal{S})$
return $\mathcal{S}, \mathcal{M}_{\text{final}}$

M.2 NORMALITY MODELING AND SCORING

Algorithm 2: normal Normality Modeling and Scoring

// normal Normality Model Fitting
// Part 1: Training-Pair Construction
Input: $\mathbf{A}^{(s)}$: FC matrix of a training control; U_k: candidate; $C_k^{(s)}$: its context; $d_y = 14$: graph-feature dimension; K: number of candidates
Output: $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{21+K} \times \mathbb{R}^{d_y}$: one regression pair
// Internal state of the subgraph (regression target)
$\mathbf{y} \leftarrow \textsc{GraphFeatures}(\mathbf{A}^{(s)}, U_k) \in \mathbb{R}^{d_y}$
// Context features and subgraph–context interaction
$\mathbf{z}^{\text{ctx}} \leftarrow \textsc{GraphFeatures}(\mathbf{A}^{(s)}, C_k^{(s)}) \in \mathbb{R}^{d_y}$
$\mathbf{z}^{\text{cross}} \leftarrow \textsc{CrossFeatures}(\mathbf{A}^{(s)}, U_k, C_k^{(s)}) \in \mathbb{R}^6$
// Identity and size control variables
$\mathbf{1}_k \leftarrow$ one-hot encoding of candidate index $k \in \{0, 1\}^K$
$\mathbf{x} \leftarrow [\, \mathbf{z}^{\text{ctx}} \parallel \mathbf{z}^{\text{cross}} \parallel |U_k|/N \parallel \mathbf{1}_k \,]$ *// $(21 + K)$*
return $(\mathbf{x}, \mathbf{y})$

// Part 2: normal Normality Model – Fitting
Input: $\{\mathbf{A}^{(s)} : y^{(s)} = 0\}$: training controls; $\mathcal{P}$ (or $\mathcal{S}$): candidate set; Λ: ridge grid; s_Σ: covariance shrinkage; ϵ: jitter
Output: $\mathcal{M} = \{\boldsymbol{\mu}_x, \boldsymbol{\sigma}_x, \boldsymbol{\mu}_y, \boldsymbol{\sigma}_y, \hat{\mathbf{W}}, \hat{\mathbf{b}}, \alpha^\star, \boldsymbol{\Sigma}^{-1}, \log\det \boldsymbol{\Sigma}\}$
// Assemble all control $\times$ candidate pairs into one design matrix
$\mathbf{X} \leftarrow [\,]$; $\mathbf{Y} \leftarrow [\,]$
// design and target matrices, accumulated over all control $\times$ candidate pairs
foreach training control s **and** candidate $U_k \in \mathcal{P}$ **do**
  $(\mathbf{x}, \mathbf{y}) \leftarrow \textsc{TrainingPair}(\mathbf{A}^{(s)}, U_k)$ *// Alg. 2*
  $\mathbf{X}$.append($\mathbf{x}$); $\mathbf{Y}$.append($\mathbf{y}$)
end
// Column-wise standardisation (frozen statistics)
$\boldsymbol{\mu}_x, \boldsymbol{\sigma}_x \leftarrow \text{mean, std}(\mathbf{X}, \text{axis} = 0)$
$\mathbf{X}_z \leftarrow (\mathbf{X} - \boldsymbol{\mu}_x)/\boldsymbol{\sigma}_x$
$\boldsymbol{\mu}_y, \boldsymbol{\sigma}_y \leftarrow \text{mean, std}(\mathbf{Y}, \text{axis} = 0)$
$\mathbf{Y}_z \leftarrow (\mathbf{Y} - \boldsymbol{\mu}_y)/\boldsymbol{\sigma}_y$
// Multi-output ridge regression, α by generalised cross-validation
$\alpha^\star \leftarrow \arg\min_{\alpha \in \Lambda}\ \text{GCV}\big(\mathbf{X}_z, \mathbf{Y}_z, \alpha\big)$
$\hat{\mathbf{W}} \leftarrow (\mathbf{X}_z^\top \mathbf{X}_z + \alpha^\star \mathbf{I})^{-1} \mathbf{X}_z^\top \mathbf{Y}_z$
$\hat{\mathbf{b}} \leftarrow \text{intercept}(\mathbf{X}_z, \mathbf{Y}_z, \alpha^\star)$
// Residual covariance (shrunk, jittered)
$\mathbf{E} \leftarrow \mathbf{Y}_z - \mathbf{X}_z \hat{\mathbf{W}} - \hat{\mathbf{b}}$
$\boldsymbol{\Sigma} \leftarrow s_\Sigma \,\text{cov}(\mathbf{E}) + (1 - s_\Sigma)\,\text{diag}(\text{cov}(\mathbf{E})) + \epsilon\,\mathbf{I}$
if $\text{slogdet}(\boldsymbol{\Sigma}) \le 0$ **then**
  $\boldsymbol{\Sigma} \leftarrow \boldsymbol{\Sigma} + 10^{-4}\mathbf{I}$
end if
$\boldsymbol{\Sigma}^{-1} \leftarrow \text{pinv}(\boldsymbol{\Sigma})$
$\log\det \boldsymbol{\Sigma} \leftarrow \text{slogdet}(\boldsymbol{\Sigma})$
return $\mathcal{M}$

// Normality-Model Scoring
Input: $\mathcal{M}$: frozen normality model; $\mathbf{A}^{(s)}$: FC matrix of *any* subject; $\mathcal{S} = \{U_k\}_{k=1}^K$: candidates
Output: $a_k, \mathbf{y}_k, \hat{\mathbf{y}}_k, \mathbf{r}_k, \mathbf{z}_k^{\text{ctx}}, \mathbf{z}_k^{\text{cross}}$ for $k = 1..K$
foreach $k \leftarrow 1$ **to** K **do**
  if $|U_k| < 3$ **then**
    $a_k \leftarrow$ NaN; **continue** *// invalid candidate*
  end if
  $C_k^{(s)} \leftarrow \textsc{Context}(\mathbf{A}^{(s)}, U_k)$ *// Alg. 1*
  $(\mathbf{x}, \mathbf{y}) \leftarrow \textsc{TrainingPair}(\mathbf{A}^{(s)}, U_k, C_k^{(s)})$ *// Alg. 2*
  $\mathbf{x}_z \leftarrow (\mathbf{x} - \boldsymbol{\mu}_x)/\boldsymbol{\sigma}_x$
  $\mathbf{y}_z \leftarrow (\mathbf{y} - \boldsymbol{\mu}_y)/\boldsymbol{\sigma}_y$

$\hat{\mathbf{y}}_z \leftarrow \hat{\mathbf{W}}^\top \mathbf{x}_z + \hat{\mathbf{b}}$
$\mathbf{r}_z \leftarrow \mathbf{y}_z - \hat{\mathbf{y}}_z$
$a_k \leftarrow \frac{1}{2}\big(\mathbf{r}_z^\top \boldsymbol{\Sigma}^{-1} \mathbf{r}_z + \log\det \boldsymbol{\Sigma} + d_y \log 2\pi\big)$
$\hat{\mathbf{y}}_k \leftarrow \boldsymbol{\sigma}_y \odot \hat{\mathbf{y}}_z + \boldsymbol{\mu}_y$ *// back to original units*
$\mathbf{r}_k \leftarrow \mathbf{y} - \hat{\mathbf{y}}_k$ *// signed deviation direction*
$(\mathbf{z}_k^{\text{ctx}}, \mathbf{z}_k^{\text{cross}}) \leftarrow$ components of $\mathbf{x}$
end
foreach k with $a_k = \text{NaN}$ **do**
    $a_k \leftarrow \text{mean}\{a_{k'} : a_{k'} \neq \text{NaN}\}$ *// NaN fill-in*
end
return evidence bundle $\{a_k, \mathbf{y}_k, \hat{\mathbf{y}}_k, \mathbf{r}_k, \mathbf{z}_k^{\text{ctx}}, \mathbf{z}_k^{\text{cross}}\}_{k=1}^K$

M.3 Subgraph Anomaly Tokenization

Algorithm 3: Subgraph Evidence and Token Construction

// Part 1: Role Membership Construction (in / ctx / bdry)
Input: $\mathbf{A}^{(s)}$: FC matrix; $\mathcal{S}$: retained candidates; C: context size
Output: $\mathbf{M}^{(s)} \in \{0,1\}^{N\times K\times 3}$: role masks [in, ctx, bdry]
$\mathbf{M}^{(s)} \leftarrow \mathbf{0}_{N\times K\times 3}$
foreach $k \leftarrow 1$ **to** K **do**
    $\mathbf{M}^{(s)}[U_k, k, 0] \leftarrow 1$ *// internal role*
    $C_k^{(s)} \leftarrow \textsc{Context}(\mathbf{A}^{(s)}, U_k)$ *// Alg. 1*
    $\mathbf{M}^{(s)}[C_k^{(s)}, k, 1] \leftarrow 1$ *// context role*
    $\mathbf{B} \leftarrow |\mathbf{A}^{(s)}|[U_k, C_k^{(s)}]$ *// $|U_k| \times C$ block*
    $\mathbf{M}^{(s)}[\{i \in U_k : \sum_j \mathbf{B}_{ij} > 0\}, k, 2] \leftarrow 1$ *// boundary: subgraph side*
    $\mathbf{M}^{(s)}[\{j \in C_k^{(s)} : \sum_i \mathbf{B}_{ij} > 0\}, k, 2] \leftarrow 1$ *// boundary: context side*
end
return $\mathbf{M}^{(s)}$

// Part 2: Control-Calibrated Evidence and Gate Features
Input: $\{a_k^{(s)}\}$: anomaly scores; $\{\mathbf{r}_k^{(s)}\}$: residuals; $\{a_k^{(s)} : y^{(s)} = 0\}$: scores of training controls; τ: support threshold
Output: $c_k^{(s)}$: calibrated evidence; $\mathbf{u}^{(s)} \in \mathbb{R}^6$: gate features
$\mathcal{D}_{\text{ctl}} \leftarrow \{a_k^{(s)} : y^{(s)} = 0\}$
foreach subject s **and** candidate k **do**
    $p_k^{(s)} \leftarrow \dfrac{1 + \#\{a \in \mathcal{D}_{\text{ctl}} : a \geq a_k^{(s)}\}}{|\mathcal{D}_{\text{ctl}}| + 1}$
    // empirical upper-tail probability
    $c_k^{(s)} \leftarrow -\log(p_k^{(s)} + \epsilon)$ *// rarity evidence*
end
$\mathbf{u}^{(s)} \leftarrow \big[\ \text{mean}_k(c_k^{(s)}) \parallel \text{top3mean}_k(c_k^{(s)}) \parallel \text{std}_k(c_k^{(s)}) \parallel$
$\max_k(c_k^{(s)}) \parallel \text{mean}_k\|\mathbf{r}_k^{(s)}\| \parallel \text{mean}_k \mathbb{I}[p_k^{(s)} < \tau]\ \big]$
return $\{c_k^{(s)}\}, \{\mathbf{u}^{(s)}\}$

// Part 3: Subgraph Token Construction
Input: evidence bundle $\{\mathbf{y}_k, \mathbf{z}_k^{\text{ctx}}, \hat{\mathbf{y}}_k, \mathbf{r}_k, a_k\}_{k=1}^K$; $d_y = 14$; $d_c = 8$: identity-embedding size; T: token dimension
Output: $\mathbf{t} \in \mathbb{R}^{B\times K\times T}$: subgraph anomaly tokens
$\mathbf{E} \leftarrow \text{Embedding}(K, d_c)$ *// learnable candidate identities*
foreach $k \leftarrow 1$ **to** K **do**
    $\mathbf{f}_k \leftarrow [\ \mathbf{y}_k \parallel \hat{\mathbf{y}}_k \parallel \mathbf{r}_k \parallel \mathbf{z}_k^{\text{ctx}} \parallel$
    $\mathbf{z}_k^{\text{cross}} \parallel a_k \parallel \mathbf{E}_k\ ]$ *// $4d_y + 6 + 1 + d_c = 71$*
    // $4d_y + 6 + 1 + d_c = 71$
    $\mathbf{t}_k \leftarrow \text{Linear}_2\big(\text{Dropout}\big(\text{ReLU}\big(\text{LayerNorm}(\text{Linear}_1(\mathbf{f}_k))\big)\big)\big)$
    // $71 \to 2T \to T$
    $\mathbf{t}_k \leftarrow \text{LayerNorm}(\mathbf{t}_k)$
end

32. **return** $\mathbf{t}$

M.4 Anomaly-Conditioned Signed Graph Attention Network

Algorithm 4: Anomaly-Guided Signed Graph Encoding

1. **// Structural Guidance and Conditioning**
2. **// Part 1: Structural Backbone (shared GCN stem)**
   Input: $\mathbf{X} \in \mathbb{R}^{B\times N\times D}$: node features; $\mathbf{A} \in \mathbb{R}^{B\times N\times N}$: signed FC; H: hidden width
   Output: $\mathbf{H}_{\text{shared}} \in \mathbb{R}^{B\times N\times H}$
3. $\tilde{\mathbf{A}} \leftarrow \mathbf{A} + \mathbf{I}_N$
4. $\tilde{\mathbf{D}} \leftarrow \text{Diag}\big(\sum_j |\tilde{\mathbf{A}}_{:,j}|\big)$
5. $\hat{\mathbf{A}} \leftarrow \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$ *// symmetric normalisation*
6. $\mathbf{H}_{\text{shared}} \leftarrow \text{Dropout}\big(\text{ReLU}(\hat{\mathbf{A}}\,\mathbf{X}\mathbf{W})\big)$
7. **//** $\mathbf{W} \in \mathbb{R}^{D\times H}$
8. **return** $\mathbf{H}_{\text{shared}}$

9. **// Part 2: Role-Aware Subgraph-to-ROI Projection**
   Input: $\mathbf{t} \in \mathbb{R}^{B\times K\times T}$: tokens; $\mathbf{H}_{\text{shared}}$; $\mathbf{M} \in \{0,1\}^{B\times N\times K\times 3}$: role masks; $\{a_k\}$ or $\{c_k\}$: anomaly evidence; $\{|U_k|\}$: candidate sizes; $\mathcal{R} = \{\text{in}, \text{ctx}, \text{bdry}\}$
   Output: $\mathbf{G} \in \mathbb{R}^{B\times N\times H}$: ROI-level anomaly guidance
10. $\mathbf{a}^+ \leftarrow c_k$ **if** calibration enabled **else** $\text{softplus}(a_k)$ *// non-negative evidence*
11. **foreach** role $r \in \mathcal{R}$ **do**
12. $\quad \mathbf{s}^r_{ik} \leftarrow (\mathbf{W}^r_h \mathbf{h}_i)^\top (\mathbf{W}^r_t \mathbf{t}_k) + \gamma_r\, \mathbf{a}^+_k$
13. $\quad\quad -\delta_r \log|U_k| + b_r$
14. $\quad \mathbf{s}^r_{ik} \leftarrow \mathbf{s}^r_{ik} + (1 - \mathbf{M}^r_{ik}) \cdot (-10^9)$ *// mask uncovered candidates*
15. $\quad \boldsymbol{\alpha}^r_{ik} \leftarrow \text{softmax}_k(\mathbf{s}^r_{ik})$
16. $\quad \mathbf{G} \leftarrow \mathbf{G} + \sum_k \boldsymbol{\alpha}^r_{ik} \mathbf{M}^r_{ik}\, \mathbf{W}^r_{\text{out}} \mathbf{t}_k$
17. $\quad$ **// accumulate over roles**
18. **end**
19. $\text{coverage}_i \leftarrow \mathbb{I}\big[\sum_{r,k} \mathbf{M}^r_{ik} > 0\big]$
20. $\mathbf{G} \leftarrow \mathbf{G} \odot \text{coverage}$ *// uncover ROIs receive nothing*
21. **return** $\mathbf{G}$

22. **// Part 3: Edge-Level Anomaly Fields**
   Input: $\mathbf{t} \in \mathbb{R}^{B\times K\times T}$; $\mathbf{M} \in \{0,1\}^{B\times N\times K\times 3}$
   Output: $\mathbf{R}^{\text{in}}, \mathbf{R}^{\text{bdry}} \in \mathbb{R}^{B\times N\times N}$
23. $\mathbf{q} \leftarrow \text{MLP}_q(\mathbf{t}) \in \mathbb{R}^{B\times K}$ *// scalar modulation strength (signed)*
24. $\mathbf{m}^{\text{in}} \leftarrow \mathbf{M}[:,:,:,0]; \quad \mathbf{m}^{\text{ctx}} \leftarrow \mathbf{M}[:,:,:,1]$
25. $\mathbf{A}^{\text{in}} \leftarrow \mathbf{m}^{\text{in}} \odot \mathbf{q}$
26. $\mathbf{A}^{\text{ctx}} \leftarrow \mathbf{m}^{\text{ctx}} \odot \mathbf{q}$ *// broadcast q_k over ROIs*
27. $\mathbf{R}^{\text{in}} \leftarrow \mathbf{A}^{\text{in}}(\mathbf{m}^{\text{in}})^\top$
28. $\mathbf{R}^{\text{bdry}} \leftarrow \mathbf{A}^{\text{in}}(\mathbf{m}^{\text{ctx}})^\top$
29. $\quad\quad + \mathbf{A}^{\text{ctx}}(\mathbf{m}^{\text{in}})^\top$
30. **return** $\mathbf{R}^{\text{in}}, \mathbf{R}^{\text{bdry}}$

31. **// Part 4: FiLM Conditioning via Anomaly Guidance**
   Input: $\mathbf{G}$: ROI-level guidance; $\mathbf{H}_{\text{shared}}$; H_{film}: conditioner width
   Output: $\mathbf{H}_{\text{cond}} \in \mathbb{R}^{B\times N\times H}$
32. $[\boldsymbol{\gamma}, \boldsymbol{\beta}] \leftarrow \text{MLP}_{\text{cond}}(\mathbf{G})$ *// $\rightarrow \mathbb{R}^{B\times N\times 2H}$*
33. **//** $\rightarrow \mathbb{R}^{B\times N\times 2H}$
34. $\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma}[:H]$
35. $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}[H:]$ *// split into gain and shift*
36. $\mathbf{H}_{\text{cond}} \leftarrow (1 + \boldsymbol{\gamma}) \odot \mathbf{H}_{\text{shared}} + \boldsymbol{\beta}$
37. **return** $\mathbf{H}_{\text{cond}}$

38. **// Signed Graph Attention with Field Injection**
   Input: $\mathbf{H}_{\text{cond}}$; $\mathbf{G}$; $\mathbf{A}$: signed FC; $\mathbf{R}^{\text{in}}, \mathbf{R}^{\text{bdry}}$; n_{heads}; H_{att}: attention width
   Output: $\mathbf{H}_{\text{guided}} \in \mathbb{R}^{B\times N\times H}$
39. **// Signed decomposition of the connectivity matrix**

$\mathbf{A}^{+} \leftarrow \mathrm{ReLU}(\mathbf{A}); \quad \mathbf{A}^{-} \leftarrow \mathrm{ReLU}(-\mathbf{A})$
foreach channel $\pm \in \{+, -\}$ **do**
  // Edge attention scores (two independent MLPs)
  $\mathbf{e}_{ij}^{\pm} \leftarrow f_{\pm}\big([\,\mathbf{h}_i \,\|\, \mathbf{h}_j \,\|\, \mathbf{g}_i \,\|\, \mathbf{g}_j$
    $\|\, \mathbf{A}_{ij}^{\pm}\,]\big)$
    $+\phi_{\text{in}}^{\pm}(\mathbf{R}_{ij}^{\text{in}}) + \phi_{\text{bdry}}^{\pm}(\mathbf{R}_{ij}^{\text{bdry}})$
  $\mathbf{e}_{ij}^{\pm} \leftarrow \mathbf{e}_{ij}^{\pm} + \big(1 - \mathbb{I}[\mathbf{A}_{ij}^{\pm} > 0]\big) \cdot (-10^{9})$
  // mask non-edges
  $\boldsymbol{\alpha}_{ij}^{\pm} \leftarrow \mathrm{softmax}_j(\mathbf{e}_{ij}^{\pm})$
  $\boldsymbol{\alpha}^{\pm} \leftarrow \mathrm{Dropout}(\boldsymbol{\alpha}^{\pm})$
  // Message aggregation
  $\mathbf{V}^{\pm} \leftarrow \mathbf{W}^{\pm}\mathbf{H}_{\text{cond}}$
  $\mathbf{m}_i^{\pm} \leftarrow \sum_j \boldsymbol{\alpha}_{ij}^{\pm}\mathbf{V}_j^{\pm}$
end
// Signed residual update and normalisation
$\mathbf{H}_{\text{guided}} \leftarrow \mathrm{LayerNorm}\big(\mathbf{H}_{\text{cond}} + \beta^{+}\mathbf{m}^{+} - \beta^{-}\mathbf{m}^{-}\big)$
return $\mathbf{H}_{\text{guided}}$

M.5 Calibrated Evidence Fusion and Classification

Algorithm 5: Evidence Fusion and Classification

// Part 1: Statistical Readouts and Residual Fusion
Input: $\mathbf{H}_{\text{shared}}, \mathbf{H}_{\text{guided}}$; $\mathbf{A}$: for the valid-ROI mask; $\mathbf{u}$: gate features; $\{a_k\}$: anomaly scores
Output: $\mathbf{z} \in \mathbb{R}^{B\times 2H}$; summary statistics $\mathbf{p} \in \mathbb{R}^{B\times 3}$
$\mathbf{v}_i \leftarrow \mathbb{I}\big[\sum_j |\mathbf{A}_{ij}| > \epsilon\big]; \quad n_v \leftarrow \sum_i \mathbf{v}_i$ *// valid ROIs*
$\mathbf{z}_{\text{base}} \leftarrow \big[\, \frac{1}{n_v}\sum_{i:\mathbf{v}_i=1} \mathbf{h}_i^{\text{shared}} \,\big\|$
  $\max_{i:\mathbf{v}_i=1} \mathbf{h}_i^{\text{shared}}\,\big]$
$\mathbf{z}_{\text{guided}} \leftarrow \big[\, \frac{1}{n_v}\sum_{i:\mathbf{v}_i=1} \mathbf{h}_i^{\text{guided}} \,\big\|$
  $\max_{i:\mathbf{v}_i=1} \mathbf{h}_i^{\text{guided}}\,\big]$
$\boldsymbol{\rho} \leftarrow \sigma\big(\mathrm{MLP}_{\rho}(\mathbf{u})\big)$ *// control-calibrated reliability gate*
$\mathbf{z} \leftarrow \mathbf{z}_{\text{base}} + \boldsymbol{\rho} \odot \big(\mathbf{W}_g\mathbf{z}_{\text{guided}}\big)$ *// residual fusion*
$\mathbf{p} \leftarrow [\,\mathrm{mean}_k(a_k) \,\|\, \mathrm{top3mean}_k(a_k) \,\|\, \max_k(a_k)\,]$
return $\mathbf{z}, \mathbf{p}$

// Part 2: Guided-Primary Classifier
Input: $\mathbf{z} \in \mathbb{R}^{B\times 2H}$: fused representation; $\mathbf{p} \in \mathbb{R}^{B\times 3}$: NLL statistics
Output: $\hat{s} \in \mathbb{R}^{B}$: class logit
$\mathbf{z}_{\text{in}} \leftarrow [\,\mathbf{z} \,\|\, \mathbf{p}\,]$ *// $(2H+3)$*
$\mathbf{h} \leftarrow \mathrm{Dropout}\big(\mathrm{ReLU}\big(\mathrm{BN}\big(\mathrm{Linear}_1(\mathbf{z}_{\text{in}})\big)\big)\big)$
// $2H{+}3 \to 64$
$\mathbf{h} \leftarrow \mathrm{Dropout}\big(\mathrm{ReLU}\big(\mathrm{BN}\big(\mathrm{Linear}_2(\mathbf{h})\big)\big)\big)$ *// $64 \to 32$*
$\hat{s} \leftarrow \mathrm{Linear}_{\text{out}}(\mathbf{h})$ *// $32 \to 1$*
return $\hat{s}$

M.6 Training and Evaluation

Algorithm 6: Training and Cross-Validation Protocol

// Training One Outer Fold
Input: $\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{val}}$: training/validation loaders (balanced sampling, batch size B); γ_{focal}; $\eta_{\max}$; $E_{\max}$; P: patience; $E_{\min}$: minimum epochs; metric $\in \{\text{AUC}, \text{ACC}\}$
Output: best checkpoint $\theta^{\star}$; validation metric value
$\theta \leftarrow$ initialise (FiLM conditioner last layer zero-initialised, so $\boldsymbol{\gamma} = \boldsymbol{\beta} = \mathbf{0}$)
$m_{\text{best}} \leftarrow -\infty; \quad \text{bad} \leftarrow 0$
for $e \leftarrow 1$ **to** $E_{\max}$ **do**
  foreach mini-batch $(\mathcal{B})$ drawn by the balanced sampler **do**

$\hat{s} \leftarrow \textsc{Forward}(\mathcal{B}; \theta)$ *// Alg. 3–5*
BCE $\leftarrow -\big[y \log \sigma(\hat{s}) + (1-y)\log(1-\sigma(\hat{s}))\big]$
$\mathcal{L} \leftarrow \text{mean}\big[(1-e^{-\text{BCE}})^{\gamma_{\text{focal}}}\,\text{BCE}\big]$ *// focal loss*
$\theta \leftarrow \text{AdamW}\big(\theta, \nabla_\theta \mathcal{L}, \eta_e\big)$ with $\|\nabla_\theta \mathcal{L}\|_2 \le 5$
end
$\eta_e \leftarrow \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\big(1 + \cos(\pi e / E_{\max})\big)$ *// cosine annealing*
$m \leftarrow \textsc{Evaluate}(\theta, \mathcal{D}_{\text{val}}, \text{metric})$
if $m > m_{\text{best}}$ **then**
$m_{\text{best}} \leftarrow m$; $\theta^\star \leftarrow \text{clone}(\theta)$; bad $\leftarrow 0$
else
bad $\leftarrow$ bad $+ 1$
end if
if $e \ge E_{\min}$ **and** bad $\ge P$ **then break** *// early stopping*
end for
$\theta \leftarrow \theta^\star$ *// restore the best checkpoint*
return $\theta^\star, m_{\text{best}}$

// Threshold Selection and Complete Cross-Validation
// Part 1: Threshold Selection and Evaluation
Input: $\theta^\star$; $\mathcal{D}_{\text{val}}, \mathcal{D}_{\text{te}}$: validation/test folds; mode $\in \{\text{ACC}, \text{BalACC}\}$
Output: AUC, ACC, BalACC, SEN, SPEC on the test fold
$\{\hat{p}^{(s)}\}_{s \in \mathcal{D}_{\text{val}}} \leftarrow \sigma\big(\textsc{Forward}(\cdot; \theta^\star)\big)$
// validation probabilities
$\mathcal{T} \leftarrow$ 200 equally spaced thresholds in $[\min_s \hat{p}^{(s)}, \max_s \hat{p}^{(s)}]$
$\tau^\star \leftarrow \arg\max_{\tau \in \mathcal{T}} \text{metric}_{\text{mode}}\big(\{\hat{p}^{(s)} \ge \tau\}, \{y^{(s)}\}\big)$
$\{\hat{p}^{(s)}\}_{s \in \mathcal{D}_{\text{te}}} \leftarrow \sigma\big(\textsc{Forward}(\cdot; \theta^\star)\big)$
$\hat{y}^{(s)} \leftarrow \mathbb{I}[\hat{p}^{(s)} \ge \tau^\star]$ *// test predictions*
return AUC, ACC, BalACC, SEN, SPEC computed on $\{\hat{y}^{(s)}\}, \{y^{(s)}\}$

// Part 2: Complete Cross-Validation Protocol
Input: $\mathcal{D}$: cohort (FC matrices, labels, site tags); $S = 4$: folds; seed
Output: per-fold and averaged metrics
$\{(\mathcal{D}_{\text{tr}}^{(i)}, \mathcal{D}_{\text{val}}^{(i)}, \mathcal{D}_{\text{te}}^{(i)})\}_{i=1}^{S}$
$\leftarrow \textsc{StratifiedFold}(\mathcal{D}, S, \text{seed})$
// val ratio 0.25 **inside the training part**
for $i \leftarrow 1$ **to** S **do**
$\mathcal{P} \leftarrow \textsc{SiteEdgeMining}(\mathcal{D}_{\text{tr}}^{(i)})$ *// Alg. 1*
$\mathcal{S}, \mathcal{M}_{\text{final}} \leftarrow \textsc{TwoStageSelection}(\mathcal{D}_{\text{tr}}^{(i)}, \mathcal{P})$
// Alg. 1
evidence $\leftarrow \textsc{NormalityScore}(\mathcal{M}_{\text{final}}, \cdot)$ for all subjects *// Alg. 2*
$\mathbf{M} \leftarrow \textsc{RoleMembership}(\cdot)$
$\{c_k\}, \{\mathbf{u}\} \leftarrow \textsc{CalibratedEvidence}(\cdot)$ *// Alg. 3*
$\theta^\star \leftarrow \textsc{Train}(\mathcal{D}_{\text{tr}}^{(i)}, \mathcal{D}_{\text{val}}^{(i)})$ *// Alg. 6*
$\text{metrics}_i \leftarrow \textsc{ThresholdAndEvaluate}(\theta^\star,$
$\mathcal{D}_{\text{val}}^{(i)}, \mathcal{D}_{\text{te}}^{(i)})$ *// Alg. 6*
end for
return $\{\text{metrics}_i\}_{i=1}^{S}$ and their mean $\pm$ std