POPULATION-REGULARIZED LEARNING OF INDIVIDUALIZED BRAIN EFFECTIVE CONNECTIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain effective connectivity (EC) characterizes directed interactions among brain regions, but individualized EC estimation remains challenging with limited and noisy resting-state functional magnetic resonance imaging (rs-fMRI) data. Although population-level connectivity patterns shared among related subjects can provide stable guidance, existing population-informed methods rely largely on global priors that overlook pairwise inter-subject relevance. We propose **PR-EC**, a population-regularized framework for individualized EC estimation. PR-EC constructs a multimodal similarity graph that encodes pairwise inter-subject relevance, and aggregates the EC estimates of each target's most relevant neighbors into a subject-specific population reference. This reference provides stable population guidance and is used to refine each target subject's EC estimate. Because data-quality variations can affect EC edges differently, PR-EC further introduces a quality-control (QC)-aware refinement that adaptively modulates the contribution of the population reference to each edge. Theoretical analysis characterizes the trade-off between leveraging cross-subject information and preserving individual-specific connectivity. Extensive experiments on three datasets show that PR-EC improves EC-based disease classification. Cross-cohort analyses further reveal reproducible autism spectrum disorder (ASD)-related EC alterations as biomarkers, including bidirectionally enhanced thalamo-temporal connectivity.

1 INTRODUCTION

Brain effective connectivity (EC) characterizes the directed influence that one brain region exerts on another (Friston et al., 2003). Estimating EC from neuroimaging time series data, particularly resting-state functional magnetic resonance imaging (rs-fMRI) (Friston et al., 2014), helps reveal how distributed brain regions interact to support human cognition and behavior. Disease-associated EC patterns have also been investigated across neurological (Kahan et al., 2014) and psychiatric disorders (Zhou et al., 2018), highlighting their potential as biomarkers for early detection and treatment, particularly in conditions such as Alzheimer's disease (Huang et al., 2022).

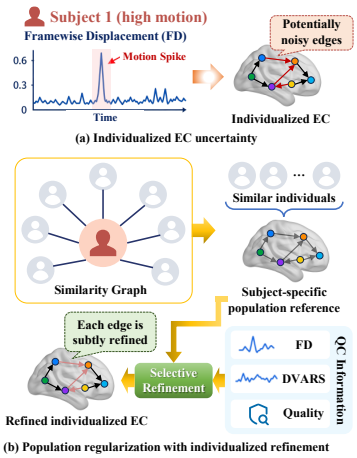


Figure 1: Motivation for PR-EC.

Existing EC estimators broadly include model-based approaches, such as dynamic causal modeling, and data-driven approaches, ranging from Granger causality (Roebroeck et al., 2005) to recent deep and spatiotemporal models (Lu et al., 2022; Zou et al., 2024; Zhang et al., 2024; Xiong et al., 2025). Despite promising results, most existing methods infer subject-specific EC solely from an individual's rs-fMRI time series, without leveraging common brain connectivity patterns shared across subjects. Such subject-only estimation becomes unstable when neuroimaging scans are short or noisy, as limited reliable observations make directed dependencies difficult to recover. In severe cases like data corruption, e.g., head-motion spikes in framewise displacement (FD), imaging signals often distort and cause subject-only methods to infer spurious EC edges (Fig. 1(a)). Existing neuroimaging studies have identified reproducible **population-level commonalities in brain connectivity** (Yeo et al., 2011; Seeley et al., 2009). For example, compared

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.

with age-matched healthy controls, patients with mild Alzheimer's disease commonly exhibit disrupted hippocampal connectivity with default-mode regions, particularly the posterior cingulate and medial prefrontal cortices (Greicius et al., 2004; Wang et al., 2006). Such population-level regularities suggest that brain connectivity information from related subjects can provide a stable reference for subject-specific EC estimation. We refer to brain connectivity information derived from subjects in the population as **population connectivity information**, which motivates us to incorporate it into individualized EC estimation.

Prior work has made initial attempts to incorporate population connectivity information into EC estimation, using hierarchical Bayesian models or empirical population priors (Chiang et al., 2017; Zeidman et al., 2019). However, these global priors summarize cohort-wide connectivity patterns without explicitly modeling pairwise subject similarity, limiting their ability to select the most relevant information to each target. Population similarity graphs offer a natural alternative by representing subjects as nodes and their similarities as edges, yet remain underexplored for subject-specific EC estimation. The effectiveness of the similarity graph depends critically on its ability to accurately capture each subject's relevance to a given target. However, phenotypic information has been shown to be associated with brain connectivity (Tsvetanov et al., 2016; Lund et al., 2020). Non-diagnostic phenotypic information, such as demographic attributes (e.g., age and sex) and cognitive measures (e.g., memory), provides a complementary view of subject similarity beyond rs-fMRI-derived functional connectivity (FC) alone (Jun et al., 2020; Lin et al., 2023). We therefore construct a multimodal population similarity graph by integrating FC with non-diagnostic phenotypic information. Using this multimodal population similarity graph, we aggregate EC estimates from subjects similar to each target into a **subject-specific population reference**, which is then used to refine the target's EC estimate (Fig. 1(b)). However, uniformly applying this reference across connectivity edges may be suboptimal, as data corruption can affect individual edges differently. Quality-control (QC) measures, such as framewise displacement (FD) and derivative of root mean square variance over voxels (DVARS), characterize data-quality variation, and their associations with connectivity estimates can vary across edges (Power et al., 2012; 2014; Ciric et al., 2017). We therefore use QC information to adaptively modulate the contribution of the population reference to each edge, yielding a QC-guided selective refinement process (Fig. 1(b)).

We propose **PR-EC**, a **P**opulation-**R**egularized framework for individualized EC estimation that combines a population reference inferred from a multimodal similarity graph with QC-guided selective refinement. We provide a theoretical analysis of the trade-off between incorporating population connectivity information and preserving individual connectivity. Under an idealized model, we show that appropriately attenuating data-quality-related perturbations can reduce EC estimation error.

Contributions. Methodologically, we make the first attempt at selectively leveraging population connectivity information for individualized EC estimation under limited rs-fMRI time series. From a neuroscientific perspective, we identify reproducible autism spectrum disorder (ASD)-related alterations in directed brain connectivity across cohorts. Specifically, **(i)** we propose a population-regularized framework (PR-EC) that constructs an inter-subject similarity graph from multimodal information to derive a subject-specific population reference; **(ii)** we introduce a QC-aware self-supervised refinement network that performs edge-wise selective refinement guided by quality information; **(iii)** extensive experiments on three datasets demonstrate the downstream utility of PR-EC-derived individualized EC representations; **(iv)** cross-cohort analyses of ASD-related datasets reveal reproducible ASD-related directed connectivity alterations, most notably bidirectionally enhanced thalamo-temporal connectivity consistent with prior neuroimaging findings.

2 Related Work

EC Estimation. Conventional EC estimation broadly follows model-based and data-driven paradigms. Model-based methods largely rely on predefined dynamical models, where dynamic causal modeling (DCM) is a representative method (Friston et al., 2003) for inferring directed brain connectivity. Subsequent extensions of DCM have broadened its applicability to resting-state (Razi et al., 2015), large-scale (Razi et al., 2017), and computationally scalable EC inference (Frässle et al., 2017; 2021). Data-driven methods, instead, directly infer directed connectivity from observed time series, with Granger causality as a classical example (Roebroeck et al., 2005). Recently, deep learning has been increasingly applied to the estimation and pattern discovery of ECs, including

approaches based on deep reinforcement learning (Lu et al., 2022), spatiotemporal graph convolution (Zou et al., 2024), meta-reinforcement learning (Zhang et al., 2024) and Fourier spatiotemporal attention (Xiong et al., 2025). Despite their different formulations, these methods primarily derive each individualized EC estimate from the target subject's own time-series observations, without explicitly incorporating population connectivity information from related individuals. In contrast, our work leverages EC information from related subjects to construct a subject-specific population reference for regularizing individualized EC estimation.

Population-Informed Connectivity Modeling. An important direction in neuroimaging is to leverage *cross-subject information*, i.e., information derived from observations or relationships across multiple subjects, to improve individualized modeling. In EC analysis, hierarchical Bayesian and empirical Bayes approaches jointly model subject-specific connectivity parameters and shared population structure (Friston et al., 2016; Zeidman et al., 2019), allowing population connectivity information to regularize individual parameter estimates (Chiang et al., 2017). These studies demonstrate the effectiveness of such regularization for individualized EC estimation. Another line of work exploits inter-subject similarity for brain disorder prediction. Population-graph methods integrate imaging information with phenotypic or other non-imaging subject characteristics to construct or learn inter-subject similarity graphs, enabling information sharing across subjects (Parisot et al., 2018; Pan et al., 2022; Chen et al., 2024). These works highlight the value of subject similarity for identifying relevant cross-subject information. These findings motivate us to use inter-subject similarity to guide selective EC aggregation, thereby incorporating relevant population connectivity information into individualized EC estimation.

Quality-Aware Connectivity Refinement. QC measures such as FD and DVARS are commonly used to characterize data-quality-related variation in rs-fMRI (Power et al., 2012). Connectivity estimates from rs-fMRI are sensitive to data quality, with head motion known to introduce systematic biases in functional connectivity (Van Dijk et al., 2012). Beyond conventional preprocessing and motion correction, recent studies have explored learning-based strategies for improving robustness to noise. BrainMAE employs self-supervised masked reconstruction to learn robust representations from fMRI time series (Yang et al., 2024), whereas BDHT incorporates denoising diffusion with structural connectivity guidance for EC estimation (Zuo et al., 2025). These studies highlight the value of explicitly accounting for data quality and noise. Moreover, the associations between QC measures and connectivity can vary between edges, reflecting the edge variation associated with QC (Ciric et al., 2017). However, these methods do not consider such variation in conjunction with population connectivity information for individualized EC refinement.

3 Problem Formulation

To address the aforementioned limitations, we formulate population-informed and quality-aware EC estimation, where EC information from similar subjects is selectively aggregated into a subject-specific reference EC and used to refine connectivity edges according to their sensitivity to data-quality variation measured by QC. Formally, consider a population of N subjects. For each subject i, we observe a d-dimensional multivariate time series $X_i \in \mathbb{R}^{T_i \times d}$ and an individual-level covariate vector $c_i \in \mathbb{R}^p$, with the goal of estimating an individualized directed structure $A_i^\star \in \mathbb{R}^{d \times d}$. In the rs-fMRI setting considered in this work, X_i denotes the region of interest (ROI) time series, c_i contains phenotypic covariates, including demographic and cognitive measures, and $A_i^\star$ represents the ideal individualized EC. Let $\mathcal{E}_\theta$ denote a shared structure estimator. For subject i, the initial individualized EC is obtained from its own time series as $\widetilde{A}_i = \mathcal{E}_\theta(X_i)$. When T_i is limited or the observations are noisy, $\widetilde{A}_i$ may exhibit substantial uncertainty (Varoquaux et al., 2010; Ma et al., 2024; Frässle & Stephan, 2022). To incorporate population connectivity information, we additionally consider phenotypic and functional information from the population. Specifically, we define the population descriptor $z_i = [c_i, \phi(X_i)]$, where c_i represents non-diagnostic phenotypic covariates (e.g., age and sex) and $\phi(X_i)$ represents vectorized Pearson functional connectivity matrix computed from the observed ROI time series X_i (Finn et al., 2015; Smith et al., 2015). The resulting descriptor z_i is used to identify subjects similar to subject i in the population.

We assume that the ideal ECs of neighboring subjects satisfy a local smoothness condition in the descriptor space (Belkin et al., 2006).

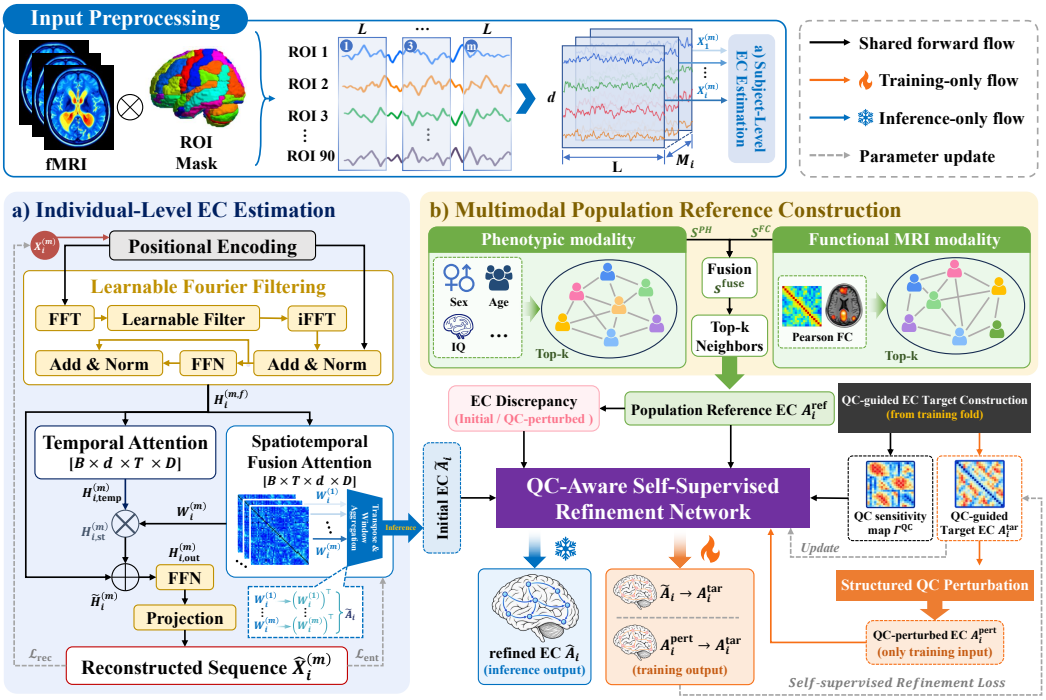


Figure 2: Overview of the PR-EC framework.

Assumption 3.1 (Local Structural Smoothness)**.** *For the local neighborhood $\mathcal{N}_K(i)$ of subject i, there exists a constant $L_A > 0$ such that*

$$\left\|A_i^\star - A_j^\star\right\|_F \le L_A\, \rho(z_i, z_j), \qquad j \in \mathcal{N}_K(i), \tag{1}$$

where $A_i^\star$ and $A_j^\star$ denote the ideal individualized ECs of subjects i and j; z_i and z_j denote their population descriptors; $\rho(\cdot,\cdot)$ is a nonnegative dissimilarity measure in the descriptor space; and $L_A > 0$ is a local smoothness constant relating descriptor dissimilarity to EC differences.

This assumption allows subject-specific structures to remain distinct while imposing local regularity among similar subjects in the descriptor space. Further theoretical motivation is provided in Appendix A.1. Accordingly, we study **population-regularized individualized structure learning** under this local smoothness principle.

4 Method

In this section, we introduce PR-EC (Fig. 2), a population-regularized framework for individualized EC estimation from multivariate rs-fMRI time series. PR-EC consists of three core modules: (1) **Individual-Level EC Estimation** for deriving an initial individualized EC; (2) **Multimodal Population Reference Construction** for constructing a subject-specific population reference; and (3) **QC-Aware Self-Supervised Refinement** for selectively refining the initial individualized EC using QC information and the population reference. We detail these modules below.

4.1 Individual-Level EC Estimation

Windowed Input Preprocessing. For the ROI time series X_i of subject i, we first independently apply z-score normalization (Zhao et al., 2022) to each ROI along the temporal dimension to reduce differences in signal baseline and scale. Since the effective sequence length T_i may vary across subjects in multi-site data, we partition the sequence into overlapping temporal windows of length L with stride H (Allen et al., 2014) to obtain fixed-length local dynamic segments for subsequent modeling. The m-th window is defined as $X_i^{(m)} = X_i[(m-1)H + 1 : (m-1)H + L, :] \in \mathbb{R}^{L\times d}$, where $m = 1, \ldots, M_i$ and $M_i = \lfloor (T_i - L)/H \rfloor + 1$ denotes the number of windows for subject i. These fixed-length local segments are then used as inputs for Individual-Level EC Estimation.

Spatiotemporal Estimator. Given the fixed-length temporal windows $\{X_i^{(m)}\}_{m=1}^{M_i}$ of subject i, we instantiate the shared estimator $\mathcal{E}_\theta$ in Sec. 3 as a self-supervised Fourier-enhanced spatiotemporal estimator (Fig. 2(a)), inspired by the spatiotemporal modeling strategy (Xiong et al., 2025).

For the m-th window, feature embedding and temporal positional encoding first produce $H_i^{(m,0)} \in \mathbb{R}^{L\times d\times D}$. We then apply learnable Fourier filtering to obtain $H_i^{(m,f)} = \mathcal{F}^{-1}[\Phi \odot \mathcal{F}(H_i^{(m,0)})]$, where Φ is a learnable complex-valued frequency response. Next, temporal self-attention produces $H_{i,\text{temp}}^{(m)} = \text{TempAttn}(H_i^{(m,f)})$, capturing intra-ROI temporal dependencies. In parallel, spatiotemporal fusion attention (STFA) captures asymmetric inter-regional relations. Let $S_{i,t,h}^{(m)}$ denote its attention matrix at temporal position t and head h. We aggregate these relations as $W_i^{(m)} = (LH_a)^{-1}\sum_{t=1}^{L}\sum_{h=1}^{H_a} S_{i,t,h}^{(m)}$, where H_a is the number of attention heads. Detailed attention computations underlying temporal self-attention and STFA are provided in Appendix B.1. The temporal representation and learned spatial relations are then fused for sequence reconstruction, where $\text{FFN}(\cdot)$ denotes a feed-forward network and $\text{Proj}(\cdot)$ denotes the output projection:

$$\widehat{X}_i^{(m)} = \text{Proj}\left[\text{FFN}\left(H_i^{(m,f)} + \text{Fuse}\left(H_{i,\text{temp}}^{(m)}, W_i^{(m)}\right)\right)\right]. \tag{2}$$

Self-supervised Pretraining: During pretraining, $W_i^{(m)}$ serves as an internal representation of inter-regional relations and participates in sequence reconstruction. The estimator is trained with

$$\mathcal{L}_{\text{rec}} = \frac{1}{Ld\sum_i M_i}\sum_i\sum_{m=1}^{M_i}\left\|\widehat{X}_i^{(m)} - X_i^{(m)}\right\|_F^2, \qquad \mathcal{L}_{\text{EC}} = \mathcal{L}_{\text{rec}} + \alpha\mathcal{L}_{\text{ent}}. \tag{3}$$

The parameters are optimized as $\theta^\star = \arg\min_\theta \mathcal{L}_{\text{EC}}$, where $\mathcal{L}_{\text{rec}}$ trains the inter-regional relations through their contribution to sequence reconstruction, while $\mathcal{L}_{\text{ent}}$ discourages diffuse attention and promotes more selective relations underlying $W_i^{(m)}$ through entropy regularization (Zhang et al., 2019), with α controlling its strength.

EC Inference: After pretraining, the estimator is fixed and applied to all temporal windows of each subject. Following previous work (Xiong et al., 2025), the resulting relation matrix for each window is transposed as $A_i^{(m)} = (W_i^{(m)})^\top$, and then aggregated across windows to obtain the initial individualized EC: $\widetilde{A}_i = \frac{1}{M_i}\sum_{m=1}^{M_i} A_i^{(m)}$. Equivalently, as defined in Sec. 3, $\widetilde{A}_i = \mathcal{E}_{\theta^\star}(X_i)$, with its diagonal entries set to zero.

4.2 Multimodal Population Reference Construction

Motivated by Assumption 3.1, we introduce a multimodal population reference (MPR) by retrieving locally related subjects. Specifically, we characterize inter-subject similarity from two complementary views: phenotypic covariates (PH) view and functional connectivity (FC) view (Parisot et al., 2018; Rakhimberdina et al., 2020). The **phenotypic affinity** is derived from non-diagnostic covariates c_i, while the **functional affinity** is derived from the functional descriptor $\phi(X_i)$ in Sec. 3. These two views operationalize inter-subject similarity in the descriptor space defined by $z_i = [c_i, \phi(X_i)]$.

Let s_{ij}^{PH} and s_{ij}^{FC} denote phenotypic and functional affinities, respectively (Fig. 2(b)). For each view, we retain the top-K affinities and perform row normalization to obtain a sparse subject-to-subject affinity graph, with larger values indicating greater similarity. We then fuse the two view-specific affinities as $s_{ij}^{\text{fuse}} = (1-\beta)s_{ij}^{\text{PH}} + \beta s_{ij}^{\text{FC}}$, where $\beta \in [0,1]$ controls their relative contributions, and define the corresponding dissimilarity measure in Assumption 3.1 as $\rho(z_i, z_j) = 1 - s_{ij}^{\text{fuse}}$. For each subject i, we further retain the K largest fused affinities and normalize them to obtain the final reference weights:

$$w_{ij} = \frac{s_{ij}^{\text{fuse}}}{\sum_{k\in\mathcal{N}_K(i)} s_{ik}^{\text{fuse}}}, \qquad A_i^{\text{ref}} = \sum_{j\in\mathcal{N}_K(i)} w_{ij}\widetilde{A}_j, \qquad w_{ij}\geq 0, \quad \sum_{j\in\mathcal{N}_K(i)} w_{ij} = 1. \tag{4}$$

The reference A_i^{ref} aggregates locally related EC estimates as complementary connectivity evidence. Further details are provided in Appendix B.2. To capture this reference, we analyze its risk under Assumption 3.1. Conditioning on the retrieved neighborhood and weights, assume $\widetilde{A}_j = A_j^\star + \varepsilon_j$, where $\mathbb{E}[\varepsilon_j] = 0$, $\mathbb{E}\|\varepsilon_j\|_F^2 \leq \sigma_j^2$, and $\mathbb{E}\langle\varepsilon_j, \varepsilon_k\rangle_F = 0$ for $j \neq k$. Then:

Proposition 4.1 (Risk Bound for Local Population Reference)**.** *Under the local structural smoothness assumption, the population reference satisfies*

$$\mathbb{E}\|A_i^{\mathrm{ref}} - A_i^{\star}\|_F^2 \le L_A^2 \left(\sum_{j \in \mathcal{N}_K(i)} w_{ij} \rho(z_i, z_j) \right)^2 + \sum_{j \in \mathcal{N}_K(i)} w_{ij}^2 \sigma_j^2. \tag{5}$$

This bound reveals a **structural mismatch–variance trade-off** in population borrowing, where less similar neighbors introduce greater structural mismatch. The first term increases with this mismatch, whereas the second can be reduced through aggregation; for K equally weighted neighbors with common noise level σ^2, the variance term reduces to σ^2/K. This trade-off motivates the construction of a local population reference weighted by similarity rather than averaging indiscriminately across subjects. A detailed proof is provided in Appendix A.2. Accordingly, we use A_i^{ref} as complementary connectivity evidence to refine, rather than replace, the subject-specific EC estimate.

4.3 QC-Aware Self-Supervised Refinement

While A_i^{ref} provides reference EC information, it does not explicitly account for QC-associated variation. As discussed in Sec. 1, data-quality variation can differentially affect individual EC edges. We therefore introduce **QC**-Aware **S**elf-Supervised **R**efinement (QSR) to learn QC-sensitive edge-wise refinement. The overall QSR pipeline is illustrated in Fig. 5 and is detailed below. Implementation and optimization details are provided in Appendices B.3 and B.4.

QC-guided EC Target Construction. We use three individual-level QC measures ($M_q = 3$): mean FD, DVARS (Power et al., 2012; 2014), and the fraction of low-motion frames with FD $\le$ 0.2 mm (Quality). Within each cross-validation fold, missing values are imputed using training-fold medians, and each measure is log-transformed and robustly normalized by the training-fold median and interquartile range (IQR), with negative normalized values clipped to zero, yielding the QC profile $q_i = [q_{i,1}, \ldots, q_{i,M_q}]^\top$. For each directed edge (u, v), we fit ridge regression (Hoerl & Kennard, 1970) over the training subjects, using $\widetilde{A}_{i,uv}$ as the response and q_i together with non-diagnostic confounders as predictors. The QC-associated coefficients form population-level patterns $\mathcal{B} = \{B_r\}_{r=1}^{M_q}$, where B_r captures the confound-adjusted association between QC measure r and edge-wise EC variation. We then define the subject-specific QC-associated EC component C_i^{QC}, the QC-guided target A_i^{tar}, and the population-level QC sensitivity map Γ^{QC} as

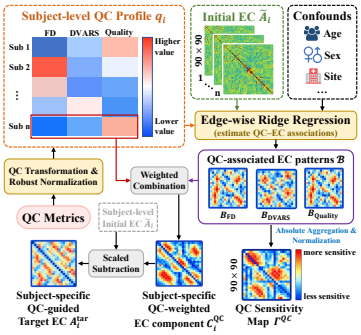


Figure 3: QC-guided EC target construction.

$$C_i^{\mathrm{QC}} = \sum_{r=1}^{M_q} q_{i,r} B_r, \qquad A_i^{\mathrm{tar}} = \widetilde{A}_i - \eta C_i^{\mathrm{QC}}, \qquad \Gamma_{uv}^{\mathrm{QC}} = \mathbb{I}(u \neq v) \frac{\sum_{r=1}^{M_q} |B_{r,uv}|}{\max_{a \neq b} \sum_{r=1}^{M_q} |B_{r,ab}|}. \tag{6}$$

η controls the QC-guided adjustment, while Γ^{QC} measures edge-wise QC sensitivity from the absolute coefficient magnitudes, irrespective of association direction. The procedure is shown in Fig. 3.

Proposition 4.2 (**Idealized Error Reduction under Additive QC Variation**)**.** *Treating C_i^{QC} as fixed for subject i, consider the idealized additive model $\widetilde{A}_i = A_i^{\star} + C_i^{\mathrm{QC}} + \varepsilon_i$, where $\mathbb{E}[\varepsilon_i] = 0$.*

$$\begin{aligned} \mathbb{E}\left\|A_i^{\mathrm{tar}} - A_i^{\star}\right\|_F^2 &= (1-\eta)^2 \left\|C_i^{\mathrm{QC}}\right\|_F^2 + \mathbb{E}\left\|\varepsilon_i\right\|_F^2, \\ \mathbb{E}\left\|\widetilde{A}_i - A_i^{\star}\right\|_F^2 - \mathbb{E}\left\|A_i^{\mathrm{tar}} - A_i^{\star}\right\|_F^2 &= \eta(2-\eta) \left\|C_i^{\mathrm{QC}}\right\|_F^2. \end{aligned} \tag{7}$$

A detailed proof of Proposition 4.2 is provided in Appendix A.3. This proposition shows that, under the idealized additive model with $0 < \eta < 2$ and $C_i^{\mathrm{QC}} \neq 0$, the QC-guided adjustment reduces the expected squared error. This provides an idealized theoretical justification for the construction of the QC-guided target A_i^{tar} and supports its use as the target for subsequent self-supervised refinement.

Structured QC Perturbation. Since the true QC-free EC is not directly observable, we construct structured self-supervised perturbations using $\mathcal{B}$ (Fig. 4). Inspired by corruption-and-restoration strategies in denoising representation learning (Vincent et al., 2008), we randomly sample two distinct QC profiles q_a and q_b from the training population. The QC-perturbed EC A_i^{pert} is defined as follows:

$$\Delta q = q_a - q_b, \qquad A_i^{\text{pert}} = A_i^{\text{tar}} + \gamma_{\text{qc}} \sum_{r=1}^{M_q} \Delta q_r B_r, \tag{8}$$

where γ_{qc} controls perturbation strength. Unlike unstructured random noise, the perturbation generates QC-consistent deviations constrained to the subspace spanned by the learned QC-associated patterns $\mathcal{B}$.

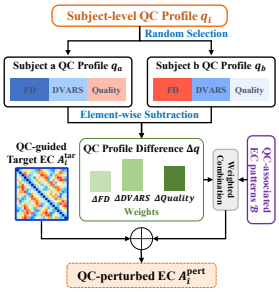


Figure 4: Structured QC perturbation generation.

QSR Network. Given a current EC A_i^{cur}, QSR uses A_i^{cur}, A_i^{ref}, their absolute discrepancy, and Γ^{QC} to predict an edge-wise gate G_i^{cur} and direction map D_i^{cur} (Fig. 5). The refined EC is given by $\mathcal{R}_{\theta_{\text{QSR}}}(A_i^{\text{cur}}) = A_i^{\text{cur}} + G_i^{\text{cur}} \odot D_i^{\text{cur}} \odot \left(A_i^{\text{ref}} - A_i^{\text{cur}}\right)$, where G_i^{cur} controls the correction magnitude, while D_i^{cur} modulates the signed adjustment relative to the population reference. The detailed network architecture is provided in Appendix B.4. **Self-supervised Training:** QSR is optimized through two paths. The initial EC refinement path takes $\widetilde{A}_i$ as input, while the QC-perturbation restoration path takes A_i^{pert} as input; both are trained toward the QC-guided target A_i^{tar}. The overall objective is $\mathcal{L}_{\text{QSR}} = \mathcal{L}_{\text{refine}} + \mathcal{L}_{\text{restore}} + \lambda_g \mathcal{L}_{\text{gate}} + \lambda_v \mathcal{L}_{\text{var}}$, where the first two terms supervise the two training paths, while the remaining terms regularize edge-wise corrections and preserve inter-subject structural variability. **EC Inference:** No QC perturbation or subject-level QC measurements are required. We set $A_i^{\text{cur}} = \widetilde{A}_i$ and apply the trained refiner to obtain the final individualized EC: $\widehat{A}_i = \mathcal{R}_{\theta_{\text{QSR}}}(\widetilde{A}_i)$, with Γ^{QC} fixed to the map estimated from the training fold.

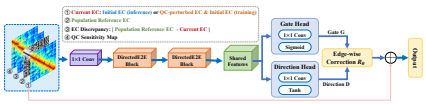


Figure 5: QSR network architecture.

5 Experiments

5.1 Experimental Setup

i) Datasets: We evaluated PR-EC on three publicly available multi-site rs-fMRI datasets: Autism Brain Imaging Data Exchange I (ABIDE-I; Di Martino et al., 2014), Autism Brain Imaging Data Exchange II (ABIDE-II; Di Martino et al., 2017)[1], and Attention Deficit Hyperactivity Disorder 200 (ADHD200; The ADHD-200 Consortium, 2012)[2]. For all datasets, the brain was parcellated using the Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002), with the first 90 ROIs retained for subsequent analysis. **ii) Baseline Methods:** We compared PR-EC with representative undirected and directed connectivity methods. The undirected baselines include Partial-FC (Marrelec et al., 2006; Ledoit & Wolf, 2004) and VarCoNet (Lamprou et al., 2026), while the directed baselines include Sparse-VAR (Basu & Michailidis, 2015), GVAR (Marcinkevičs & Vogt, 2021), CR-VAE (Li et al., 2023), NPI-MLP (Luo et al., 2025), and FSTA-EC (Xiong et al., 2025). **iii) Implementation and Evaluation:** Disease classification was evaluated using accuracy (ACC), sensitivity (SEN), area under the receiver operating characteristic curve (AUC), precision, and F1-score. For threshold-dependent metrics, thresholds were selected on the validation set using Youden's index (Youden, 1950) and fixed for testing. Experiments were implemented in PyTorch on an NVIDIA RTX 5090 GPU with a random seed of 42. Individual-Level EC Estimation used an FFN hidden dimension of 64 and two attention heads; Multimodal population reference (MPR) used $K = 20$ and a fusion coefficient of 0.6; and QC-Aware Self-Supervised Refinement (QSR) used 8 hidden channels. The same Directed BrainNetCNN (Kawahara et al., 2017) classifier was used for all methods. We used stratified 10-fold cross-validation, with one fold for testing and 20% of the remaining subjects for validation. Individual-Level EC Estimation was pretrained on all subjects, while all subsequent data-dependent procedures were fitted only on the training subset. Results are reported as mean ± standard deviation over the ten test folds. Additional details are provided in Appendix C.

[1]ABIDE: https://fcon_1000.projects.nitrc.org/indi/abide/
[2]ADHD200: https://fcon_1000.projects.nitrc.org/indi/adhd200/

5.2 PERFORMANCE

Table 1: Disease classification performance on three datasets. **Bold** denotes the best result and underline denotes the second-best result. All metrics are reported in percentages (%).

Dataset	Metric	Undirected		Directed					
		Partial-FC	VarCoNet	Sparse-VAR	GVAR	CR-VAE	NPI-MLP	FSTA-EC	PR-EC
ABIDE-I	ACC ↑	52.12 ± 4.39	57.54 ± 6.77	50.51 ± 4.31	47.76 ± 4.41	49.72 ± 5.39	50.62 ± 5.61	**60.86 ± 5.14**	60.39 ± 5.12
	SEN ↑	62.52 ± 31.57	65.45 ± 16.63	61.73 ± 30.73	48.63 ± 27.89	50.10 ± 29.49	56.40 ± 28.33	66.66 ± 14.10	**70.50 ± 13.83**
	AUC ↑	53.55 ± 5.84	59.89 ± 8.01	50.83 ± 5.99	50.65 ± 5.91	50.39 ± 8.99	56.65 ± 7.31	66.50 ± 4.15	**66.85 ± 4.32**
	Prec. ↑	44.43 ± 15.13	54.15 ± 6.09	48.55 ± 4.96	40.88 ± 10.09	51.20 ± 17.30	49.74 ± 6.43	**57.56 ± 6.49**	55.83 ± 4.87
	F1 ↑	49.38 ± 20.41	58.08 ± 7.55	49.26 ± 15.46	42.12 ± 17.68	43.01 ± 18.07	47.91 ± 12.69	60.58 ± 6.87	**61.73 ± 6.89**
ABIDE-II	ACC ↑	52.52 ± 3.76	55.39 ± 2.87	49.64 ± 4.70	50.57 ± 3.32	48.51 ± 5.55	56.52 ± 4.26	54.99 ± 5.11	**60.62 ± 5.14**
	SEN ↑	48.53 ± 23.11	54.13 ± 15.73	61.83 ± 29.94	69.31 ± 15.69	50.56 ± 24.88	57.17 ± 17.94	53.01 ± 15.37	**69.69 ± 7.88**
	AUC ↑	54.71 ± 4.19	60.27 ± 4.20	50.61 ± 4.23	51.01 ± 4.68	48.20 ± 7.46	59.84 ± 5.98	58.83 ± 5.27	**66.83 ± 3.92**
	Prec. ↑	56.27 ± 15.66	53.38 ± 4.18	49.55 ± 9.34	48.97 ± 3.00	47.99 ± 8.14	54.06 ± 4.99	53.62 ± 5.24	**57.52 ± 5.33**
	F1 ↑	45.63 ± 14.97	52.40 ± 7.66	48.53 ± 19.79	56.34 ± 5.36	44.64 ± 15.64	54.10 ± 9.76	51.69 ± 7.90	**62.61 ± 3.91**
ADHD200	ACC ↑	54.78 ± 10.50	53.68 ± 7.85	47.73 ± 8.24	46.68 ± 9.25	51.39 ± 7.32	53.28 ± 8.82	55.19 ± 8.50	**58.62 ± 6.36**
	SEN ↑	38.93 ± 27.78	53.57 ± 18.90	57.14 ± 29.06	55.00 ± 26.16	49.29 ± 31.70	53.21 ± 24.14	**58.93 ± 18.02**	57.50 ± 20.17
	AUC ↑	52.57 ± 7.55	55.35 ± 6.18	49.20 ± 8.05	49.16 ± 6.27	50.47 ± 7.16	54.43 ± 5.00	60.37 ± 4.92	**61.21 ± 5.85**
	Prec. ↑	35.96 ± 13.67	40.33 ± 4.73	36.32 ± 6.49	31.16 ± 10.81	36.18 ± 9.44	41.61 ± 10.04	43.58 ± 8.22	**45.83 ± 6.98**
	F1 ↑	33.46 ± 15.93	43.82 ± 7.01	40.47 ± 12.66	38.65 ± 14.19	37.32 ± 14.67	41.88 ± 12.91	47.39 ± 6.02	**47.78 ± 11.11**

Results. Across the three datasets, our proposed PR-EC shows consistently strong classification performance (Table 1). It achieves the best SEN, AUC, and F1 on ABIDE-I and the best results across all metrics on ABIDE-II, with a similar trend on ADHD200. Interestingly, directed EC baselines (e.g., Sparse-VAR and CR-VAE) do not consistently outperform undirected FC methods, which may reflect the difficulty of estimating directional dependencies from rs-fMRI. Despite this challenge, PR-EC outperforms all undirected baselines, suggesting that its population-informed, QC-guided refinement yields directed connectivity patterns that are more informative for downstream prediction. Moreover, its directed connectivity estimates enhance interpretability by revealing the direction of interregional influences.

Table 2: Ablation study on three datasets. Shaded cells indicate the best results.

Dataset	MPR	QSR	ACC	AUC	SEN	F1
ABIDE-I	✗	✗	59.47 ± 5.35	66.20 ± 4.32	67.74 ± 14.90	60.14 ± 7.01
	✓	✗	60.39 ± 4.93	65.79 ± 4.89	68.99 ± 14.68	61.10 ± 6.49
	✓	✓	60.39 ± 5.12	66.85 ± 4.32	70.50 ± 13.83	61.73 ± 6.89
ABIDE-II	✗	✗	60.00 ± 4.66	66.53 ± 3.68	67.51 ± 7.85	61.48 ± 3.33
	✓	✗	61.96 ± 3.34	65.39 ± 3.34	67.96 ± 9.88	62.61 ± 4.21
	✓	✓	60.62 ± 5.14	66.83 ± 3.92	69.69 ± 7.88	62.61 ± 3.91
ADHD200	✗	✗	58.37 ± 6.53	60.68 ± 6.19	56.79 ± 20.91	47.27 ± 11.46
	✓	✗	58.50 ± 5.11	60.87 ± 6.02	56.79 ± 23.71	46.68 ± 13.29
	✓	✓	58.62 ± 6.36	61.21 ± 5.85	57.50 ± 20.17	47.78 ± 11.11



Figure 6: Fine-grained ablation on ABIDE-I.

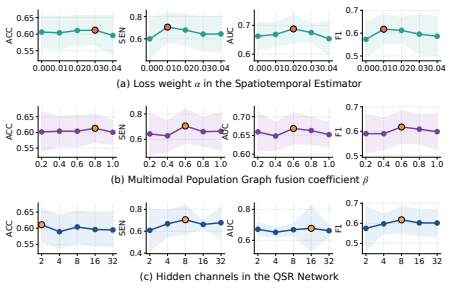


Figure 7: Sensitivity analysis of three key hyperparameters on ABIDE-I.

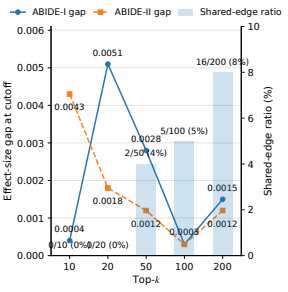


Figure 8: Reproducibility across top-k selections.

Ablation Study. We conduct ablation studies on all three datasets to evaluate the contributions of MPR and QSR (Table 2). For variants without QSR, we use a lightweight edge-wise gating module as the refinement backbone. Introducing MPR yields dataset-dependent gains. On ABIDE-II, ACC increases from 60.00% to 61.96%, while the changes on ABIDE-I and ADHD200 are smaller or

mixed. Fine-grained ablations on ABIDE-I (Fig. 6) show that combining PH and FC improves SEN, AUC, and F1 over either source alone, indicating their complementary contributions (Parisot et al., 2018; Pan et al., 2022). Adding QSR further improves AUC and SEN across all three datasets and benefits most other metrics. These results support the effectiveness of QSR.

Sensitivity Analysis. Fig. 7 presents a post-hoc sensitivity analysis of three key hyperparameters on the test folds, with the default settings fixed before this analysis and one hyperparameter varied at a time. For the entropy regularization weight α, the adopted $\alpha = 0.01$ achieves the best SEN and F1 while maintaining competitive ACC and AUC. The adopted fusion coefficient $\beta = 0.6$ yields the best SEN, AUC, and F1, while 8 hidden channels provide the strongest overall performance for QSR, with larger capacities showing no consistent benefit. Overall, the results indicate stable performance around the adopted configurations.

Cross-cohort ASD Connectivity Analysis. Based on the model-estimated EC matrices, we assess edge-wise differences between the ASD and control groups using two-sided Welch's t-tests (Welch, 1947) and quantify effect sizes using Hedges' g (Hedges, 1981). The resulting p-values are corrected for false discovery rate (FDR) within each cohort (Benjamini & Hochberg, 1995), yielding FDR-adjusted p-values (hereafter referred to as q-values), while connections are ranked by $|g|$ for the top-k analysis. The consistently small boundary gaps and low cross-cohort overlap (0–8%) across top-k selections (Fig. 8) indicate limited cross-cohort reproducibility of fixed top-k selection, consistent with previous reports of poor cross-site reproducibility of ASD-related functional connectivity alterations (King et al., 2019; He et al., 2020).

We therefore define replicated EC alterations as connections with $q < 0.05$ in both ABIDE-I and ABIDE-II and consistent effect directions. This criterion identifies 213 replicated EC alterations across 68 regions, including 111 ASD-enhanced and 102 ASD-reduced connections (Fig. 9). Regional involvement counts further highlight the right middle temporal gyrus (MTG.R) as one of the most involved regions, with prominent alterations involving the thalamus. Visualization of these replicated alterations using BrainNet Viewer (Xia et al., 2013) reveals a particularly consistent pattern: all eight possible directed connections between the bilateral MTG and the thalamus are enhanced by ASD (Fig. 10), revealing a consistent bidirectional pattern of thalamo-temporal enhancement. This finding is consistent with previous reports of increased thalamo-temporal functional connectivity in ASD (Woodward et al., 2017; Baran et al., 2023), while further revealing its directional organization from an EC perspective.

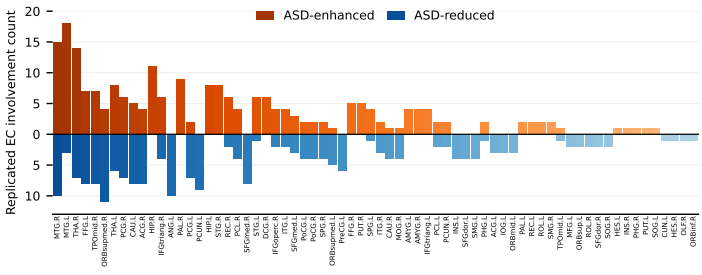


Figure 9: Regional involvement counts of replicated EC alterations, where both source and target regions are counted for each connection ($213 \times 2 = 426$).

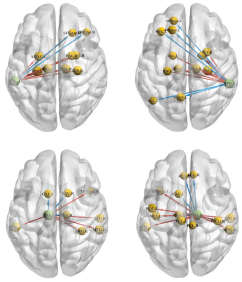


Figure 10: Directed EC alterations involving bilateral MTG and THA regions.

6 Conclusion

This work investigates how population connectivity information can improve individualized EC estimation from limited and noisy rs-fMRI observations. Experiments on three datasets show that population information complements individual observations, while QC-aware selective refinement yields EC representations more informative for downstream prediction. Theoretical analysis and ablation studies further support balancing population regularities with individual specificity. Cross-cohort analysis on ABIDE also identifies replicated ASD-related EC alterations, highlighting the potential of individualized EC for characterizing directional brain-network alterations.

7 AI USE STATEMENT

A large language model (LLM) was used solely for language editing, including improving the grammar, clarity, and readability of the manuscript. It was not used to generate scientific ideas, methodologies, or experimental results.

8 REPRODUCIBILITY STATEMENT

Appendices B and C provide the details required for reproducibility. Specifically, Appendix B provides detailed implementation of the proposed method, while Appendix C describes data preprocessing, training configurations, hyperparameter settings, and evaluation protocols. Our source code is available in the supplementary material.

9 ETHICS STATEMENT

Our work develops a population-regularized framework for individualized brain effective connectivity estimation from rs-fMRI, with the aim of advancing neuroscience research and facilitating the study of connectivity alterations associated with neurological and psychiatric conditions. We acknowledge potential ethical concerns associated with the analysis of neuroimaging and phenotypic data, including privacy risks, population biases, and inappropriate clinical interpretation. The proposed framework is intended solely for research purposes and is not a validated clinical diagnostic tool; any use in clinical decision-making would require rigorous clinical validation and appropriate expert oversight. All experiments in this study use publicly available, anonymized datasets, and we follow the data-use requirements of the respective datasets. Future applications involving sensitive or identifiable clinical data should follow appropriate privacy safeguards and institutional ethical review.

REFERENCES

Elena A. Allen, Eswar Damaraju, Sergey M. Plis, Erik B. Erhardt, Tom Eichele, and Vince D. Calhoun. Tracking whole-brain connectivity dynamics in the resting state. *Cerebral Cortex*, 24 (3):663–676, 2014. doi: 10.1093/cercor/bhs352.

Bengi Baran, Quynh Trang Huong Nguyen, Dimitrios Mylonas, Susan L. Santangelo, and Dara S. Manoach. Increased resting-state thalamocortical functional connectivity in children and young adults with autism spectrum disorder. *Autism Research*, 16(2):271–279, 2023. doi: 10.1002/aur.2875.

Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015. doi: 10.1214/15-AOS1315.

Yashar Behzadi, Khaled Restom, Joy Liau, and Thomas T. Liu. A component based noise correction method (compcor) for bold and perfusion based fmri. *NeuroImage*, 37(1):90–101, 2007. doi: 10.1016/j.neuroimage.2007.04.042.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Wanyi Chen, Jianjun Yang, Zhongquan Sun, Xiang Zhang, Guangyu Tao, Yuan Ding, Jingjun Gu, Jiajun Bu, and Haishuai Wang. DeepASD: a deep adversarial-regularized graph learning method for ASD diagnosis with multimodal data. *Translational Psychiatry*, 14(1):375, 2024. doi: 10.1038/s41398-024-02972-2.

Sharon Chiang, Michele Guindani, Hsiang J. Yeh, Zulfi Haneef, John M. Stern, and Marina Vannucci. Bayesian vector autoregressive model for multi-subject effective connectivity inference using multi-modal neuroimaging data. *Human Brain Mapping*, 38(3):1311–1332, 2017. doi: 10.1002/hbm.23456.

Rastko Ciric, Daniel H. Wolf, Jonathan D. Power, David R. Roalf, Graham L. Baum, Kosha Ruparel, Russell T. Shinohara, Mark A. Elliott, Simon B. Eickhoff, Christos Davatzikos, Ruben C. Gur, Raquel E. Gur, Danielle S. Bassett, and Theodore D. Satterthwaite. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*, 154:174–187, 2017. doi: 10.1016/j.neuroimage.2017.03.020.

Cameron Craddock, Sharad Sikka, Brian Cheung, Ranjeet Khanuja, Satrajit S. Ghosh, Chaogan Yan, Qingyang Li, Daniel Lurie, Joshua Vogelstein, Randal Burns, Stanley Colcombe, Maarten Mennes, Clare Kelly, Adriana Di Martino, Francisco X. Castellanos, and Michael Milham. Towards automated analysis of connectomes: The Configurable Pipeline for the Analysis of Connectomes (C-PAC). In *Neuroinformatics 2013*, 2013. doi: 10.3389/conf.fninf.2013.09.00042. Conference Abstract.

Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X. Castellanos, Kaat Alaerts, Jeffrey S. Anderson, Michal Assaf, Susan Y. Bookheimer, Mirella Dapretto, Ben Deen, Sonja Delmonte, Ilan Dinstein, Birgit Ertl-Wagner, Damien A. Fair, Louise Gallagher, Daniel P. Kennedy, Christopher L. Keown, Christian Keysers, Janet E. Lainhart, Catherine Lord, Beatriz Luna, Vinod Menon, Nancy Minshew, Christopher S. Monk, Sophia Mueller, Ralph-Axel Müller, Mary Beth Nebel, Joel T. Nigg, Kirsten O'Hearn, Kevin A. Pelphrey, Scott J. Peltier, Jeffrey D. Rudie, Stefan Sunaert, Marc Thioux, J. Michael Tyszka, Lucina Q. Uddin, Judith S. Verhoeven, Nicole Wenderoth, Jillian L. Wiggins, Stewart H. Mostofsky, and Michael P. Milham. The Autism Brain Imaging Data Exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, 2014. doi: 10.1038/mp.2013.78.

Adriana Di Martino, David O'Connor, Bosi Chen, Kaat Alaerts, Jeffrey S. Anderson, Michal Assaf, Joshua H. Balsters, Leslie Baxter, Anita Beggiato, Sylvie Bernaerts, Laura M. E. Blanken, Susan Y. Bookheimer, B. Blair Braden, Lisa Byrge, F. Xavier Castellanos, Mirella Dapretto, Richard Delorme, Damien A. Fair, Inna Fishman, Jacqueline Fitzgerald, Louise Gallagher, R. Joanne Jao Keehn, Daniel P. Kennedy, Janet E. Lainhart, Beatriz Luna, Stewart H. Mostofsky, Ralph-Axel Müller, Mary Beth Nebel, Joel T. Nigg, Kirsten O'Hearn, Marjorie Solomon, Roberto Toro, Chandan J. Vaidya, Nicole Wenderoth, Tonya White, R. Cameron Craddock, Catherine Lord, Bennett Leventhal, and Michael P. Milham. Enhancing studies of the connectome in autism using the Autism Brain Imaging Data Exchange II. *Scientific Data*, 4:170010, 2017. doi: 10.1038/sdata.2017.10.

Oscar Esteban, Christopher J. Markiewicz, Ross W. Blair, Craig A. Moodie, A. Ilkay Isik, Asier Erramuzpe, James D. Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit S. Ghosh, Jessey Wright, Joke Durnez, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16: 111–116, 2019. doi: 10.1038/s41592-018-0235-4.

Emily S. Finn, Xilin Shen, Dustin Scheinost, Monica D. Rosenberg, Jessica Huang, Marvin M. Chun, Xenophon Papademetris, and R. Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, 18(11):1664–1671, 2015. doi: 10.1038/nn.4135.

Stefan Frässle and Klaas E. Stephan. Test-retest reliability of regression dynamic causal modeling. *Network Neuroscience*, 6(1):135–160, 2022. doi: 10.1162/netn_a_00215.

Stefan Frässle, Ekaterina I. Lomakina, Adeel Razi, Karl J. Friston, Joachim M. Buhmann, and Klaas E. Stephan. Regression dcm for fmri. *NeuroImage*, 155:406–421, 2017. doi: 10.1016/j.neuroimage.2017.02.090.

Stefan Frässle, Samuel J. Harrison, Jakob Heinzle, Brett A. Clementz, Carol A. Tamminga, John A. Sweeney, Elliot S. Gershon, Matcheri S. Keshavan, Godfrey D. Pearlson, Albert Powers, and Klaas E. Stephan. Regression dynamic causal modeling for resting-state fmri. *Human Brain Mapping*, 42(7):2159–2180, 2021. doi: 10.1002/hbm.25357.

Karl J. Friston, Steven Williams, Robert Howard, Richard S. J. Frackowiak, and Robert Turner. Movement-related effects in fmri time-series. *Magnetic Resonance in Medicine*, 35(3):346–355, 1996. doi: 10.1002/mrm.1910350312.

Karl J. Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *NeuroImage*, 19(4): 1273–1302, 2003. doi: 10.1016/S1053-8119(03)00202-7.

Karl J. Friston, Joshua Kahan, Bharat Biswal, and Adeel Razi. A DCM for resting state fMRI. *NeuroImage*, 94:396–407, 2014. doi: 10.1016/j.neuroimage.2013.12.009.

Karl J. Friston, Vladimir Litvak, Ashwini Oswal, Adeel Razi, Klaas E. Stephan, Bernadette C. M. van Wijk, Gabriel Ziegler, and Peter Zeidman. Bayesian model reduction and empirical bayes for group (DCM) studies. *NeuroImage*, 128:413–431, 2016. doi: 10.1016/j.neuroimage.2015.11.015.

Krzysztof J. Gorgolewski, Tibor Auer, Vince D. Calhoun, R. Cameron Craddock, Samir Das, Eugene P. Duff, Guillaume Flandin, Satrajit S. Ghosh, Tristan Glatard, Yaroslav O. Halchenko, Daniel A. Handwerker, Michael Hanke, David Keator, Xiangrui Li, Zachary Michael, Camille Maumet, B. Nolan Nichols, Thomas E. Nichols, John Pellman, Jean-Baptiste Poline, Ariel Rokem, Gunnar Schaefer, Vanessa Sochat, William Triplett, Jessica A. Turner, Gaël Varoquaux, and Russell A. Poldrack. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3:160044, 2016. doi: 10.1038/sdata.2016.44.

Michael D. Greicius, Gaurav Srivastava, Allan L. Reiss, and Vinod Menon. Default-mode network activity distinguishes alzheimer's disease from healthy aging: evidence from functional mri. *Proceedings of the National Academy of Sciences of the United States of America*, 101(13): 4637–4642, 2004. doi: 10.1073/pnas.0308627101.

Ye He, Lisa Byrge, and Daniel P. Kennedy. Nonreplication of functional connectivity differences in autism spectrum disorder across multiple sites and denoising strategies. *Human Brain Mapping*, 41(5):1334–1350, 2020. doi: 10.1002/hbm.24879.

Larry V. Hedges. Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128, 1981. doi: 10.3102/10769986006002107.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

J. Huang, J.-Y. Jung, and C. S. Nam. Estimating effective connectivity in alzheimer's disease progression: A dynamic causal modeling study. *Frontiers in Human Neuroscience*, 16:1060936, 2022. doi: 10.3389/fnhum.2022.1060936.

Eunji Jun, Kyoung-Sae Na, Wooyoung Kang, Jiyeon Lee, Heung-Il Suk, and Byung-Joo Ham. Identifying resting-state effective connectivity abnormalities in drug-naïve major depressive disorder diagnosis via graph convolutional networks. *Human Brain Mapping*, 41(17):4997–5014, 2020. doi: 10.1002/hbm.25175.

Joshua Kahan, Maren Urner, Rosalyn Moran, Guillaume Flandin, Andre Marreiros, Laura Mancini, Mark White, John Thornton, Tarek Yousry, Ludvic Zrinzo, Marwan Hariz, Patricia Limousin, Karl Friston, and Tom Foltynie. Resting state functional MRI in Parkinson's disease: the impact of deep brain stimulation on 'effective' connectivity. *Brain*, 137(4):1130–1144, 2014. doi: 10.1093/brain/awu027.

Jeremy Kawahara, Colin J. Brown, Steven P. Miller, Brian G. Booth, Vann Chau, Ruth E. Grunau, Jill G. Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017. doi: 10.1016/j.neuroimage.2016.09.046.

Jace B. King, Molly B. D. Prigge, Carolyn K. King, Jubel Morgan, Fiona Weathersby, J. Chancellor Fox, Douglas C. Dean, III, Abigail Freeman, Joaquin Alfonso M. Villaruz, Karen L. Kane, Erin D. Bigler, Andrew L. Alexander, Nicholas Lange, Brandon Zielinski, Janet E. Lainhart, and Jeffrey S. Anderson. Generalizability and reproducibility of functional connectivity in autism. *Molecular Autism*, 10:27, 2019. doi: 10.1186/s13229-019-0273-5.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

Charalampos Lamprou, Aamna Alshehhi, Leontios J. Hadjileontiadis, and Mohamed L. Seghier. Varconet: A variability-aware self-supervised framework for functional connectome extraction from resting-state fmri. *Human Brain Mapping*, 47(4):e70469, 2026. doi: 10.1002/hbm.70469.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Hongming Li, Shujian Yu, and José C. Príncipe. Causal recurrent variational autoencoder for medical time series generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8562–8570, 2023. doi: 10.1609/aaai.v37i7.26031.

Lan Lin, Min Xiong, Ge Zhang, Wenjie Kang, Shen Sun, Shuicai Wu, and Alzheimer's Disease Neuroimaging Initiative. A convolutional neural network and graph convolutional network based framework for AD classification. *Sensors*, 23(4):1914, 2023. doi: 10.3390/s23041914.

Yilin Lu, Jinduo Liu, Junzhong Ji, Han Lv, and Mengdi Huai. Brain effective connectivity learning with deep reinforcement learning. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1664–1667. IEEE, 2022. doi: 10.1109/BIBM55620.2022.9995284.

Martina J. Lund, Dag Alnæs, Simon Schwab, Dennis van der Meer, Ole A. Andreassen, Lars T. Westlye, and Tobias Kaufmann. Differences in directed functional brain connectivity related to age, sex and mental health. *Human Brain Mapping*, 41(15):4173–4186, 2020. doi: 10.1002/hbm.25116.

Zixiang Luo, Kaining Peng, Zhichao Liang, Shengyuan Cai, Chenyu Xu, Dan Li, Yu Hu, Changsong Zhou, and Quanying Liu. Mapping effective connectivity by virtually perturbing a surrogate brain. *Nature Methods*, 22:1376–1385, 2025. doi: 10.1038/s41592-025-02654-x.

Liangsuo Ma, Sarah E. Braun, Joel L. Steinberg, James M. Bjork, Caitlin E. Martin, Larry D. Keen, II, and F. Gerard Moeller. Effect of scanning duration and sample size on reliability in resting state fMRI dynamic causal modeling analysis. *NeuroImage*, 292:120604, 2024. doi: 10.1016/j.neuroimage.2024.120604.

Ričards Marcinkevičs and Julia E. Vogt. Interpretable models for granger causality using self-explaining neural networks. In *International Conference on Learning Representations*, 2021.

Christopher J. Markiewicz, Oscar Esteban, Mathias Goncalves, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI, 2026. Version 25.2.5.

Guillaume Marrelec, Alexandre Krainik, Hugues Duffau, Mélanie Pélégrini-Issac, Stéphane Lehéricy, Julien Doyon, and Habib Benali. Partial correlation for functional brain interactivity investigation in functional mri. *NeuroImage*, 32(1):228–237, 2006. doi: 10.1016/j.neuroimage.2005.12.057.

Jiacheng Pan, Haocai Lin, Yihong Dong, Yu Wang, and Yunxin Ji. Mamf-gcn: Multi-scale adaptive multi-channel fusion deep graph convolutional network for predicting mental disorder. *Computers in Biology and Medicine*, 148:105823, 2022. doi: 10.1016/j.compbiomed.2022.105823.

Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero, Ben Glocker, and Daniel Rueckert. Disease prediction using graph convolutional networks: Application to autism spectrum disorder and alzheimer's disease. *Medical Image Analysis*, 48:117–130, 2018. doi: 10.1016/j.media.2018.06.001.

Jonathan D. Power, Kelly A. Barnes, Abraham Z. Snyder, Bradley L. Schlaggar, and Steven E. Petersen. Spurious but systematic correlations in functional connectivity mri networks arise from subject motion. *NeuroImage*, 59(3):2142–2154, 2012. doi: 10.1016/j.neuroimage.2011.10.018.

Jonathan D. Power, Anish Mitra, Timothy O. Laumann, Abraham Z. Snyder, Bradley L. Schlaggar, and Steven E. Petersen. Methods to detect, characterize, and remove motion artifact in resting state fmri. *NeuroImage*, 84:320–341, 2014. doi: 10.1016/j.neuroimage.2013.08.048.

Zarina Rakhimberdina, Xin Liu, and Tsuyoshi Murata. Population graph-based multi-model ensemble method for diagnosing autism spectrum disorder. *Sensors*, 20(21):6001, 2020. doi: 10.3390/s20216001.

Adeel Razi, Joshua Kahan, Geraint Rees, and Karl J. Friston. Construct validation of a dcm for resting state fmri. *NeuroImage*, 106:1–14, 2015. doi: 10.1016/j.neuroimage.2014.11.027.

Adeel Razi, Mohamed L. Seghier, Yuan Zhou, Peter McColgan, Peter Zeidman, Hae-Jeong Park, Olaf Sporns, Geraint Rees, and Karl J. Friston. Large-scale dcms for resting-state fmri. *Network Neuroscience*, 1(3):222–241, 2017. doi: 10.1162/NETN_a_00015.

Alard Roebroeck, Elia Formisano, and Rainer Goebel. Mapping directed influence over the brain using granger causality and fmri. *NeuroImage*, 25(1):230–242, 2005. doi: 10.1016/j.neuroimage.2004.11.017.

William W. Seeley, Rachel K. Crawford, Jie Zhou, Bruce L. Miller, and Michael D. Greicius. Neurodegenerative diseases target large-scale human brain networks. *Neuron*, 62(1):42–52, 2009. doi: 10.1016/j.neuron.2009.03.024.

Stephen M. Smith, Thomas E. Nichols, Diego Vidaurre, Anderson M. Winkler, Timothy E. J. Behrens, Matthew F. Glasser, Kamil Ugurbil, Deanna M. Barch, David C. Van Essen, and Karla L. Miller. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience*, 18(11):1565–1567, 2015. doi: 10.1038/nn.4125.

The 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. doi: 10.3389/fnsys.2012.00062.

Kamen A. Tsvetanov, Richard N. A. Henson, Lorraine K. Tyler, Adeel Razi, Linda Geerligs, Timothy E. Ham, James B. Rowe, and Cambridge Centre for Ageing and Neuroscience. Extrinsic and intrinsic brain network connectivity maintains cognition across the lifespan despite accelerated decay of regional brain activation. *The Journal of Neuroscience*, 36(11):3115–3126, 2016. doi: 10.1523/JNEUROSCI.2733-15.2016.

Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Octave Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage*, 15(1):273–289, 2002. doi: 10.1006/nimg.2001.0978.

Koene R. A. Van Dijk, Mert R. Sabuncu, and Randy L. Buckner. The influence of head motion on intrinsic functional connectivity mri. *NeuroImage*, 59(1):431–438, 2012. doi: 10.1016/j.neuroimage.2011.07.044.

Gaël Varoquaux, Alexandre Gramfort, Jean-Baptiste Poline, and Bertrand Thirion. Brain covariance selection: better individual functional connectivity models using population prior. In *Advances in Neural Information Processing Systems*, volume 23, pp. 2334–2342. Curran Associates, Inc., 2010.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103. ACM, 2008. doi: 10.1145/1390156.1390294.

Liang Wang, Yufeng Zang, Yong He, Meng Liang, Xinqing Zhang, Lixia Tian, Tao Wu, Tianzi Jiang, and Kuncheng Li. Changes in hippocampal connectivity in the early stages of alzheimer's disease: evidence from resting state fmri. *NeuroImage*, 31(2):496–504, 2006. doi: 10.1016/j.neuroimage.2005.12.033.

B. L. Welch. The generalization of Student's problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947. doi: 10.1093/biomet/34.1-2.28.

Neil D. Woodward, Monica Giraldo-Chica, Baxter Rogers, and Carissa J. Cascio. Thalamocortical dysconnectivity in autism spectrum disorder: An analysis of the autism brain imaging data exchange. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2(1):76–84, 2017. doi: 10.1016/j.bpsc.2016.09.002.

Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: A network visualization tool for human brain connectomics. *PLoS ONE*, 8(7):e68910, 2013. doi: 10.1371/journal.pone.0068910.

Wen Xiong, Jinduo Liu, Junzhong Ji, and Fenglong Ma. Brain effective connectivity estimation via fourier spatiotemporal attention. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, volume 1, pp. 1657–1668. Association for Computing Machinery, 2025. doi: 10.1145/3690624.3709226.

Yifan Yang, Yutong Mao, Xufu Liu, and Xiao Liu. BrainMAE: A region-aware self-supervised learning framework for brain signals. *arXiv preprint arXiv:2406.17086*, 2024. doi: 10.48550/arXiv.2406.17086.

B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zöllei, Jonathan R. Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, 2011. doi: 10.1152/jn.00338.2011.

William J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. doi: 10.1002/1097-0142(1950)3:1⟨32::AID-CNCR2820030106⟩3.0.CO;2-3.

Peter Zeidman, Amirhossein Jafarian, Mohamed L. Seghier, Vladimir Litvak, Hayriye Cagnan, Cathy J. Price, and Karl J. Friston. A guide to group effective connectivity analysis, part 2: Second level analysis with PEB. *NeuroImage*, 200:12–25, 2019. doi: 10.1016/j.neuroimage.2019.06.032.

Jiajun Zhang, Yang Zhao, Haoran Li, and Chengqing Zong. Attention with sparsity regularization for neural machine translation and summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(3):507–518, 2019. doi: 10.1109/TASLP.2018.2883740.

Zuozhen Zhang, Junzhong Ji, and Jinduo Liu. MetaRLEC: Meta-reinforcement learning for discovery of brain effective connectivity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):10261–10269, 2024. doi: 10.1609/aaai.v38i9.28892.

Chongyue Zhao, Liang Zhan, Paul M. Thompson, and Heng Huang. Predicting spatio-temporal human brain response using fmri. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 336–345. Springer, 2022. doi: 10.1007/978-3-031-16431-6_32.

Yuan Zhou, Peter Zeidman, Shihao Wu, Adeel Razi, Cheng Chen, Liuqing Yang, Jilin Zou, Gaohua Wang, Huiling Wang, and Karl J. Friston. Altered intrinsic and extrinsic connectivity in schizophrenia. *NeuroImage: Clinical*, 17:704–716, 2018. doi: 10.1016/j.nicl.2017.12.006.

Aixiao Zou, Junzhong Ji, Minglong Lei, Jinduo Liu, and Yongduan Song. Exploring brain effective connectivity networks through spatiotemporal graph convolutional models. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7871–7883, 2024. doi: 10.1109/TNNLS.2022.3221617.

Qiankun Zuo, Ling Chen, Yanyan Shen, Michael Kwok-Po Ng, Baiying Lei, and Shuqiang Wang. BDHT: Generative ai enables causality analysis for mild cognitive impairment. *IEEE Transactions on Automation Science and Engineering*, 22:5601–5613, 2025. doi: 10.1109/TASE.2024.3425949.

A Theoretical Analysis

This section provides additional theoretical analysis of the proposed framework, including the motivation for the local structural smoothness assumption, the risk analysis of the population reference, and the error-reduction property of QC-guided adjustment.

A.1 Theoretical Motivation of Assumption 3.1

The proposed population regularization relies on the intuition that subjects with similar population descriptors tend to exhibit related latent EC structures, following the general local smoothness principle underlying manifold regularization (Belkin et al., 2006). Specifically, the descriptor $z_i = [c_i, \phi(X_i)]$ integrates complementary phenotypic and functional connectivity, which characterizes individual-level similarity from both non-imaging and imaging perspectives. Therefore, subjects that are close in the descriptor space are expected to have smaller differences in their underlying EC.

Assumption 3.1 formalizes this intuition through a local smoothness condition. Importantly, this assumption does not require identical EC structures among neighboring subjects. Instead, it allows individual-specific variation while constraining the magnitude of EC differences according to their dissimilarity in the population descriptor space:

$$\|A_i^\star - A_j^\star\|_F \le L_A \rho(z_i, z_j), \tag{9}$$

where L_A controls the maximum rate of structural variation. Thus, the assumption describes a locally smooth relationship between population characteristics and latent directed connectivity structures.

Under this assumption, the subject-specific population reference in equation 4 can be interpreted as a local regularized estimator of the target EC. Specifically, by aggregating EC estimates from neighboring subjects,

$$A_i^{\mathrm{ref}} = \sum_{j \in \mathcal{N}_K(i)} w_{ij} \widetilde{A}_j, \tag{10}$$

the estimation uncertainty of individual EC estimates can be reduced through information sharing, while the structural mismatch introduced by neighboring subjects is controlled by their dissimilarity in the population descriptor space. This establishes a **bias–variance trade-off**: selecting closer neighbors reduces structural bias, whereas aggregating multiple neighbors reduces estimation variance. More formally, consider the decomposition

$$\widetilde{A}_j = A_j^\star + \varepsilon_j, \tag{11}$$

where ε_j denotes estimation noise. The deviation of the population reference from the target structure can be decomposed as

$$A_i^{\mathrm{ref}} - A_i^\star = \sum_{j \in \mathcal{N}_K(i)} w_{ij}(A_j^\star - A_i^\star) + \sum_{j \in \mathcal{N}_K(i)} w_{ij}\varepsilon_j. \tag{12}$$

The first term represents the structural discrepancy caused by local neighbor approximation, which is bounded by the smoothness assumption. The second term corresponds to the aggregated estimation uncertainty, which decreases when reliable neighboring estimates are combined. Therefore, the assumption provides the theoretical foundation for using a similarity-weighted population reference as complementary information rather than replacing the individualized EC estimate. The resulting bias–variance characterization is further formalized in Proposition 4.1.

A.2 Proof of Proposition 4.1

Proof. All expectations below are taken with respect to the estimation noise, conditioning on the retrieved neighborhood $\mathcal{N}_K(i)$ and the associated weights $\{w_{ij}\}_{j \in \mathcal{N}_K(i)}$. Thus, the neighborhood and weights are treated as fixed. The latent EC matrices $A_i^\star$ and $\{A_j^\star\}_{j \in \mathcal{N}_K(i)}$ are also treated as fixed.

Recall that

$$A_i^{\mathrm{ref}} = \sum_{j \in \mathcal{N}_K(i)} w_{ij} \widetilde{A}_j, \tag{13}$$

where $w_{ij} \ge 0$ and $\sum_{j \in \mathcal{N}_K(i)} w_{ij} = 1$. Substituting $\widetilde{A}_j = A_j^\star + \varepsilon_j$ into the above definition gives

$$\begin{aligned} A_i^{\mathrm{ref}} - A_i^\star &= \sum_{j \in \mathcal{N}_K(i)} w_{ij}(A_j^\star + \varepsilon_j) - A_i^\star \\ &= \sum_{j \in \mathcal{N}_K(i)} w_{ij}(A_j^\star - A_i^\star) + \sum_{j \in \mathcal{N}_K(i)} w_{ij}\varepsilon_j, \end{aligned} \tag{14}$$

where the second equality follows from $\sum_{j\in\mathcal{N}_K(i)} w_{ij} = 1$. Define

$$B_i = \sum_{j\in\mathcal{N}_K(i)} w_{ij}(A_j^\star - A_i^\star), \qquad E_i = \sum_{j\in\mathcal{N}_K(i)} w_{ij}\varepsilon_j. \tag{15}$$

Then,

$$A_i^{\mathrm{ref}} - A_i^\star = B_i + E_i. \tag{16}$$

Here, B_i represents the connectivity discrepancy induced by neighbor aggregation, whereas E_i represents the aggregated estimation noise.

For B_i, since the weights are nonnegative, the triangle inequality and Assumption 3.1 give

$$\begin{aligned}\|B_i\|_F &= \left\| \sum_{j\in\mathcal{N}_K(i)} w_{ij}(A_j^\star - A_i^\star) \right\|_F \\ &\le \sum_{j\in\mathcal{N}_K(i)} w_{ij}\|A_j^\star - A_i^\star\|_F \\ &\le L_A \sum_{j\in\mathcal{N}_K(i)} w_{ij}\rho(z_i, z_j).\end{aligned} \tag{17}$$

Therefore,

$$\|B_i\|_F^2 \le L_A^2 \left(\sum_{j\in\mathcal{N}_K(i)} w_{ij}\rho(z_i, z_j) \right)^2. \tag{18}$$

For the aggregated noise term, $\mathbb{E}[\varepsilon_j] = 0$ implies $\mathbb{E}[E_i] = \sum_{j\in\mathcal{N}_K(i)} w_{ij}\mathbb{E}[\varepsilon_j] = 0$. Its second moment satisfies

$$\begin{aligned}\mathbb{E}\|E_i\|_F^2 &= \mathbb{E}\left\langle \sum_{j\in\mathcal{N}_K(i)} w_{ij}\varepsilon_j, \sum_{k\in\mathcal{N}_K(i)} w_{ik}\varepsilon_k \right\rangle_F \\ &= \sum_{j,k\in\mathcal{N}_K(i)} w_{ij}w_{ik}\mathbb{E}\langle\varepsilon_j, \varepsilon_k\rangle_F \\ &= \sum_{j\in\mathcal{N}_K(i)} w_{ij}^2\mathbb{E}\|\varepsilon_j\|_F^2 + \sum_{\substack{j,k\in\mathcal{N}_K(i)\\ j\neq k}} w_{ij}w_{ik}\mathbb{E}\langle\varepsilon_j, \varepsilon_k\rangle_F \\ &= \sum_{j\in\mathcal{N}_K(i)} w_{ij}^2\mathbb{E}\|\varepsilon_j\|_F^2 \\ &\le \sum_{j\in\mathcal{N}_K(i)} w_{ij}^2\sigma_j^2.\end{aligned} \tag{19}$$

The penultimate equality follows from $\mathbb{E}\langle\varepsilon_j, \varepsilon_k\rangle_F = 0$ for $j \neq k$. Thus, pairwise uncorrelated estimation errors are sufficient for this variance bound; independence is not required.

Moreover, since $\mathbb{E}[E_i] = 0$, the conditional bias of the population reference is $\mathbb{E}[A_i^{\mathrm{ref}} - A_i^\star] = B_i$. Because B_i is fixed under the above conditioning, $\mathbb{E}\langle B_i, E_i\rangle_F = \langle B_i, \mathbb{E}[E_i]\rangle_F = 0$. Hence,

$$\begin{aligned}\mathbb{E}\|A_i^{\mathrm{ref}} - A_i^\star\|_F^2 &= \mathbb{E}\|B_i + E_i\|_F^2 \\ &= \|B_i\|_F^2 + 2\mathbb{E}\langle B_i, E_i\rangle_F + \mathbb{E}\|E_i\|_F^2 \\ &= \|B_i\|_F^2 + \mathbb{E}\|E_i\|_F^2 \\ &\le L_A^2 \left(\sum_{j\in\mathcal{N}_K(i)} w_{ij}\rho(z_i, z_j) \right)^2 + \sum_{j\in\mathcal{N}_K(i)} w_{ij}^2\sigma_j^2.\end{aligned} \tag{20}$$

This is exactly the bound in equation 5, which completes the proof. □

A.3 Proof of Proposition 4.2

Proof. For a fixed subject i, let C_i^{QC} be treated as deterministic, with the expectation taken over ε_i. Under the assumed additive model $\widetilde{A}_i = A_i^\star + C_i^{\mathrm{QC}} + \varepsilon_i$, we have

$$\begin{aligned} A_i^{\mathrm{tar}} &= \widetilde{A}_i - \eta C_i^{\mathrm{QC}} \\ &= A_i^\star + (1-\eta) C_i^{\mathrm{QC}} + \varepsilon_i. \end{aligned} \tag{21}$$

Consequently, the error remaining after the adjustment is $A_i^{\mathrm{tar}} - A_i^\star = (1-\eta) C_i^{\mathrm{QC}} + \varepsilon_i$.

Its expected squared Frobenius norm can be evaluated directly as

$$\begin{aligned} \mathbb{E} \left\| A_i^{\mathrm{tar}} - A_i^\star \right\|_F^2 &= \mathbb{E} \left\| (1-\eta) C_i^{\mathrm{QC}} + \varepsilon_i \right\|_F^2 \\ &= (1-\eta)^2 \left\| C_i^{\mathrm{QC}} \right\|_F^2 + \mathbb{E} \left\| \varepsilon_i \right\|_F^2 \\ &\quad + 2(1-\eta) \mathbb{E} \left\langle C_i^{\mathrm{QC}}, \varepsilon_i \right\rangle_F . \end{aligned} \tag{22}$$

Because C_i^{QC} is fixed and $\mathbb{E}[\varepsilon_i] = 0$, the last term is zero, since $\mathbb{E}\langle C_i^{\mathrm{QC}}, \varepsilon_i \rangle_F = \langle C_i^{\mathrm{QC}}, \mathbb{E}[\varepsilon_i] \rangle_F = 0$. It follows that

$$\mathbb{E} \left\| A_i^{\mathrm{tar}} - A_i^\star \right\|_F^2 = (1-\eta)^2 \left\| C_i^{\mathrm{QC}} \right\|_F^2 + \mathbb{E} \left\| \varepsilon_i \right\|_F^2 . \tag{23}$$

We next compare this quantity with the error before QC-guided adjustment. The additive model gives $\widetilde{A}_i - A_i^\star = C_i^{\mathrm{QC}} + \varepsilon_i$ and, by the same zero-mean property,

$$\mathbb{E} \left\| \widetilde{A}_i - A_i^\star \right\|_F^2 = \left\| C_i^{\mathrm{QC}} \right\|_F^2 + \mathbb{E} \left\| \varepsilon_i \right\|_F^2 . \tag{24}$$

The change in expected squared error produced by the QC-guided adjustment is therefore

$$\begin{aligned} &\mathbb{E} \left\| \widetilde{A}_i - A_i^\star \right\|_F^2 - \mathbb{E} \left\| A_i^{\mathrm{tar}} - A_i^\star \right\|_F^2 \\ &= \left[1 - (1-\eta)^2 \right] \left\| C_i^{\mathrm{QC}} \right\|_F^2 \\ &= \eta(2-\eta) \left\| C_i^{\mathrm{QC}} \right\|_F^2 . \end{aligned} \tag{25}$$

If $0 < \eta < 2$, then $\eta(2-\eta) > 0$. Hence, for any nonzero C_i^{QC}, the above difference is strictly positive, implying

$$\mathbb{E} \left\| A_i^{\mathrm{tar}} - A_i^\star \right\|_F^2 < \mathbb{E} \left\| \widetilde{A}_i - A_i^\star \right\|_F^2 . \tag{26}$$

Thus, within the idealized additive model, the QC-guided adjustment reduces the expected squared estimation error for $0 < \eta < 2$ whenever the QC-associated component is nonzero. □

B Supplementary Methodological Details

B.1 Spatiotemporal Attention and Directed EC Readout

This subsection provides the detailed formulations of the compact operations introduced in Sec. 4.1 for the individual-level EC estimator. Starting from the frequency-enhanced representation $H_i^{(m,f)}$, we detail the temporal self-attention and spatiotemporal fusion attention (STFA), their integration for self-supervised sequence reconstruction, the conversion of learned spatial relations into source-to-target directed EC, and the spatial-attention entropy regularization used during training.

Temporal Self-Attention. We first model within-window temporal dependencies independently for each ROI using shared attention parameters across ROIs. For ROI r, its frequency-enhanced representation is

$$H_{i,:,r}^{(m,f)} \in \mathbb{R}^{L \times D}, \tag{27}$$

where the L rows correspond to temporal positions within the m-th window. In our implementation, temporal self-attention and STFA reuse the same multi-head self-attention module (Vaswani

et al., 2017), with the attention operation applied along different dimensions. Accordingly, we denote the shared query, key, value, and output projection parameters by W_h^Q, W_h^K, W_h^V, and W^O, respectively.

For the h-th attention head, the query, key, and value representations along the temporal dimension are computed as

$$Q_{i,r,h}^{(m,T)} = H_{i,:,r}^{(m,f)} W_h^Q, \qquad K_{i,r,h}^{(m,T)} = H_{i,:,r}^{(m,f)} W_h^K, \qquad V_{i,r,h}^{(m,T)} = H_{i,:,r}^{(m,f)} W_h^V, \tag{28}$$

where

$$W_h^Q, W_h^K \in \mathbb{R}^{D \times D_k}, \qquad W_h^V \in \mathbb{R}^{D \times D_v}, \tag{29}$$

and therefore

$$Q_{i,r,h}^{(m,T)}, K_{i,r,h}^{(m,T)} \in \mathbb{R}^{L \times D_k}, \qquad V_{i,r,h}^{(m,T)} \in \mathbb{R}^{L \times D_v}. \tag{30}$$

The temporal attention weights are obtained using scaled dot-product attention (Vaswani et al., 2017):

$$P_{i,r,h}^{(m,T)} = \mathrm{Softmax}\left(\frac{Q_{i,r,h}^{(m,T)} \left(K_{i,r,h}^{(m,T)} \right)^\top}{\sqrt{D_k}} \right) \in \mathbb{R}^{L \times L}, \tag{31}$$

where the softmax is applied over the key temporal dimension. Thus, $\left[P_{i,r,h}^{(m,T)} \right]_{ts}$ quantifies the attention assigned by temporal position t to position s within the same ROI. During training, dropout is applied to the attention weights before value aggregation, giving

$$Z_{i,r,h}^{(m,T)} = \mathrm{Dropout}\left(P_{i,r,h}^{(m,T)} \right) V_{i,r,h}^{(m,T)} \in \mathbb{R}^{L \times D_v}. \tag{32}$$

The outputs of all H_a attention heads are concatenated and projected back to the model dimension:

$$O_{i,r}^{(m,T)} = \mathrm{Dropout}\left(\mathrm{Concat}_{h=1}^{H_a} \left(Z_{i,r,h}^{(m,T)} \right) W^O \right) \in \mathbb{R}^{L \times D}, \tag{33}$$

where $W^O \in \mathbb{R}^{H_a D_v \times D}$. A residual connection followed by layer normalization gives

$$\overline{H}_{i,:,r}^{(m,T)} = \mathrm{LN}\left(H_{i,:,r}^{(m,f)} + O_{i,r}^{(m,T)} \right). \tag{34}$$

The resulting representation is further processed by a position-wise feed-forward network (FFN):

$$\mathrm{FFN}(X) = \mathrm{ReLU}\left(X W_1 + b_1 \right) W_2 + b_2, \tag{35}$$

where $W_1 \in \mathbb{R}^{D \times D_{\mathrm{ff}}}$ and $W_2 \in \mathbb{R}^{D_{\mathrm{ff}} \times D}$. A second residual connection followed by layer normalization produces the temporally contextualized representation of ROI r:

$$H_{i,:,r,\mathrm{temp}}^{(m)} = \mathrm{LN}\left(\overline{H}_{i,:,r}^{(m,T)} + \mathrm{Dropout}\left(\mathrm{FFN}\left(\overline{H}_{i,:,r}^{(m,T)} \right) \right) \right) \in \mathbb{R}^{L \times D}. \tag{36}$$

Finally, stacking the temporally contextualized representations over all ROIs gives

$$H_{i,\mathrm{temp}}^{(m)} = \mathrm{Stack}_{r=1}^{d} \left(H_{i,:,r,\mathrm{temp}}^{(m)} \right) \in \mathbb{R}^{L \times d \times D}. \tag{37}$$

Hence, $H_{i,\mathrm{temp}}^{(m)}$ captures intra-ROI temporal dependencies within the current window while preserving the ROI dimension for subsequent spatiotemporal feature fusion.

Spatiotemporal Fusion Attention. Following the spatiotemporal attention strategy in Xiong et al. (2025), STFA derives time-resolved inter-regional relations from the same frequency-enhanced representation $H_i^{(m,f)}$. Specifically, at temporal position t, we consider the ROI-wise representation

$$H_{i,t,:}^{(m,f)} \in \mathbb{R}^{d \times D}, \tag{38}$$

where each row corresponds to the representation of one ROI at the t-th temporal position. Unlike temporal self-attention, which applies the shared attention operator along the temporal dimension of each ROI, STFA applies the same operator along the ROI dimension at each temporal position.

For the h-th attention head, the spatial query and key representations are computed using the shared projection matrices W_h^Q and W_h^K defined in the temporal self-attention module:

$$Q_{i,t,h}^{(m,S)} = H_{i,t,:}^{(m,f)} W_h^Q, \qquad K_{i,t,h}^{(m,S)} = H_{i,t,:}^{(m,f)} W_h^K, \tag{39}$$

where

$$Q_{i,t,h}^{(m,S)}, K_{i,t,h}^{(m,S)} \in \mathbb{R}^{d \times D_k}. \tag{40}$$

Here, the superscript (S) indicates that the shared attention operator is applied along the spatial, i.e., ROI, dimension rather than denoting a separate set of attention parameters.

The corresponding ROI-to-ROI attention matrix is obtained using scaled dot-product attention:

$$S_{i,t,h}^{(m)} = \mathrm{Softmax}\left(\frac{Q_{i,t,h}^{(m,S)}\left(K_{i,t,h}^{(m,S)}\right)^{\top}}{\sqrt{D_k}}\right) \in \mathbb{R}^{d \times d}, \tag{41}$$

where the softmax is applied over the key ROI dimension. During training, attention dropout is applied to the spatial attention weights before subsequent aggregation. We denote the resulting attention matrix by

$$\widetilde{S}_{i,t,h}^{(m)} = \mathrm{Dropout}\left(S_{i,t,h}^{(m)}\right). \tag{42}$$

At inference, dropout is disabled, and therefore $\widetilde{S}_{i,t,h}^{(m)} = S_{i,t,h}^{(m)}$.

To make the directional convention explicit, let $q_{i,t,h,v}^{(m,S)} \in \mathbb{R}^{D_k}$ denote the query representation of ROI v, and let $k_{i,t,h,u}^{(m,S)} \in \mathbb{R}^{D_k}$ denote the key representation of ROI u. Before attention dropout, the (v,u)-th attention coefficient is

$$\left[S_{i,t,h}^{(m)}\right]_{vu} = \frac{\exp\left(\left(q_{i,t,h,v}^{(m,S)}\right)^{\top} k_{i,t,h,u}^{(m,S)}/\sqrt{D_k}\right)}{\sum_{u'=1}^{d}\exp\left(\left(q_{i,t,h,v}^{(m,S)}\right)^{\top} k_{i,t,h,u'}^{(m,S)}/\sqrt{D_k}\right)}. \tag{43}$$

Under the target-by-source orientation adopted for the internal attention representation, the row index v corresponds to the target/query ROI, whereas the column index u corresponds to the source/key ROI. Consequently, $\left[S_{i,t,h}^{(m)}\right]_{vu}$ represents the model-derived directed relation from source ROI u to target ROI v at temporal position t and attention head h.

Because the query and key representations are obtained through distinct learned projections, the pairwise scores are not constrained to be symmetric, and therefore

$$\left[S_{i,t,h}^{(m)}\right]_{vu} \neq \left[S_{i,t,h}^{(m)}\right]_{uv} \tag{44}$$

in general. STFA can thus retain asymmetric inter-regional relations rather than enforcing an undirected connectivity structure.

For individual-level EC extraction, each temporal window is processed independently with dropout disabled. We average the spatial attention matrices across temporal positions and attention heads to obtain the window-level relation matrix

$$W_i^{(m)} = \frac{1}{LH_a}\sum_{t=1}^{L}\sum_{h=1}^{H_a} S_{i,t,h}^{(m)} \in \mathbb{R}^{d \times d}, \tag{45}$$

where H_a denotes the number of attention heads. The aggregation preserves the target-by-source orientation, such that $W_{i,vu}^{(m)}$ summarizes the model-derived relation from source ROI u to target ROI v over the m-th temporal window.

During minibatch training, the implementation additionally averages the dropout-perturbed spatial attention matrices across the windows in the minibatch. For a minibatch $\mathcal{I}$, the relation matrix used for feature fusion is therefore

$$W_{\mathcal{I}} = \frac{1}{|\mathcal{I}|LH_a}\sum_{(i,m)\in\mathcal{I}}\sum_{t=1}^{L}\sum_{h=1}^{H_a}\widetilde{S}_{i,t,h}^{(m)} \in \mathbb{R}^{d \times d}. \tag{46}$$

Spatiotemporal Feature Fusion. The minibatch-level inter-regional relation $W_{\mathcal{I}}$ is subsequently combined with the temporally contextualized representations. For each window $(i,m) \in \mathcal{I}$ and

temporal position t, the implementation performs feature fusion as

$$H^{(m)}_{i,t,\mathrm{st}} = W_{\mathcal{I}}^{\top} H^{(m)}_{i,t,:,\mathrm{temp}} \in \mathbb{R}^{d\times D}, \tag{47}$$

where $H^{(m)}_{i,t,:,\mathrm{temp}} \in \mathbb{R}^{d\times D}$ contains the temporally contextualized representations of all ROIs at temporal position t.

Equivalently, for ROI u,

$$\left[H^{(m)}_{i,t,\mathrm{st}}\right]_{u,:} = \sum_{v=1}^{d} [W_{\mathcal{I}}]_{vu} \left[H^{(m)}_{i,t,:,\mathrm{temp}}\right]_{v,:}. \tag{48}$$

Thus, the transpose converts the internally stored target-by-source relation into the orientation used for feature propagation, yielding a source-oriented fused representation. This operation couples the intra-ROI temporal dependencies learned by temporal self-attention with the asymmetric inter-regional relations derived by STFA.

Applying the same operation over all temporal positions gives

$$H^{(m)}_{i,\mathrm{st}} \in \mathbb{R}^{L\times d\times D}. \tag{49}$$

The fused representation is then combined with the frequency-enhanced representation through a residual connection:

$$R^{(m)}_i = H^{(m,f)}_i + H^{(m)}_{i,\mathrm{st}} \in \mathbb{R}^{L\times d\times D}. \tag{50}$$

The resulting representation is processed by the shared position-wise feed-forward network, followed by a residual connection and layer normalization:

$$H^{(m)}_{i,\mathrm{out}} = \mathrm{LN}\left(R^{(m)}_i + \mathrm{Dropout}\left(\mathrm{FFN}\left(R^{(m)}_i\right)\right)\right) \in \mathbb{R}^{L\times d\times D}. \tag{51}$$

Finally, a point-wise output projection maps the hidden representation back to the original ROI signal space:

$$\widehat{X}^{(m)}_i = \mathrm{Proj}\left(H^{(m)}_{i,\mathrm{out}}\right) \in \mathbb{R}^{L\times d}, \tag{52}$$

where $\mathrm{Proj} : \mathbb{R}^D \to \mathbb{R}$ denotes the shared output projection applied independently to each ROI and temporal position. The reconstructed window $\widehat{X}^{(m)}_i$ is used in the self-supervised reconstruction objective in equation 3.

Directed EC Readout. The window-level relation matrix $W^{(m)}_i$ obtained above follows the internal target-by-source orientation, where $W^{(m)}_{i,vu}$ represents the model-derived directed relation from source ROI u to target ROI v. To express the connectivity matrix under the source-to-target convention adopted throughout this work, we transpose the internal relation matrix and define the window-level directed EC as

$$A^{(m)}_i = \left(W^{(m)}_i\right)^{\top} \in \mathbb{R}^{d\times d}. \tag{53}$$

Accordingly,

$$A^{(m)}_{i,uv} = W^{(m)}_{i,vu}, \tag{54}$$

such that the row index u of $A^{(m)}_i$ denotes the source ROI and the column index v denotes the target ROI. Hence, $A^{(m)}_{i,uv}$ represents the estimated directed EC from ROI u to ROI v within the m-th temporal window.

We emphasize that the transpose in equation 53 only converts the matrix from the internal target-by-source orientation to the source-to-target convention used for reporting EC; it does not itself introduce directionality. The directed structure is inherited from the STFA relation matrix, whose query–key scores are not constrained to be symmetric.

To obtain an individual-level EC estimate, the window-level directed EC matrices are averaged across all M_i valid temporal windows of subject i:

$$\widetilde{A}_i = \frac{1}{M_i}\sum_{m=1}^{M_i} A^{(m)}_i = \left(\frac{1}{M_i}\sum_{m=1}^{M_i} W^{(m)}_i\right)^{\top} \in \mathbb{R}^{d\times d}. \tag{55}$$

This aggregation preserves the source-to-target orientation while reducing window-specific variability in the estimated directed relations.

Finally, self-connections are excluded from the individual-level EC matrix by setting its diagonal entries to zero:

$$\left[\widetilde{A}_i\right]_{uu} \leftarrow 0, \qquad u = 1, \ldots, d. \tag{56}$$

The resulting $\widetilde{A}_i$ is the individualized directed EC matrix used in the subsequent population-level regularization and refinement stages.

Spatial-Attention Entropy Regularization. To discourage overly diffuse inter-regional attention during self-supervised training, we impose an entropy regularizer (Zhang et al., 2019) on the minibatch-level spatial relation matrix $W_{\mathcal{I}}$ defined in equation 46. Since attention dropout is applied before aggregation, the rows of $W_{\mathcal{I}}$ are not necessarily normalized to sum to one. We therefore first clamp its entries for numerical stability and then renormalize each row:

$$\overline{W}_{\mathcal{I},vu} = \max\left([W_{\mathcal{I}}]_{vu}, \epsilon\right), \qquad \Pi_{\mathcal{I},vu} = \frac{\overline{W}_{\mathcal{I},vu}}{\sum_{u'=1}^{d} \overline{W}_{\mathcal{I},vu'}}. \tag{57}$$

Here, $\epsilon > 0$ is a small numerical constant ($\epsilon = 10^{-8}$ in our implementation). Thus, $\Pi_{\mathcal{I},v,:}$ defines a valid probability distribution over source/key ROIs for target/query ROI v.

The spatial-attention entropy regularizer is defined as the mean row-wise entropy normalized by its maximum value $\log d$:

$$\mathcal{L}_{\text{ent}} = -\frac{1}{d \log d} \sum_{v=1}^{d} \sum_{u=1}^{d} \Pi_{\mathcal{I},vu} \log \Pi_{\mathcal{I},vu}. \tag{58}$$

The normalization by $\log d$ removes the dependence of the entropy scale on the number of ROIs, with a uniform row distribution attaining the maximum normalized entropy. Since $\mathcal{L}_{\text{ent}}$ is minimized jointly with the reconstruction loss, it discourages diffuse spatial attention and encourages each target ROI to distribute its relational mass more selectively across source ROIs.

The resulting regularizer is incorporated into the self-supervised training objective as

$$\mathcal{L}_{\text{EC}} = \mathcal{L}_{\text{rec}} + \alpha \mathcal{L}_{\text{ent}}, \tag{59}$$

where $\mathcal{L}_{\text{rec}}$ denotes the window-level reconstruction loss and α controls the strength of the entropy regularization, as specified in equation 3. The entropy term is used only during training and does not modify the individual-level EC readout at inference.

B.2 Multimodal Affinity Computation and Reference Assembly

This subsection details the construction of the multimodal population reference (MPR) introduced in Sec. 4.2. For each target subject, phenotypic (PH) and functional-connectivity (FC) information are used to construct two complementary similarity graphs, following the population-graph paradigm used in multimodal neuroimaging analysis (Parisot et al., 2018; Rakhimberdina et al., 2020). The two graphs are independently sparsified and normalized, fused into a multimodal affinity graph, and further sparsified to obtain the final reference neighborhood and aggregation weights.

Reference population. Let $\mathcal{T}$ denote the training population in the current cross-validation fold. For each target subject i, the candidate reference set is defined as

$$\mathcal{C}_i = \begin{cases} \mathcal{T} \setminus \{i\}, & i \in \mathcal{T}, \\ \mathcal{T}, & i \notin \mathcal{T}. \end{cases} \tag{60}$$

Thus, training subjects exclude themselves from reference retrieval, whereas validation and test subjects retrieve neighbors exclusively from the training population. The same candidate set is used for both affinity modalities, and diagnostic labels are not involved in the construction.

Phenotypic affinity. Let $c_i = [c_i^{\text{cont}}, c_i^{\text{cat}}]$ denote the non-diagnostic phenotypic covariates of subject i, where c_i^{cont} and c_i^{cat} collect the continuous and categorical covariates, respectively. Continuous covariates are median-imputed and standardized using statistics estimated from the training

population, whereas missing categorical values are treated as a separate category. For $j \in \mathcal{C}_i$, the phenotypic distance is defined as

$$d_{ij}^{\mathrm{PH}} = \sum_{r=1}^{p_c} \lambda_r (\widetilde{c}_{i,r} - \widetilde{c}_{j,r})^2 + \lambda_{\mathrm{cat}} \sum_{\ell=1}^{p_g} \mathbb{I}(c_{i,\ell}^{\mathrm{cat}} \neq c_{j,\ell}^{\mathrm{cat}}), \tag{61}$$

where p_c and p_g denote the numbers of continuous and categorical covariates, respectively, $\widetilde{c}_{i,r}$ denotes the standardized r-th continuous covariate, λ_r controls its contribution, and λ_{cat} controls the categorical mismatch penalty.

We retain the $K_i = \min(K, |\mathcal{C}_i|)$ candidates with the smallest phenotypic distances, denoted by $\mathcal{N}_K^{\mathrm{PH}}(i)$. Their affinities are computed using the radial basis function (RBF) kernel, $a_{ij}^{\mathrm{PH}} = \exp(-d_{ij}^{\mathrm{PH}}/\tau_{\mathrm{PH}})$, where $\tau_{\mathrm{PH}} > 0$ denotes the bandwidth, and row-normalized to obtain

$$s_{ij}^{\mathrm{PH}} = \begin{cases} \dfrac{a_{ij}^{\mathrm{PH}}}{\sum_{k \in \mathcal{N}_K^{\mathrm{PH}}(i)} a_{ik}^{\mathrm{PH}}}, & j \in \mathcal{N}_K^{\mathrm{PH}}(i), \\ 0, & \text{otherwise.} \end{cases} \tag{62}$$

Functional affinity. The functional view is derived directly from the observed ROI time series $X_i \in \mathbb{R}^{T_i \times d}$. We construct a functional descriptor from the Pearson correlation matrix (Finn et al., 2015; Smith et al., 2015) as

$$R_i = \mathrm{Corr}(X_i) \in \mathbb{R}^{d \times d}, \qquad \phi(X_i) \quad = \mathrm{vec}_{\mathrm{ut}}(R_i), \qquad \bar{\phi}_i = \frac{\phi(X_i)}{\max(\|\phi(X_i)\|_2, \epsilon_{\mathrm{FC}})}. \tag{63}$$

Here, $\mathrm{vec}_{\mathrm{ut}}(\cdot)$ vectorizes the strict upper-triangular entries of the symmetric correlation matrix. For $j \in \mathcal{C}_i$, functional similarity is measured using nonnegative cosine similarity:

$$a_{ij}^{\mathrm{FC}} = \mathrm{clip}\left(\bar{\phi}_i^\top \bar{\phi}_j,\, 0,\, 1\right). \tag{64}$$

We retain the K_i candidates with the largest functional similarities, denoted by $\mathcal{N}_K^{\mathrm{FC}}(i)$. The corresponding sparse functional affinities are normalized as

$$s_{ij}^{\mathrm{FC}} = \begin{cases} \dfrac{a_{ij}^{\mathrm{FC}}}{\sum_{k \in \mathcal{N}_K^{\mathrm{FC}}(i)} a_{ik}^{\mathrm{FC}}}, & j \in \mathcal{N}_K^{\mathrm{FC}}(i), \\ 0, & \text{otherwise.} \end{cases} \tag{65}$$

If the selected functional affinities have zero total mass, uniform weights are assigned to the selected neighbors.

Multimodal fusion and reference assembly. The two sparse affinity graphs are fused as

$$s_{ij}^{\mathrm{fuse}} = (1-\beta) s_{ij}^{\mathrm{PH}} + \beta s_{ij}^{\mathrm{FC}}, \qquad \beta \in [0,1]. \tag{66}$$

A second top-K operation is applied to the fused affinities to obtain the final neighborhood $\mathcal{N}_K(i)$, which is determined by the combined multimodal similarities rather than either modality alone. The retained affinities are normalized to define

$$w_{ij} = \begin{cases} \dfrac{s_{ij}^{\mathrm{fuse}}}{\sum_{k \in \mathcal{N}_K(i)} s_{ik}^{\mathrm{fuse}}}, & j \in \mathcal{N}_K(i), \\ 0, & \text{otherwise,} \end{cases} \qquad \sum_{j \in \mathcal{N}_K(i)} w_{ij} = 1. \tag{67}$$

If the retained fused affinities have zero total mass, uniform weights are assigned over the selected neighborhood. The descriptor dissimilarity in Assumption 3.1 is defined as $\rho(z_i, z_j) = 1 - s_{ij}^{\mathrm{fuse}}$, where $z_i = [c_i, \phi(X_i)]$.

The subject-specific population reference is then assembled as

$$A_i^{\mathrm{ref}} = \sum_{j \in \mathcal{N}_K(i)} w_{ij} \widetilde{A}_j. \tag{68}$$

Since $w_{ij} \geq 0$ and the weights sum to one, A_i^{ref} is a convex combination of the selected individualized EC estimates.

All population-dependent preprocessing statistics are estimated exclusively from the training population within each cross-validation fold. For validation and test subjects, their own phenotypic and functional descriptors are used only to compute affinities with training subjects, while all reference neighbors and their initial EC estimates are retrieved exclusively from the corresponding training population. Diagnostic labels are not used in affinity computation or reference assembly.

B.3 QC Pattern Estimation and Structured Perturbation

This section provides additional details on the estimation of QC-associated EC and the generation of structured QC perturbations used for self-supervised refinement.

QC profile construction. For each subject i, we collect three individual-level QC measures: mean framewise displacement (FD), DVARS, and the proportion of low-motion frames (FD ≤ 0.2 mm) (Power et al., 2012; 2014). Missing values for each QC measure are imputed using the median estimated from the training subset of the corresponding fold, and the same median is then applied to the held-out subjects. Since these measures have different scales and distributions, we subsequently apply a logarithmic transformation:

$$z_{i,m} = \log(1 + \max(Q_{i,m}, 0)), \tag{69}$$

where $Q_{i,m}$ denotes the original value of QC measure m. The transformed values are then normalized using the median and interquartile range estimated from the training population:

$$q_{i,m} = \max\left(\frac{z_{i,m} - \text{median}_{j\in\mathcal{T}}(z_{jm})}{\text{IQR}_{j\in\mathcal{T}}(z_{jm})}, 0\right). \tag{70}$$

The resulting vector $q_i = [q_{i,1}, ..., q_{i,M_q}]^\top$ represents the QC profile of subject i.

QC-associated EC pattern estimation. To characterize how QC variations are associated with EC deviations, we estimate edge-specific QC effects from the training population, following the QC–connectivity association perspective used in motion-confound analysis (Ciric et al., 2017). For each directed edge (u, v), ridge regression (Hoerl & Kennard, 1970) is performed:

$$\widetilde{A}_{i,uv} = \beta^0_{uv} + (\beta^q_{uv})^\top q_i + (\beta^c_{uv})^\top c_i + \epsilon_{i,uv}, \tag{71}$$

where c_i denotes non-diagnostic confounders. The ridge penalty is applied to stabilize estimation under the high-dimensional edge space.

The coefficients corresponding to QC variables are collected into population-level QC patterns:

$$[B_m]_{uv} = [\beta^q_{uv}]_m, \tag{72}$$

where $B_m \in \mathbb{R}^{d\times d}$ represents the EC variation associated with QC factor m. Therefore, the QC-associated component of subject i can be estimated as:

$$C_i^{\text{QC}} = \sum_{m=1}^{M_q} q_{i,m} B_m. \tag{73}$$

Structured QC perturbation generation. Because the true QC-free EC is unavailable, we construct self-supervised perturbations within the learned QC variation subspace, following the corruption-and-restoration principle of denoising representation learning (Vincent et al., 2008). Specifically, two QC profiles q_a and q_b are randomly sampled from the training population and their difference is calculated: $\Delta q = q_a - q_b$. The perturbation matrix is then generated as:

$$A_i^{\text{tar}} = \sum_{m=1}^{M_q} \Delta q_m B_m, \tag{74}$$

and added to the QC-guided target:

$$A_i^{\text{pert}} = A_i^{\text{tar}} + \gamma_{\text{qc}} A_i^{\text{tar}}. \tag{75}$$

Unlike random Gaussian noise, this perturbation preserves the population-observed relationship between QC variations and EC changes. Therefore, the restoration task encourages QSR to learn cor-

rection patterns associated with realistic QC-induced deviations. The perturbation is not designed to simulate arbitrary corruption, but to sample plausible EC deviations lying in the empirically estimated QC-associated subspace.

B.4 QSR Architecture Details

This section provides additional details on the architecture, training objectives, and inference procedure of the QC-aware self-supervised refinement (QSR) module. QSR uses the population reference A_i^{ref} as complementary connectivity evidence and selectively refines $\widetilde{A}_i$ through edge-wise QC-aware residual adjustments.

QSR architecture. QSR operates on a current EC matrix A_i^{cur}, where $A_i^{\mathrm{cur}} = \widetilde{A}_i$ or A_i^{pert} for the two training paths, respectively. Given A_i^{ref} and Γ^{QC}, we first define the reference discrepancy as

$$\Delta_i^{\mathrm{cur}} = A_i^{\mathrm{ref}} - A_i^{\mathrm{cur}}. \tag{76}$$

The QSR input is then constructed as

$$U_i^{\mathrm{cur}} = \mathrm{Stack}\left(A_i^{\mathrm{cur}}, A_i^{\mathrm{ref}}, |\Delta_i^{\mathrm{cur}}|, \Gamma^{\mathrm{QC}}\right) \in \mathbb{R}^{4\times d\times d}. \tag{77}$$

The input is first projected by a 1×1 convolution and then processed by two DirectedE2E blocks, adapted from the edge-to-edge operation of BrainNetCNN (Kawahara et al., 2017):

$$H_i^{\mathrm{cur}} = \mathrm{E2E}_2\left(\mathrm{E2E}_1\left(\mathrm{Proj}_{1\times1}(U_i^{\mathrm{cur}})\right)\right), \tag{78}$$

where DirectedE2E aggregates edge features according to the source and target ROIs of each directed connection. The hidden representation contains 8 channels in our implementation.

Two edge-wise heads are applied to H_i^{cur} to obtain the refinement gate G_i^{cur} and direction map D_i^{cur}:

$$G_i^{\mathrm{cur}} = h_g(H_i^{\mathrm{cur}}), \qquad D_i^{\mathrm{cur}} = h_d(H_i^{\mathrm{cur}}). \tag{79}$$

The refined EC is given by

$$\mathcal{R}_{\theta_{\mathrm{QSR}}}\left(A_i^{\mathrm{cur}}\right) = A_i^{\mathrm{cur}} + G_i^{\mathrm{cur}} \odot D_i^{\mathrm{cur}} \odot \Delta_i^{\mathrm{cur}}, \tag{80}$$

where $\odot$ denotes element-wise multiplication. Here, G_i^{cur} controls the edge-wise refinement magnitude, while D_i^{cur} determines its direction. Thus, A_i^{ref} provides a reference for refinement rather than directly replacing A_i^{cur}, allowing QSR to selectively adjust individual EC edges.

Dual-path self-supervised training. QSR is trained through two complementary paths that share the same parameters θ_{QSR} and supervision target A_i^{tar}.

For the *initial EC refinement path*, we set $A_i^{\mathrm{cur}} = \widetilde{A}_i$, and obtain $\widehat{A}_i^{\mathrm{refine}} = \mathcal{R}_{\theta_{\mathrm{QSR}}}\left(\widetilde{A}_i\right)$. The corresponding refinement loss is

$$\mathcal{L}_{\mathrm{refine}} = \ell\left(\widehat{A}_i^{\mathrm{refine}}, A_i^{\mathrm{tar}}\right), \tag{81}$$

where $\ell(\cdot,\cdot)$ denotes the edge-wise reconstruction loss. This path directly learns to refine the initial EC $\widetilde{A}_i$ toward A_i^{tar}.

For the *QC-perturbation restoration path*, we instead set $A_i^{\mathrm{cur}} = A_i^{\mathrm{pert}}$, where A_i^{pert} is generated from A_i^{tar} using the structured QC perturbation defined in Appendix B.3. The corresponding QSR output and restoration loss are defined as:

$$\widehat{A}_i^{\mathrm{restore}} = \mathcal{R}_{\theta_{\mathrm{QSR}}}\left(A_i^{\mathrm{pert}}\right), \qquad \mathcal{L}_{\mathrm{restore}} = \ell\left(\widehat{A}_i^{\mathrm{restore}}, A_i^{\mathrm{tar}}\right). \tag{82}$$

Since A_i^{pert} deviates from A_i^{tar} within the estimated QC-associated subspace, this path explicitly trains QSR to recover EC from structured QC-related perturbations.

The two paths serve complementary roles: $\mathcal{L}_{\mathrm{refine}}$ directly supervises refinement from the same initial EC input used at inference, whereas $\mathcal{L}_{\mathrm{restore}}$ trains QSR to correct structured deviations along the QC-associated subspace.

To discourage unnecessary edge-wise adjustments, we regularize the gates predicted by both paths as

$$\mathcal{L}_{\mathrm{gate}} = \frac{1}{2}\left[\mathrm{Mean}\left(\left|G_i^{\mathrm{refine}}\right|\right) + \mathrm{Mean}\left(\left|G_i^{\mathrm{restore}}\right|\right)\right], \tag{83}$$

where $\mathrm{Mean}(\cdot)$ is taken over subjects and directed edges.

To prevent excessive homogenization across subjects, we further penalize the refined EC $\widehat{A}$ when its average edge-wise inter-subject variance falls below a fraction τ_{var} of that of the initial EC $\widetilde{A}$:

$$\mathcal{L}_{\mathrm{var}} = \left[\tau_{\mathrm{var}} \operatorname{Mean}_{u,v} \operatorname{Var}_i \left(\widetilde{A}_{i,uv}\right) - \operatorname{Mean}_{u,v} \operatorname{Var}_i \left(\widehat{A}^{\mathrm{refine}}_{i,uv}\right)\right]_+, \tag{84}$$

where the variance is computed across training subjects within each fold, $[x]_+ = \max(x, 0)$, and τ_{var} specifies the minimum retained variance ratio. The complete QSR objective is the following.

$$\mathcal{L}_{\mathrm{QSR}} = \mathcal{L}_{\mathrm{refine}} + \mathcal{L}_{\mathrm{restore}} + \lambda_g \mathcal{L}_{\mathrm{gate}} + \lambda_v \mathcal{L}_{\mathrm{var}}, \tag{85}$$

where λ_g and λ_v control the strengths of the gate and the regularization of variance-retention, respectively.

Inference. At inference, neither A_i^{tar} nor A_i^{pert} is required. For each validation or test subject, we set $A_i^{\mathrm{cur}} = \widetilde{A}_i$ and obtain the final EC as

$$\widehat{A}_i = \mathcal{R}_{\theta_{\mathrm{QSR}}} \left(\widetilde{A}_i\right), \tag{86}$$

where A_i^{ref} is constructed from the corresponding training pool and Γ^{QC} is fixed to the map estimated from the training fold. Accordingly, subject-level QC profiles are only used during training to construct A_i^{tar} and A_i^{pert}. No QC measurements from validation or test subjects are required for QSR inference.

C Experimental Details

C.1 Data Preprocessing

We used three publicly available datasets: ABIDE-I (Di Martino et al., 2014), ABIDE-II (Di Martino et al., 2017), and ADHD200 (The ADHD-200 Consortium, 2012). The demographic and site characteristics of the subjects included in the final analysis are summarized in Table 3. For ABIDE-I, we included all subjects available in the publicly released preprocessed dataset. For ABIDE-II and ADHD200, the complete datasets were initially downloaded; however, some subjects could not be successfully processed because of missing or incomplete imaging data or other preprocessing-related issues. Therefore, only subjects with the required imaging data for which preprocessing was successfully completed were retained for subsequent analyses.

Table 3: Demographic information of the three datasets.

Dataset	N	Case/Control	Age range (years)	Sex (M/F)	Sites
ABIDE-I	871	403/468	6.47–58.00	727/144	20
ABIDE-II	975	462/513	5.13–64.00	743/232	16
ADHD200	788	280/508	7.09–21.83	486/302	8

Preprocessing Pipeline. For ABIDE-I, we used the C-PAC-preprocessed derivatives (Craddock et al., 2013). For ABIDE-II and ADHD200, raw anatomical and resting-state fMRI data were organized according to the Brain Imaging Data Structure (BIDS) specification (Gorgolewski et al., 2016) and processed using fMRIPrep v25.2.5 (Esteban et al., 2019; Markiewicz et al., 2026), with FreeSurfer-based surface reconstruction disabled. Using each subject's T1-weighted image as the anatomical reference, fMRIPrep performed brain extraction, tissue segmentation, blood-oxygen-level-dependent (BOLD) reference generation, head-motion correction, functional-to-anatomical registration, and spatial normalization. The preprocessed BOLD images were transformed into the MNI152NLin6Asym standard space.

Nuisance Regression and Temporal Filtering. For ABIDE-II and ADHD200, the confound regressors generated by fMRIPrep were used for further denoising. The six rigid-body motion parameters were expanded into a Friston-24 motion model (Friston et al., 1996), together with the first five available anatomical CompCor (aCompCor) components (Behzadi et al., 2007). The resulting

time series were linearly detrended and band-pass filtered between 0.01 and 0.10 Hz. Global signal regression was not applied, and no frames were removed based on FD or DVARS.

ROI Time-Series Extraction. For all datasets, the preprocessed BOLD signals were parcellated using the Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002). The mean BOLD signal within each region was extracted, and the first 90 cerebral ROIs were retained. Each subject was consequently represented as an ROI-level time-series matrix $X_i \in \mathbb{R}^{T_i \times 90}$, which served as the common input for subsequent connectivity estimation.

C.2 Baseline Methods and Parameter Settings

To further illustrate the comparison methods used in the main experiment, this section provides a brief introduction to each baseline method.

Undirected Connectivity Methods.

(i) **Partial-FC** (Marrelec et al., 2006): Partial correlation estimates the conditional association between two ROIs while accounting for the influence of the remaining regions. We use Ledoit–Wolf shrinkage (Ledoit & Wolf, 2004) to obtain a regularized covariance estimate before computing the corresponding partial correlations.

(ii) **VarCoNet** (Lamprou et al., 2026): It is a self-supervised framework for estimating subject-specific functional connectivity from rs-fMRI. The model combines a 1D convolutional neural network with a Transformer encoder to learn representations from segmented ROI time series and uses contrastive learning to preserve subject-specific functional variability. Functional connectivity is subsequently derived from the similarity between the learned representations.

Directed Connectivity Methods.

(i) **Sparse-VAR** (Basu & Michailidis, 2015): Sparse vector autoregression models directed temporal dependencies among brain regions through lagged interactions. Sparsity constraints are imposed on the autoregressive coefficients to suppress redundant connections and obtain a sparse directed connectivity representation.

(ii) **GVAR** (Marcinkevičs & Vogt, 2021): Generalized Vector Autoregression extends Granger-based modeling to nonlinear multivariate time series using neural networks. It learns nonlinear temporal dependencies among variables and derives directed interactions from the learned predictive dynamics.

(iii) **CR-VAE** (Li et al., 2023): It is a recurrent variational autoencoder designed for modeling multivariate time series. It learns latent temporal dynamics from the observed sequences and uses the resulting dependency structure to characterize directed interactions among variables.

(iv) **NPI-MLP** (Luo et al., 2025): Neural Perturbational Inference estimates directed interactions through virtual perturbations of a learned surrogate model. An MLP is used as the surrogate, and the influence of one variable on another is quantified by measuring the response induced by perturbing the corresponding input variable.

(v) **FSTA-EC** (Xiong et al., 2025): This model combines Fourier representations with spatiotemporal attention for subject-specific effective connectivity estimation. It jointly models temporal dynamics and inter-regional interactions in rs-fMRI time series to derive directed connectivity patterns between brain regions.

The key parameter settings of all baseline methods are summarized in Table 4.

Table 4: Key parameter settings of the baseline methods.

Type	Method	Key Parameters
Undirected	Partial-FC	Ledoit–Wolf shrinkage, diagonal = 0
	VarCoNet	layers = 1, min length = 30, heads = 2, FFN = 512, $\tau = 0.0541$
Directed	Sparse-VAR	lag = 1, $\alpha = 0.03$, ℓ_1 ratio = 1.0
	GVAR	order = 1, hidden size = 16, hidden layers = 1, $\lambda = 0.01$, $\gamma = 0$
	CR-VAE	context = 20, hidden size = 64, $\lambda = 0.1$, lr = 0.05
	NPI-MLP	input steps = 3, perturbation strength = 1.0
	FSTA-EC	batch-size = 32, $\alpha_{\text{sp}} = 0.8$, $d_{\text{model}} = 16$, heads = 2

C.3 Evaluation Protocol

We adopt stratified 10-fold cross-validation. Individual-Level EC Estimation is first pretrained in a self-supervised manner using the rs-fMRI time series of all subjects, without access to diagnostic labels. The resulting estimator is then used to derive the initial individualized ECs for downstream evaluation.

For each cross-validation fold, one fold (approximately 10% of the subjects) is held out for testing, while 20% of the remaining subjects are reserved for validation. All subsequent data-dependent procedures, including phenotypic normalization, missing-value imputation, MPR, QSR, and downstream classification, are performed using only the training subset. Validation and test subjects retrieve population references exclusively from the corresponding training pool. Model selection and classification-threshold selection use only the validation set, while test labels are used solely for final performance evaluation.

C.4 Downstream Classification

For each connectivity method, we independently train the same Directed BrainNetCNN classifier to evaluate the disease-discriminative capacity of the resulting connectivity representation. Within each cross-validation fold, each connectivity edge is standardized using the mean and standard deviation estimated from the training subset, and the diagonal entries are set to zero. Given a standardized connectivity matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, we construct a two-channel input

$$\mathbf{X}_{\text{cls}} = [\mathbf{A}, \mathbf{A}^\top] \in \mathbb{R}^{2\times d\times d}. \tag{87}$$

For asymmetric EC, the two channels encode the original and reverse connectivity orientations, respectively, whereas they are identical for symmetric FC.

The Directed BrainNetCNN is adapted from BrainNetCNN (Kawahara et al., 2017) and consists of two directed edge-to-edge (E2E) layers, followed by an edge-to-node (E2N) layer, a node-to-graph (N2G) layer, and a fully connected classification head. The two E2E layers use 4 and 8 output channels, respectively. In the E2E and E2N layers, row-wise and column-wise convolutions separately aggregate outgoing and incoming connectivity patterns, preserving directional information during feature extraction. The N2G layer then aggregates node-level features into a graph-level representation for binary classification.

The classifier is optimized using binary cross-entropy with logits and Adam (Kingma & Ba, 2015) with a learning rate of 10^{-3}, weight decay of 10^{-4}, and a batch size of 32. LeakyReLU with a negative slope of 0.1 is used for activation, and dropout with a rate of 0.3 is applied in the classification head. Model selection is based on validation loss with early stopping. For threshold-dependent metrics, the classification threshold is selected on the validation set by maximizing Youden's index (Youden, 1950),

$$t^\star = \arg\max_t \left[\text{TPR}(t) - \text{FPR}(t)\right], \tag{88}$$

and is then fixed when evaluating the corresponding test fold. Test labels are used only for final performance evaluation.

C.5 Evaluation Metrics

We evaluate classification performance using accuracy (ACC), sensitivity (SEN), precision, F1-score, and area under the receiver operating characteristic curve (AUC). Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The threshold-dependent metrics are defined as

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN}, \quad \mathrm{SEN} = \frac{TP}{TP + FN}, \ \mathrm{Precision} \quad = \frac{TP}{TP + FP}, \mathrm{F1} \quad = \frac{2, \mathrm{Precision} \times \mathrm{SEN}}{\mathrm{Precision} + \mathrm{SEN}}. \tag{89}$$

The true-positive rate (TPR) and false-positive rate (FPR) are defined as

$$\mathrm{TPR} = \frac{TP}{TP + FN} = \mathrm{SEN}, \qquad \mathrm{FPR} = \frac{FP}{FP + TN}. \tag{90}$$

AUC summarizes discriminative performance across classification thresholds.

C.6 Implementation Details

The main hyperparameters are summarized in Table 5.

Table 5: Implementation details of PR-EC. Values separated by "/" correspond to ABIDE-I, ABIDE-II, and ADHD200, respectively.

Module	Hyperparameter	Setting
Individual-Level EC Estimation	Window length L	78 / 80 / 76
	Stride H	39 / 40 / 25
	Model dimension D	16
	FFN hidden dimension	64
	Attention heads H_a	2
	Training epochs	101 / 81 / 141
Multimodal Population Reference Construction	Neighbors K	20
	Fusion coefficient β	0.6
	PH RBF bandwidth τ_{PH}	2.0
	Categorical mismatch penalty λ_{cat}	4.0
QC-Aware Self-Supervised Refinement	Hidden channels	8
	Training epochs	80 / 100 / 100
	Learning rate	3×10^{-3}
	QC target coefficient η	0.15
	Perturbation strength γ_{qc}	0.5
	Gate regularization λ_g	10^{-3}
	Variance regularization λ_v	0.1

For Individual-Level EC Estimation, the spatial-attention entropy coefficient α was set to 0.01, and dropout was set to 0.2. For QC-guided EC target construction, the ridge regularization coefficient was set to 10^{-3}. QSR used a maximum gate value of 0.5. All other hyperparameters were shared across the three datasets.