Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Prediction-Coupled Self-Explaining Functional Connectivity Learning Framework for rs-fMRI-Based Brain Disorder Classification

Mingyang HUANG

huang@ustb.edu.cn

University of Science and Technology Beijing

Research Article

Keywords: rs-fMRI, functional connectivity, self-explaining model, faithfulness, reproducibility

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11285494/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Existing resting-state functional magnetic resonance imaging (rs-fMRI) functional connectivity (FC) models can achieve useful predictive performance, but their explanations are often generated after prediction and are not systematically evaluated for faithfulness and reproducibility. We propose SE-FC, a prediction-coupled self-explaining functional connectivity framework for interpretable FC learning. The explanation pathway learns signed coefficients over FC edges, whose interaction with standardized FC values produces signed edge evidence that directly contributes to the final prediction. Deletion, insertion, and average output probability change (AOPC) analyses showed that SE-FC-derived evidence was more strongly associated with model output than random edge rankings. Cross-run analysis further quantified explanation reproducibility by identifying recurrent consensus patterns while characterizing substantial variability in exact edge selection across independent training runs. Predictive capability was evaluated as supporting evidence under strict leave-one-site-out validation on ABIDE (435 participants, 10 sites) and SRPBS (925 participants, six sites), with three independent repeated training runs for each dataset. SE-FC achieved AUC values of 0.6133 ± 0.0044 on ABIDE and 0.6127 ± 0.0051 on SRPBS, indicating competitive cross-site predictive capability. Overall, SE-FC provides prediction-coupled connectivity evidence with quantitatively assessed model-level faithfulness and cross-run reproducibility while retaining useful predictive performance.

Highlights

Prediction-coupled signed FC evidence contributes directly to inference.

LOSO evaluation covers two multi-site rs-fMRI disorder datasets.

Deletion, insertion, and AOPC assess model-level explanation faithfulness.

Cross-run consensus analysis quantifies explanation reproducibility.

1. Introduction

Resting-state functional magnetic resonance imaging (rs-fMRI) functional connectivity (FC) provides a structured biomedical signal for learning distributed statistical relationships among brain regions [1, 2, 19]. FC-based models have been studied for brain-disorder classification, but a class prediction alone does not identify which observed connections influenced the decision. Explanation is difficult because AAL116 yields 6670 undirected edges, while available multi-site cohorts remain modest and heterogeneous in acquisition and population characteristics [3, 6, 7]. These conditions permit multiple correlated predictive patterns and increase sensitivity to site shifts. We therefore treat classification as a methodological evaluation of FC learning rather than an individually deployable diagnostic system.

FC classifiers commonly learn from vectorized edges or graph representations. Edge-vector models capture distributed nonlinear combinations of connections, whereas graph methods preserve regional

topology through message passing or connectome-specific operations [8–10, 16, 22, 28]. BrainIB and related interpretable graph approaches seek compact task-relevant representations [11]. Recent Biomedical Signal Processing and Control studies have also examined multi-site graph learning, attention, and supervised FC construction [28–30]. These developments improve predictive representation and can identify salient components. However, an attention weight, compressed graph, or selected component does not necessarily quantify how an observed participant-level FC value contributes to a particular prediction [26]. Representation interpretability and prediction-level evidence are therefore related but distinct objectives.

Many model explanations are generated after prediction. Saliency maps, local surrogate methods, attention visualizations, and gradient-based attribution can highlight plausible FC patterns, yet their displayed scores may remain external to the forward decision [14, 23, 24, 26]. Visual plausibility therefore does not establish that a highlighted edge materially contributed to the model output. Prediction-coupled self-explanation addresses this limitation by placing learned explanatory evidence within inference rather than attaching an attribution after training [13]. For FC analysis, this design should retain the sign and observed magnitude of each connection so that the explanation describes evidence used by the model, not only an abstract importance coefficient.

Prediction coupling must be evaluated rather than assumed. Model-level faithfulness asks whether removing or restoring highly ranked evidence changes the output of the trained model. Deletion, insertion, and AOPC provide perturbation-based tests against random edge orderings [15, 25]. Explanation reliability also requires a separate assessment of reproducibility. High-dimensional FC contains correlated or partially substitutable connections, so independent optimization runs may recover different high-ranked edges despite similar predictive performance [27]. Cross-run overlap and recurrence can expose this uncertainty instead of presenting one initialization as definitive. Faithfulness measures model dependence, whereas reproducibility measures agreement across training processes; neither property establishes biological validity or causality.

We propose SE-FC, a prediction-coupled self-explaining FC framework that learns signed edge evidence contributing to the final decision. A hybrid backbone captures complementary global and graph-structured FC information, while a self-explanation pathway combines learned coefficients with standardized participant FC values. Faithfulness is evaluated through deletion, insertion, and AOPC, and reproducibility through Top-30 overlap, Jaccard similarity, run frequency, and representative connectivity patterns. Strict leave-one-site-out (LOSO) evaluation on ABIDE [4] and SRPBS [5] assesses whether explanation learning retains predictive utility under site shifts [17, 18]. The resulting patterns are interpreted as model-derived predictive evidence with explicit cross-run uncertainty, not as biological markers (Fig. 1).

Contribution Statement

The methodological contributions are:

1. A prediction-coupled self-explaining FC framework in which learned signed edge evidence contributes directly to inference rather than serving only as a post-hoc attribution.

2. Model-level faithfulness evaluation using deletion, insertion, AOPC, and repeated random rankings to test whether high-ranked evidence affects model output.

3. Cross-run reproducibility analysis using Top-30 overlap, Jaccard similarity, run frequency, and representative FC patterns to report recurrence and exact-edge variability under LOSO evaluation.

2. Method

SE-FC combines hybrid FC representation learning with a self-explanation pathway whose signed edge evidence contributes to prediction. The framework supports model-level interpretation without claiming biological validity.

2.1 Problem Formulation

For participant i, the input is a resting-state functional connectivity (FC) matrix $F_i \in \mathbb{R}^{116\times 116}$ constructed according to the AAL116 parcellation. The matrix is symmetric and has a zero diagonal. Its strict upper triangle is vectorized to obtain the undirected edge representation

$$x_i = vec_u\left(F_i\right) \in \mathbb{R}^E, E = \frac{116 \times 115}{2} = 6670$$

1

Within each leave-one-site-out (LOSO) fold, the mean $\backslash \mathrm{varvec}\mu$ and standard deviation $\backslash \mathrm{varvec}\sigma$ of every edge are estimated from the source-site training participants only. The standardized edge vector is

$$z_i = clip\left(\frac{x_i - \backslash \mathrm{varvec}\mu}{\backslash \mathrm{varvec}\sigma}, -6{,}6\right)$$

2

The standardized vector is reconstructed as a symmetric matrix $\widetilde{F}_i$ for backbone processing. The class label is $y_i \in \{0{,}1\}$. Conventional FC classifiers primarily learn a mapping from $\widetilde{F}_i$ to a predicted label. SE-FC extends this objective by learning both a prediction and edge-level explanatory evidence:

$$(\hat{y}_i, e_i) = \mathrm{SEFC}\left(\widetilde{F}_i\right)$$

3

The evidence vector $e_i \in \mathbb{R}^{6670}$ is participant-specific and signed. It describes how standardized FC values contribute within the prediction pathway. This definition concerns model-derived predictive

evidence and does not establish biological validity for individual connections.

2.2 Overall Framework of SE-FC

SE-FC follows five operations: construct the standardized FC matrix and 6670-edge vector; learn complementary global and graph representations; generate an edge coefficient conditioned on FC and backbone logits; multiply each coefficient by the observed standardized edge value; and combine the aggregated evidence with backbone logits for final prediction.

Let $\mathscr{B}_\phi$ denote the hybrid backbone and $\mathcal{G}_\psi$ denote the self-explanation pathway. The main data flow can be written as

$$(o_i, h_i) = B_\phi\ \left(\widetilde{F}_i\right)$$

4

$$(\backslash\mathrm{varvec}\theta\ _i, q_i) = G_\psi\ \ (sg(\ z_i), sg(o_i\ ))$$

5

$$\ell\ _i = sg\,(o_i) + \alpha\, q_i,$$
$$\hat{\mathrm{y}}_i = argmax_c[\ell\ _i]_c$$

6

where $o_i \in\ \mathbb{R}^2$ contains the backbone logits, h_i is the hybrid representation, $q_i \in\ \mathbb{R}^2$ contains the explanation logits, $\backslash\mathrm{varvec}\ell\ _i$ contains the fused logits, and $\alpha\ = 0.25$ controls the contribution of the explanation pathway.The stop-gradient operator $sg(\cdot)$ preserves the forward values while blocking gradients. Consequently, the explanation objectives train the explanation pathway without modifying the backbone feature extractor (Fig. 2).

2.3 Prediction-Coupled Self-Explanation Learning

The explanation pathway produces inference-time coefficients from separately encoded, detached FC edges and backbone logits:

$$u_i = f_{\mathrm{edge}}\left(sg(\ z_i\ )\right), v_i = f_{\mathrm{logit}}\left(sg(\ o_i\ )\right)$$

7

The encodings are projected to a bounded low-rank context:

$$c_i = \tanh\left(f_{\mathrm{ctx}}\left([\ u_i; v_i\ ]\right)\right) \in\ \mathbb{R}^R, R = 48$$

8

Three terms determine the edge coefficients: a subject-conditioned low-rank component, a class-conditioned template, and a shared edge bias. Let $B \in \mathbb{R}^{R \times E}$ denote the learned edge basis, $T \in \mathbb{R}^{2 \times E}$ denote the class templates, and $b \in \mathbb{R}^{E}$ denote the edge bias. With $p_i = softmax\left(sg(o_i)\right)$, the unscaled coefficient vector is

$$r_i = c_i B + p_i^{\top} T + b$$

9

The final coefficients are

$$\varvec{\theta}_i = \tanh\left(softplus\left(s\right) * r_i\right)$$

10

where s is a learned global scale. Each coefficient $\theta_{i,m} \in [-1,1]$ is a learned signed modulation term for edge m. It is neither a gradient nor an attention probability. Its sign and magnitude parameterize the explanation pathway, but the coefficient alone is not interpreted as the final explanatory quantity.

Signed edge evidence incorporates the participant FC value:

$$e_{i,m} = \theta_{i,m} z_{i,m}, e_i = \varvec{\theta}_i \odot z_i$$

11

Here, $z_{i,m}$ contains participant-specific standardized FC information, and $e_{i,m}$ is the quantity that enters the explanatory prediction. Positive and negative evidence can therefore be retained for directional interpretation. For ranking analyses, the magnitude $|e_{i,m}|$ is used because it measures the strength of the contribution independently of direction.

Unlike post-hoc attribution, this signed evidence directly enters the prediction pathway.

Evidence is aggregated into a scalar correction:

$$\delta_i = \frac{1}{\sqrt{E}} \sum_{m=1}^{E} e_{i,m} + b_\delta, q_i = [-\delta_i, \delta_i]$$

12

where b_δ is learned. The final logits are $\varvec{\ell}_i = sg\left(o_i\right) + \alpha\, q_i$. Thus, the learned evidence contributes directly to inference. This prediction coupling distinguishes the explanation pathway from an external visualization applied after model training.

2.4 Hybrid Functional Connectivity Representation Learning

The hybrid backbone combines an edge-vector branch for global FC patterns with a graph branch for sparse regional interactions; it is a representation component rather than a new graph architecture.

In the edge-vector branch, z_i is normalized and projected from 6670 to 256 dimensions, followed by a Gaussian error linear unit and dropout. Two residual multilayer perceptron blocks then refine the representation. Each residual block applies layer normalization, an expansion to twice the hidden dimension, a nonlinear transformation, dropout, a projection back to 256 dimensions, and a residual connection. The resulting embedding is denoted by $h_i^{\text{edge}} \in \mathbb{R}^{256}$.

For the graph branch, the row of $\widetilde{F}_i$ associated with each AAL region is used as a 116-dimensional node feature. A participant-specific adjacency matrix is constructed from the absolute FC magnitudes. For each node, the 20 strongest neighbors are retained, the directed selections are combined with their transpose to obtain a symmetric graph, and self-loops are added. If A_i is the resulting weighted adjacency and D_i is its degree matrix, the normalized adjacency is

$$\widehat{A}_i = D_i^(-1/2) A_i D_i^(-1/2)$$

13

Two residual graph layers with a hidden dimension of 64 are applied. For layer l, the update is

$$H_i^{(l+1)} = Dropout\left[GELU(\widehat{A}_i H_i^l W^l + b^l)\right] + S^l\left(H_i^l\right)$$

14

where $\mathcal{S}^{(l)}$ is an identity or linear skip projection. Mean, maximum, and minimum pooling across nodes are concatenated after the final graph layer, producing $h_i^{\text{graph}} \in \mathbb{R}^{192}$. The edge and graph embeddings are then concatenated:

$$h_i = [h_i^{\text{edge}}; h_i^{\text{graph}}]$$

15

A fusion classifier maps h_i to the backbone logits o_i. This design enables the backbone to summarize dense global edge patterns and sparse graph interactions before the self-explanation pathway constructs edge-level evidence.

2.5 Optimization Objective

The objective contains two weighted cross-entropy terms and a coefficient sparsity term.

The backbone objective is

$$\mathcal{L}_{\text{base}} = CE_w(o_i, y_i)$$

16
This term directly supervises the hybrid backbone.

The fused prediction objective is

$$\mathscr{L}_{\text{fused}} = CE_w(\varvec{\ell}_i, y_i)$$

17
This term supervises signed evidence through the fused logits without a separate consistency divergence.

The sparsity objective penalizes mean absolute coefficient magnitude:

$$\mathscr{L}_{\text{sparse}} = \frac{1}{N_b E} \sum_{i=1}^{N_b} \sum_{m=1}^{E} |\theta_{i,m}|$$

18
where N_b is the number of participants in the optimization batch. This weak L_1 term discourages uniformly large or diffuse coefficients and favors a more compact allocation of explanation weight. It is a soft preference rather than a hard edge-selection rule.

The complete objective is

$$\mathscr{L} = \mathscr{L}_{\text{base}} + \lambda_{\text{fused}} \mathscr{L}_{\text{fused}} + \lambda_{\text{sparse}} \mathscr{L}_{\text{sparse}}$$

19
with $\lambda_{\text{fused}} = 0.5$ and $\lambda_{\text{sparse}} = 10^{-5}$ in the final SE-FC configuration. Gradient isolation creates an asymmetric optimization scheme. The base objective updates the hybrid backbone, whereas the fused and sparse objectives update the self-explanation pathway. The forward prediction still combines both pathways, but explanation regularization cannot reshape the backbone representation. This routing preserves an independently optimized backbone pathway while retaining trainable prediction-coupled self-explanation.

2.6 Explanation Generation, Faithfulness, and Cross-Run Reproducibility

After training, the participant-level explanatory ranking is computed from signed edge evidence rather than from the coefficient alone. For participant i, edges are ranked in descending order of $|e_{i,m}| = |\theta_{i,m} z_{i,m}|$. This ranking reflects both the learned contribution coefficient and the FC value observed for the participant. The sign of $e_{i,m}$ is retained in exported evidence for directional inspection, while the absolute value is used for ranking.

Faithfulness evaluation. Deletion, insertion, and average output probability change (AOPC) are used to evaluate whether highly ranked edges affect model output. The participant-specific ranking is computed once from the unperturbed input and is shared across all perturbation ratios $\mathscr{R} = \{0, 0.01, 0.02, 0.05, 0.10, 0.20, 0.30, 0.50\}$. For deletion, the highest-ranked FC connections at each ratio are set to zero in the original FC space, matrix symmetry is preserved, the fold-specific standardization is reapplied, and inference is repeated. For insertion, inference starts from a zero FC matrix and progressively restores the highest-ranked original connections.

Let $\hat{c}_i$ be the class predicted from the unperturbed input, let $p_i^{(0)}$ be its probability, and let $p_i^{\mathrm{del}}(r)$ be the probability after deletion ratio r. Deletion AOPC is

$$AOPC_{\mathrm{del}} = \frac{1}{|\mathscr{R}_+|} \sum_{r \in \mathscr{R}_+} \left[p_i^{(0)} - p_i^{\mathrm{del}}(r) \right], \qquad \mathscr{R}_+ = \mathscr{R} \setminus \{0\}$$

20

A larger positive deletion AOPC indicates a stronger decrease in the original prediction after high-ranked evidence is removed. A normalized value can additionally be obtained by dividing each probability change by $p_i^{(0)}$. Random edge orderings are repeated 10 times per participant and averaged to provide a reference. These perturbation analyses evaluate model-level explanation faithfulness. They do not establish biological validity or causality.

Cross-run explanation reproducibility evaluation. Reproducibility is evaluated across the independent repeated training runs described in Section 3.1. Within each run, absolute evidence is aggregated equally across held-out LOSO folds, and the 30 highest-ranked edges form the set $\mathcal{T}_s^{30}$. For two runs a and b, overlap and Jaccard similarity are defined as

$$O(a,b) = |\mathcal{T}_a^{30} \cap \mathcal{T}_b^{30}|, J(a,b) = \frac{\left|\mathcal{T}_a^{30} \cap \mathcal{T}_b^{30}\right|}{\left|\mathcal{T}_a^{30} \cup \mathcal{T}_b^{30}\right|}$$

21

The run frequency of edge m is

$$f_m = \sum_{s \in \mathcal{S}} \mathbb{I}(m \in \mathcal{T}_s^{30}) \in \{1, 2, 3\}$$

22

Run frequency summarizes recurrence, while mean absolute evidence summarizes magnitude. These measures support reproducibility analysis and qualitative inspection of model-derived predictive patterns.

3. Experiments

The experimental protocol evaluated three complementary properties: whether signed edge evidence affects model output, whether high-ranked explanations recur after independent training, and whether prediction-coupled self-explanation retains predictive utility under held-out-site distribution shifts. Classification was treated as supporting evaluation rather than the only objective.

3.1 Datasets and Evaluation Protocol

ABIDE [4] included 435 participants from 10 sites (209 ASD and 226 controls), and SRPBS [5] included 925 participants from six sites (458 MDD and 467 controls). AAL116 FC matrices yielded 6670 undirected edges per participant.

Strict LOSO evaluation was applied separately to each dataset, producing 10 ABIDE folds and six SRPBS folds. In each fold, one site was held out exclusively for testing. The remaining source participants were divided 80:20 into training and validation subsets, stratified by class and site when feasible. Standardization, model selection, and early stopping used source data only; no held-out-site information entered development.

All methods used identical cohorts, splits, labels, and evaluation procedures. Each dataset underwent three independent training runs. Metrics were averaged across LOSO folds within each run and then reported as the mean and sample standard deviation across run-level results. The same runs supported explanation reproducibility analysis as shown in Table 1, which is listed as follows.

Table 1
Datasets and formal reporting protocol.

Dataset	Task	Participants	Class counts	Sites
ABIDE	ASD vs. control	435	209 / 226	10
SRPBS	MDD vs. control	925	458 / 467	6

Class counts are disorder / control. Results are aggregated over three independent training runs.

3.2 Baselines

Three representative baselines covered complementary FC learning paradigms. The MLP learned nonlinear combinations directly from the 6670-edge vector without graph topology. The GCN [12] represented AAL regions as nodes and FC values as weighted edges, providing a standard regional message-passing comparison. BrainIB [11] represented interpretable graph learning through an information-bottleneck formulation. All baselines used the same retained cohorts, LOSO splits, source-only validation, repeated runs, labels, and metric aggregation as SE-FC. This common protocol isolated methodological differences from data-split or reporting differences.

3.3 Implementation Details

The final model used the hybrid backbone and prediction-coupled explanation pathway described in Section 2. Detailed architectural dimensions, graph-neighborhood settings, dropout, explanation rank, loss coefficients, and additional implementation settings are provided in the Supplementary Material.

Training used AdamW with a learning rate of 0.001 and weight decay of 0.0001, class-weighted cross-entropy, and at most 160 epochs. Early stopping and model selection were based exclusively on source-site validation AUC, while the held-out site remained inaccessible. A fixed threshold of 0.5 was used for final classification.

The objective combined backbone classification, fused classification, and coefficient sparsity. The fused term supervised evidence participating in the final decision, while sparsity discouraged diffuse coefficients. Gradient isolation allowed explanation-specific terms to update the explanation pathway without modifying the backbone feature extractor.

Ablations removed fusion, sparsity, or the explanation pathway one at a time while preserving all data splits, optimization settings, and evaluation procedures.

3.4 Evaluation Metrics

Model-level explanation faithfulness. Faithfulness was evaluated without retraining. For each held-out participant, FC edges were ranked once according to the magnitude of the signed edge evidence, |e_i,m|, obtained from the unperturbed input. The same participant-specific ranking was used at all perturbation ratios. Deletion removed the highest-ranked original FC connections, whereas insertion progressively restored them from a zero FC matrix. Matrix symmetry was preserved and the fold-specific standardization was reapplied before inference. Perturbation ratios of 0, 0.01, 0.02, 0.05, 0.10, 0.20, 0.30, and 0.50 were evaluated. Random edge orderings were repeated 10 times per participant as the reference condition.

Deletion and insertion curves report changes in the probability of the class predicted from the unperturbed input. Deletion AOPC summarizes the mean probability decrease over the nonzero deletion ratios, with larger positive values indicating stronger dependence of the model output on the ranked evidence.

$$AOPC_{\mathrm{del}} = \frac{1}{|\mathscr{R}_{+}|} \sum_{r \in \mathscr{R}_{+}} \left[p_i^{(0)} - p_i^{\mathrm{del}}(r) \right]$$

23

Cross-run explanation reproducibility. For each independent repeated training run, absolute signed edge evidence was averaged equally across LOSO folds, and the 30 highest-ranked connections formed a run-level Top-30 set. Pairwise overlap counted identical AAL edges shared by two run-level sets, Jaccard similarity normalized this overlap by the size of their union, and run frequency recorded the number of repeated training runs in which an edge appeared in the Top-30 set.

$$O(a,b) = |\mathcal{T}_a^{30} \cap \mathcal{T}_b^{30}|, J(a,b) = \frac{\left|\mathcal{T}_a^{30} \cap \mathcal{T}_b^{30}\right|}{\left|\mathcal{T}_a^{30} \cup \mathcal{T}_b^{30}\right|}$$

24

$$f_m = \sum_{s \in \mathcal{S}} \mathbb{I}(m \in \mathcal{T}_s^{30})$$

25

Mean absolute evidence was retained only as a descriptive magnitude summary for qualitative inspection of representative connectivity patterns and was not treated as a reproducibility metric.

Predictive capability. Accuracy (ACC), balanced accuracy (BAC), and area under the receiver operating characteristic curve (AUC) were used to assess predictive capability under cross-site evaluation. Sensitivity and specificity were additionally reported to describe class-specific performance. BAC was computed as the mean of sensitivity and specificity, whereas AUC evaluated score ranking across decision thresholds.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

26

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \mathrm{BAC} = \frac{\mathrm{Sensitivity} + \mathrm{Specificity}}{2}$$

27

Accordingly, deletion, insertion, and AOPC assess model-level explanation faithfulness; Top-30 overlap, Jaccard similarity, and run frequency assess cross-run explanation reproducibility; and ACC, BAC, and AUC assess predictive capability. These evaluations characterize model behavior and do not establish biological validity or causal relevance.

4. Results

Results are reported in the order of faithfulness, reproducibility, representative connectivity patterns, predictive capability, and component ablation.

4.1 Model-Level Explanation Faithfulness

Removing explanation-ranked edges reduced the original predicted-class probability faster than random deletion in both datasets (Fig. 3). For example, at 10% deletion the ABIDE probability was 0.7769 for SE-FC versus 0.8391 for random ranking; the corresponding SRPBS values were 0.7553 and 0.9181.

Insertion recovered prediction confidence more rapidly for SE-FC in the sparse-to-moderate regime. At 10%, probabilities were 0.6667 versus 0.6110 on ABIDE and 0.5810 versus 0.4115 on SRPBS. At higher ratios, particularly on SRPBS, random insertion partially exceeded SE-FC, indicating that the learned ranking was most informative for compact evidence subsets.

Deletion AOPC was larger for evidence-based than random rankings: 0.1123 ± 0.0317 versus 0.0433 ± 0.0200 on ABIDE, and 0.2362 ± 0.0462 versus 0.0951 ± 0.0056 on SRPBS (Table 2; Fig. 4).

These perturbations support model-level faithfulness but do not imply physiological importance or causality.

Table 2
Deletion AOPC comparison between evidence-based and random rankings.

Dataset	Ranking	AOPC	Random repeats
ABIDE	$\|\theta \times FC\|$	0.1123 ± 0.0317	10 per participant
ABIDE	Random	0.0433 ± 0.0200	10 per participant
SRPBS	$\|\theta \times FC\|$	0.2362 ± 0.0462	10 per participant
SRPBS	Random	0.0951 ± 0.0056	10 per participant

AOPC values are reported as mean ± standard deviation across independent repeated training runs. Random rankings were repeated 10 times per participant.

4.2 Cross-Run Explanation Reproducibility

Cross-run analysis quantified agreement and uncertainty rather than treating one explanation ranking as definitive.

Exact Top-30 agreement was low (Fig. 5; Table 3). Jaccard similarity ranged from 0.0169 to 0.0345 on ABIDE and from 0.0169 to 0.0526 on SRPBS. Each dataset nevertheless contained one connection present in all three runs, with additional two-run recurrence.

The recurrent subset is therefore interpreted as a consensus pattern across repeated training, whereas variability in the remaining edges represents explanation uncertainty. This complements faithfulness, which evaluates dependence of a particular trained model on its evidence.

Table 3
Cross-run Top-30 explanation reproducibility.

Dataset	Pairwise overlaps	Jaccard range	3/3 edges	2/3 edges	1/3 edges
ABIDE	1, 2, 1	0.0169–0.0345	1	1	85
SRPBS	1, 3, 1	0.0169–0.0526	1	2	83

4.3 Model-Derived Predictive Connectivity Patterns

Table 4 and Figs. 6–7 provide qualitative examples of recurrent or relatively high-magnitude predictive connectivity evidence. They are not an exhaustive ranking or a validated biological signature. For visualization in Figs. 6 and 7, edge color was scaled by the product of run frequency and mean absolute evidence; this quantity was used only as a visualization weight and was not treated as a reproducibility metric.

In ABIDE, Frontal_Inf_Orb_R–Parietal_Inf_L recurred in all three runs (mean absolute evidence 0.1854), illustrating recurrence across independent training.

In SRPBS, Cuneus_R–Pallidum_R recurred in all three runs (0.3587). Other listed patterns span frontal-temporal, occipital-subcortical, and parietal-related systems but remain model-derived rather than disease-specific mechanisms.

These representative patterns summarize connectivity-level evidence while retaining the uncertainty indicated by limited exact-edge overlap; anatomical interpretation requires independent validation.

Table 4
Representative model-derived predictive connectivity patterns identified by SE-FC.

Dataset	AAL region pair	Run frequency	Mean absolute evidence
ABIDE	Frontal_Inf_Orb_R – Parietal_Inf_L	3	0.1854
ABIDE	Fusiform_L – Vermis_10	2	0.1688
ABIDE	Cingulum_Mid_L – Hippocampus_R	1	0.2503
SRPBS	Cuneus_R – Pallidum_R	3	0.3587
SRPBS	Precuneus_R – Temporal_Sup_L	2	0.3743
SRPBS	Frontal_Inf_Oper_L – Temporal_Pole_Sup_R	2	0.3523

Run frequency is the number of independent runs in which an edge entered the run-level Top-30 set. Mean absolute evidence summarizes model-derived magnitude. Listed connections are qualitative examples, not an exhaustive or biologically validated ranking.

4.4 Predictive Capability under Cross-Site Evaluation

Table 5 compares predictive capability under common cross-site LOSO evaluation, where each test fold contains an unseen acquisition site.

On ABIDE, the AUC of SE-FC (0.6133 ± 0.0044) was comparable to the MLP (0.6132 ± 0.0066); ACC and BAC were slightly lower than MLP but higher than GCN and BrainIB. Thus, explanation coupling preserved competitive capability without consistently improving the strongest edge-vector baseline.

On SRPBS, SE-FC achieved the highest ACC, BAC, and AUC among evaluated methods, including an AUC of 0.6127 ± 0.0051.

These dataset-dependent results support coexistence of prediction-coupled explanation and useful cross-site prediction, not universal superiority.

Table 5
Predictive performance comparison under cross-site LOSO evaluation.

Dataset	Method	ACC	BAC	AUC
ABIDE	MLP	0.5729 ± 0.0219	0.5772 ± 0.0241	0.6132 ± 0.0066
ABIDE	GCN	0.5300 ± 0.0091	0.5290 ± 0.0036	0.5477 ± 0.0117
ABIDE	BrainIB	0.5160 ± 0.0146	0.5239 ± 0.0178	0.5464 ± 0.0282
ABIDE	SE-FC	0.5623 ± 0.0306	0.5612 ± 0.0290	0.6133 ± 0.0044
SRPBS	MLP	0.5757 ± 0.0315	0.5744 ± 0.0331	0.5987 ± 0.0127
SRPBS	GCN	0.5706 ± 0.0161	0.5683 ± 0.0157	0.5887 ± 0.0089
SRPBS	BrainIB	0.5262 ± 0.0304	0.5314 ± 0.0254	0.5813 ± 0.0053
SRPBS	SE-FC	0.6019 ± 0.0170	0.5988 ± 0.0157	0.6127 ± 0.0051

Values are reported as mean ± standard deviation across independent repeated training runs, with each run-level value obtained by averaging over the corresponding LOSO folds.

4.5 Component Ablation

Ablations isolated the predictive roles of fusion, sparsity, and the explanation pathway under the same LOSO protocol (Table 6).

On SRPBS, removing the explanation pathway reduced AUC from 0.6127 to 0.5961, and removing the fused objective reduced it to 0.5964. Removing sparsity produced a smaller decrease to 0.6094, suggesting that fusion contributed more directly to prediction while sparsity mainly constrained explanation complexity.

On ABIDE, full, no-fusion, and no-sparsity AUC values were nearly identical, and removing the explanation pathway slightly increased AUC while lowering other metrics relative to some variants. Explanation-

related components therefore provided no consistent predictive advantage on this dataset.

Overall, the predictive contribution of the explanation pathway was clearer on SRPBS than ABIDE. Sparsity did not consistently improve classification, and these ablations do not establish changes in faithfulness or reproducibility.

Table 6
Component ablation under cross-site LOSO evaluation.

Dataset	Variant	ACC	BAC	AUC
ABIDE	Full SE-FC	0.5623 ± 0.0306	0.5612 ± 0.0290	0.6133 ± 0.0044
ABIDE	w/o fused objective	0.5693 ± 0.0341	0.5638 ± 0.0309	0.6134 ± 0.0025
ABIDE	w/o sparsity regularization	0.5623 ± 0.0306	0.5612 ± 0.0290	0.6132 ± 0.0043
ABIDE	w/o explanation pathway	0.5529 ± 0.0411	0.5539 ± 0.0318	0.6151 ± 0.0074
SRPBS	Full SE-FC	0.6019 ± 0.0170	0.5988 ± 0.0157	0.6127 ± 0.0051
SRPBS	w/o fused objective	0.5842 ± 0.0173	0.5809 ± 0.0183	0.5964 ± 0.0243
SRPBS	w/o sparsity regularization	0.5966 ± 0.0081	0.5935 ± 0.0067	0.6094 ± 0.0051
SRPBS	w/o explanation pathway	0.5850 ± 0.0186	0.5797 ± 0.0204	0.5961 ± 0.0087

Values are reported as mean ± standard deviation across independent repeated training runs, with each run-level value averaged over the corresponding LOSO folds.

5. Discussion

SE-FC integrated signed edge evidence into prediction and retained competitive cross-site capability. Perturbation supported model-level faithfulness, whereas repeated training revealed substantial exact-edge variability and only a small recurrent subset. These findings make the distinction among faithfulness, reproducibility, and biological validity central to interpretation. They also show why explanation quality cannot be inferred from classification performance or visual plausibility alone.

5.1 Main Findings

The prediction-coupled design makes explanation part of inference, while gradient isolation separates backbone representation learning from explanation-specific optimization. Deletion and insertion showed stronger model dependence on learned rankings than on random orderings, indicating that the evidence pathway was not merely decorative. However, faithful explanations varied across independent runs. The recurrent connections should therefore be viewed as a consensus-oriented summary accompanied by substantial uncertainty rather than as a deterministic signature. Predictive and ablation results add a further qualification: explanation coupling preserved useful classification, but its predictive contribution

was clearer on SRPBS than on ABIDE. The framework is consequently most appropriately evaluated by the combined evidence for coupling, faithfulness, reproducibility, and predictive utility rather than by any single metric.

5.2 Comparison with Existing FC Learning Methods

Edge-vector and graph classifiers primarily optimize predictive representations. Vector models can capture distributed connectivity combinations, while graph methods introduce regional topology and message passing [8–10, 12, 16, 22, 28]. Interpretable graph approaches such as BrainIB emphasize compact task-relevant components [11]. Attention and post-hoc attribution offer additional interpretive views, but they do not necessarily make observed FC evidence part of inference or demonstrate that a displayed score controls model output [23, 24, 26, 29]. SE-FC addresses a complementary objective: learned coefficients interact with standardized participant FC values to form signed evidence that enters the fused prediction [13]. The framework then tests whether this evidence affects output and whether high-ranked patterns recur after retraining. Its contribution is therefore the joint treatment of prediction coupling, model-level faithfulness, and cross-run reproducibility, rather than a claim of a new graph architecture or universally superior classifier. This distinction is relevant to recent FC-learning work that emphasizes graph, attention, and supervised representation design [28, 30].

5.3 Model-Level Explanation Faithfulness and Cross-Run Reproducibility

Faithfulness and reproducibility characterize different properties. Faithfulness asks whether ranked evidence affects the output of a particular trained model; reproducibility asks whether similar evidence is recovered after independent optimization. Neither follows automatically from predictive accuracy. In high-dimensional FC space, correlated or partially substitutable connections may support similar decisions, so an explanation can be faithful to one model while exact edge identities vary across runs. Conversely, a repeatedly selected edge is not necessarily influential unless perturbing it changes model output. Evaluating both properties therefore provides a more complete and appropriately conservative account of explanation reliability.

Evidence-based deletion and low-to-moderate insertion support faithfulness in this model-specific sense. Removing ranked evidence caused a faster decline in the original prediction than random removal, whereas restoring compact ranked subsets generally recovered confidence more rapidly. The insertion advantage was not uniform at high ratios, especially on SRPBS, and this negative result limits claims about a globally optimal edge ordering. In addition, deletion and insertion create synthetic FC matrices that may lie outside the empirical distribution even when symmetry and fold-specific standardization are preserved. Perturbation therefore demonstrates dependence of model output on ranked evidence, not physiological validity, biological importance, or causal relevance [14, 15].

Low Top-30 overlap in both datasets provides a conservative view of reproducibility. High dimensionality, correlations among connections, modest sample size relative to the feature space, and stochastic

optimization may contribute to the observed variability, although the present experiments did not isolate these factors [27]. The small recurrent subsets remain useful as auditable consensus summaries, but they should be reported together with exact-edge uncertainty. Region- or network-level aggregation and larger ensembles may prove more stable than matching atlas edges individually. Most importantly, model-level faithfulness, cross-run recurrence, and biological validity remain separate. Neither perturbation effects nor repeated selection establishes disease mechanisms or clinically validated signatures; both require independent neurobiological and external-cohort validation.

5.4 Limitations and Future Work

Several limitations constrain interpretation. The retained public multi-site cohorts are heterogeneous in acquisition and participant characteristics, and their sample sizes remain modest relative to 6670 FC edges [4–7]. Strict LOSO evaluation reduces direct information leakage and tests site shifts, but it cannot remove residual effects of scanner protocol, motion, demographics, medication, or other unmeasured confounders. The study did not include an independent external cohort or a complete confound analysis. In addition, three repeated runs provide only an initial estimate of training-related uncertainty, and the baseline comparison covers representative rather than exhaustive methods.

Faithfulness was assessed through synthetic deletion and insertion. These manipulations test dependence of the trained model on ranked edges but may produce connectomes that are not physiologically realizable. Exact-edge reproducibility also remained limited, so the listed connections should not be interpreted as fixed biological signatures. Future work should evaluate larger independent cohorts, broader repeated-run ensembles, and region- or network-level consensus. Network-level aggregation should also be tested for faithfulness in future independent studies rather than accepted solely because it produces higher apparent stability. Harmonization and explicit control of motion, demographic, clinical, and acquisition variables are needed to separate predictive structure from site-related variation. Such analyses should precede claims about neurobiological relevance or translational utility.

6. Conclusions

SE-FC is a prediction-coupled self-explaining framework for rs-fMRI functional connectivity learning. Signed edge evidence enters inference, enabling model-level faithfulness to be tested rather than assumed. Deletion, insertion, and AOPC supported dependence of model output on the learned ranking, while repeated training revealed substantial exact-edge uncertainty and a small recurrent subset. Under strict cross-site LOSO evaluation, SE-FC retained competitive prediction: performance was comparable to the strongest evaluated ABIDE baseline and highest among evaluated SRPBS methods. These findings show that prediction-coupled explanation can coexist with useful cross-site classification, although its predictive contribution remains dataset dependent. Synthetic perturbation, limited repeated runs, multi-site heterogeneity, and modest reproducibility constrain interpretation. The reported connections are model-derived predictive patterns, not validated biological signatures. External cohorts,

broader repeated-run analysis, and stronger network-level consensus are required to assess generalizability and biological relevance.

Declarations

Ethics statement

This study involved secondary analysis of de-identified data obtained from existing neuroimaging datasets. No new participants were recruited for the present study. Ethical approval and informed consent were obtained by the original investigators at the participating sites in accordance with the protocols reported for the ABIDE and SRPBS datasets [4,5]. The present study used only de-identified data released for research purposes under the corresponding data-use terms.

Data availability

The ABIDE data analyzed in this study are available through the Autism Brain Imaging Data Exchange and the Preprocessed Connectomes Project. The SRPBS data are available through the Strategic Research Program for Brain Sciences database subject to the corresponding access procedures and data-use terms. Redistribution of processed participant-level data is subject to the conditions specified by the original data providers.

Code availability

The code used for model training, evaluation, explanation analysis, and visualization is available from the corresponding author upon reasonable request.

CRediT authorship contribution statement

Mingyang Huang: Conceptualization, Supervision, Project administration, Validation, Writing – review and editing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT and Codex (OpenAI) to assist with language editing, manuscript organization, and formatting. After using these tools, the authors reviewed and edited the content as needed, verified the scientific content, numerical results, and references, and take full responsibility for the content of the publication.

References

1. Matthews PM, Jezzard P. Functional magnetic resonance imaging. J Neurol Neurosurg Psychiatry. 2004;75:6–12. https://doi.org/10.1136/jnnp.2003.032714.
2. Biswal BB, Mennes M, Zuo XN et al. Toward discovery science of human brain function, Proc. Natl. Acad. Sci. U.S.A. 107 (2010) 4734–4739. https://doi.org/10.1073/pnas.0911855107
3. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage. 2002;15:273–89. https://doi.org/10.1006/nimg.2001.0978.
4. Di Martino A, Yan CG, Li Q, et al. The autism brain imaging data exchange: toward a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry. 2014;19:659–67. https://doi.org/10.1038/mp.2013.78.
5. Tanaka SC, Yamashita A, Yahata N, et al. A multi-site, multi-disorder resting-state magnetic resonance image database. Sci Data. 2021;8:227. https://doi.org/10.1038/s41597-021-01004-8.
6. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. NeuroImage. 2012;59:2142–54. https://doi.org/10.1016/j.neuroimage.2011.10.018.
7. Yu M, Linn KA, Cook PA, et al. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data, Hum. Brain Mapp. 2018;39:4213–27. https://doi.org/10.1002/hbm.24241.
8. Kawahara J, Brown CJ, Miller SP, et al. BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage. 2017;146:1038–49. https://doi.org/10.1016/j.neuroimage.2016.09.046.
9. Ktena SI, Parisot S, Ferrante E, et al. Metric learning with spectral graph convolutions on brain connectivity networks. NeuroImage. 2018;169:431–42. https://doi.org/10.1016/j.neuroimage.2017.12.052.

10. Li X, Zhou Y, Dvornek N, et al. BrainGNN: interpretable brain graph neural network for fMRI analysis, Med. Image Anal. 2021;74:102233. https://doi.org/10.1016/j.media.2021.102233.
11. Zheng K, Yu S, Li B, Jenssen R, Chen B. BrainIB: interpretable brain network-based psychiatric diagnosis with graph information bottleneck. IEEE Trans Neural Netw Learn Syst. 2025;36:13066–79. https://doi.org/10.1109/TNNLS.2024.3449419.
12. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, 2017.
13. Alvarez-Melis D, Jaakkola TS. Towards robust interpretability with self-explaining neural networks. Adv Neural Inf Process Syst. 2018;31:7775–84.
14. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. Adv Neural Inf Process Syst. 2018;31:9505–15.
15. Petsiuk V, Das A, Saenko K. RISE: randomized input sampling for explanation of black-box models, in: British Machine Vision Conference, 2018. https://doi.org/10.5244/C.32.151
16. Heinsfeld AS, Franco AR, Craddock RC, Buchweitz A, Meneguzzi F. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. NeuroImage Clin. 2018;17:16–23. https://doi.org/10.1016/j.nicl.2017.08.017.
17. Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars, NeuroImage 180 (2018) 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061
18. Scheinost D, Noble S, Horien C, et al. Ten simple rules for predictive modeling of individual differences in neuroimaging. NeuroImage. 2019;193:35–45. https://doi.org/10.1016/j.neuroimage.2019.02.057.
19. Fox MD, Raichle ME. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat Rev Neurosci. 2007;8:700–11. https://doi.org/10.1038/nrn2201.
20. Nielsen JA, Zielinski BA, Fletcher PT, et al. Multisite functional connectivity MRI classification of autism: ABIDE results. Front Hum Neurosci. 2013;7:599. https://doi.org/10.3389/fnhum.2013.00599.
21. Abraham A, Milham MP, Di Martino A, Craddock RC, Samaras D, Thirion B, Varoquaux G. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. NeuroImage. 2017;147:736–45. https://doi.org/10.1016/j.neuroimage.2016.10.045.
22. Parisot S, Ktena SI, Ferrante E, Lee MCH, Guerrero R, Glocker B, Rueckert D. Disease prediction using graph convolutional networks: Application to autism spectrum disorder and Alzheimer's disease, Med. Image Anal. 2018;48:117–30. https://doi.org/10.1016/j.media.2018.06.001.
23. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778
24. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks, in: Proc. 34th Int. Conf. Mach. Learn., PMLR 70, 2017, pp. 3319–3328.

25. Samek W, Binder A, Montavon G, Lapuschkin S, Müller K-R. Evaluating the visualization of what a deep neural network has learned. IEEE Trans Neural Netw Learn Syst. 2017;28:2660–73. https://doi.org/10.1109/TNNLS.2016.2599820.
26. Jain S, Wallace BC. Attention is not explanation, in: Proc. NAACL-HLT 2019, vol. 1, 2019, pp. 3543–3556. https://doi.org/10.18653/v1/N19-1357
27. Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile, Proc. AAAI Conf. Artif. Intell. 33 (2019) 3681–3688. https://doi.org/10.1609/aaai.v33i01.33013681
28. Cao M, Yang M, Qin C, Zhu X, Chen Y, Wang J, Liu T. Using DeepGCN to identify the autism spectrum disorder from multi-site resting-state data. Biomed Signal Process Control. 2021;70:103015. https://doi.org/10.1016/j.bspc.2021.103015.
29. Yeh C-K, Kim J, Yen IE-H, Ravikumar P. On the (In)fidelity and Sensitivity of Explanations. Adv Neural Inf Process Syst. 2019;32. https://doi.org/10.48550/arXiv.1901.09392.
30. Pang M, Liu L, Zhao F, Fu B, Zhang L, Qiao L. Learning discriminative functional connectivity networks for identifying autism spectrum disorder. Biomed Signal Process Control. 2025;104:107466. https://doi.org/10.1016/j.bspc.2024.107466.

Figures

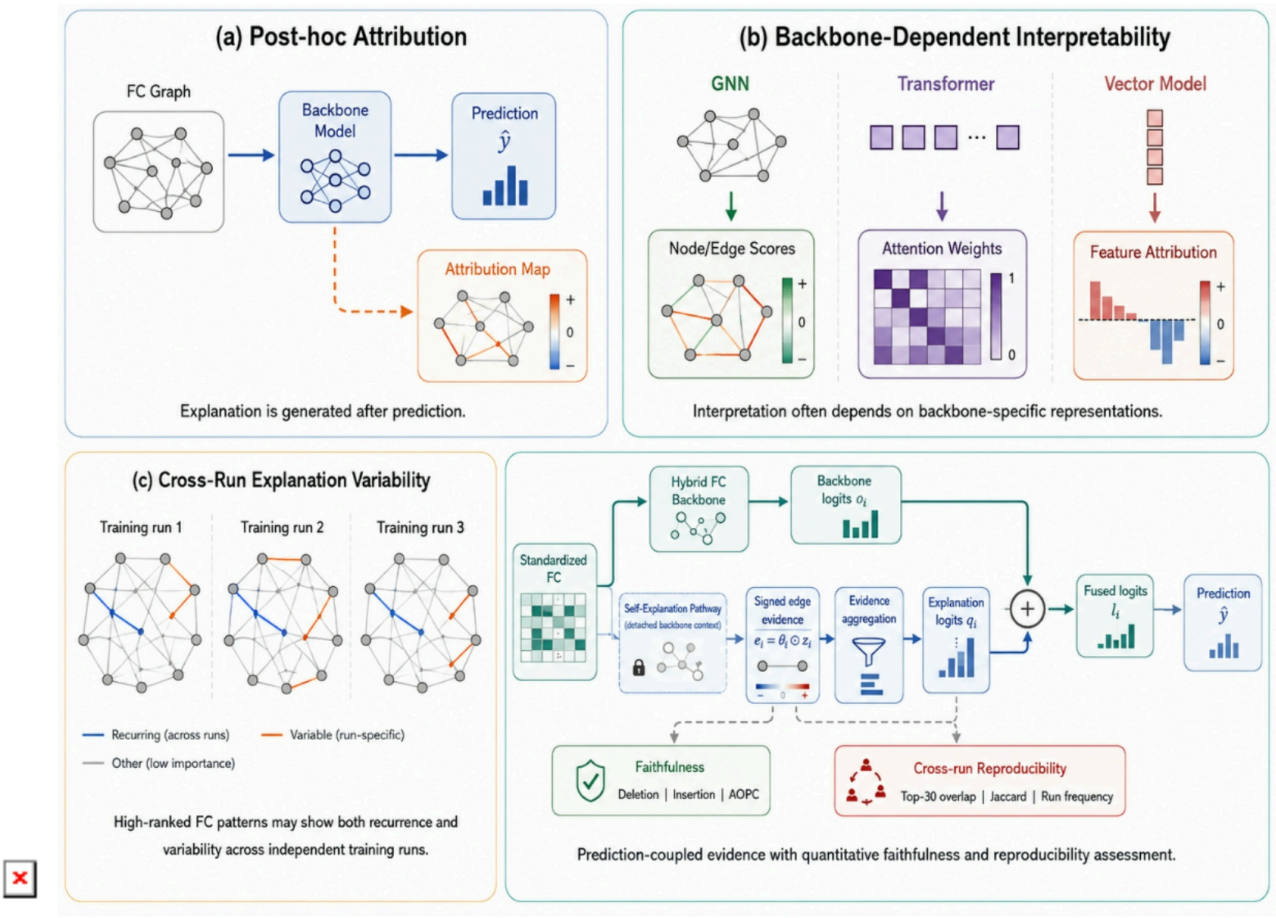


Figure 1

Conceptual motivation of the proposed SE-FC framework. (a) Post-hoc explanations are generated after prediction. (b) Interpretability mechanisms may depend on backbone-specific representations. (c) High-ranked FC patterns can vary across independent training runs. (d) SE-FC integrates signed edge evidence into prediction and separately evaluates model-level faithfulness and cross-run reproducibility.

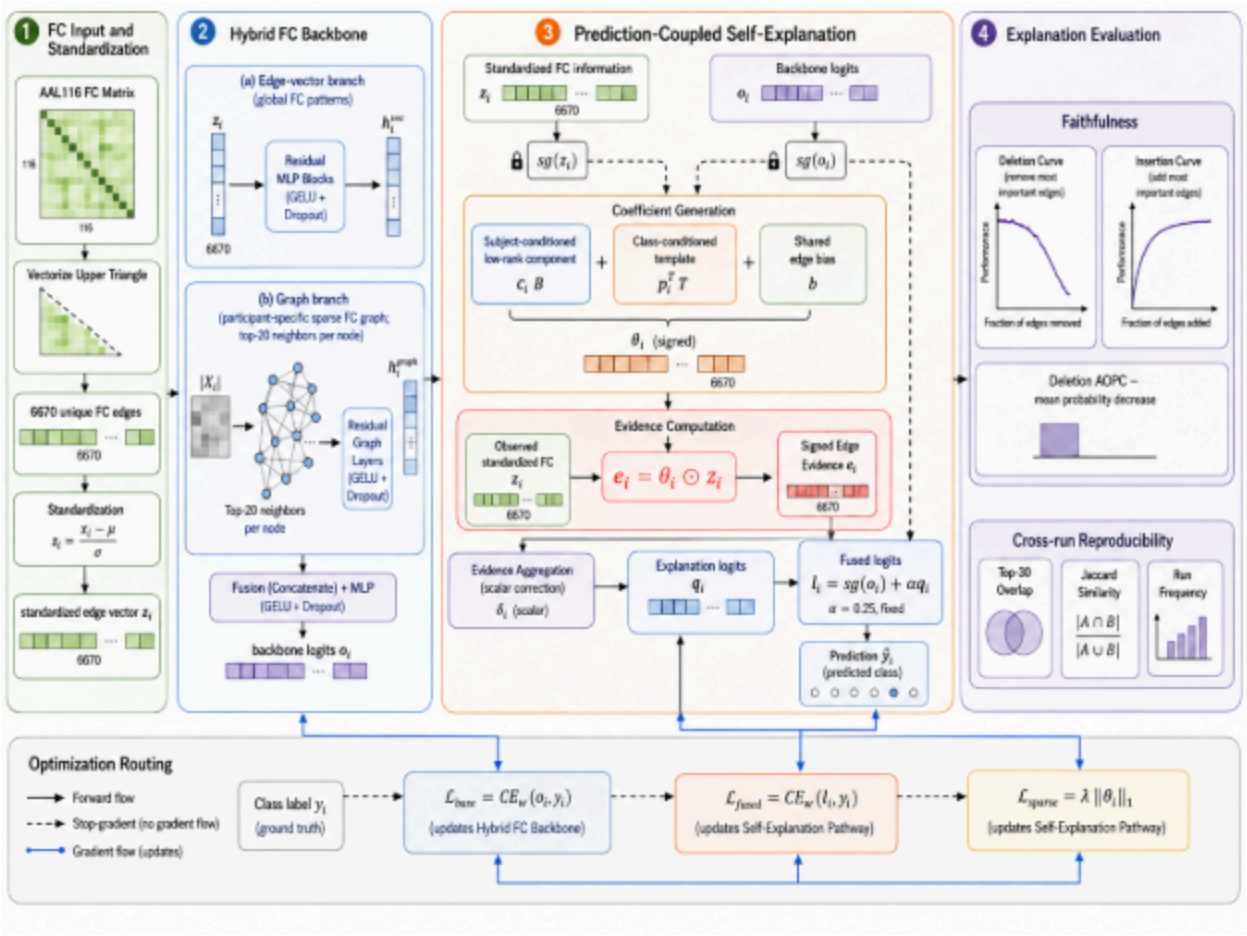


Figure 2

Overall architecture of SE-FC. Signed edge evidence is generated by the self-explanation pathway and fused with backbone logits for prediction, with separate evaluation of faithfulness and cross-run reproducibility.

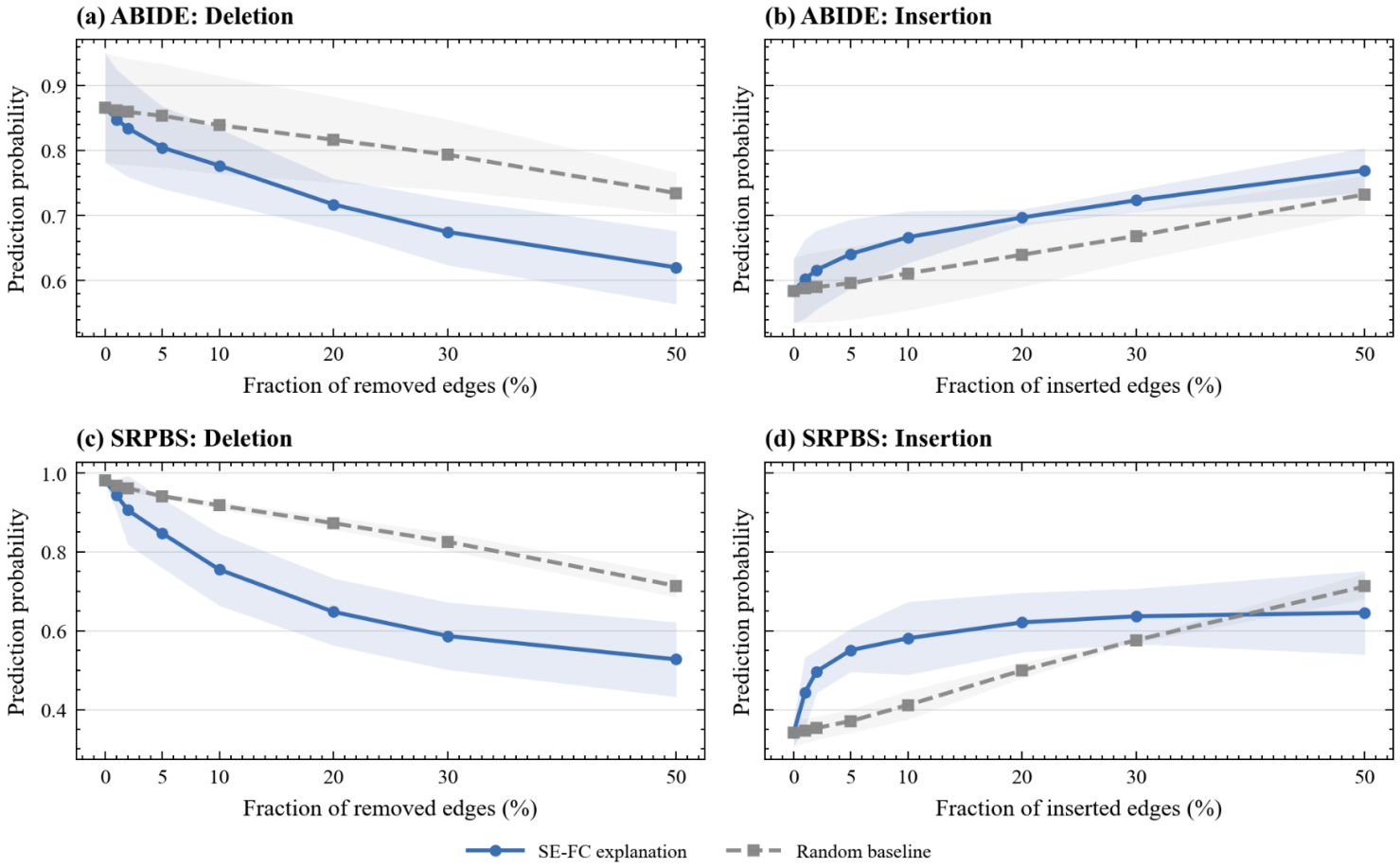


Figure 3

Dataset-separated deletion and insertion curves. Shaded regions show mean ± standard deviation across independent repeated training runs.

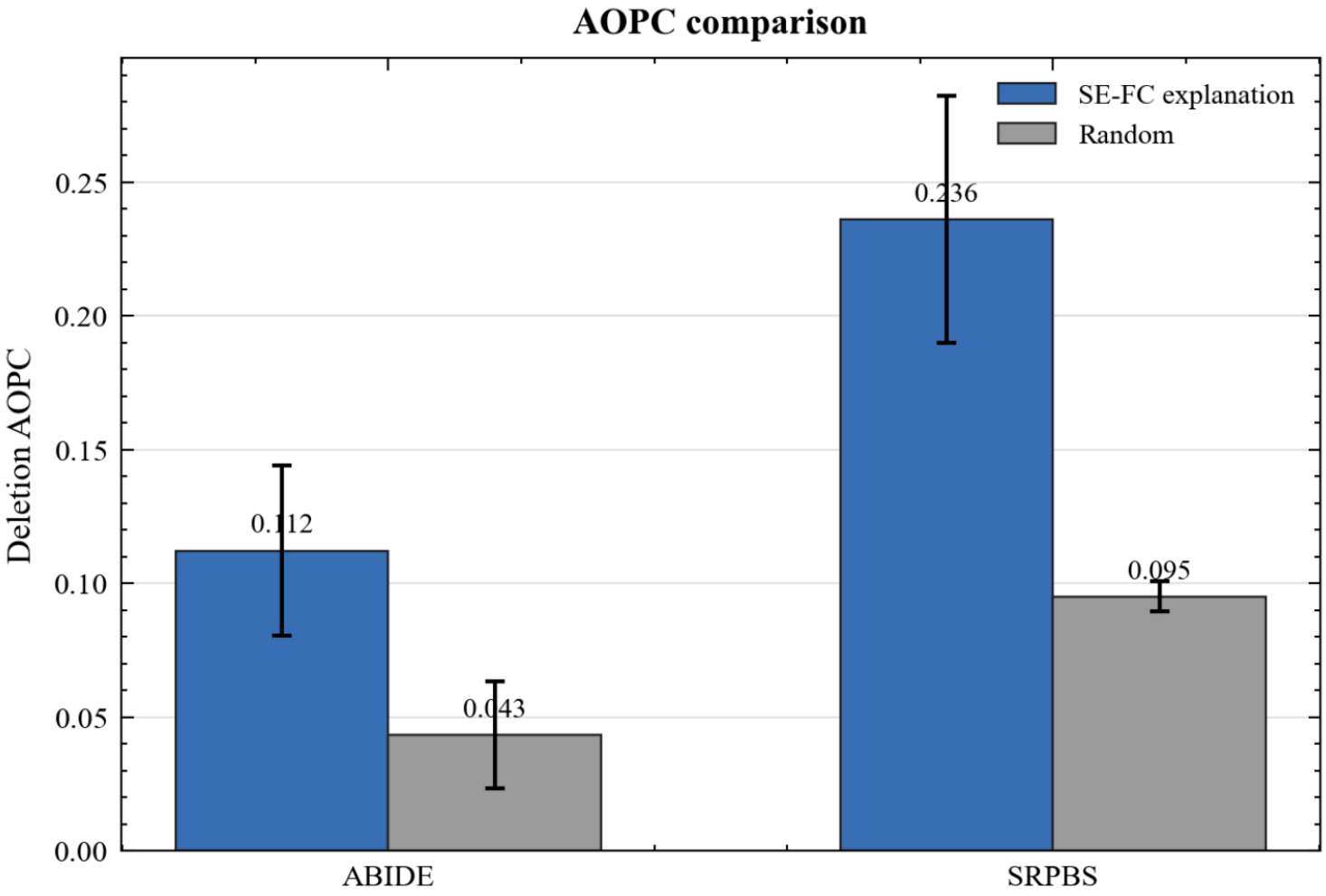


Figure 4

Deletion AOPC for evidence-based and random rankings. Error bars show standard deviation across independent repeated training runs.

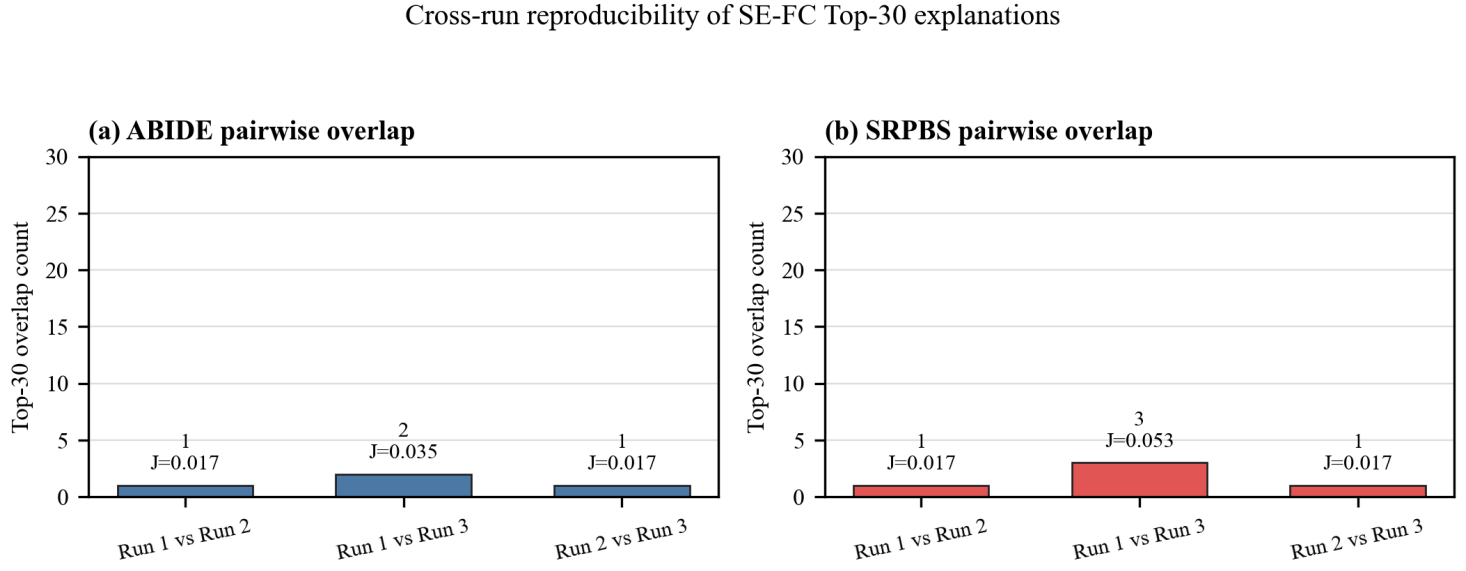


Figure 5

Cross-run reproducibility of SE-FC Top-30 explanations. Pairwise overlap and Jaccard similarity reveal a small consensus subset together with substantial exact-edge variability.

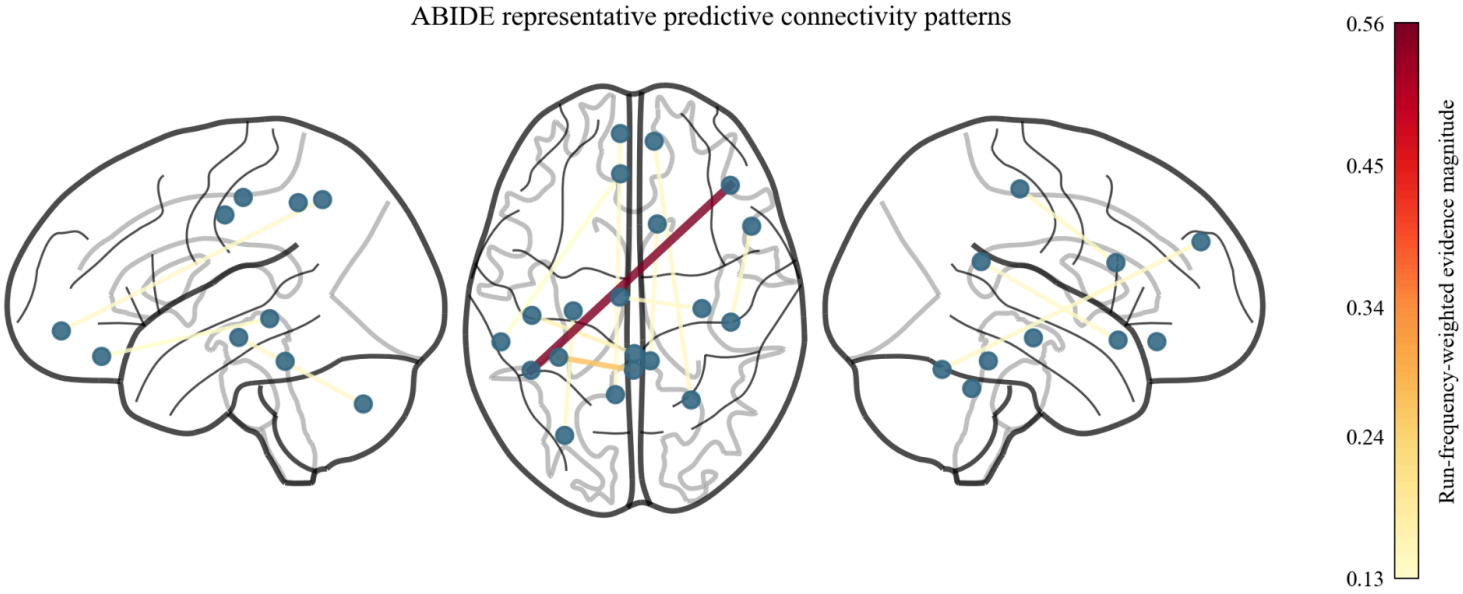


Figure 6

ABIDE representative predictive connectivity patterns. Edge color represents run frequency × mean absolute evidence and is used only for visualization; only nodes involved in the displayed connections are shown.

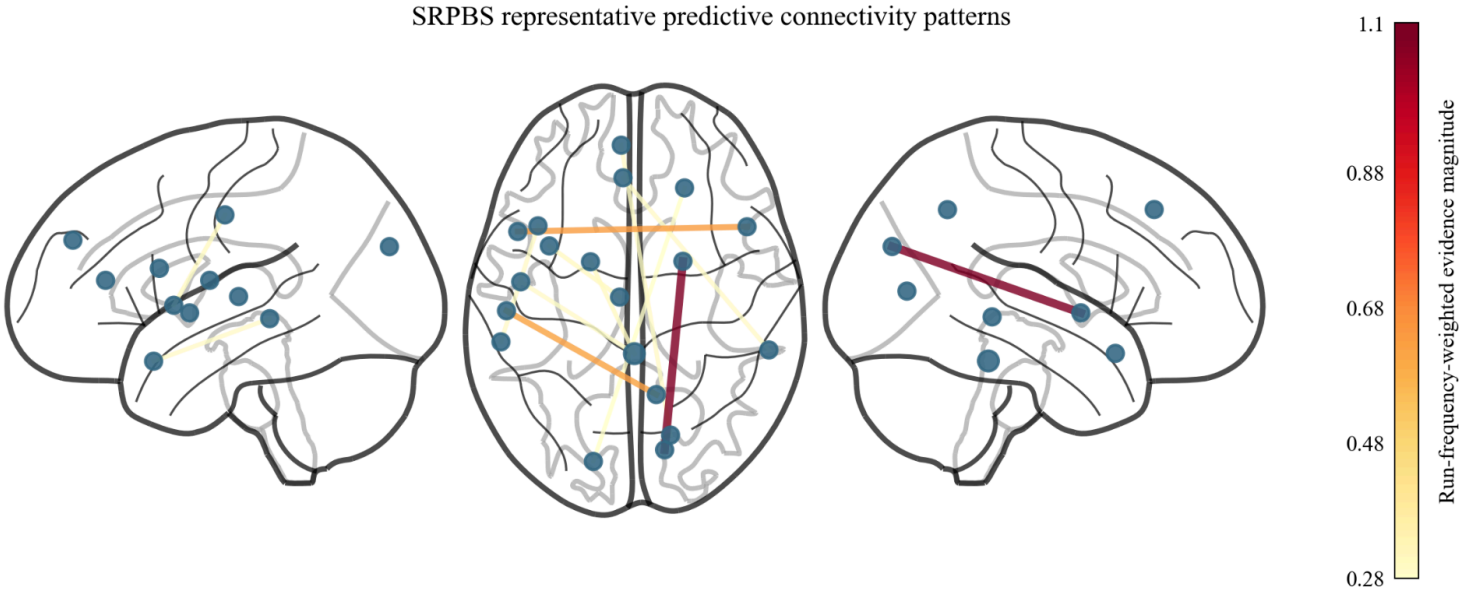


Figure 7

SRPBS representative predictive connectivity patterns. Edge color represents run frequency × mean absolute evidence and is used only for visualization; only nodes involved in the displayed connections are shown.