BEYOND SITE AGREEMENT: RE-ESTIMATION FOR BRAIN NETWORK GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-site out-of-distribution (OOD) generalization in resting-state functional magnetic resonance imaging (rs-fMRI) often relies on learning task-discriminative representations from full-scan functional connectivity (FC) graphs and promoting invariance across source sites. However, FC graphs are estimated from finite, temporally correlated blood-oxygen-level-dependent (BOLD) sequences. Cross-site agreement therefore does not necessarily imply that predictive evidence remains supported under FC re-estimation within the same scan. In this paper, we propose **B**rain Network **R**e-estimation-**I**nformed **O**OD Learning (BRIO), a framework that uses within-scan FC re-estimation to guide cross-site alignment. BRIO maps full-scan graphs and their re-estimates into consistently indexed connectome factors, enabling comparisons of their predictive contributions. It assesses re-estimation support from changes in these contributions relative to within-class subject variability and class separation. For each source-site pair and class, this task-calibrated support from both sites is combined with predictive relevance to form pairwise qualifications, which determine relative factor weights and overall alignment strength. Leave-one-site-out experiments on four real-world datasets (ABIDE, REST-meta-MDD, SRPBS, and ABCD) show that BRIO consistently outperforms competitive baselines, with relative improvements of up to 3.8% in accuracy. These gains also persist under an alternative brain parcellation on ABIDE.

1 INTRODUCTION

Resting-state functional magnetic resonance imaging (rs-fMRI) enables non-invasive characterization of macroscale functional interactions among brain regions Finn et al. (2015); Bessadok et al. (2022). A standard pipeline parcellates the brain into regions of interest (ROIs) and constructs subject-level functional connectivity (FC) graphs, where nodes represent ROIs and edges encode statistical dependencies estimated from regional blood-oxygen-level-dependent (BOLD) time series Kan et al. (2022); Yang et al. (2026). To leverage these interregional dependencies for predicting neurological and psychiatric disorders, graph neural networks (GNNs) aggregate information across connected ROIs to learn subject-level predictive representations Li et al. (2021); Gu et al. (2025).

Despite this progress, brain network models often generalize poorly to unseen sites because differences in scanners, acquisition protocols, and cohort composition induce systematic shifts in FC distributions Qiu et al. (2025); Yamashita et al. (2025); Yu et al. (2018). To address these shifts, many cross-site out-of-distribution (OOD) methods seek FC graph representations that are both task-discriminative and invariant across source sites Li et al. (2025); Xu et al. (2025); Yu et al. (2025). They pursue this goal through site-confounder removal, invariant feature or subgraph learning, and class-conditional alignment Wang et al. (2026b); Chen et al. (2022); Wang et al. (2024). However, each underlying FC graph is a finite-sample estimate derived from a temporally correlated BOLD sequence Birn et al. (2013); Noble et al. (2019); Pakravan (2026). Re-estimating FC using different temporal samples from the same scan may change the estimated connectivity patterns and their contributions to model predictions Yang et al. (2026). These within-subject changes are not directly characterized by the discrepancy between representation distributions across sites. Consequently, low cross-site discrepancy among full-scan representations does not by itself reveal whether the predictive evidence used for alignment remains supported under within-scan FC re-estimation.

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.

In this paper, we revisit cross-site OOD generalization in brain networks by accounting for variability from finite-sample FC estimation. This raises three sequential challenges: (1) *How can FC re-estimation variability be probed from a single rs-fMRI scan?* A full-scan FC graph provides only one finite-sample estimate. Probing this variability involves varying the temporal sample composition while retaining local temporal dependence and cross-ROI temporal correspondence Bellec et al. (2008); Kudela et al. (2017). (2) *How can changes in predictive evidence under FC re-estimation be represented and compared consistently?* A GNN with graph-level readout aggregates information from multiple connectivity patterns into a single predictive representation Cui et al. (2022); Han et al. (2026). Comparing full-scan and re-estimated representations as a whole does not directly reveal changes in individual patterns' predictive contributions. Resolving these changes requires consistent pattern correspondence across subjects, sites, and FC estimates. (3) *How can within-subject re-estimation variability be interpreted for cross-site alignment?* FC re-estimation probes changes within subjects, whereas cross-site alignment compares representation distributions across sites. The same contribution change can have different implications depending on within-class subject variability and class separation, which may vary across sites Yamashita et al. (2025). Thus, contribution changes need to be interpreted in the task context of both sites to assess a pattern's suitability for alignment.

To address these challenges, we propose **B**rain Network **R**e-estimation-**I**nformed **O**OD Learning (BRIO), a unified framework using within-scan FC re-estimation to guide cross-site alignment. It comprises three components: (i) *Correspondence-Preserving Connectome Factorization* constructs FC re-estimates from each BOLD sequence and maps full-scan and re-estimated graphs into consistently indexed connectome factors, enabling comparisons of predictive contributions. (ii) *Task-Calibrated Re-estimation Qualification* evaluates each factor's predictive relevance and assesses its re-estimation support for each site and class by calibrating contribution changes under FC re-estimation against within-class subject variability and class separation. (iii) *Qualification-Guided Cross-Site Alignment* combines predictive relevance with support from both sites to form factor qualifications for each source-site pair and class. Normalized qualifications determine relative factor weights, while total qualification modulates class-conditional alignment strength between full-scan representations. Leave-one-site-out (LOSO) experiments on real-world datasets (ABIDE, REST-meta-MDD, SRPBS, and ABCD) show that BRIO consistently outperforms competitive baselines, achieving relative accuracy gains of up to 3.8% and maintaining its advantage under the alternative brain parcellation.

Our contributions are summarized as follows: (1) We revisit cross-site brain network OOD generalization by considering within-scan re-estimation variability, consistent comparisons of pattern-level predictive contributions, and interpretation of re-estimation variability for cross-site alignment. (2) We propose BRIO, which learns consistently indexed connectome factors and combines predictive relevance with task-calibrated re-estimation support to determine factor weights and class-conditional alignment strength. (3) We conduct leave-one-site-out experiments on four real-world datasets, showing consistent gains over competitive baselines, including under the alternative brain parcellation.

2 Related Work

Cross-Site OOD Generalization in Brain Networks. Graph OOD generalization aims to learn robust predictors under environment shifts through invariant subgraph learning, separation of invariant and spurious information, and representation alignment Li et al. (2025); Chen et al. (2022); Wang et al. (2024). In multi-site neuroimaging, harmonization methods reduce site effects in feature distributions and covariance Yu et al. (2018); Chen et al. (2021); Yang et al. (2026), while related analyses characterize how subject, scanner, and acquisition effects influence predictive biomarkers Yamashita et al. (2025). To learn representations that generalize to unseen sites, brain network OOD methods employ transferable explanations, feature selection, and invariant subgraph learning Qiu et al. (2025); Xu et al. (2025). Other approaches use causal augmentation Yu et al. (2025) or combine site-aware deconfounding with transferable connectivity dynamics Wang et al. (2026b). BRIO uses within-scan FC re-estimation support to qualify predictive factors for cross-site alignment.

Functional Connectome Estimation and Reliability. Functional connectomes are estimated from finite BOLD sequences, with scan duration, temporal sampling, and autocorrelation affecting their reliability and statistical properties Birn et al. (2013); Huotari et al. (2019). Test–retest studies examine how acquisition and processing choices affect FC reliability Noble et al. (2019), while methodological comparisons show how connectivity estimators shape network structure Smith et al. (2011). Statistical approaches include empirical Bayes normalization of connectivity metrics and bootstrap procedures

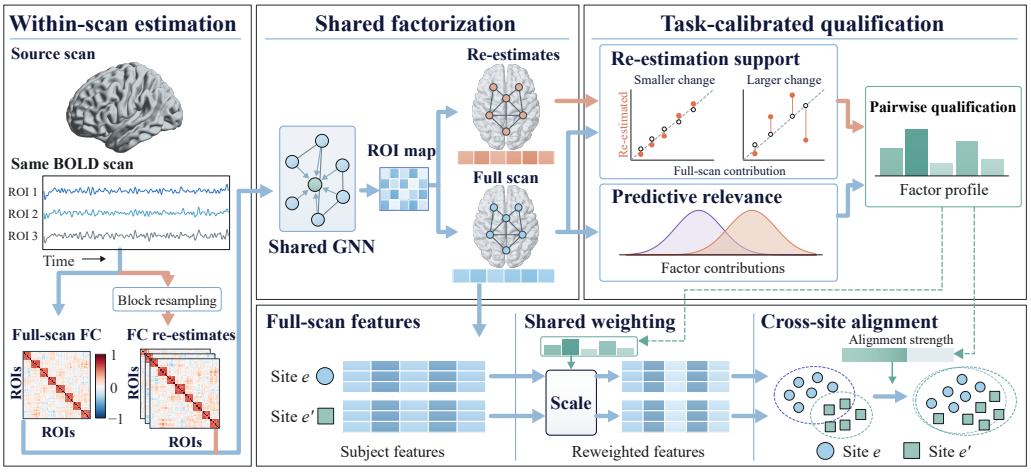


Figure 1: Overview of BRIO. Block resampling of the same BOLD scan produces FC re-estimates. Shared factorization uses a shared GNN and ROI map to obtain consistently indexed factors from full-scan and re-estimated graphs. Predictive relevance and task-calibrated re-estimation support from both sites form pairwise qualifications for each site pair and class, which control shared weighting and alignment strength for cross-site alignment of reweighted full-scan features.

for inference and uncertainty quantification Chen et al. (2015); Bellec et al. (2008); Kudela et al. (2017). Recent work develops learnable connectome representations from BOLD signals Yang et al. (2026), uses temporal segments for contrastive representation learning Lamprou et al. (2026), and explicitly models connectivity uncertainty Pakravan (2026). BRIO assesses task-calibrated re-estimation support through changes in each connectome factor's predictive contribution.

3 Methodology

Problem Setup. We study cross-site OOD generalization in brain network analysis with imaging sites as environments. Let $\mathcal{E}_{\text{tr}} = \{1, \ldots, E\}$, with $E \geq 2$, and $\mathcal{E}_{\text{te}}$ denote disjoint source and unseen test environments. Each environment e has a joint distribution $\mathbb{P}^e$ over BOLD sequences and labels. For each source site $e \in \mathcal{E}_{\text{tr}}$, we observe $\mathcal{D}^e = \{(\boldsymbol{X}_i^e, Y_i^e)\}_{i=1}^{N_e} \sim (\mathbb{P}^e)^{N_e}$, where $\boldsymbol{X}_i^e \in \mathbb{R}^{T_i^e \times P}$ is the preprocessed BOLD sequence with T_i^e valid time points and P ROIs, and $Y_i^e \in \{0, 1\}$ is the label. Sites share the label space, preprocessing pipeline, parcellation, and FC estimator, while $\mathbb{P}^e$ may differ. The shared FC estimator Ψ yields full-scan graphs $\boldsymbol{A}_i^e = \Psi(\boldsymbol{X}_i^e) \in \mathbb{R}^{P \times P}$, whose entries encode FC between ROIs. Let f_θ denote a predictor operating on FC graphs. We train it using only source data and evaluate its average risk on unseen sites:

$$\mathcal{R}_{\text{OOD}}(f_\theta) = \frac{1}{|\mathcal{E}_{\text{te}}|} \sum_{e \in \mathcal{E}_{\text{te}}} \mathbb{E}_{(\boldsymbol{X}, Y) \sim \mathbb{P}^e} \left[\ell(f_\theta(\Psi(\boldsymbol{X})), Y) \right], \tag{1}$$

where ℓ denotes the prediction loss.

Overview. As shown in Fig. 1, BRIO uses within-scan FC re-estimation to guide cross-site alignment through three components: (i) **Correspondence-Preserving Connectome Factorization** constructs FC re-estimates from each BOLD sequence and maps full-scan and re-estimated graphs into consistently indexed connectome factors using a shared signed GNN and shared ROI assignment. (ii) **Task-Calibrated Re-estimation Qualification** evaluates source-balanced predictive relevance and assesses re-estimation support for each site and class by calibrating changes in factor-wise predictive contributions against within-class subject variability and between-class separation. (iii) **Qualification-Guided Cross-Site Alignment** combines predictive relevance with support from both sites to form qualifications for each source-site pair and class. These qualifications determine shared factor weights and the overall strength of class-conditional alignment between reweighted full-scan representations.

3.1 Correspondence-Preserving Connectome Factorization

FC estimated from a finite BOLD sequence can vary with the temporal samples used Kudela et al. (2017), potentially changing the predictive contributions of connectivity patterns. To compare these

changes at the pattern level, we construct within-scan FC re-estimates and map them together with full-scan graphs into consistently indexed connectome factors.

(i) Within-Scan Re-estimation and Shared Connectome Encoding. BOLD signals exhibit temporal dependence and share a common time axis across ROIs. With $\boldsymbol{A}_i^{e,(0)} = \boldsymbol{A}_i^e$ denoting the full-scan graph, we construct K FC re-estimates by jointly resampling temporal blocks:

$$\boldsymbol{X}_i^{e,(k)} = \mathcal{R}_k(\boldsymbol{X}_i^e), \quad \boldsymbol{A}_i^{e,(k)} = \Psi\Big(\boldsymbol{X}_i^{e,(k)}\Big), \quad k = 1, \dots, K, \tag{2}$$

where $\mathcal{R}_k$ denotes joint circular moving-block resampling with a source-estimated block length ℓ_b Shao & Yu (1993); Bellec et al. (2008). Resampling retains local temporal dependence within sampled blocks and preserves cross-ROI temporal correspondence through shared sampling indices.

To encode all FC estimates over the same ROI pairs while retaining their estimate-specific weights, we construct a fixed sparse propagation support $\boldsymbol{S} \in \{0,1\}^{P\times P}$ solely from source full-scan graphs. For $u \neq v$, we compute

$$\overline{a}_{uv} = \mathrm{median}_{e\in\mathcal{E}_{\mathrm{tr}}} \, \mathrm{median}_{1\le i\le N_e} \left| \left[\boldsymbol{A}_i^{e,(0)}\right]_{uv} \right|, \quad \boldsymbol{S} = \mathrm{SymTop}_m\big(\overline{\boldsymbol{A}}\big), \tag{3}$$

where $\overline{\boldsymbol{A}} = [\overline{a}_{uv}]$ has zero diagonal and $m = \lceil \rho_g(P-1) \rceil$, with $\rho_g \in (0,1]$. The operator SymTop_m selects the m largest off-diagonal entries per row, sets $S_{uv} = S_{vu} = 1$ if either ROI selects the other, and sets all remaining entries to zero. For $k = 0, \dots, K$, we retain each FC estimate's weights on this support and separate positive weights from negative-weight magnitudes:

$$\boldsymbol{G}_i^{e,(k)} = \boldsymbol{S} \odot \boldsymbol{A}_i^{e,(k)}, \quad \boldsymbol{G}_i^{e,(k),+} = \Big[\boldsymbol{G}_i^{e,(k)}\Big]_+, \quad \boldsymbol{G}_i^{e,(k),-} = \Big[-\boldsymbol{G}_i^{e,(k)}\Big]_+, \tag{4}$$

where $\odot$ denotes elementwise multiplication and $[\cdot]_+$ takes the elementwise positive part.

The shared parcellation provides consistent ROI labels across graphs, while local connectivity on $\boldsymbol{S}$ varies across subjects and FC estimates. Let $\boldsymbol{e}_u^{\mathrm{ROI}} \in \mathbb{R}^{d_e}$ denote the learnable identity embedding of ROI u. We summarize its local connectivity using the descriptor $\boldsymbol{c}_{i,u}^{e,(k)}$:

$$\boldsymbol{c}_{i,u}^{e,(k)} = \left[\frac{1}{d_u} \sum_v G_{i,uv}^{e,(k),+}, \frac{1}{d_u} \sum_v G_{i,uv}^{e,(k),-}, \sqrt{\frac{1}{d_u} \sum_v \Big(G_{i,uv}^{e,(k)}\Big)^2} \right]^\top, \quad d_u = \sum_v S_{uv}. \tag{5}$$

Here, $G_{i,uv}^{e,(k)}$ and $G_{i,uv}^{e,(k),\pm}$ are entries of the corresponding adjacency matrices, and d_u is the degree of ROI u in $\boldsymbol{S}$. We combine ROI identity and connectivity to initialize node features, then encode them with the shared signed GNN:

$$\boldsymbol{h}_{i,u}^{e,(k),[0]} = \phi_{\mathrm{in}}\Big(\boldsymbol{e}_u^{\mathrm{ROI}}, \boldsymbol{c}_{i,u}^{e,(k)}\Big), \quad \boldsymbol{H}_i^{e,(k)} = \mathcal{G}_{\theta_g}\Big(\boldsymbol{H}_i^{e,(k),[0]}, \boldsymbol{G}_i^{e,(k),+}, \boldsymbol{G}_i^{e,(k),-}\Big), \tag{6}$$

where $\phi_{\mathrm{in}} : \mathbb{R}^{d_e} \times \mathbb{R}^3 \to \mathbb{R}^{d_h}$ initializes node features, $\boldsymbol{H}_i^{e,(k),[0]} \in \mathbb{R}^{P\times d_h}$ stacks them row-wise, and $\mathcal{G}_{\theta_g}$ is the shared signed GNN. Its output $\boldsymbol{H}_i^{e,(k)} \in \mathbb{R}^{P\times d_h}$ has u-th row $(\boldsymbol{h}_{i,u}^{e,(k)})^\top$.

(ii) Shared Connectome Factorization and Factor-Wise Scoring. Predictive connectivity patterns may involve multiple ROIs Li et al. (2021). We therefore pool the encoded ROI representations into connectome factors using shared ROI weighting profiles, so that each factor can be compared consistently across subjects, sites, and FC estimates.

Let $\boldsymbol{E}^{\mathrm{ROI}} = [\boldsymbol{e}_1^{\mathrm{ROI}}, \dots, \boldsymbol{e}_P^{\mathrm{ROI}}]^\top \in \mathbb{R}^{P\times d_e}$ stack the shared ROI identity embeddings, and let $\boldsymbol{Q} = [\boldsymbol{q}_1, \dots, \boldsymbol{q}_J]^\top$ denote J learnable factor queries. The shared assignment $\boldsymbol{\Pi} \in \mathbb{R}_+^{J\times P}$ is defined as

$$\boldsymbol{\Pi} = \mathcal{A}_\omega\left(\boldsymbol{Q}, \boldsymbol{E}^{\mathrm{ROI}}\right), \quad \boldsymbol{\Pi}\mathbf{1}_P = \frac{1}{J}\mathbf{1}_J, \quad \boldsymbol{\Pi}^\top \mathbf{1}_J = \frac{1}{P}\mathbf{1}_P, \tag{7}$$

where $\mathcal{A}_\omega$ is a learnable mapping that computes a balanced assignment from similarities between factor queries and ROI identity embeddings, and $\mathbf{1}_n$ denotes the length-n all-ones vector. Since $\boldsymbol{\Pi}$ depends only on shared queries and ROI identity embeddings, each factor uses the same ROI weighting profile across subjects, sites, and FC estimates. Pooling the encoded ROI representations under these profiles gives the connectome factors:

$$\boldsymbol{r}_{i,j}^{e,(k)} = \mathcal{F}_\phi\left(J \sum_{u=1}^{P} \Pi_{j,u} \boldsymbol{h}_{i,u}^{e,(k)}\right) \in \mathbb{S}^{d_t-1}, \quad j = 1, \dots, J, \tag{8}$$

where $\mathcal{F}_\phi : \mathbb{R}^{d_h} \to \mathbb{S}^{d_t-1}$ is a shared mapping with unit-norm output.

To retain the individual factor representations, we concatenate them in a fixed order:

$$\boldsymbol{s}_i^{e,(k)} = \frac{1}{\sqrt{J}} \left[\boldsymbol{r}_{i,1}^{e,(k)} \parallel \cdots \parallel \boldsymbol{r}_{i,J}^{e,(k)} \right]. \tag{9}$$

Here, $\parallel$ denotes concatenation. To attribute prediction changes to individual factors, we use a linear head on the concatenated representation, yielding the prediction logit and additive factor scores:

$$z_i^{e,(k)} = \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \boldsymbol{w}_j^\top \boldsymbol{r}_{i,j}^{e,(k)} + b, \quad a_{i,j}^{e,(k)} = \boldsymbol{w}_j^\top \boldsymbol{r}_{i,j}^{e,(k)}, \tag{10}$$

where $\boldsymbol{w}_j$ is the classifier block for factor j, b is the bias, and $a_{i,j}^{e,(k)}$ is the factor's additive contribution to the logit before the common $1/\sqrt{J}$ scaling.

3.2 Task-Calibrated Re-estimation Qualification

The factor-wise contributions may distinguish classes in full-scan graphs while varying across FC re-estimates of the same subject. Within-class subject variability and between-class separation provide a reference for assessing these changes at each site. We therefore evaluate source-balanced predictive relevance and task-calibrated re-estimation support.

To assess predictive relevance, we measure class separation in full-scan contributions relative to within-class variability. For source site e and class $c \in \{0, 1\}$, let $\mathcal{I}_{e,c} = \{i \in \{1, \ldots, N_e\} : Y_i^e = c\}$ and $N_{e,c} = |\mathcal{I}_{e,c}|$, with both classes represented at each source site. Each qualification evaluation computes the full-scan and re-estimated contributions of all subjects in $\mathcal{I}_{e,c}$ under the same fixed model snapshot. From the full-scan contributions, we define the empirical mean, within-class variance, and class contrast as

$$\mu_{e,c,j} = \frac{1}{N_{e,c}} \sum_{i \in \mathcal{I}_{e,c}} a_{i,j}^{e,(0)}, \quad V_{e,c,j}^{\mathrm{sub}} = \frac{1}{N_{e,c}} \sum_{i \in \mathcal{I}_{e,c}} \left(a_{i,j}^{e,(0)} - \mu_{e,c,j} \right)^2, \quad \Delta_{e,j} = \mu_{e,1,j} - \mu_{e,0,j}. \tag{11}$$

We aggregate these within-site variances and signed class contrasts with equal site weights so that larger cohorts do not dominate the relevance estimate:

$$V_{c,j}^{\mathrm{pool}} = \frac{1}{E} \sum_{e \in \mathcal{E}_{\mathrm{tr}}} V_{e,c,j}^{\mathrm{sub}}, \quad \Delta_j = \frac{1}{E} \sum_{e \in \mathcal{E}_{\mathrm{tr}}} \Delta_{e,j}. \tag{12}$$

Opposing class contrasts offset one another before Δ_j is squared. We measure relative class separation by normalizing the squared aggregate contrast by within-class variability. To account for contribution scale, we further weight this ratio by the relative squared classifier-block norm β_j, since unit-norm factors satisfy $|a_{i,j}^{e,(k)}| \le \|\boldsymbol{w}_j\|_2$. The resulting relevance scores are normalized across factors:

$$\beta_j = \frac{\|\boldsymbol{w}_j\|_2^2}{\sum_{j'=1}^{J} \|\boldsymbol{w}_{j'}\|_2^2 + \epsilon_w}, \quad R_j^{\mathrm{task}} = \frac{\beta_j \Delta_j^2}{V_{0,j}^{\mathrm{pool}} + V_{1,j}^{\mathrm{pool}} + \epsilon_s}, \quad \pi_j^{\mathrm{task}} = \frac{R_j^{\mathrm{task}}}{\sum_{j'=1}^{J} R_{j'}^{\mathrm{task}}}, \tag{13}$$

where $\epsilon_w, \epsilon_s > 0$ are numerical stabilizers. For $\sum_{j=1}^{J} R_j^{\mathrm{task}} > 0$, the normalized scores π_j^{task} form a probability distribution over factors.

We next assess within-scan re-estimation changes by averaging the squared differences between each subject's re-estimated and full-scan contributions over subjects and re-estimates within each site and class:

$$D_{e,c,j}^{\mathrm{re}} = \frac{1}{N_{e,c} K} \sum_{i \in \mathcal{I}_{e,c}} \sum_{k=1}^{K} \left(a_{i,j}^{e,(k)} - a_{i,j}^{e,(0)} \right)^2. \tag{14}$$

To interpret these deviations at each site, we combine full-scan within-class variability and class separation into a local reference scale:

$$M_{e,j}^{\mathrm{task}} = \frac{1}{4} \Delta_{e,j}^2, \quad T_{e,c,j} = V_{e,c,j}^{\mathrm{sub}} + M_{e,j}^{\mathrm{task}}, \tag{15}$$

where $M^{\text{task}}_{e,j}$ is the squared half-separation between the two class means at site e. Consequently, $T_{e,c,j}$ equals the empirical mean squared distance of full-scan contributions in class c from the midpoint of these means. We use this reference scale to define re-estimation support for each site and class:

$$q^{\text{re}}_{e,c,j} = \frac{T_{e,c,j} + \epsilon_q}{T_{e,c,j} + D^{\text{re}}_{e,c,j} + \epsilon_q} \in (0, 1], \tag{16}$$

where $\epsilon_q > 0$ is a numerical stabilizer. Higher support indicates smaller mean squared contribution changes relative to the local full-scan reference scale.

Proposition 1 (Class-Contrast Stability under FC Re-estimation) *Fix the model snapshot, a source site e, and a factor j, with $N_{e,0}, N_{e,1} > 0$ and $K \geq 1$. Let $\Delta^{\text{re}}_{e,j}$ be the class contrast of contributions averaged over subjects and FC re-estimates, and let $\kappa_{e,c,j} = \Delta^2_{e,j} / \big(4(V^{\text{sub}}_{e,c,j} + \epsilon_q)\big)$ be the full-scan separation ratio of class c. Then*

$$\left|\Delta^{\text{re}}_{e,j} - \Delta_{e,j}\right| \leq \sum_{c \in \{0,1\}} \sqrt{(T_{e,c,j} + \epsilon_q)\left(\frac{1}{q^{\text{re}}_{e,c,j}} - 1\right)}, \tag{17}$$

and $\Delta^{\text{re}}_{e,j}$ has the same sign as $\Delta_{e,j}$ whenever $q^{\text{re}}_{e,c,j} > (1+\kappa_{e,c,j})/(1+2\kappa_{e,c,j})$ for both $c \in \{0, 1\}$.

Proposition 1 bounds the change of class contrast under FC re-estimation by the re-estimation variation, and shows that the support needed to preserve the sign of the contrast decreases with full-scan class separation, from one for weakly separated factors to one half for strongly separated ones. A factor with high relevance but low support, or low relevance but high support, can thus flip its class contrast under re-estimation, so we qualify factors by assessing the two criteria jointly.

3.3 Qualification-Guided Cross-Site Alignment

Predictive relevance combines class separation with contribution scale across source sites, whereas re-estimation support varies by site and class. For class-conditional alignment, we combine predictive relevance with support from both sites to determine relative factor weights and alignment strength.

For two distinct source sites $e, e' \in \mathcal{E}_{\text{tr}}$, class $c \in \{0, 1\}$, and factor j, we combine re-estimation support from both sites and weight it by predictive relevance:

$$q^{\text{pair}}_{e,e',c,j} = \sqrt{q^{\text{re}}_{e,c,j} q^{\text{re}}_{e',c,j}}, \quad \alpha_{e,e',c,j} = \pi^{\text{task}}_j q^{\text{pair}}_{e,e',c,j}. \tag{18}$$

The geometric mean treats the two sites symmetrically and decreases when support at either site decreases. We normalize the qualifications to determine relative factor weights and retain their sum as the overall qualification for scaling the alignment loss:

$$g_{e,e',c} = \sum_{j=1}^{J} \alpha_{e,e',c,j}, \quad \overline{\alpha}_{e,e',c,j} = \frac{\alpha_{e,e',c,j}}{g_{e,e',c}}, \quad \sum_{j=1}^{J} \overline{\alpha}_{e,e',c,j} = 1, \tag{19}$$

where $0 < g_{e,e',c} \leq 1$.

Proposition 2 (Variational Form of Pairwise Qualification) *For a probability vector $\boldsymbol{\pi}$ over J factors and supports $q_j \in (0, 1]$, define $\alpha_j = \pi_j q_j$, $g = \sum_{j=1}^{J} \alpha_j$, and $\overline{\alpha}_j = \alpha_j / g$. Let $\mathcal{V}_\pi$ be the set of probability vectors $\boldsymbol{v}$ with $v_j = 0$ whenever $\pi_j = 0$, and define*

$$\mathcal{J}(\boldsymbol{v}) = \text{KL}(\boldsymbol{v} \,\|\, \boldsymbol{\pi}) + \sum_{j=1}^{J} v_j \log \frac{1}{q_j}. \tag{20}$$

Then $\overline{\boldsymbol{\alpha}}$ is the unique minimizer of $\mathcal{J}$ over $\mathcal{V}_\pi$, with minimum value $-\log g$, and

$$\overline{\alpha}_j = \frac{\partial \log g}{\partial \log q_j}, \quad -\log g \leq \sum_{j=1}^{J} \pi_j \log \frac{1}{q_j}, \tag{21}$$

where equality holds if and only if q_j is constant over the factors with $\pi_j > 0$.

With $\boldsymbol{\pi} = \boldsymbol{\pi}^{\text{task}}$ and $q_j = q^{\text{pair}}_{e,e',c,j}$, Proposition 2 recovers $\overline{\boldsymbol{\alpha}}_{e,e',c}$ and $g_{e,e',c}$. The normalized qualifications are the weights closest to predictive relevance under a penalty for weak support, each equal to the sensitivity of alignment strength to its factor's support. Weights and strength thus derive from one objective, and reweighting lowers the support penalty below relevance-only weighting.

We use the normalized qualifications derived from predictive contributions to weight full-scan factor representations identically at both sites. For $d \in \{e, e'\}$ and subject $i \in \mathcal{I}_{d,c}$, the weighted representation is

$$\boldsymbol{\varphi}^{Q,d}_{i|e,e',c} = \frac{1}{\sqrt{2}} \left[\sqrt{\overline{\alpha}_{e,e',c,1}}\, \boldsymbol{r}^{d,(0)}_{i,1} \parallel \cdots \parallel \sqrt{\overline{\alpha}_{e,e',c,J}}\, \boldsymbol{r}^{d,(0)}_{i,J} \right]. \tag{22}$$

For subjects $i \in \mathcal{I}_{e,c}$ and $h \in \mathcal{I}_{e',c}$, their squared distance in the common weighted space is

$$\left\| \boldsymbol{\varphi}^{Q,e}_{i|e,e',c} - \boldsymbol{\varphi}^{Q,e'}_{h|e,e',c} \right\|_2^2 = \sum_{j=1}^{J} \overline{\alpha}_{e,e',c,j} \left[1 - \left(\boldsymbol{r}^{e,(0)}_{i,j} \right)^{\top} \boldsymbol{r}^{e',(0)}_{h,j} \right]. \tag{23}$$

Each factor's cosine dissimilarity is weighted by its relative qualification. Since the factors have unit norm and the weights sum to one, the squared distance lies in $[0, 2]$. We then construct the empirical class-conditional distributions in this weighted space. Given nonempty source subsets $\mathcal{B}_{e,c} \subseteq \mathcal{I}_{e,c}$ and $\mathcal{B}_{e',c} \subseteq \mathcal{I}_{e',c}$, we define

$$\widehat{P}^{Q[e,e']}_{e,c} = \frac{1}{|\mathcal{B}_{e,c}|} \sum_{i \in \mathcal{B}_{e,c}} \delta_{\boldsymbol{\varphi}^{Q,e}_{i|e,e',c}}, \quad \widehat{P}^{Q[e,e']}_{e',c} = \frac{1}{|\mathcal{B}_{e',c}|} \sum_{h \in \mathcal{B}_{e',c}} \delta_{\boldsymbol{\varphi}^{Q,e'}_{h|e,e',c}}, \tag{24}$$

where $\delta_{\boldsymbol{x}}$ denotes a point mass at $\boldsymbol{x}$. We compare these distributions using the debiased Sinkhorn divergence $\mathcal{S}_\varepsilon$ with the squared Euclidean ground cost and entropic regularization parameter $\varepsilon > 0$ Cuturi (2013); Feydy et al. (2019). Details of debiased Sinkhorn divergence $\mathcal{S}_\varepsilon$ are provided in Appendix D. Weighting each divergence by $g_{e,e',c}$ and averaging over source-site pairs and classes yields the alignment loss:

$$\mathcal{L}^{Q}_{\text{env}} = \frac{2}{E(E-1)} \sum_{\substack{e,e' \in \mathcal{E}_{\text{tr}} \\ e < e'}} \frac{1}{2} \sum_{c \in \{0,1\}} g_{e,e',c} \mathcal{S}_\varepsilon \left(\widehat{P}^{Q[e,e']}_{e,c}, \widehat{P}^{Q[e,e']}_{e',c} \right). \tag{25}$$

3.4 Learning Objective

For each full-scan graph, the predicted probability is $\widehat{p}^{\,e,(0)}_i = \sigma\left(z^{e,(0)}_i \right)$, where σ denotes the sigmoid function. We define the classification loss with equal weights across source sites and classes:

$$\mathcal{L}_{\text{cls}} = \frac{1}{2E} \sum_{e \in \mathcal{E}_{\text{tr}}} \sum_{c \in \{0,1\}} \frac{1}{N_{e,c}} \sum_{i \in \mathcal{I}_{e,c}} \ell_{\text{BCE}} \left(\widehat{p}^{\,e,(0)}_i, Y^e_i \right), \tag{26}$$

where ℓ_{BCE} denotes binary cross-entropy. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{env}} \mathcal{L}^{Q}_{\text{env}}, \tag{27}$$

where $\lambda_{\text{env}} \geq 0$ controls the strength of cross-site alignment.

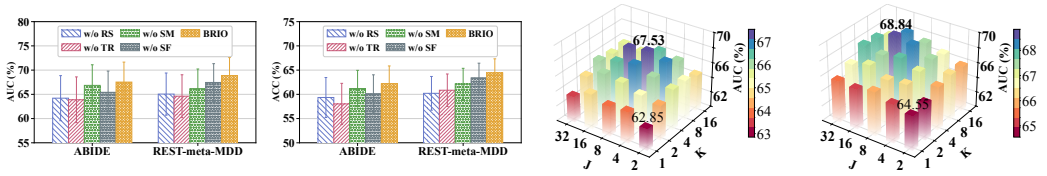
4 Experiments

Datasets. To evaluate the effectiveness of BRIO, we conduct LOSO experiments on four real-world fMRI datasets: ABIDE Di Martino et al. (2014), REST-meta-MDD Yan et al. (2019), SRPBS Tanaka et al. (2021), and ABCD Casey et al. (2018). These datasets cover prediction tasks related to autism spectrum disorder (ASD), major depressive disorder (MDD), and attention-deficit/hyperactivity disorder (ADHD), with ABCD providing a developmental cohort. We additionally include ABIDE (CC200), constructed using the CC200 functional parcellation, to examine the effect of atlas choice. More details of these datasets are provided in Appendix E.

Baselines. We compare BRIO with competitive baselines as follows: (1) General GNNs: GCN Kipf & Welling (2016), GAT Veličković et al. (2017), and GIN Xu et al. (2018); (2) General OOD methods: CORAL Sun & Saenko (2016) and IRM Arjovsky et al. (2019); (3) Graph OOD methods: GSAT Miao et al. (2022), DisC Fan et al. (2022), CEPG Wang et al. (2026a), and DiSCO Sun et al. (2026); (4) Brain network methods: BrainNetTF Kan et al. (2022), XG-GNN Qiu et al. (2024), AGMGC Noman et al. (2025), FC-HGNN Gu et al. (2025), BrainOOD Xu et al. (2025), DeCI Yu et al. (2026) and CORE Wang et al. (2026b). More details of baselines are provided in Appendix F.

Table 1: Results on ABIDE, ABIDE (CC200), REST-meta-MDD, SRPBS, and ABCD (ADHD-task) over LOSO sites. **Bold** results indicate the best performance.

Type	Method	ABIDE		ABIDE (CC200)		REST-meta-MDD		SRPBS		ABCD (ADHD-task)	
		AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
GNN	GCN	61.91$_{\pm 10.32}$	56.12$_{\pm 9.66}$	50.21$_{\pm 8.86}$	48.38$_{\pm 8.55}$	60.58$_{\pm 4.18}$	56.26$_{\pm 2.87}$	65.24$_{\pm 19.47}$	61.68$_{\pm 19.17}$	71.75$_{\pm 6.52}$	63.14$_{\pm 9.15}$
	GAT	58.21$_{\pm 8.69}$	52.31$_{\pm 6.01}$	57.68$_{\pm 9.37}$	54.69$_{\pm 7.74}$	59.97$_{\pm 9.38}$	56.17$_{\pm 6.72}$	66.68$_{\pm 15.58}$	60.86$_{\pm 14.38}$	70.37$_{\pm 11.32}$	63.32$_{\pm 11.20}$
	GIN	57.73$_{\pm 5.21}$	52.08$_{\pm 3.10}$	52.40$_{\pm 8.83}$	47.84$_{\pm 10.23}$	56.55$_{\pm 6.72}$	53.46$_{\pm 4.72}$	63.74$_{\pm 21.68}$	60.19$_{\pm 18.08}$	68.97$_{\pm 8.84}$	60.73$_{\pm 7.03}$
OOD	CORAL	55.48$_{\pm 6.12}$	50.23$_{\pm 5.49}$	53.75$_{\pm 10.69}$	50.06$_{\pm 7.42}$	60.37$_{\pm 8.97}$	57.18$_{\pm 7.78}$	63.88$_{\pm 15.48}$	59.86$_{\pm 12.68}$	69.43$_{\pm 7.70}$	62.55$_{\pm 10.46}$
	IRM	58.56$_{\pm 6.12}$	52.98$_{\pm 3.85}$	51.77$_{\pm 10.64}$	50.84$_{\pm 6.99}$	58.56$_{\pm 4.38}$	54.18$_{\pm 3.35}$	60.98$_{\pm 21.87}$	59.27$_{\pm 16.78}$	70.37$_{\pm 6.60}$	59.06$_{\pm 7.57}$
Graph OOD	GSAT	58.48$_{\pm 7.62}$	53.61$_{\pm 7.98}$	50.81$_{\pm 7.05}$	49.05$_{\pm 6.27}$	57.38$_{\pm 5.77}$	54.76$_{\pm 4.52}$	62.76$_{\pm 18.97}$	59.18$_{\pm 15.78}$	71.27$_{\pm 8.73}$	61.16$_{\pm 7.48}$
	DisC	56.46$_{\pm 7.85}$	53.62$_{\pm 8.68}$	52.85$_{\pm 6.48}$	50.75$_{\pm 6.41}$	56.19$_{\pm 4.42}$	53.27$_{\pm 3.88}$	65.18$_{\pm 19.75}$	61.46$_{\pm 16.79}$	69.59$_{\pm 7.70}$	58.83$_{\pm 9.45}$
	CEPG	50.83$_{\pm 8.99}$	47.51$_{\pm 6.12}$	44.82$_{\pm 9.74}$	45.45$_{\pm 8.93}$	53.89$_{\pm 5.27}$	52.96$_{\pm 4.46}$	55.34$_{\pm 8.64}$	50.58$_{\pm 6.28}$	63.11$_{\pm 11.06}$	57.07$_{\pm 9.59}$
	DiSCO	56.57$_{\pm 6.72}$	52.96$_{\pm 4.83}$	51.37$_{\pm 8.45}$	49.35$_{\pm 6.21}$	57.92$_{\pm 3.73}$	54.08$_{\pm 2.84}$	65.68$_{\pm 18.16}$	59.12$_{\pm 15.46}$	67.51$_{\pm 9.83}$	61.98$_{\pm 7.44}$
Brain Networks	BrainNetTF	62.64$_{\pm 6.20}$	58.38$_{\pm 7.06}$	60.48$_{\pm 8.12}$	54.25$_{\pm 7.14}$	61.48$_{\pm 7.38}$	58.17$_{\pm 5.58}$	53.04$_{\pm 17.02}$	51.89$_{\pm 10.78}$	68.28$_{\pm 4.54}$	56.91$_{\pm 10.43}$
	XG-GNN	60.20$_{\pm 7.73}$	51.61$_{\pm 4.09}$	55.94$_{\pm 9.85}$	53.64$_{\pm 6.42}$	60.08$_{\pm 7.23}$	56.24$_{\pm 6.49}$	73.28$_{\pm 15.19}$	66.17$_{\pm 17.27}$	71.20$_{\pm 5.46}$	55.66$_{\pm 9.78}$
	AGMGC	58.47$_{\pm 8.68}$	48.97$_{\pm 5.42}$	59.32$_{\pm 6.54}$	53.17$_{\pm 5.08}$	59.06$_{\pm 5.07}$	54.74$_{\pm 4.12}$	74.98$_{\pm 15.31}$	67.34$_{\pm 16.85}$	70.86$_{\pm 8.67}$	63.02$_{\pm 11.29}$
	FC-HGNN	55.05$_{\pm 7.90}$	49.71$_{\pm 4.58}$	58.76$_{\pm 7.21}$	54.82$_{\pm 5.16}$	58.69$_{\pm 6.69}$	56.78$_{\pm 5.71}$	48.78$_{\pm 10.58}$	45.28$_{\pm 6.38}$	71.08$_{\pm 5.74}$	59.10$_{\pm 12.34}$
	BrainOOD	63.61$_{\pm 6.73}$	57.14$_{\pm 9.70}$	66.15$_{\pm 4.12}$	58.21$_{\pm 3.84}$	63.22$_{\pm 6.29}$	58.45$_{\pm 5.20}$	81.48$_{\pm 7.31}$	69.18$_{\pm 15.21}$	72.77$_{\pm 8.07}$	64.36$_{\pm 7.24}$
	DeCI	55.81$_{\pm 8.91}$	58.26$_{\pm 2.89}$	62.18$_{\pm 3.75}$	59.48$_{\pm 3.52}$	54.38$_{\pm 2.95}$	56.18$_{\pm 2.89}$	82.05$_{\pm 12.28}$	76.69$_{\pm 13.66}$	55.03$_{\pm 6.79}$	62.81$_{\pm 5.46}$
	CORE	65.32$_{\pm 6.88}$	60.97$_{\pm 5.90}$	68.42$_{\pm 6.35}$	59.65$_{\pm 5.42}$	67.44$_{\pm 8.20}$	62.81$_{\pm 6.04}$	82.61$_{\pm 8.25}$	77.65$_{\pm 9.84}$	74.31$_{\pm 8.34}$	67.26$_{\pm 7.53}$
	BRIO	**67.53**$_{\pm 4.12}$	**62.18**$_{\pm 3.68}$	**70.86**$_{\pm 4.87}$	**61.92**$_{\pm 3.94}$	**68.84**$_{\pm 3.75}$	**64.42**$_{\pm 2.91}$	**83.91**$_{\pm 5.63}$	**78.82**$_{\pm 5.28}$	**75.98**$_{\pm 4.39}$	**68.74**$_{\pm 3.82}$



(a) AUC (b) ACC (c) ABIDE (d) REST-meta-MDD

Figure 2: Ablation results in (a,b) and sensitivity results in (c,d) on ABIDE and REST-meta-MDD.

4.1 Performance Comparison

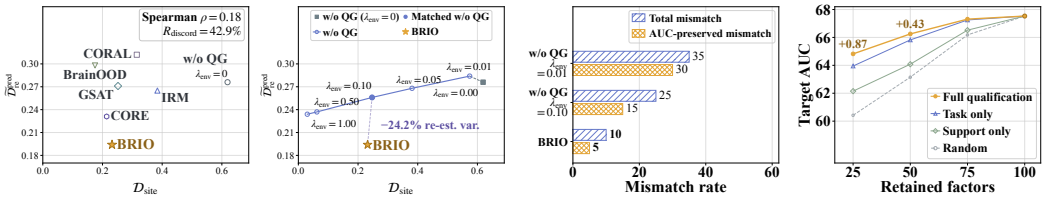
We compare BRIO with baseline methods under the LOSO protocol on ABIDE, REST-meta-MDD, SRPBS, and ABCD in Table 1. We observe that: (1) General OOD methods do not consistently outperform general GNNs across datasets, suggesting that generic robustness objectives offer limited gains in cross-site brain network generalization. (2) The best-performing brain network methods consistently outperform graph OOD methods, while the gains from other brain network models vary across datasets. This suggests that cross-site generalization benefits not only from modeling brain connectivity but also from learning task-relevant patterns that transfer across sites. (3) BRIO consistently outperforms all baselines across datasets. These gains can be attributed to two complementary mechanisms: (i) shared connectome factorization enables consistent comparisons of predictive contributions across FC estimates, while task-calibrated qualification identifies task-relevant factors supported under re-estimation; (ii) qualification-guided alignment prioritizes factors supported by both sites and adjusts the overall alignment strength accordingly, limiting the influence of weakly supported factors on cross-site matching. We further evaluate performance on ABIDE using CC200 (Craddock et al., 2012) instead of AAL. Although baseline rankings vary between the two parcellations, BRIO outperforms all baselines on both metrics, indicating that its gains persist across parcellations.

4.2 Ablation Study

We evaluate four ablation variants: (1) BRIO w/o RS uses task relevance alone to weight factors and fixes the strength multiplier to one; (2) BRIO w/o TR replaces explicit task relevance weights with uniform weights while retaining task-calibrated re-estimation support; (3) BRIO w/o SM retains qualification-based factor weights but fixes the strength multiplier to one; and (4) BRIO w/o SF uses task relevance alone to weight factors while retaining qualification-based strength modulation. As shown in Fig. 2(a,b), removing RS reduces performance, indicating that assessing predictive contributions under FC re-estimation provides benefits beyond full-scan task relevance. Removing TR degrades performance, highlighting the value of prioritizing discriminative factors rather than relying on re-estimation support alone. The performance drops without SM and SF support the complementary roles of qualification in modulating alignment strength for each site pair and class and prioritizing supported factors in cross-site matching. More results are shown in Appendix J.4.

4.3 Sensitivity Study

We examine the sensitivity of BRIO to the number of FC re-estimates K and connectome factors J. Fig. 2(c,d) shows that increasing K from small values generally improves AUC on ABIDE and



(a) Cross-method (b) Alignment (c) Mismatch (d) Factor retention

Figure 3: Cross-method relationships (a), alignment trajectories (b), mismatch rates (c), and factor retention (d) on ABIDE.

REST-meta-MDD, whereas further increases yield no consistent gains. This suggests that additional re-estimates help characterize variation in predictive contributions, with diminishing benefits at larger budgets. We vary J with the total representation dimension fixed. Intermediate values perform better than very small or large values on both datasets, suggesting a trade-off between separating connectivity patterns and preserving sufficient capacity within each factor. Too few factors may combine patterns with distinct predictive roles and re-estimation responses, whereas too many reduce each factor's dimension and may limit its discriminative capacity. More results are shown in Appendix J.5.

4.4 Case Study

Site Discrepancy and Re-estimation Variation. To examine whether lower source-site discrepancy implies lower predictive variation under FC re-estimation, we conduct cross-method comparisons and a controlled alignment-strength sweep, the latter using BRIO w/o QG, which removes qualification guidance by assigning equal weights to all factors and disabling qualification-based strength modulation, trained independently for each tested λ_{env}. We measure class-conditional full-scan site discrepancy $\mathcal{D}_{\text{site}}$ in Eq. (83) and normalized predictive re-estimation variation $\widetilde{\mathcal{D}}^{\text{pred}}_{\text{re}}$ in Eq. (84), and summarize their relation across methods by Spearman ρ, the discordance rate R_{discord}, and the total and AUC-preserved mismatch rates defined in Appendix J.2. Fig. 3(a) shows that lower $\mathcal{D}_{\text{site}}$ does not consistently correspond to lower $\widetilde{\mathcal{D}}^{\text{pred}}_{\text{re}}$ across methods, so cross-site agreement alone does not imply stable within-subject predictions. In the controlled sweep in Fig. 3(b), stronger alignment reduces both quantities, while the full BRIO achieves lower $\widetilde{\mathcal{D}}^{\text{pred}}_{\text{re}}$ than the w/o QG setting at comparable $\mathcal{D}_{\text{site}}$, indicating that qualification-guided weighting and strength modulation offer benefits beyond simply increasing alignment strength. Beyond these average trends, Fig. 3(c) shows that BRIO achieves lower total and AUC-preserved mismatch rates than the evaluated w/o QG settings, with fewer mismatches even when predictive performance is preserved.

Qualification and Factor Transferability. To examine whether source-side qualification identifies transferable factors, we conduct a factor-retention experiment on ABIDE using a fixed, trained BRIO model. We rank factors by full qualification, task relevance alone, or re-estimation support alone, with random ranking as a reference. Qualification and support scores are averaged over source-site pairs and classes before ranking. For each ranking, we retain varying proportions of the highest-ranked factors and evaluate AUC on held-out sites using the original classifier without retraining, allowing direct comparisons of factor selection strategies. Fig. 3(d) shows that full qualification yields the highest AUC at partial retention, with larger gains when fewer factors are retained. Its advantage over task relevance alone suggests that re-estimation support provides complementary information about factor transferability, while the weaker performance of support alone highlights the importance of considering both criteria jointly. As more factors are retained, the performance gaps narrow and disappear at full retention, where all rankings recover the original model predictions.

5 Conclusion

To address finite-sample FC estimation variability in cross-site brain network generalization, we proposed BRIO. BRIO qualifies predictive evidence under within-scan FC re-estimation before aligning it across sites. Consistently indexed connectome factors enable comparison of predictive contributions across estimates. Calibrated changes in these contributions, combined with predictive relevance, determine which factors to align and how strongly. Experiments on four real-world datasets show consistent gains over competitive baselines. These findings support estimation stability as a criterion for what to align rather than only how much to align, and motivate future research on adaptive re-estimation budgets and dynamic or multimodal connectomes.

AI USE STATEMENT

Large language models, including ChatGPT, were only used during manuscript revision for language polishing. The authors reviewed and revised the AI-assisted content and take full responsibility for the research ideas, experimental design, data analysis, and scientific conclusions.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Pierre Bellec, Guillaume Marrelec, and Habib Benali. A bootstrap test to investigate changes in brain connectivity for functional mri. *Statistica Sinica*, pp. 1253–1268, 2008.

Alaa Bessadok, Mohamed Ali Mahjoub, and Islem Rekik. Graph neural networks in network neuroscience. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5833–5848, 2022.

Rasmus M Birn, Erin K Molloy, Rémi Patriat, Taurean Parker, Timothy B Meier, Gregory R Kirk, Veena A Nair, M Elizabeth Meyerand, and Vivek Prabhakaran. The effect of scan length on the reliability of resting-state fmri connectivity estimates. *Neuroimage*, 83:550–558, 2013.

Betty Jo Casey, Tariq Cannonier, May I Conley, Alexandra O Cohen, Deanna M Barch, Mary M Heitzeg, Mary E Soules, Theresa Teslovich, Danielle V Dellarco, Hugh Garavan, et al. The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, 32:43–54, 2018.

Andrew A. Chen, Joanne C. Beer, N. Tustison, Philip A Cook, Russell T. Shinohara, and Haochang Shou. Mitigating site effects in covariance for machine learning in neuroimaging data. *Human Brain Mapping*, 43:1179 – 1195, 2021.

Shuo Chen, Jian Kang, and Guoqing Wang. An empirical bayes normalization method for connectivity metrics in resting state fmri. *Frontiers in neuroscience*, 9:316, 2015.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. *Proceedings of the Conference on Neural Information Processing Systems*, 35:22131–22148, 2022.

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

Hejie Cui, Wei Dai, Yanqiao Zhu, Xiaoxiao Li, Lifang He, and Carl Yang. Interpretable graph neural networks for connectome-based brain disorder analysis. In *International conference on medical image computing and computer-assisted intervention*, pp. 375–385. Springer, 2022.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Proceedings of the Conference on Neural Information Processing Systems*, 26, 2013.

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, et al. 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.

Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. Debiasing graph neural networks via learning disentangled causal substructure. *Proceedings of the Conference on Neural Information Processing Systems*, 35:24934–24946, 2022.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690. PMLR, 2019.

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.

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Yuheng Gu, Shoubo Peng, Yaqin Li, Linlin Gao, and Yihong Dong. Fc-hgnn: A heterogeneous graph neural network based on brain functional connectivity for mental disorder identification. *Information Fusion*, 113:102619, 2025.

Keqi Han, Yao Su, Lifang He, Liang Zhan, Sergey Plis, Vince Calhoun, and Carl Yang. Rethinking functional brain connectome analysis: do graph deep learning models help. *npj Artificial Intelligence*, 2(1):19, 2026.

Niko Huotari, Lauri Raitamaa, Heta Helakari, Janne Kananen, Ville Raatikainen, Aleksi Rasila, Timo Tuovinen, Jussi Kantola, Viola Borchardt, Vesa J Kiviniemi, et al. Sampling rate effects on resting state fmri metrics. *Frontiers in neuroscience*, 13:279, 2019.

Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. Brain network transformer. *Proceedings of the Conference on Neural Information Processing Systems*, 35:25586–25599, 2022.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Maria Kudela, Jaroslaw Harezlak, and Martin A Lindquist. Assessing uncertainty in dynamic functional connectivity. *NeuroImage*, 149:165–177, 2017.

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.

Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. Out-of-distribution generalization on graphs: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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

Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *Proceedings of the International Conference on Machine Learning*, pp. 15524–15543. PMLR, 2022.

Stephanie Noble, Dustin Scheinost, and R Todd Constable. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *Neuroimage*, 203:116157, 2019.

Fuad Noman, Raphaël C-W Phan, Hernando Ombao, and Chee-Ming Ting. Adaptive graph learning with multi-graph convolutions for brain disorder classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 56–65. Springer, 2025.

Mansooreh Pakravan. Uncertainty-aware type-ii fuzzy graph modeling of resting-state fmri uncovers robust sex differences. *Journal of Neuroscience Methods*, pp. 110745, 2026.

Xinmei Qiu, Fan Wang, Yongheng Sun, Chunfeng Lian, and Jianhua Ma. Towards graph neural networks with domain-generalizable explainability for fmri-based brain disorder diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 454–464. Springer, 2024.

Xinmei Qiu, Yongheng Sun, Yilin Shi, Xujun Duan, Fan Wang, and Jianhua Ma. Metaexplainer: Revisit domain generalization of functional connectome analyses from the perspective of explainability. *Medical Image Analysis*, 105:103664, 2025.

Qi-Man Shao and Hao Yu. Bootstrapping the sample means for stationary mixing sequences. *Stochastic Processes and their Applications*, 48(1):175–190, 1993.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV Workshops*, 2016.

Jerry Sun, Mohamed Abubakr Hassan, Yaoyu Zhang, Wanying Zhang, and Chi-Guhn Lee. Diverse and sparse mixture-of-experts for causal subgraph–based out-of-distribution graph learning. In *Proceedings of the International Conference on Learning Representations*, volume 2026, pp. 79573–79597, 2026.

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

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.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Qixun Wang, Yifei Wang, Yisen Wang, and Xianghua Ying. Dissecting the failure of invariant learning on graphs. *Proceedings of the Conference on Neural Information Processing Systems*, 37: 80383–80438, 2024.

Shuo Wang, Mingchen Sun, Qiang Huang, and Ying Wang. Environment promoted invariant information learning for graph out-of-distribution generalization. *Artificial Intelligence*, pp. 104522, 2026a.

Yingxu Wang, Kunyu Zhang, Yanwu Yang, Thomas Wolfers, Yujie Wu, Siyang Gao, and Nan Yin. When brain networks travel: Learning beyond site. *arXiv preprint arXiv:2605.06050*, 2026b.

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

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Okito Yamashita, Ayumu Yamashita, Yuji Takahara, Yuki Sakai, Yasumasa Okamoto, Go Okada, Masahiro Takamura, Motoaki Nakamura, Takashi Itahashi, Takashi Hanakawa, et al. Computational mechanisms of neuroimaging biomarkers uncovered by multicenter resting-state fmri connectivity variation profile. *Molecular Psychiatry*, 30(11):5463–5474, 2025.

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

Liang Yang, Shuai Zhai, Ziyi Ma, Jiaming Zhuo, Di Jin, Chuan Wang, Zhen Wang, and Xiaochun Cao. Brain networks should be learned, not constructed. In *Proceedings of the International Conference on Machine Learning*, 2026.

Guoqi Yu, Xiaowei Hu, Angelica I Aviles-Rivero, Anqi Qiu, and Shujun Wang. Moving beyond functional connectivity: Time-series modeling for fmri-based brain disorder classification. *IEEE Transactions on Medical Imaging*, 2026.

Meichen Yu, Kristin A Linn, Philip A Cook, Mary L Phillips, Melvin McInnis, Maurizio Fava, Madhukar H Trivedi, Myrna M Weissman, Russell T Shinohara, and Yvette I Sheline. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fmri data. *Human brain mapping*, 39(11):4213–4227, 2018.

Minqi Yu, Jinduo Liu, and Junzhong Ji. Causal invariance-aware augmentation for brain graph contrastive learning. In *Proceedings of the International Conference on Machine Learning*, 2025.

Table 2: Summary of key notations.

Symbol	Description
$\mathcal{E}_{\mathrm{tr}}, \mathcal{E}_{\mathrm{te}}, E$	Source and unseen test site sets, and the number of source sites.
$\boldsymbol{X}_i^e, Y_i^e$	Preprocessed BOLD sequence and binary label of subject i at site e.
T_i^e, P	Numbers of valid time points and ROIs.
$\mathcal{I}_{e,c}, N_{e,c}$	Subject index set for site e and class c, and its size.
$\mathcal{B}_{e,c}$	Training mini-batch subset for site e and class c.
$\Psi, \mathcal{R}_k, K$	FC estimator, joint circular moving-block resampling operator, and number of FC re-estimates per subject.
$\ell_b, \ell_{b,i}^e$	Fold-level block length estimated from training subjects, and its per-sequence clipped value, in time points.
$\tau_{\mathrm{ac}}, \ell_{\min}, \ell_{\max}$	Autocorrelation threshold and clipping bounds of the block-length rule.
$\boldsymbol{A}_i^{e,(0)}, \boldsymbol{A}_i^{e,(k)}$	Full-scan FC graph and its k-th re-estimate, with $k \geq 1$.
$\boldsymbol{S}, \rho_g$	Fixed source-derived binary propagation support and per-ROI neighbor selection ratio.
$\boldsymbol{G}_i^{e,(k)}, \boldsymbol{G}_i^{e,(k),\pm}$	Supported FC adjacency and its positive weights $(+)$ and negative-weight magnitudes $(-)$.
$\boldsymbol{E}^{\mathrm{ROI}}, \boldsymbol{H}_i^{e,(k)}$	Shared ROI identity embeddings and ROI representations produced by the shared signed GNN.
$J, \boldsymbol{Q}$	Number of connectome factors and learnable factor queries.
$\boldsymbol{\Pi}, \overline{\boldsymbol{\Pi}}$	Shared balanced factor–ROI assignment and its cached value fixed after supervised warm-up.
$\boldsymbol{r}_{i,j}^{e,(k)}$	Unit-norm representation of factor j for subject i at site e and FC estimate k.
$\boldsymbol{s}_i^{e,(k)}$	Graph representation formed by scaled factor concatenation in a fixed order.
$\boldsymbol{w}_j, b$	Classifier weight block for factor j and the prediction bias.
$a_{i,j}^{e,(k)}$	Factor score $\boldsymbol{w}_j^\top \boldsymbol{r}_{i,j}^{e,(k)}$ before the common $1/\sqrt{J}$ scaling.
$z_i^{e,(k)}, \widehat{p}_i^{e,(0)}$	Prediction logit for FC estimate k and full-scan prediction probability.
$\mu_{e,c,j}$	Mean full-scan contribution of factor j for site e and class c.
$V_{e,c,j}^{\mathrm{sub}}, V_{c,j}^{\mathrm{pool}}$	Within-class contribution variance at a site and its equal-site average.
$\Delta_{e,j}, \Delta_j$	Full-scan class-mean contrast at a site and its signed equal-site average.
$\beta_j, R_j^{\mathrm{task}}, \pi_j^{\mathrm{task}}$	Relative squared classifier-block norm, unnormalized predictive relevance, and its normalization across factors.
$D_{e,c,j}^{\mathrm{re}}$	Mean squared difference between paired re-estimated and full-scan contributions.
$M_{e,j}^{\mathrm{task}}, T_{e,c,j}$	Squared half-separation $\Delta_{e,j}^2/4$ and local reference scale $V_{e,c,j}^{\mathrm{sub}} + M_{e,j}^{\mathrm{task}}$.
$q_{e,c,j}^{\mathrm{re}}, \kappa_{e,c,j}$	Task-calibrated re-estimation support and full-scan separation ratio for factor j at site e and class c.
$q_{e,e',c,j}^{\mathrm{pair}}$	Geometric mean of factor j's re-estimation supports at sites e and e' for class c.
$\alpha_{e,e',c,j}, \overline{\alpha}_{e,e',c,j}$	Unnormalized pairwise qualification and its normalization determining the relative factor weight in alignment.
$g_{e,e',c}$	Total pairwise qualification scaling alignment for sites e, e' and class c.
$\boldsymbol{\varphi}_{i\mid e,e',c}^{Q,d}$	Pair-specific graph representation at site $d \in \{e, e'\}$, formed from reweighted full-scan factors.
$\widehat{P}_{e,c}^{Q[e,e']}$	Empirical class-conditional distribution at site e in the common pairwise weighted space.
$\mathcal{S}_\varepsilon, \varepsilon$	Debiased Sinkhorn divergence and its entropic regularization parameter.

A Notation Summary

As shown in the Table 2, we summarize the key notations of this paper.

B Proof of Proposition 1

Proposition 1 (Class-Contrast Stability under FC Re-estimation) *Fix the model snapshot, a source site e, and a factor j, with $N_{e,0}, N_{e,1} > 0$ and $K \geq 1$. Let $\Delta_{e,j}^{\mathrm{re}}$ be the class contrast of contributions averaged over subjects and FC re-estimates, and let $\kappa_{e,c,j} = \Delta_{e,j}^2 / \big(4(V_{e,c,j}^{\mathrm{sub}} + \epsilon_q)\big)$ be the full-scan*

separation ratio of class c. Then

$$\left|\Delta^{\mathrm{re}}_{e,j} - \Delta_{e,j}\right| \le \sum_{c\in\{0,1\}} \sqrt{(T_{e,c,j}+\epsilon_q)\left(\frac{1}{q^{\mathrm{re}}_{e,c,j}} - 1\right)}, \tag{28}$$

and $\Delta^{\mathrm{re}}_{e,j}$ has the same sign as $\Delta_{e,j}$ whenever $q^{\mathrm{re}}_{e,c,j} > (1+\kappa_{e,c,j})/(1+2\kappa_{e,c,j})$ for both $c \in \{0,1\}$.

Proof.

For each class $c \in \{0,1\}$, the mean re-estimated contribution and the resulting class contrast are

$$\mu^{\mathrm{re}}_{e,c,j} = \frac{1}{N_{e,c}K} \sum_{i\in\mathcal{I}_{e,c}} \sum_{k=1}^{K} a^{e,(k)}_{i,j}, \quad \Delta^{\mathrm{re}}_{e,j} = \mu^{\mathrm{re}}_{e,1,j} - \mu^{\mathrm{re}}_{e,0,j}. \tag{29}$$

Let $\xi^{e,(k)}_{i,j} = a^{e,(k)}_{i,j} - a^{e,(0)}_{i,j}$ denote the contribution change between the k-th FC re-estimate and the full-scan estimate. Repeating each full-scan contribution across the K paired comparisons gives

$$\mu^{\mathrm{re}}_{e,c,j} - \mu_{e,c,j} = \frac{1}{N_{e,c}K} \sum_{i\in\mathcal{I}_{e,c}} \sum_{k=1}^{K} \xi^{e,(k)}_{i,j}. \tag{30}$$

Applying the Cauchy–Schwarz inequality to these $N_{e,c}K$ paired changes yields

$$\begin{aligned} \left|\mu^{\mathrm{re}}_{e,c,j} - \mu_{e,c,j}\right|^2 &= \frac{1}{(N_{e,c}K)^2}\left(\sum_{i\in\mathcal{I}_{e,c}} \sum_{k=1}^{K} \xi^{e,(k)}_{i,j}\right)^2 \\ &\le \frac{1}{N_{e,c}K} \sum_{i\in\mathcal{I}_{e,c}} \sum_{k=1}^{K} \left(\xi^{e,(k)}_{i,j}\right)^2 = D^{\mathrm{re}}_{e,c,j}, \end{aligned} \tag{31}$$

where the final equality follows from Eq. (14), and equality holds if and only if $\xi^{e,(k)}_{i,j}$ is constant over $i \in \mathcal{I}_{e,c}$ and k, that is, when re-estimation shifts all class-c contributions by a common constant. Since the change in the class contrast is the difference between the two class-mean shifts, taking square roots and applying the triangle inequality gives

$$\begin{aligned} \left|\Delta^{\mathrm{re}}_{e,j} - \Delta_{e,j}\right| &= \left|\left(\mu^{\mathrm{re}}_{e,1,j} - \mu_{e,1,j}\right) - \left(\mu^{\mathrm{re}}_{e,0,j} - \mu_{e,0,j}\right)\right| \\ &\le \left|\mu^{\mathrm{re}}_{e,1,j} - \mu_{e,1,j}\right| + \left|\mu^{\mathrm{re}}_{e,0,j} - \mu_{e,0,j}\right| \le \sum_{c\in\{0,1\}} \sqrt{D^{\mathrm{re}}_{e,c,j}}. \end{aligned} \tag{32}$$

To express this bound through re-estimation support, let $m_{e,j} = (\mu_{e,0,j} + \mu_{e,1,j})/2$ denote the midpoint of the full-scan class means, so that $\mu_{e,c,j} - m_{e,j} = (2c-1)\Delta_{e,j}/2$ for either class. Expanding the squared distance from this midpoint around the class mean yields

$$\begin{aligned} \frac{1}{N_{e,c}} \sum_{i\in\mathcal{I}_{e,c}} \left(a^{e,(0)}_{i,j} - m_{e,j}\right)^2 &= \frac{1}{N_{e,c}} \sum_{i\in\mathcal{I}_{e,c}} \left[\left(a^{e,(0)}_{i,j} - \mu_{e,c,j}\right) + \left(\mu_{e,c,j} - m_{e,j}\right)\right]^2 \\ &= V^{\mathrm{sub}}_{e,c,j} + \frac{1}{4}\Delta^2_{e,j} = T_{e,c,j}, \end{aligned} \tag{33}$$

where the cross term vanishes because $\sum_{i\in\mathcal{I}_{e,c}} (a^{e,(0)}_{i,j} - \mu_{e,c,j}) = 0$. Thus $T_{e,c,j} \ge 0$ is the empirical mean squared distance of class-c full-scan contributions from the midpoint of the two class means. Together with $D^{\mathrm{re}}_{e,c,j} \ge 0$ and $\epsilon_q > 0$, this ensures $0 < q^{\mathrm{re}}_{e,c,j} \le 1$, and rearranging Eq. (16) gives

$$(T_{e,c,j}+\epsilon_q)\left(\frac{1}{q^{\mathrm{re}}_{e,c,j}} - 1\right) = (T_{e,c,j}+\epsilon_q)\frac{D^{\mathrm{re}}_{e,c,j}}{T_{e,c,j}+\epsilon_q} = D^{\mathrm{re}}_{e,c,j}. \tag{34}$$

Substituting Eq. (34) into Eq. (32) establishes Eq. (17).

For sign preservation, we first show that the support threshold is equivalent to a bound on the re-estimation variation. Since $4T_{e,c,j} = 4V^{\mathrm{sub}}_{e,c,j} + \Delta^2_{e,j}$ by Eq. (33),

$$\frac{1+\kappa_{e,c,j}}{1+2\kappa_{e,c,j}} = \frac{4\left(V^{\mathrm{sub}}_{e,c,j}+\epsilon_q\right)+\Delta^2_{e,j}}{4\left(V^{\mathrm{sub}}_{e,c,j}+\epsilon_q\right)+2\Delta^2_{e,j}} = \frac{T_{e,c,j}+\epsilon_q}{T_{e,c,j}+\epsilon_q+\Delta^2_{e,j}/4}, \tag{35}$$

so by Eq. (16),

$$q^{\mathrm{re}}_{e,c,j} > \frac{1+\kappa_{e,c,j}}{1+2\kappa_{e,c,j}} \iff \frac{T_{e,c,j}+\epsilon_q}{T_{e,c,j}+\epsilon_q+D^{\mathrm{re}}_{e,c,j}} > \frac{T_{e,c,j}+\epsilon_q}{T_{e,c,j}+\epsilon_q+\Delta^2_{e,j}/4} \iff D^{\mathrm{re}}_{e,c,j} < \frac{\Delta^2_{e,j}}{4}. \tag{36}$$

Suppose the threshold condition holds for both classes. Then $\sqrt{D^{\mathrm{re}}_{e,c,j}} < |\Delta_{e,j}|/2$ for each c, hence $\sum_c \sqrt{D^{\mathrm{re}}_{e,c,j}} < |\Delta_{e,j}|$ and $\Delta_{e,j} \neq 0$. Using Eq. (32), we obtain

$$\begin{aligned}\Delta_{e,j}\Delta^{\mathrm{re}}_{e,j} &= \Delta^2_{e,j} + \Delta_{e,j}\left(\Delta^{\mathrm{re}}_{e,j} - \Delta_{e,j}\right) \\ &\geq \Delta^2_{e,j} - |\Delta_{e,j}|\left|\Delta^{\mathrm{re}}_{e,j} - \Delta_{e,j}\right| \geq |\Delta_{e,j}|\Big(|\Delta_{e,j}| - \sum_{c\in\{0,1\}}\sqrt{D^{\mathrm{re}}_{e,c,j}}\Big) > 0.\end{aligned} \tag{37}$$

Hence $\Delta^{\mathrm{re}}_{e,j}$ and $\Delta_{e,j}$ have the same sign. The threshold $(1+\kappa_{e,c,j})/(1+2\kappa_{e,c,j})$ lies in $(1/2, 1]$ and decreases in $\kappa_{e,c,j}$, so strongly separated factors preserve the sign of their class contrast under moderate support, whereas weakly separated factors require support close to one.

C Proof of Proposition 2

Proposition 2 (Variational Form of Pairwise Qualification) *For a probability vector $\boldsymbol{\pi}$ over J factors and supports $q_j \in (0,1]$, define $\alpha_j = \pi_j q_j$, $g = \sum_{j=1}^J \alpha_j$, and $\overline{\alpha}_j = \alpha_j/g$. Let $\mathcal{V}_\pi$ be the set of probability vectors $\boldsymbol{v}$ with $v_j = 0$ whenever $\pi_j = 0$, and define*

$$\mathcal{J}(\boldsymbol{v}) = \mathrm{KL}(\boldsymbol{v} \,\|\, \boldsymbol{\pi}) + \sum_{j=1}^J v_j \log\frac{1}{q_j}. \tag{38}$$

Then $\overline{\boldsymbol{\alpha}}$ is the unique minimizer of $\mathcal{J}$ over $\mathcal{V}_\pi$, with minimum value $-\log g$, and

$$\overline{\alpha}_j = \frac{\partial \log g}{\partial \log q_j}, \quad -\log g \leq \sum_{j=1}^J \pi_j \log\frac{1}{q_j}, \tag{39}$$

where equality holds if and only if q_j is constant over the factors with $\pi_j > 0$.

Proof.

Let $\mathcal{I}_\pi = \{j \in \{1,\dots,J\} : \pi_j > 0\}$ denote the support of $\boldsymbol{\pi}$, which is nonempty because $\boldsymbol{\pi}$ is a probability vector. Every $\boldsymbol{v} \in \mathcal{V}_\pi$ is supported on $\mathcal{I}_\pi$ and satisfies $\sum_{j\in\mathcal{I}_\pi} v_j = 1$. Throughout the proof, we use the convention $0\log(0/b) = 0$ for $b > 0$.

Since $q_j \in (0,1]$, α_j is strictly positive on $\mathcal{I}_\pi$ and zero outside it. Moreover,

$$0 < g = \sum_{j\in\mathcal{I}_\pi} \pi_j q_j \leq \sum_{j\in\mathcal{I}_\pi} \pi_j = 1. \tag{40}$$

Thus, $\overline{\boldsymbol{\alpha}}$ is well defined, belongs to $\mathcal{V}_\pi$, and has strictly positive entries on $\mathcal{I}_\pi$.

For every $\boldsymbol{v} \in \mathcal{V}_\pi$, combining the two terms of $\mathcal{J}$ and substituting $\alpha_j = g\,\overline{\alpha}_j$ gives

$$\begin{aligned}\mathcal{J}(\boldsymbol{v}) &= \sum_{j\in\mathcal{I}_\pi} v_j \log\frac{v_j}{\pi_j} + \sum_{j\in\mathcal{I}_\pi} v_j \log\frac{1}{q_j} = \sum_{j\in\mathcal{I}_\pi} v_j \log\frac{v_j}{\alpha_j} \\ &= \sum_{j\in\mathcal{I}_\pi} v_j \log\frac{v_j}{\overline{\alpha}_j} - \left(\sum_{j\in\mathcal{I}_\pi} v_j\right)\log g = \mathrm{KL}(\boldsymbol{v} \,\|\, \overline{\boldsymbol{\alpha}}) - \log g,\end{aligned} \tag{41}$$

where the last equality uses $\sum_{j\in\mathcal{I}_\pi} v_j = 1$. The objective is therefore a KL divergence to $\overline{\boldsymbol{\alpha}}$ plus a term independent of $\boldsymbol{v}$.

Let $\mathcal{I}_{\boldsymbol{v}} = \{j \in \mathcal{I}_\pi : v_j > 0\}$. Applying $\log t \le t - 1$ for $t > 0$ gives

$$\mathrm{KL}(\boldsymbol{v} \,\|\, \overline{\boldsymbol{\alpha}}) = -\sum_{j\in\mathcal{I}_{\boldsymbol{v}}} v_j \log \frac{\overline{\alpha}_j}{v_j} \ge \sum_{j\in\mathcal{I}_{\boldsymbol{v}}} (v_j - \overline{\alpha}_j) = \sum_{j\in\mathcal{I}_\pi\setminus\mathcal{I}_{\boldsymbol{v}}} \overline{\alpha}_j \ge 0. \tag{42}$$

If $\mathcal{I}_{\boldsymbol{v}} \neq \mathcal{I}_\pi$, the final sum is strictly positive because $\overline{\alpha}_j > 0$ on $\mathcal{I}_\pi$. Otherwise, equality holds precisely when $\overline{\alpha}_j / v_j = 1$ for every $j \in \mathcal{I}_\pi$, by the equality condition of $\log t \le t - 1$. Hence the KL divergence vanishes if and only if $\boldsymbol{v} = \overline{\boldsymbol{\alpha}}$, and Eq. (41) yields

$$\mathcal{J}(\boldsymbol{v}) \ge -\log g = \mathcal{J}(\overline{\boldsymbol{\alpha}}), \qquad \boldsymbol{v} \in \mathcal{V}_\pi, \tag{43}$$

with equality if and only if $\boldsymbol{v} = \overline{\boldsymbol{\alpha}}$. Since $\overline{\boldsymbol{\alpha}}$ is feasible, it is the unique minimizer, and the minimum value is $-\log g$.

For the sensitivity identity, $g = \sum_{j'\in\mathcal{I}_\pi} \pi_{j'} q_{j'}$ is linear in each q_j with $\partial g / \partial q_j = \pi_j$, so

$$\frac{\partial \log g}{\partial \log q_j} = \frac{q_j}{g}\frac{\partial g}{\partial q_j} = \frac{\pi_j q_j}{g} = \overline{\alpha}_j. \tag{44}$$

For the inequality, $\boldsymbol{\pi} \in \mathcal{V}_\pi$ and $\mathrm{KL}(\boldsymbol{\pi}\|\boldsymbol{\pi}) = 0$, so Eq. (43) evaluated at $\boldsymbol{v} = \boldsymbol{\pi}$ gives

$$-\log g \le \mathcal{J}(\boldsymbol{\pi}) = \sum_{j\in\mathcal{I}_\pi} \pi_j \log \frac{1}{q_j}, \tag{45}$$

with equality if and only if $\boldsymbol{\pi} = \overline{\boldsymbol{\alpha}}$, that is, $\pi_j = \pi_j q_j / g$ for every $j \in \mathcal{I}_\pi$, which holds if and only if $q_j = g$ for every $j \in \mathcal{I}_\pi$. Equivalently, Eq. (45) is Jensen's inequality $\log \sum_j \pi_j q_j \ge \sum_j \pi_j \log q_j$ for the concave logarithm, with the same equality condition. □

With $\boldsymbol{\pi} = \boldsymbol{\pi}^{\mathrm{task}}$ and $q_j = q^{\mathrm{pair}}_{e,e',c,j}$, the objective specializes to pairwise qualification. By Eq. (16), the penalty expands as

$$\log \frac{1}{q^{\mathrm{pair}}_{e,e',c,j}} = \frac{1}{2}\log\left(1 + \frac{D^{\mathrm{re}}_{e,c,j}}{T_{e,c,j} + \epsilon_q}\right) + \frac{1}{2}\log\left(1 + \frac{D^{\mathrm{re}}_{e',c,j}}{T_{e',c,j} + \epsilon_q}\right), \tag{46}$$

so $\mathcal{J}_{e,e',c}$ trades proximity to $\boldsymbol{\pi}^{\mathrm{task}}$ against re-estimation deviations measured on the local reference scales of both sites. Eq. (44) becomes

$$\frac{\partial \log g_{e,e',c}}{\partial \log q^{\mathrm{pair}}_{e,e',c,j}} = \overline{\alpha}_{e,e',c,j}, \quad \mathrm{d}\log g_{e,e',c} = \sum_{j=1}^{J} \overline{\alpha}_{e,e',c,j}\, \mathrm{d}\log q^{\mathrm{pair}}_{e,e',c,j}, \tag{47}$$

so a relative change in the support of factor j moves the alignment strength by its normalized weight times that change. Evaluating Eq. (41) at $\boldsymbol{v} = \boldsymbol{\pi}^{\mathrm{task}}$ gives the exact reduction in support penalty obtained by reweighting,

$$\sum_{j=1}^{J} \pi^{\mathrm{task}}_j \log \frac{1}{q^{\mathrm{pair}}_{e,e',c,j}} - (-\log g_{e,e',c}) = \mathrm{KL}\big(\boldsymbol{\pi}^{\mathrm{task}} \,\big\|\, \overline{\boldsymbol{\alpha}}_{e,e',c}\big) \ge 0, \tag{48}$$

which is strictly positive unless $q^{\mathrm{pair}}_{e,e',c,j}$ is constant over the factors with $\pi^{\mathrm{task}}_j > 0$.

D Debiased Sinkhorn Divergence

Optimal transport compares empirical distributions by minimizing a ground cost over couplings that preserve their probability masses Gabriel & Marco (2019). Entropic regularization enables computation through Sinkhorn scaling, and a self-transport correction removes the regularization bias at identical inputs Cuturi (2013); Feydy et al. (2019). This part gives the construction for the qualification-weighted distributions in Eq. (25) and derives how the qualifications enter the alignment gradient.

Fix a source-site pair (e, e'), class c, and the associated qualification profile. Write $\mu = \widehat{P}_{e,c}^{Q[e,e']}$ and $\nu = \widehat{P}_{e',c}^{Q[e,e']}$ for the empirical distributions in Eq. (24), and let $\boldsymbol{x}_i$ and $\boldsymbol{y}_h$ denote the weighted representations from Eq. (22). Then

$$\begin{aligned} \mu &= \sum_{i=1}^{n} \eta_i \delta_{\boldsymbol{x}_i}, \quad \eta_i = \frac{1}{n}, \quad n = |\mathcal{B}_{e,c}|, \\ \nu &= \sum_{h=1}^{n'} \zeta_h \delta_{\boldsymbol{y}_h}, \quad \zeta_h = \frac{1}{n'}, \quad n' = |\mathcal{B}_{e',c}|, \end{aligned} \tag{49}$$

with $n, n' > 0$, so the marginal vectors $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$ have positive entries and unit total mass. Both sites use the same normalized qualifications, so the squared Euclidean ground cost coincides with the qualification-weighted dissimilarity in Eq. (23):

$$C_{ih} = \|\boldsymbol{x}_i - \boldsymbol{y}_h\|_2^2, \quad \|\boldsymbol{x}_i\|_2^2 = \|\boldsymbol{y}_h\|_2^2 = \frac{1}{2}, \quad 0 \le C_{ih} \le 2. \tag{50}$$

Let $\boldsymbol{C} = [C_{ih}] \in \mathbb{R}^{n \times n'}$. The admissible couplings are

$$\mathcal{U}(\boldsymbol{\eta}, \boldsymbol{\zeta}) = \left\{ \boldsymbol{\Gamma} \in \mathbb{R}_+^{n \times n'} : \boldsymbol{\Gamma} \mathbf{1}_{n'} = \boldsymbol{\eta}, \quad \boldsymbol{\Gamma}^\top \mathbf{1}_n = \boldsymbol{\zeta} \right\}, \tag{51}$$

and the unregularized transport cost is

$$\mathrm{OT}_0(\mu, \nu) = \min_{\boldsymbol{\Gamma} \in \mathcal{U}(\boldsymbol{\eta}, \boldsymbol{\zeta})} \langle \boldsymbol{\Gamma}, \boldsymbol{C} \rangle, \quad \langle \boldsymbol{\Gamma}, \boldsymbol{C} \rangle = \sum_{i=1}^{n} \sum_{h=1}^{n'} \Gamma_{ih} C_{ih}, \tag{52}$$

which equals the squared Wasserstein-2 distance $W_2^2(\mu, \nu)$ for the squared Euclidean ground cost Gabriel & Marco (2019).

We regularize the coupling relative to the independent coupling $\boldsymbol{\eta}\boldsymbol{\zeta}^\top$:

$$\mathrm{OT}_\varepsilon(\mu, \nu) = \min_{\boldsymbol{\Gamma} \in \mathcal{U}(\boldsymbol{\eta}, \boldsymbol{\zeta})} \left\{ \langle \boldsymbol{\Gamma}, \boldsymbol{C} \rangle + \varepsilon \, \mathrm{KL}\big(\boldsymbol{\Gamma} \,\big\|\, \boldsymbol{\eta}\boldsymbol{\zeta}^\top\big) \right\}, \quad \varepsilon > 0, \tag{53}$$

where the generalized KL divergence is

$$\mathrm{KL}\big(\boldsymbol{\Gamma} \,\big\|\, \boldsymbol{\eta}\boldsymbol{\zeta}^\top\big) = \sum_{i,h} \left[\Gamma_{ih} \log \frac{\Gamma_{ih}}{\eta_i \zeta_h} - \Gamma_{ih} + \eta_i \zeta_h \right], \tag{54}$$

with $0 \log(0/b) = 0$ for $b > 0$. For feasible couplings, the linear terms cancel and the marginal constraints give

$$\mathrm{KL}\big(\boldsymbol{\Gamma} \,\big\|\, \boldsymbol{\eta}\boldsymbol{\zeta}^\top\big) = \sum_{i,h} \Gamma_{ih} \log \Gamma_{ih} - \sum_i \eta_i \log \eta_i - \sum_h \zeta_h \log \zeta_h, \tag{55}$$

so this formulation and the negative-entropy formulation share the same optimal coupling and differ by marginal-dependent constants. The feasible set is compact and the regularizer strictly convex, so the minimizer exists, is unique, and has strictly positive entries Cuturi (2013).

Introduce dual potentials $\boldsymbol{\phi} \in \mathbb{R}^n$ and $\boldsymbol{\psi} \in \mathbb{R}^{n'}$ for the marginal constraints. The Lagrangian is

$$\mathcal{L}_{\mathrm{OT}}(\boldsymbol{\Gamma}, \boldsymbol{\phi}, \boldsymbol{\psi}) = \langle \boldsymbol{\Gamma}, \boldsymbol{C} \rangle + \varepsilon \, \mathrm{KL}\big(\boldsymbol{\Gamma} \,\big\|\, \boldsymbol{\eta}\boldsymbol{\zeta}^\top\big) + \boldsymbol{\phi}^\top (\boldsymbol{\eta} - \boldsymbol{\Gamma} \mathbf{1}_{n'}) + \boldsymbol{\psi}^\top (\boldsymbol{\zeta} - \boldsymbol{\Gamma}^\top \mathbf{1}_n), \tag{56}$$

and stationarity in each coupling entry yields

$$C_{ih} + \varepsilon \log \frac{\Gamma_{ih}^\star}{\eta_i \zeta_h} - \phi_i^\star - \psi_h^\star = 0, \quad \Gamma_{ih}^\star = \eta_i \zeta_h \exp\left(\frac{\phi_i^\star + \psi_h^\star - C_{ih}}{\varepsilon} \right). \tag{57}$$

Minimizing the Lagrangian over $\boldsymbol{\Gamma}$ for fixed potentials gives the dual problem

$$\mathrm{OT}_\varepsilon(\mu, \nu) = \max_{\boldsymbol{\phi}, \boldsymbol{\psi}} \left\{ \boldsymbol{\eta}^\top \boldsymbol{\phi} + \boldsymbol{\zeta}^\top \boldsymbol{\psi} - \varepsilon \sum_{i,h} \eta_i \zeta_h \left[\exp\left(\frac{\phi_i + \psi_h - C_{ih}}{\varepsilon} \right) - 1 \right] \right\}, \tag{58}$$

where strong duality holds because the strictly positive coupling $\boldsymbol{\eta}\boldsymbol{\zeta}^\top$ is feasible. At optimality, the exponential term recovers the optimal coupling, whose total mass is one, so

$$\mathrm{OT}_\varepsilon(\mu,\nu) = \boldsymbol{\eta}^\top \boldsymbol{\phi}^\star + \boldsymbol{\zeta}^\top \boldsymbol{\psi}^\star. \tag{59}$$

The exponential form leads to Sinkhorn scaling. With

$$(\boldsymbol{K}_\varepsilon)_{ih} = \exp(-C_{ih}/\varepsilon), \quad u_i = \eta_i \exp(\phi_i/\varepsilon), \quad v_h = \zeta_h \exp(\psi_h/\varepsilon), \tag{60}$$

the coupling is $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{u})\boldsymbol{K}_\varepsilon \mathrm{diag}(\boldsymbol{v})$, and alternately enforcing the two marginal constraints gives

$$\boldsymbol{u}^{(t+1)} = \boldsymbol{\eta} \oslash (\boldsymbol{K}_\varepsilon \boldsymbol{v}^{(t)}), \quad \boldsymbol{v}^{(t+1)} = \boldsymbol{\zeta} \oslash (\boldsymbol{K}_\varepsilon^\top \boldsymbol{u}^{(t+1)}), \tag{61}$$

where $\oslash$ denotes elementwise division. For a positive kernel and positive marginals, positive initialization converges to the unique regularized optimum Sinkhorn & Knopp (1967); Cuturi (2013). Substituting Eq. (57) into the marginal constraints expresses the same updates in the dual potentials,

$$\phi_i^{(t+1)} = -\varepsilon\,\mathrm{LSE}_h\left(\log\zeta_h + \frac{\psi_h^{(t)} - C_{ih}}{\varepsilon}\right), \quad \psi_h^{(t+1)} = -\varepsilon\,\mathrm{LSE}_i\left(\log\eta_i + \frac{\phi_i^{(t+1)} - C_{ih}}{\varepsilon}\right), \tag{62}$$

where $\mathrm{LSE}_k(z_k) = \log\sum_k \exp(z_k)$. These are alternating maximizations of the dual objective and are evaluated with stable log-sum-exp reductions Feydy et al. (2019).

Entropic regularization makes the self-transport cost positive. For μ with at least two distinct support points, the two terms of the self-transport objective vanish at different couplings,

$$\begin{gathered}\langle \boldsymbol{\Gamma}, \boldsymbol{C}^{xx}\rangle = 0 \iff \boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\eta}), \quad \mathrm{KL}\big(\boldsymbol{\Gamma} \,\big\|\, \boldsymbol{\eta}\boldsymbol{\eta}^\top\big) = 0 \iff \boldsymbol{\Gamma} = \boldsymbol{\eta}\boldsymbol{\eta}^\top, \\ \mathrm{KL}\big(\mathrm{diag}(\boldsymbol{\eta}) \,\big\|\, \boldsymbol{\eta}\boldsymbol{\eta}^\top\big) = -\sum_i \eta_i \log\eta_i > 0,\end{gathered} \tag{63}$$

where $C^{xx}_{i\ell} = \|\boldsymbol{x}_i - \boldsymbol{x}_\ell\|_2^2$, so $\mathrm{OT}_\varepsilon(\mu,\mu) > 0$. Subtracting half of each self-transport cost removes this bias and defines the debiased Sinkhorn divergence Feydy et al. (2019),

$$\mathcal{S}_\varepsilon(\mu,\nu) = \mathrm{OT}_\varepsilon(\mu,\nu) - \frac{1}{2}\,\mathrm{OT}_\varepsilon(\mu,\mu) - \frac{1}{2}\,\mathrm{OT}_\varepsilon(\nu,\nu), \tag{64}$$

where all three terms use the same ε and the same weighted representation map, and $C^{yy}_{hk} = \|\boldsymbol{y}_h - \boldsymbol{y}_k\|_2^2$. By Eq. (55), with $H(\boldsymbol{\eta}) = -\sum_i \eta_i \log\eta_i$ and $\mathrm{OT}_\varepsilon^{\mathrm{ent}}$ the negative-entropy formulation,

$$\mathrm{OT}_\varepsilon(\mu,\nu) = \mathrm{OT}_\varepsilon^{\mathrm{ent}}(\mu,\nu) + \varepsilon H(\boldsymbol{\eta}) + \varepsilon H(\boldsymbol{\zeta}), \quad \mathrm{OT}_\varepsilon(\mu,\mu) = \mathrm{OT}_\varepsilon^{\mathrm{ent}}(\mu,\mu) + 2\varepsilon H(\boldsymbol{\eta}), \tag{65}$$

so the constants cancel in Eq. (64) and both formulations yield the same $\mathcal{S}_\varepsilon$. Transposition exchanges the marginals of a coupling without changing its cost or penalty, which gives

$$\mathrm{OT}_\varepsilon(\mu,\nu) = \mathrm{OT}_\varepsilon(\nu,\mu), \quad \mathcal{S}_\varepsilon(\mu,\nu) = \mathcal{S}_\varepsilon(\nu,\mu), \quad \mathcal{S}_\varepsilon(\mu,\mu) = 0. \tag{66}$$

For nonnegativity and definiteness, we use the optimal self-transport potentials Feydy et al. (2019). Let $\mathcal{D}_{\mu\mu}$ denote the self-transport dual objective in Eq. (58) with $\nu = \mu$, which is concave and satisfies $\mathcal{D}_{\mu\mu}(\boldsymbol{\phi},\boldsymbol{\psi}) = \mathcal{D}_{\mu\mu}(\boldsymbol{\psi},\boldsymbol{\phi})$. For an optimal pair $(\boldsymbol{\phi}^\star,\boldsymbol{\psi}^\star)$, the average $\boldsymbol{\phi}^\mu = \frac{1}{2}(\boldsymbol{\phi}^\star + \boldsymbol{\psi}^\star)$ satisfies

$$\mathcal{D}_{\mu\mu}(\boldsymbol{\phi}^\mu,\boldsymbol{\phi}^\mu) \geq \tfrac{1}{2}\mathcal{D}_{\mu\mu}(\boldsymbol{\phi}^\star,\boldsymbol{\psi}^\star) + \tfrac{1}{2}\mathcal{D}_{\mu\mu}(\boldsymbol{\psi}^\star,\boldsymbol{\phi}^\star) = \mathrm{OT}_\varepsilon(\mu,\mu), \tag{67}$$

so $(\boldsymbol{\phi}^\mu,\boldsymbol{\phi}^\mu)$ is optimal; define $\boldsymbol{\phi}^\nu$ likewise. Let $\{\boldsymbol{z}_\ell\}_{\ell=1}^{n_z}$ be the distinct points in the union of the two supports, and define

$$\Omega_{\ell r} = \exp\!\left(-\frac{\|\boldsymbol{z}_\ell - \boldsymbol{z}_r\|_2^2}{\varepsilon}\right), \quad \xi_\ell = \sum_{\{i:\boldsymbol{x}_i = \boldsymbol{z}_\ell\}} \eta_i \exp(\phi_i^\mu/\varepsilon), \quad \chi_\ell = \sum_{\{h:\boldsymbol{y}_h = \boldsymbol{z}_\ell\}} \zeta_h \exp(\phi_h^\nu/\varepsilon), \tag{68}$$

where empty sums are zero. The Gaussian Gram matrix $\boldsymbol{\Omega}$ is strictly positive definite on distinct points. Let $\boldsymbol{\Gamma}^{xx,\star}$ and $\boldsymbol{\Gamma}^{yy,\star}$ be the optimal self-couplings. Their unit total masses and Eq. (59) give

$$\begin{aligned}\boldsymbol{\xi}^\top\boldsymbol{\Omega}\boldsymbol{\xi} &= \sum_{i,\ell}\Gamma^{xx,\star}_{i\ell} = 1, & \boldsymbol{\chi}^\top\boldsymbol{\Omega}\boldsymbol{\chi} &= \sum_{h,k}\Gamma^{yy,\star}_{hk} = 1, \\ \mathrm{OT}_\varepsilon(\mu,\mu) &= 2\boldsymbol{\eta}^\top\boldsymbol{\phi}^\mu, & \mathrm{OT}_\varepsilon(\nu,\nu) &= 2\boldsymbol{\zeta}^\top\boldsymbol{\phi}^\nu.\end{aligned} \tag{69}$$

Evaluating the cross-transport dual at $(\boldsymbol{\phi}^\mu, \boldsymbol{\phi}^\nu)$ gives

$$\begin{aligned}\mathrm{OT}_\varepsilon(\mu,\nu) &\geq \boldsymbol{\eta}^\top \boldsymbol{\phi}^\mu + \boldsymbol{\zeta}^\top \boldsymbol{\phi}^\nu - \varepsilon \left[\sum_{i,h} \eta_i \zeta_h \exp\left(\frac{\phi_i^\mu + \phi_h^\nu - C_{ih}}{\varepsilon}\right) - 1\right] \\ &= \boldsymbol{\eta}^\top \boldsymbol{\phi}^\mu + \boldsymbol{\zeta}^\top \boldsymbol{\phi}^\nu + \varepsilon(1 - \boldsymbol{\xi}^\top \boldsymbol{\Omega} \boldsymbol{\chi}),\end{aligned} \tag{70}$$

and subtracting half of each self-cost from Eq. (69) yields

$$\begin{aligned}\mathcal{S}_\varepsilon(\mu,\nu) &\geq \varepsilon(1 - \boldsymbol{\xi}^\top \boldsymbol{\Omega} \boldsymbol{\chi}) = \frac{\varepsilon}{2}\left(\boldsymbol{\xi}^\top \boldsymbol{\Omega} \boldsymbol{\xi} + \boldsymbol{\chi}^\top \boldsymbol{\Omega} \boldsymbol{\chi} - 2\boldsymbol{\xi}^\top \boldsymbol{\Omega} \boldsymbol{\chi}\right) \\ &= \frac{\varepsilon}{2}(\boldsymbol{\xi} - \boldsymbol{\chi})^\top \boldsymbol{\Omega}(\boldsymbol{\xi} - \boldsymbol{\chi}) \geq 0.\end{aligned} \tag{71}$$

If $\mathcal{S}_\varepsilon(\mu,\nu) = 0$, strict positive definiteness of $\boldsymbol{\Omega}$ gives $\boldsymbol{\xi} = \boldsymbol{\chi}$, and the marginal constraints of the self-couplings recover the masses at each distinct point,

$$\mu(\{\boldsymbol{z}_\ell\}) = \sum_{\{i:\boldsymbol{x}_i = \boldsymbol{z}_\ell\}} \eta_i = \xi_\ell(\boldsymbol{\Omega}\boldsymbol{\xi})_\ell, \quad \nu(\{\boldsymbol{z}_\ell\}) = \sum_{\{h:\boldsymbol{y}_h = \boldsymbol{z}_\ell\}} \zeta_h = \chi_\ell(\boldsymbol{\Omega}\boldsymbol{\chi})_\ell, \tag{72}$$

so $\boldsymbol{\xi} = \boldsymbol{\chi}$ implies $\mu = \nu$. Together with Eq. (66), this establishes

$$\mathcal{S}_\varepsilon(\mu,\nu) \geq 0, \quad \mathcal{S}_\varepsilon(\mu,\nu) = 0 \iff \mu = \nu. \tag{73}$$

The same formulation provides the gradients used for representation learning. With fixed marginal weights, the envelope theorem gives $\partial\,\mathrm{OT}_\varepsilon / \partial C_{ih} = \Gamma_{ih}^{xy,\star}$ for the optimal cross-coupling Gabriel & Marco (2019), so

$$\nabla_{\boldsymbol{x}_i} \mathrm{OT}_\varepsilon(\mu,\nu) = 2\sum_h \Gamma_{ih}^{xy,\star}(\boldsymbol{x}_i - \boldsymbol{y}_h). \tag{74}$$

For self-transport, transposition preserves feasibility and objective, so uniqueness makes the optimal self-coupling symmetric; since each representation appears in both arguments of the self-cost,

$$\begin{aligned}\nabla_{\boldsymbol{x}_i} \mathrm{OT}_\varepsilon(\mu,\mu) &= 2\sum_\ell \Gamma_{i\ell}^{xx,\star}(\boldsymbol{x}_i - \boldsymbol{x}_\ell) + 2\sum_\ell \Gamma_{\ell i}^{xx,\star}(\boldsymbol{x}_i - \boldsymbol{x}_\ell) \\ &= 4\sum_\ell \Gamma_{i\ell}^{xx,\star}(\boldsymbol{x}_i - \boldsymbol{x}_\ell).\end{aligned} \tag{75}$$

Combining the two gives, at the converged transport solutions Feydy et al. (2019),

$$\begin{aligned}\nabla_{\boldsymbol{x}_i} \mathcal{S}_\varepsilon(\mu,\nu) &= 2\sum_h \Gamma_{ih}^{xy,\star}(\boldsymbol{x}_i - \boldsymbol{y}_h) - 2\sum_\ell \Gamma_{i\ell}^{xx,\star}(\boldsymbol{x}_i - \boldsymbol{x}_\ell), \\ \nabla_{\boldsymbol{y}_h} \mathcal{S}_\varepsilon(\mu,\nu) &= 2\sum_i \Gamma_{ih}^{xy,\star}(\boldsymbol{y}_h - \boldsymbol{x}_i) - 2\sum_k \Gamma_{hk}^{yy,\star}(\boldsymbol{y}_h - \boldsymbol{y}_k).\end{aligned} \tag{76}$$

Finally, the alignment contribution for the fixed source-site pair and class is $\ell^Q_{e,e',c} = g_{e,e',c}\mathcal{S}_\varepsilon(\mu,\nu)$. By Eq. (22), the j-th block of $\boldsymbol{x}_i$ is $\sqrt{\overline{\alpha}_{e,e',c,j}/2}\,\boldsymbol{r}_{i,j}^{e,(0)}$, so the j-th block of Eq. (76) equals $\sqrt{\overline{\alpha}_{e,e',c,j}/2}$ times the transport residual

$$\boldsymbol{\Delta}_{i,j}^{e,e',c} = \sum_h \Gamma_{ih}^{xy,\star}\left(\boldsymbol{r}_{i,j}^{e,(0)} - \boldsymbol{r}_{h,j}^{e',(0)}\right) - \sum_\ell \Gamma_{i\ell}^{xx,\star}\left(\boldsymbol{r}_{i,j}^{e,(0)} - \boldsymbol{r}_{\ell,j}^{e,(0)}\right). \tag{77}$$

Applying the chain rule with the qualifications held fixed gives

$$\begin{aligned}\nabla_{\boldsymbol{r}_{i,j}^{e,(0)}} \ell^Q_{e,e',c} &= g_{e,e',c}\,\overline{\alpha}_{e,e',c,j}\,\boldsymbol{\Delta}_{i,j}^{e,e',c} = \alpha_{e,e',c,j}\,\boldsymbol{\Delta}_{i,j}^{e,e',c}, \\ \alpha_{e,e',c,j} &= \pi_j^{\mathrm{task}} q_{e,e',c,j}^{\mathrm{pair}}, \quad \sum_{j=1}^J \alpha_{e,e',c,j} = g_{e,e',c}.\end{aligned} \tag{78}$$

The transport plans are shared across factors and set by the normalized qualifications through the ground cost, while the gradient reaching factor j scales with its unnormalized qualification, and the scales sum to the overall qualification. Averaging over unordered pairs and classes recovers Eq. (25):

$$\mathcal{L}^Q_{\mathrm{env}} = \frac{2}{E(E-1)} \sum_{e<e'} \frac{1}{2} \sum_{c\in\{0,1\}} \ell^Q_{e,e',c}. \tag{79}$$

Table 3: Statistics of the datasets.

Dataset	Task	Samples	Class counts	Parcellation	Time points
ABIDE	ASD vs. TD	1,025	488 / 537	AAL116; CC200	194.51 ± 58.58
REST-meta-MDD	MDD vs. HC	2,428	1,300 / 1,128	AAL116	207.54 ± 29.29
SRPBS	MDD vs. HC	1,000	500 / 500	AAL116	107–284
ABCD	ADHD vs. HC	425	213 / 212	AAL116	383

E Datasets and Preprocessing

E.1 Dataset description

We evaluate BRIO on four multi-site rs-fMRI datasets, retaining samples with diagnostic labels, site information, and usable ROI-level time series. Table 3 summarizes the final analysis cohorts, with class counts listed in the order of the corresponding task labels.

ABIDE. The Autism Brain Imaging Data Exchange (ABIDE) Di Martino et al. (2014) provides multi-site rs-fMRI data for studying autism spectrum disorder (ASD). We use the quality-controlled ABIDE I cohort with available phenotypic information and ROI-level time series. The final cohort contains 1,025 subjects, including 488 individuals with ASD and 537 typically developing (TD) controls from 20 sites. Acquisition sites differ in demographic composition and imaging protocols. Primary experiments use the AAL116 atlas Tzourio-Mazoyer et al. (2002), while an additional evaluation uses the CC200 parcellation Craddock et al. (2012) to examine performance under an alternative ROI definition.

REST-meta-MDD. The REST-meta-MDD consortium Yan et al. (2019) provides multi-site rs-fMRI data for studying major depressive disorder (MDD), collected across clinical centers in China. The final cohort comprises 2,428 participants from 25 acquisition sites, including 1,300 individuals with MDD and 1,128 healthy controls (HC). Each retained site includes both diagnostic groups. Site sizes range from 24 to 533 subjects, with the largest site accounting for approximately 22.0% of the cohort. Among participants with available age information, the mean age is 36.24 ± 15.07 years, ranging from 12 to 82 years, with group means of 36.23 ± 14.62 years for MDD and 36.25 ± 15.58 years for HC. Records with available sex information include 1,454 females and 925 males. All subjects are represented using AAL116, with variable numbers of available time points.

SRPBS. The Strategic Research Program for Brain Sciences (SRPBS) dataset Tanaka et al. (2021) is a multi-site neuroimaging resource collected in Japan, with demographic, diagnostic, and acquisition information. We use the MDD cohort for binary classification against HC. After quality control and cohort balancing, the analysis set contains 1,000 rs-fMRI subjects, comprising 500 MDD and 500 HC subjects from 11 acquisition environments. Ages range from 18 to 80 years, with an overall mean of 42.93 ± 13.59 years and group means of 42.66 ± 12.09 years for MDD and 43.21 ± 14.95 years for HC. The cohort includes 541 females and 459 males: 237 females and 263 males in the MDD group, and 304 females and 196 males in the HC group. ROI sequences use AAL116, with lengths ranging from 107 to 284 time points across acquisition environments.

ABCD. The Adolescent Brain Cognitive Development (ABCD) Study Casey et al. (2018) is a multi-site longitudinal study of brain development and mental health. We use baseline rs-fMRI data from `baseline_year_1_arm_1` for binary classification of attention-deficit/hyperactivity disorder (ADHD) versus HC. After quality control and site filtering, the final cohort contains 425 subjects from 10 imaging sites, including 213 individuals with ADHD and 212 HC. The mean age is 9.47 ± 0.50 years. All retained subjects are represented using AAL116 and have 383 retained time points.

E.2 Data Preprocessing

We apply a common FC estimation procedure across datasets and methods that take FC graphs as input. For each subject, frames containing non-finite values in any ROI are removed jointly across ROIs, and each retained ROI signal is standardized within subject. From the resulting sequence $\boldsymbol{X}_i^e \in \mathbb{R}^{T_i^e \times P}$, we compute pairwise Pearson correlations and apply the Fisher-z transformation to

obtain

$$\boldsymbol{A}_i^{e,(0)} = \Psi(\boldsymbol{X}_i^e) \in \mathbb{R}^{P\times P}, \tag{80}$$

where T_i^e denotes the number of retained time points. Diagonal entries are set to zero. We use $P = 116$ for AAL116 and $P = 200$ for CC200.

To encode these graphs over a common set of ROI pairs, BRIO and variants sharing its backbone use a binary propagation support $\boldsymbol{S}$ constructed exclusively from training-site full-scan FC matrices. For each ROI pair, we compute the median absolute FC weight across subjects within each training site, followed by the median across training sites. Each ROI selects the $m = \lceil \rho_g(P-1) \rceil$ strongest off-diagonal connections, with $\rho_g = 0.20$. We symmetrize the support by retaining an undirected connection whenever either endpoint selects the other. Constructed separately within each LOSO fold, $\boldsymbol{S}$ remains fixed across training, validation, and test subjects and all FC re-estimates. Each graph retains its own weights on this support, with positive weights and negative-weight magnitudes processed separately.

While full-scan graphs are used for parameter updates and inference, BRIO additionally constructs FC re-estimates from training subjects for qualification evaluation. We apply joint circular moving-block resampling with identical temporal indices across ROIs, preserving cross-ROI temporal correspondence and local temporal dependence within sampled blocks. Each resampled sequence $\boldsymbol{X}_i^{e,(k)}$ contains T_i^e time points and is processed using the same Pearson-correlation and Fisher-z procedure:

$$\boldsymbol{A}_i^{e,(k)} = \Psi\Big(\boldsymbol{X}_i^{e,(k)}\Big), \qquad k = 1, \ldots, K. \tag{81}$$

The block length is estimated once per LOSO fold from training subjects only. We draw a site- and class-balanced subset of at most 128 training subjects and, for each subject, randomly select up to 24 ROI signals and up to 48 pairwise ROI products. For each series, the autocorrelation horizon is the smallest lag at which the absolute autocorrelation stays below $\tau_{\mathrm{ac}} = 0.10$ for two consecutive lags, searched over lags up to $\min(30, \lfloor T_i^e/2 \rfloor)$. Horizons are aggregated by taking the 0.75 quantile over series within each subject, the median over subjects within each site, and the median over training sites, so that every source site contributes equally to the fold-level value ℓ_b. The block length applied to sequence i is

$$\ell_{b,i}^e = \mathrm{clip}\Big(\,\mathrm{round}(\ell_b),\ \ell_{\min},\ \min\big(\ell_{\max}, \lfloor T_i^e/2 \rfloor\big)\Big), \qquad \ell_{\min} = 2, \quad \ell_{\max} = 30. \tag{82}$$

The same estimation and clipping rules are applied across datasets.

F BASELINES

We compare BRIO with different competitive baselines across four categories:

General GNNs.

- **GCN** (Kipf & Welling, 2016) updates node representations through degree-normalized neighborhood propagation, combining local features with information from adjacent nodes.
- **GAT** (Veličković et al., 2017) computes feature-dependent attention coefficients for neighboring nodes and combines multiple attention heads to learn node representations.
- **GIN** (Xu et al., 2018) applies MLPs to summed neighborhood features and pools node embeddings to capture structural differences between graphs.

General OOD methods.

- **CORAL** (Sun & Saenko, 2016) reduces covariance differences between learned representations. In our source-only setting, this regularization is applied across training sites.
- **IRM** (Arjovsky et al., 2019) encourages a representation for which a shared classifier is simultaneously optimal across training environments, reducing reliance on environment-dependent associations.

Graph OOD methods.

- **GSAT** (Miao et al., 2022) learns probabilistic edge selection through an information-bottleneck objective, suppressing graph information that is unnecessary for prediction. The resulting edge probabilities also provide interpretable subgraph explanations.
- **DisC** (Fan et al., 2022) separately encodes causal and bias subgraphs, then recombines their representations to weaken spurious associations during training, particularly when graph labels correlate strongly with bias substructures.
- **CEPG** (Wang et al., 2026a) leverages environmental information to guide invariant graph representation learning and improve prediction under distribution shifts by emphasizing predictive features shared across different training environments.
- **DiSCO** (Sun et al., 2026) trains diverse subgraph experts and uses sparse gating to select their contributions for each graph, accommodating heterogeneous predictive structures. Different experts capture complementary causal patterns across graph instances.

Brain network methods.

- **BrainNetTF** (Kan et al., 2022) learns interactions among ROI connectivity profiles through Transformer attention and forms graph embeddings using orthonormal clustering readout, which groups ROIs according to their learned functional representations.
- **XG-GNN** (Qiu et al., 2024) constructs task-oriented nonlinear functional networks and uses meta-learning with explanation regularization to preserve diagnostic patterns across acquisition sites, jointly considering diagnostic accuracy and the consistency of explanations.
- **AGMGC** (Noman et al., 2025) combines learned connectivity with correlation-based graphs, integrating spline and multi-graph convolutions with contrastive regularization to capture local connectivity patterns and broader interregional dependencies.
- **FC-HGNN** (Gu et al., 2025) first encodes individual connectomes using hemisphere-aware convolutions, then combines imaging and non-imaging information through heterogeneous population-graph aggregation. This connects individual connectivity patterns with relationships among subjects.
- **BrainOOD** (Xu et al., 2025) filters node features and regularizes extracted graph structures through information-bottleneck and structure-consistency objectives for cross-site prediction. The consistency objective encourages similar structure selection across subjects.
- **DeCI** (Yu et al., 2026) separates ROI time series into oscillatory and slowly varying components, processes each ROI independently, and aggregates the resulting predictions, allowing temporal dynamics to inform diagnosis beyond static correlations.
- **CORE** (Wang et al., 2026b) extracts a reproducible connectivity scaffold after site-specific deconfounding and models pathway dynamics on its line graph using subject-adaptive gating to combine population-level connectivity priors with individual temporal characteristics.

G Algorithm

The overall training and inference process of the proposed BRIO is shown in Algorithm 1.

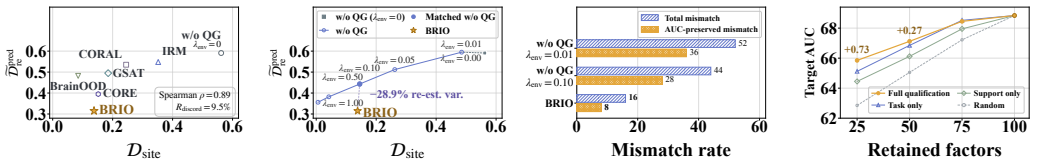
H Implementation Details

We implement BRIO in PyTorch [1]. The encoder is a two-layer signed GNN with hidden dimension 64 and 16-dimensional ROI identity embeddings. Positive and negative-magnitude adjacency matrices are symmetrically normalized separately. Each layer sums separate bias-free transformations of self-features and messages from both channels, followed by GELU, residual addition, and LayerNorm. We use $J = 4$ connectome factors with dimension $d_t = 16$. The shared ROI assignment uses 32-dimensional queries and keys, scaled dot-product similarities with temperature 0.10, and 60 log-domain Sinkhorn iterations targeting row and column masses of $1/J$ and $1/P$, respectively. The factor mapping $\mathcal{F}_\phi$ consists of a bias-free $64 \to 16$ projection followed by a $16 \to 32 \to 16$ MLP with GELU between its linear layers. Both the projection and MLP outputs are ℓ_2-normalized. We

[1] https://pytorch.org/

Table 4: ID results on ABIDE, ABIDE (CC200), REST-meta-MDD, SRPBS, and ABCD (ADHD-task) under ten-fold cross-validation stratified by site and class, reported as mean ± standard deviation across test folds. **Bold** results indicate the best performance.

Type	Method	ABIDE		ABIDE (CC200)		REST-meta-MDD		SRPBS		ABCD (ADHD-task)	
		AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
GNN	GCN	$63.45_{\pm 4.12}$	$59.21_{\pm 3.85}$	$51.88_{\pm 4.02}$	$51.05_{\pm 3.74}$	$62.15_{\pm 2.88}$	$57.02_{\pm 2.45}$	$84.12_{\pm 3.85}$	$75.34_{\pm 4.12}$	$72.18_{\pm 5.12}$	$64.05_{\pm 4.88}$
	GAT	$59.12_{\pm 3.95}$	$55.44_{\pm 3.55}$	$58.62_{\pm 4.15}$	$57.18_{\pm 3.82}$	$63.84_{\pm 3.12}$	$59.12_{\pm 2.65}$	$81.75_{\pm 4.05}$	$70.88_{\pm 4.35}$	$71.02_{\pm 5.45}$	$63.88_{\pm 5.10}$
	GIN	$58.74_{\pm 3.65}$	$54.95_{\pm 3.22}$	$53.15_{\pm 4.33}$	$50.42_{\pm 3.95}$	$59.33_{\pm 2.54}$	$56.14_{\pm 2.22}$	$82.05_{\pm 3.78}$	$71.45_{\pm 3.92}$	$69.45_{\pm 5.33}$	$61.22_{\pm 4.75}$
OOD	CORAL	$56.88_{\pm 4.05}$	$53.25_{\pm 3.75}$	$55.14_{\pm 4.45}$	$53.75_{\pm 4.05}$	$59.95_{\pm 3.15}$	$57.85_{\pm 2.85}$	$81.14_{\pm 4.12}$	$70.82_{\pm 4.05}$	$70.12_{\pm 5.25}$	$63.45_{\pm 4.85}$
	IRM	$60.15_{\pm 3.88}$	$56.78_{\pm 3.45}$	$53.42_{\pm 4.65}$	$54.12_{\pm 4.11}$	$60.75_{\pm 2.95}$	$57.44_{\pm 2.75}$	$80.22_{\pm 3.94}$	$70.15_{\pm 3.88}$	$71.25_{\pm 5.08}$	$60.14_{\pm 4.95}$
Graph OOD	GSAT	$59.85_{\pm 3.75}$	$57.14_{\pm 3.52}$	$52.45_{\pm 3.95}$	$52.12_{\pm 3.65}$	$58.45_{\pm 2.85}$	$55.75_{\pm 2.55}$	$82.45_{\pm 3.65}$	$72.15_{\pm 3.85}$	$72.05_{\pm 5.14}$	$62.45_{\pm 4.92}$
	DisC	$58.12_{\pm 4.15}$	$57.33_{\pm 3.85}$	$54.15_{\pm 4.25}$	$53.85_{\pm 4.02}$	$59.88_{\pm 3.05}$	$57.12_{\pm 2.75}$	$83.12_{\pm 3.85}$	$73.55_{\pm 4.15}$	$70.45_{\pm 5.05}$	$59.75_{\pm 4.85}$
	CEPG	$53.45_{\pm 4.55}$	$51.15_{\pm 4.12}$	$46.85_{\pm 4.85}$	$48.75_{\pm 4.45}$	$57.15_{\pm 3.25}$	$55.45_{\pm 2.95}$	$76.15_{\pm 4.25}$	$69.12_{\pm 4.35}$	$64.12_{\pm 5.85}$	$58.45_{\pm 5.15}$
	DiSCO	$58.25_{\pm 3.85}$	$56.45_{\pm 3.65}$	$53.12_{\pm 4.15}$	$52.75_{\pm 3.85}$	$60.15_{\pm 2.75}$	$57.15_{\pm 2.55}$	$85.12_{\pm 3.55}$	$74.15_{\pm 3.75}$	$68.75_{\pm 5.15}$	$62.85_{\pm 4.85}$
Brain Networks	BrainNetTF	$63.75_{\pm 3.55}$	**61.85**$_{\pm 3.25}$	$58.45_{\pm 4.15}$	$54.15_{\pm 3.85}$	$62.85_{\pm 2.55}$	$58.15_{\pm 2.35}$	$79.45_{\pm 4.15}$	$71.15_{\pm 4.35}$	$69.15_{\pm 4.85}$	$58.15_{\pm 4.55}$
	XG-GNN	$61.45_{\pm 3.75}$	$55.45_{\pm 3.45}$	$52.15_{\pm 4.55}$	$54.15_{\pm 4.15}$	$60.45_{\pm 2.95}$	$56.75_{\pm 2.65}$	$84.75_{\pm 3.65}$	$73.85_{\pm 3.85}$	$72.45_{\pm 4.85}$	$57.15_{\pm 4.65}$
	AGMGC	$59.85_{\pm 3.85}$	$52.75_{\pm 3.55}$	$57.15_{\pm 3.95}$	$52.15_{\pm 3.65}$	$58.85_{\pm 2.85}$	$55.15_{\pm 2.55}$	$86.45_{\pm 3.45}$	$77.15_{\pm 3.65}$	$71.85_{\pm 5.05}$	$64.15_{\pm 4.75}$
	FC-HGNN	$57.15_{\pm 4.15}$	$53.85_{\pm 3.85}$	$56.45_{\pm 4.25}$	$54.75_{\pm 3.95}$	$61.45_{\pm 3.05}$	$57.15_{\pm 2.75}$	$77.85_{\pm 4.05}$	$68.75_{\pm 4.25}$	$72.15_{\pm 4.95}$	$60.45_{\pm 4.85}$
	BrainOOD	$64.75_{\pm 3.45}$	$60.45_{\pm 3.15}$	$63.85_{\pm 3.65}$	$60.15_{\pm 3.35}$	$64.75_{\pm 2.75}$	$61.15_{\pm 2.45}$	$86.45_{\pm 3.25}$	$77.45_{\pm 3.45}$	$73.45_{\pm 4.75}$	$65.15_{\pm 4.55}$
	DeCI	$57.45_{\pm 4.05}$	$61.45_{\pm 3.55}$	$60.15_{\pm 3.85}$	**62.85**$_{\pm 3.45}$	$55.15_{\pm 3.15}$	$55.85_{\pm 2.85}$	$87.45_{\pm 3.35}$	$79.85_{\pm 3.55}$	$56.45_{\pm 5.25}$	$63.45_{\pm 4.95}$
	CORE	$65.45_{\pm 3.25}$	$60.85_{\pm 2.95}$	$67.45_{\pm 3.45}$	$62.15_{\pm 3.15}$	$65.85_{\pm 2.65}$	$61.45_{\pm 2.35}$	**88.75**$_{\pm 3.05}$	**81.45**$_{\pm 3.25}$	$73.15_{\pm 4.55}$	$67.45_{\pm 4.35}$
	BRIO	**66.85**$_{\pm 3.15}$	$61.55_{\pm 2.85}$	**68.75**$_{\pm 3.25}$	$62.65_{\pm 3.05}$	**66.45**$_{\pm 2.55}$	**62.15**$_{\pm 2.25}$	$88.15_{\pm 3.15}$	$80.75_{\pm 3.35}$	**76.45**$_{\pm 4.25}$	**69.15**$_{\pm 4.05}$



(a) Cross-method (b) Alignment (c) Mismatch (d) Factor retention

Figure 4: Cross-method relationships (a), alignment trajectories (b), mismatch rates (c), and factor retention (d) on REST-meta-MDD.

train the model using AdamW with learning rate 3×10^{-4} and weight decay 10^{-4}. Each LOSO fold reserves one site for testing and two for validation, while model fitting and estimation of training statistics use only the remaining sites. Results are reported as mean ± standard deviation across held-out test sites.

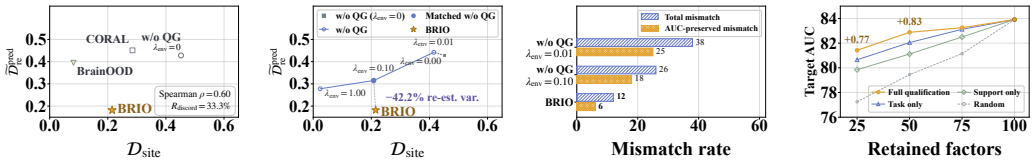
I Complexity Analysis

We analyze the per-step cost of BRIO after supervised warm-up, where the support $\boldsymbol{S}$ and assignment $\overline{\boldsymbol{\Pi}}$ are fixed. Let B be the batch size, E the number of source sites, $N_{\mathrm{src}} = \sum_e N_e$ the number of source training subjects, P the number of ROIs, T the number of time points, $s = \|\boldsymbol{S}\|_0$ the number of retained adjacency entries, J the number of factors, L the number of GNN layers, and $d = \max(d_h, d_t)$. With sparse propagation, encoding and factorization cost $C_{\mathrm{enc}} = L(sd + Pd^2) + PJd + Jd^2$ per graph. For batches balanced across sites and classes, forming pair-specific representations and cost matrices takes $\mathcal{O}(B^2Jd)$, and I_{sk} Sinkhorn iterations over all cross- and self-transport problems take $\mathcal{O}(B^2 I_{\mathrm{sk}})$. Every M steps, the qualification refresh generates K re-estimates per source subject at $C_{\mathrm{FC}} = \mathcal{O}(P^2T)$ each, encodes $K+1$ graphs per subject, and forms qualifications over site pairs in $\mathcal{O}(E^2J)$. The amortized time per step is therefore $\mathcal{O}\big((B + N_{\mathrm{src}}(K+1)/M)\, C_{\mathrm{enc}} + B^2(Jd + I_{\mathrm{sk}}) + (N_{\mathrm{src}}KC_{\mathrm{FC}} + E^2J)/M\big)$. Relative to class-conditional alignment without qualification, the overhead is confined to the $1/M$ terms. Since the refresh runs without gradients and discards re-estimates after use, peak memory is that of a training batch, $\mathcal{O}(BL(Pd+s) + B^2)$, independent of N_{src} and K.

J More Results

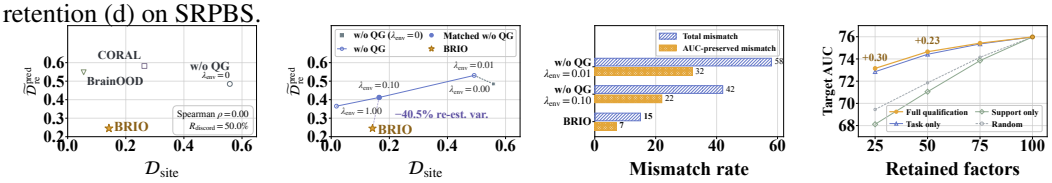
J.1 ID Performance Comparison

In this part, we assess whether BRIO also performs well under in-distribution (ID) settings. Subjects from all sites are pooled and partitioned into ten folds stratified by site and class, so that every fold preserves the site and class composition of the full cohort. Each fold in turn serves as the test set, the next two folds serve as the validation set, and the remaining seven folds form the training set, giving



(a) Cross-method (b) Alignment (c) Mismatch (d) Factor retention

Figure 5: Cross-method relationships (a), alignment trajectories (b), mismatch rates (c), and factor retention (d) on SRPBS.



(a) Cross-method (b) Alignment (c) Mismatch (d) Factor retention

Figure 6: Cross-method relationships (a), alignment trajectories (b), mismatch rates (c), and factor retention (d) on ABCD.

a 7:2:1 rotation in which every subject is tested exactly once. All sites appear in training and every training site contains both classes, so source sites continue to serve as environments for alignment. As shown in Table 4, BRIO ranks first or second on both metrics across all five settings. It achieves the highest mean AUC on ABIDE, ABIDE (CC200), REST-meta-MDD, and ABCD, and the highest mean ACC on REST-meta-MDD and ABCD. BrainNetTF and DeCI obtain the highest ACC on ABIDE and ABIDE (CC200), respectively, while CORE leads both metrics on SRPBS, with BRIO closely following in each case. These results show that the predictive value of re-estimation-informed learning extends beyond unseen-site evaluation and supports competitive ID performance alongside the cross-site OOD results.

J.2 More Site Discrepancy and Re-estimation Variation

In this part, we define the two evaluation metrics used in Sec. 4.4 and extend the cross-method comparison and alignment-strength sweep to REST-meta-MDD, SRPBS, and ABCD. The cross-method comparison includes CORAL, IRM, GSAT, BrainOOD, CORE, BRIO w/o QG with $\lambda_{\text{env}} = 0$, and BRIO on ABIDE and REST-meta-MDD, and CORAL, BrainOOD, BRIO w/o QG, and BRIO on SRPBS and ABCD. Each method uses the checkpoint selected in the main experiment and is evaluated with all trained parameters frozen and stochastic operations disabled. All methods receive the same source training subjects, full-scan FC graphs, and FC re-estimates, and each observation corresponds to one method, held-out site, and random seed.

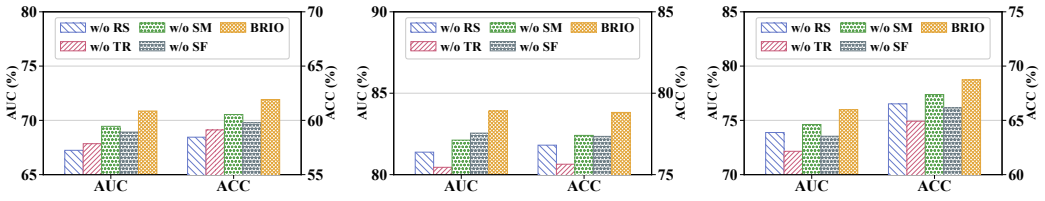
For each method, let $\widehat{P}_{e,c}$ denote the empirical distribution of unit-normalized full-scan representations from source site e and class c, taken from the layer preceding the classifier. The class-conditional site discrepancy is

$$\mathcal{D}_{\text{site}} = \frac{1}{E(E-1)} \sum_{1 \le e < e' \le E} \sum_{c \in \{0,1\}} \mathcal{S}_\varepsilon \left(\widehat{P}_{e,c}, \widehat{P}_{e',c} \right), \tag{83}$$

where E is the number of source sites. The debiased Sinkhorn divergence $\mathcal{S}_\varepsilon$ uses squared Euclidean distances and identical solver settings across methods, with the same representation in cross- and self-transport terms and without qualification-based weighting. Lower values indicate closer full-scan representations across source sites under a common geometry.

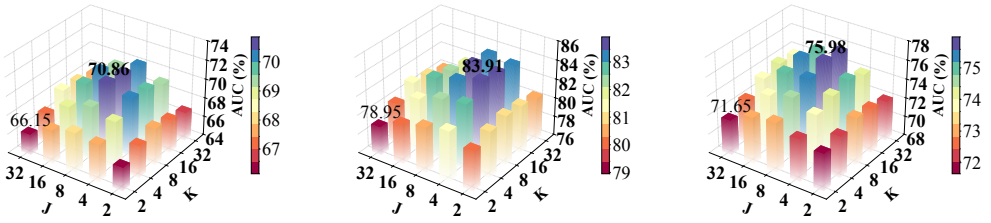
To measure predictive changes, let $\ell_i^{(k)}$ be the binary logit margin, given by the pre-sigmoid logit or the difference between the two softmax logits; a common shift of two logits leaves the predicted probability unchanged and is thus excluded. With $k = 0$ denoting the full-scan prediction, the normalized predictive re-estimation variation is

$$\widetilde{\mathcal{D}}_{\text{re}}^{\text{pred}} = \frac{1}{2E} \sum_{e=1}^{E} \sum_{c \in \{0,1\}} \frac{\widehat{\mathbb{E}}_{i,k|e,c} \left[\left(\ell_i^{(k)} - \ell_i^{(0)} \right)^2 \right]}{V_{e,c}^{\text{pred}} + \frac{1}{4} \left(\mu_{e,1}^{\text{pred}} - \mu_{e,0}^{\text{pred}} \right)^2 + \epsilon}, \tag{84}$$



(a) ABIDE (CC200) (b) SRPBS (c) ABCD

Figure 7: Ablation study on ABIDE (CC200), SRPBS, and ABCD.



(a) ABIDE (CC200) (b) SRPBS (c) ABCD

Figure 8: Sensitivity study on ABIDE (CC200), SRPBS, and ABCD.

where the empirical expectation averages over source subjects and their re-estimates with $k \geq 1$, $\mu_{e,c}^{\text{pred}}$ and $V_{e,c}^{\text{pred}}$ are the full-scan logit mean and variance within each site and class, and $\epsilon > 0$ ensures numerical stability. Numerator and denominator scale identically under a global rescaling of the logits, so the ratio measures re-estimation change relative to the within-site spread and class separation of full-scan predictions.

This output-level variation and the factor-level re-estimation support in Sec. 3.2 capture different aspects of stability. Since $z_i = J^{-1/2} \sum_j a_{i,j} + b$, a re-estimation change of the logit satisfies

$$\mathbb{E}\big[(\delta z_i)^2\big] = \frac{1}{J}\Big[\sum_j \mathbb{E}\big[(\delta a_{i,j})^2\big] + 2\sum_{j<l} \mathbb{E}[\delta a_{i,j}\, \delta a_{i,l}]\Big], \tag{85}$$

so changes of individual factors may add or cancel at the output, and the two quantities are reported separately.

The cross-method plots show method-level means, with Spearman ρ computed across individual runs. R_{discord} is the percentage of method pairs within the same fold and seed for which smaller $\mathcal{D}_{\text{site}}$ corresponds to larger $\widetilde{\mathcal{D}}_{\text{re}}^{\text{pred}}$, excluding ties. A run is a mismatch when its $\mathcal{D}_{\text{site}}$ falls below and its $\widetilde{\mathcal{D}}_{\text{re}}^{\text{pred}}$ falls above the pooled dataset medians. The total mismatch rate is the percentage of mismatches among all evaluated runs, and the AUC-preserved mismatch rate counts only mismatches whose source-validation AUC is no more than a tolerance δ_{AUC} below that of BRIO w/o QG with $\lambda_{\text{env}} = 0$, again over all evaluated runs.

For the controlled sweep, BRIO w/o QG shares the graph encoder, factor representation, classifier, and training procedure of BRIO, and removes qualification guidance by assigning equal weights to all factors and disabling qualification-based strength modulation. We train this variant independently for each tested value of λ_{env} and select its checkpoint by the same source-validation rule as the main experiment. Runs are paired by fold and seed, and the trajectories show means connected in increasing λ_{env}. The full BRIO is added as a separate reference point, and Matched w/o QG denotes the setting with positive λ_{env} selected by source validation.

Panels (a) of Figs. 4, 5, and 6 show positive rank correlations between the two metrics on REST-meta-MDD and SRPBS but zero rank correlation on ABCD, and R_{discord} is nonzero on every dataset, so lower site discrepancy does not consistently imply lower predictive variation under FC re-estimation. The controlled sweeps in panels (b) show that stronger alignment reduces both quantities, while the full BRIO achieves lower variation than BRIO w/o QG at comparable site discrepancy, indicating that qualification-guided weighting and strength modulation offer benefits beyond simply increasing alignment strength. Panels (c) further show that BRIO achieves lower total and AUC-preserved

Table 5: Sensitivity of BRIO to moving-block length ℓ_b. **Bold** indicates the highest mean.

Dataset	0.5×			1× (Default)			2×		
	ℓ_b	AUC	ACC	ℓ_b	AUC	ACC	ℓ_b	AUC	ACC
ABIDE	3 (6.0)	$65.84_{\pm 4.67}$	$60.25_{\pm 4.15}$	6 (12.0)	$\mathbf{67.53}_{\pm 4.12}$	$\mathbf{62.18}_{\pm 3.68}$	12 (24.0)	$66.71_{\pm 4.28}$	$61.54_{\pm 3.92}$
ABIDE (CC200)	3 (6.0)	$69.45_{\pm 5.25}$	$60.12_{\pm 4.58}$	6 (12.0)	$\mathbf{70.86}_{\pm 4.87}$	$\mathbf{61.92}_{\pm 3.94}$	12 (24.0)	$68.72_{\pm 5.15}$	$59.88_{\pm 4.25}$
REST-meta-MDD	4 (8.0)	$67.85_{\pm 4.22}$	$63.38_{\pm 3.42}$	7 (14.0)	$\mathbf{68.84}_{\pm 3.75}$	$\mathbf{64.42}_{\pm 2.91}$	14 (28.0)	$67.62_{\pm 3.82}$	$63.85_{\pm 3.18}$
SRPBS	3 (6.0)	$81.35_{\pm 6.35}$	$76.54_{\pm 5.82}$	6 (12.0)	$\mathbf{83.91}_{\pm 5.63}$	$\mathbf{78.82}_{\pm 5.28}$	12 (24.0)	$83.72_{\pm 5.58}$	$78.45_{\pm 5.15}$
ABCD (ADHD-task)	7 (5.6)	$72.15_{\pm 5.45}$	$65.12_{\pm 4.85}$	15 (12.0)	$\mathbf{75.98}_{\pm 4.39}$	$\mathbf{68.74}_{\pm 3.82}$	30 (24.0)	$75.45_{\pm 4.52}$	$67.92_{\pm 3.95}$

mismatch rates than the evaluated w/o QG settings on all three datasets, indicating fewer mismatches even when predictive performance is preserved.

J.3 More Qualification and Factor Transferability

In this part, we extend the factor-retention analysis to REST-meta-MDD, SRPBS, and ABCD. As shown in panels (d) of Figs. 4, 5, and 6, full qualification outperforms the other rankings when retaining 25% or 50% of factors across all three datasets. Its advantage over task relevance alone shows that re-estimation support provides complementary information for prioritizing factors with predictive value on unseen sites. Both rankings outperform support alone at partial retention, and support alone falls below random ordering on ABCD, so re-estimation support identifies transferable factors only together with task relevance. Performance gaps narrow as more factors are retained and vanish at full retention, where all rankings recover the original predictions.

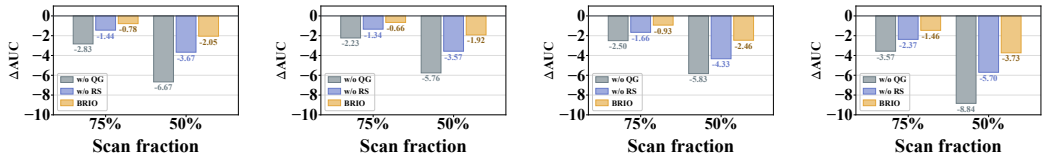
J.4 More ablation studies

In this part, we extend the ablation study to ABIDE (CC200), SRPBS, and ABCD using the same four variants as in Sec.4.2. As shown in Fig. 7, BRIO achieves the highest AUC and ACC across all three settings. BRIO w/o RS shows the largest performance drop on ABIDE (CC200), whereas BRIO w/o TR yields the lowest performance on SRPBS and ABCD. These differences suggest that the relative contributions of re-estimation support and task relevance vary across datasets, while their joint use consistently benefits cross-site prediction. BRIO w/o SM and BRIO w/o SF also underperform the full model on both metrics, supporting the complementary roles of modulating overall alignment strength and prioritizing supported factors in cross-site matching.

J.5 More sensitivity analysis

First, we extend the sensitivity analysis of the number of FC re-estimates K and connectome factors J to ABIDE (CC200), SRPBS, and ABCD. As shown in Fig. 8, increasing K from small values generally improves AUC, whereas further increases yield no consistent gains. This suggests that additional re-estimates help characterize variability in predictive contributions, with diminishing benefits at larger budgets. Intermediate values of J also tend to outperform very small or large factor counts, indicating that finer factorization does not necessarily improve cross-site prediction.

Additionally, we examine the sensitivity of BRIO to the temporal block length ℓ_b used for within-scan FC re-estimation. Whereas K controls the number of re-estimates, ℓ_b determines the length of contiguous temporal segments retained during resampling. We compare shorter, default, and longer blocks at nominal scales of $0.5\times$, $1\times$, and $2\times$: the $1\times$ setting is the fold-level estimate rounded to the nearest integer, and the $0.5\times$ and $2\times$ settings scale the unrounded estimate before rounding and are used as fixed block lengths. Table 5 reports each block length in time points together with its duration in seconds. The default block length achieves the highest mean AUC and ACC in all five settings, and the estimated defaults correspond to a similar physical duration across datasets despite their different repetition times, indicating that the autocorrelation-based estimate captures a temporal scale that transfers across acquisition protocols. Both halving and doubling the block length degrade performance, with halving being more harmful on most datasets; blocks that are too short break the temporal dependence that FC re-estimation is meant to preserve, whereas overly long blocks reduce the diversity of re-estimates.



(a) ABIDE (b) REST-meta-MDD (c) SRPBS (d) ABCD

Figure 9: Performance under reduced scan data on different datasets.

Table 6: Training time (in seconds) and GPU memory consumption (in GB) for each training epoch.

Method	ABIDE		REST-meta-MDD		SRPBS		ABCD	
	Time	Memory	Time	Memory	Time	Memory	Time	Memory
GSAT	0.68	2.1	1.52	3.3	0.48	2.0	0.34	1.7
DisC	1.05	2.4	2.38	3.8	0.74	2.2	0.51	1.9
CEPG	0.58	2.0	1.34	3.1	0.41	1.9	0.28	1.6
DiSCO	1.65	2.9	3.85	4.6	1.18	2.7	0.78	2.3
BrainNetTF	0.78	2.0	1.65	3.2	0.49	1.9	0.32	1.6
AGMGC	0.71	1.8	1.58	3.0	0.46	1.8	0.30	1.5
FC-HGNN	0.34	1.8	0.48	3.1	0.19	1.8	0.12	1.5
XG-GNN	0.63	2.1	1.31	3.4	0.38	2.0	0.25	1.7
DeCI	0.88	2.3	1.95	3.6	0.62	2.1	0.44	1.8
BrainOOD	1.45	2.5	3.32	4.1	0.88	2.3	0.58	2.0
CORE	0.35	2.0	0.52	3.7	0.22	1.9	0.12	1.6
BRIO	1.22	2.6	2.85	4.2	0.86	2.4	0.56	2.1

J.6 Effect of Scan Length on Predictive Performance

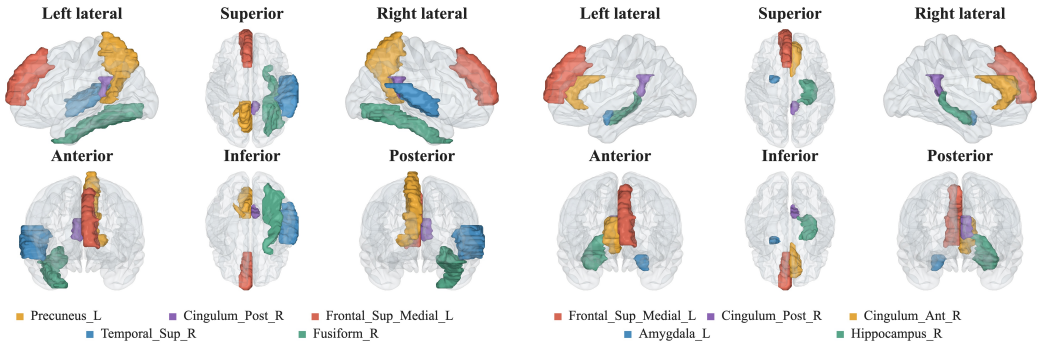
To examine how reduced temporal data for target FC estimation affects cross-site prediction, we evaluate BRIO, BRIO w/o RS, and BRIO w/o QG on ABIDE, REST-meta-MDD, SRPBS, and ABCD. We reconstruct FC graphs using 75% or 50% of each subject's preprocessed BOLD sequence, keeping all source-selected checkpoints fixed. At each fraction, AUC is computed separately for beginning, middle, and end windows and then averaged across positions. We report $\Delta\mathrm{AUC}$ relative to each model's full-scan performance. As shown in Fig. 9, all three models exhibit larger AUC reductions when fewer time points are retained, but BRIO consistently shows the smallest decline. BRIO w/o RS incurs smaller drops than BRIO w/o QG, while remaining more sensitive than the full model. The differences in degradation widen at the lower scan fraction, suggesting that re-estimation support provides benefits beyond task relevance alone. These results support the value of re-estimation-informed learning for maintaining cross-site prediction performance when target FC graphs are estimated from fewer temporal observations.

J.7 Training Efficiency and Resource Consumption

In this part, we compare the training efficiency of BRIO and representative baselines on four datasets. Table 6 reports per-epoch training time in seconds and GPU memory consumption in GB. We observe that BRIO requires more training time and GPU memory than lightweight baselines such as FC-HGNN and CORE. Compared with BrainOOD, it achieves shorter per-epoch training times while using slightly more memory, reflecting a trade-off between runtime and memory consumption. In contrast, BRIO requires less time and memory than DiSCO. These relative resource-use patterns remain consistent across all evaluated datasets.

K Interpretability Analysis

To examine how qualification-based factor weighting translates into regional pooling emphasis, we analyze the models trained on ABIDE and REST-meta-MDD using AAL116. For each model, we first summarize the normalized qualifications $\overline{\alpha}_{e,e',c,j}$ over all unordered source-site pairs and both



(a) ABIDE (b) REST-meta-MDD

Figure 10: Anatomical distribution of qualification-weighted ROI profiles on ABIDE and REST-meta-MDD.

classes:

$$\overline{\alpha}_j^{\text{global}} = \frac{1}{E(E-1)} \sum_{\substack{e,e' \in \mathcal{E}_{\text{tr}} \\ e<e'}} \sum_{c\in\{0,1\}} \overline{\alpha}_{e,e',c,j}, \qquad \sum_{j=1}^{J} \overline{\alpha}_j^{\text{global}} = 1. \tag{86}$$

This equal-weight summary captures relative factor allocation across source-site pairs and classes, separately from the overall alignment strength controlled by $g_{e,e',c}$. We then project these qualifications through the factor–ROI assignment $\overline{\mathbf{\Pi}}$ retained after supervised warm-up. For ROI u, the resulting score and its normalization are

$$S_u = \sum_{j=1}^{J} \overline{\alpha}_j^{\text{global}} \overline{\Pi}_{j,u}, \qquad \widehat{S}_u = \frac{S_u}{\sum_{v=1}^{P} S_v} = JS_u. \tag{87}$$

The assignment row masses of $1/J$ give $\sum_u S_u = 1/J$, so $\widehat{S}_u$ defines a normalized ROI profile. Higher scores indicate greater pooling emphasis in factors with higher average source qualification. Under uniform factor weights, the column-mass constraint yields $\widehat{S}_u = 1/P$ for every ROI. Regional differences therefore reflect how qualification-based factor weighting interacts with the learned ROI assignment.

To summarize these profiles across LOSO folds and random seeds, we compute $\widehat{S}_u^{(r,s)}$ separately for each fold r and seed s, using the qualifications and retained assignment of that model. Since factor indices need not correspond across independently trained models, we aggregate the scores in the common AAL116 ROI space:

$$S_u^{\text{dataset}} = \frac{1}{|\mathcal{R}||\mathcal{S}|} \sum_{r\in\mathcal{R}} \sum_{s\in\mathcal{S}} \widehat{S}_u^{(r,s)}, \qquad \sum_{u=1}^{P} S_u^{\text{dataset}} = 1, \tag{88}$$

where $\mathcal{R}$ and $\mathcal{S}$ denote the evaluated folds and seeds. This procedure preserves each model's factor–ROI correspondence during projection and yields a dataset-level summary of regional pooling emphasis.

Fig. 10 visualizes the five ROIs with the highest dataset-averaged scores. On ABIDE, these are the left precuneus, right superior temporal gyrus, right posterior cingulate gyrus, left medial superior frontal gyrus, and right fusiform gyrus. On REST-meta-MDD, the selected regions are the left medial superior frontal gyrus, left amygdala, right posterior cingulate gyrus, right hippocampus, and right anterior cingulate gyrus. The right posterior cingulate and left medial superior frontal regions appear in both profiles, while the remaining selected ROIs differ between datasets. These shared and differing regions illustrate how the factors prioritized by source-side qualification distribute pooling emphasis across the anatomical space. The profiles complement the ablation and factor-retention analyses by connecting the relative factor weights used for alignment to their regional pooling structure.

Algorithm 1 Training and Inference of BRIO

Require: Source datasets $\{\mathcal{D}^e\}_{e\in\mathcal{E}_{\rm tr}}$ with $E \geq 2$ and $N_{e,c} > 0$; unseen-site sequences $\{\boldsymbol{X}_z\}_{z=1}^{N_{\rm te}}$; Ψ, ρ_g, J, and K; block-length rule with autocorrelation threshold $\tau_{\rm ac}$ and bounds $[\ell_{\min}, \ell_{\max}]$ in time points; warm-up epochs $T_{\rm w}$; qualification interval M in optimization steps; ε, $\lambda_{\rm env}$, and a training stopping criterion.

Ensure: Unseen-site prediction probabilities $\widehat{\boldsymbol{p}}_{\rm te}$.

Cache source full-scan graphs $\boldsymbol{A}_i^{e,(0)} = \Psi(\boldsymbol{X}_i^e)$ and construct $\boldsymbol{S}$ using Eq. (3).
Estimate the fold-level block length ℓ_b from autocorrelation horizons of training-subject ROI signals with threshold $\tau_{\rm ac}$ as in Appendix E.2; for every source subject set $\ell_{b,i}^e \leftarrow \mathrm{clip}\big(\,\mathrm{round}(\ell_b),\, \ell_{\min},\, \min(\ell_{\max}, \lfloor T_i^e/2 \rfloor)\big)$.
Initialize all learnable parameters Θ; set $\Theta_{\rm asg} = \{\boldsymbol{Q}, \omega\}$ and $\mathcal{C}_Q \leftarrow \varnothing$.
Use mini-batches with equally sized nonempty subsets $\mathcal{B}_{e,c}$ for every source site and class; evaluate $\mathcal{L}_{\rm cls}$ using $\mathcal{B}_{e,c}$ and $|\mathcal{B}_{e,c}|$ in place of $\mathcal{I}_{e,c}$ and $N_{e,c}$.
Stage 1: Supervised Warm-Up and Factor Anchoring
Enable training mode and gradients.
for $t_{\rm w} = 1, \ldots, T_{\rm w}$ **do**
  for each source mini-batch $\mathcal{B}$ **do**
    Compute $\boldsymbol{\Pi}$, full-scan factors, and predictions using Eqs. (4)–(10); update Θ using $\mathcal{L}_{\rm cls}$.
  end for
end for
In evaluation mode without gradients, cache $\overline{\boldsymbol{\Pi}} \leftarrow \mathcal{A}_\omega(\boldsymbol{Q}, \boldsymbol{E}^{\rm ROI})$.
Freeze $\Theta_{\rm asg}$ and use $\overline{\boldsymbol{\Pi}}$ for subsequent pooling, while keeping ROI embeddings trainable.
Stage 2: Periodic Qualification-Guided Learning
Set $t \leftarrow 0$.
while the training stopping criterion is not met **do**
  if $t \bmod M = 0$ **then**
    Switch to evaluation mode, disable gradients, and set $\mathcal{C}_Q \leftarrow \varnothing$.
    For every source subject $i \in \mathcal{I}_{e,c}$, construct K FC re-estimates with block length $\ell_{b,i}^e$ using Eq. (2).
    Compute paired contributions $\{a_{i,j}^{e,(k)}\}_{k=0}^{K}$ using fixed $\boldsymbol{S}$ and $\overline{\boldsymbol{\Pi}}$.
    Compute $R_j^{\rm task}$ using Eqs. (11)–(13).
    if $\sum_j R_j^{\rm task} > 0$ **then**
      Compute $\pi_j^{\rm task}$ using Eq. (13) and $q_{e,c,j}^{\rm re}$ using Eqs. (14)–(16).
      Compute and cache $\mathcal{C}_Q \leftarrow \{g_{e,e',c}, \overline{\alpha}_{e,e',c,j}\}_{e<e',c,j}$ using Eqs. (18)–(19).
    end if
    Discard temporary FC re-estimates and restore training mode and gradients.
  end if
  Sample $\mathcal{B}$; compute full-scan factors, predictions, and $\mathcal{L}_{\rm cls}$ using fixed $\boldsymbol{S}$ and $\overline{\boldsymbol{\Pi}}$.
  Set $\mathcal{L}_{\rm env}^Q \leftarrow 0$.
  if $\mathcal{C}_Q \neq \varnothing$ **then**
    Construct weighted representations and empirical distributions using cached $\overline{\alpha}$ and Eqs. (22)–(24).
    Evaluate $\mathcal{L}_{\rm env}^Q$ using cached g and Eq. (25), with the same pairwise representation map and ε in all cross- and self-transport terms.
  end if
  Update $\Theta \setminus \Theta_{\rm asg}$ using $\mathcal{L}_{\rm cls} + \lambda_{\rm env}\mathcal{L}_{\rm env}^Q$, holding $\boldsymbol{S}$, $\overline{\boldsymbol{\Pi}}$, and $\mathcal{C}_Q$ fixed.
  Set $t \leftarrow t + 1$.
end while
Stage 3: Inference on Unseen Sites
Switch to evaluation mode and disable gradients.
for $z = 1, \ldots, N_{\rm te}$ **do**
  Compute $\boldsymbol{A}_z^{(0)} = \Psi(\boldsymbol{X}_z)$ and full-scan factors $\{\boldsymbol{r}_{z,j}^{(0)}\}_{j=1}^{J}$ using fixed $\boldsymbol{S}$, $\overline{\boldsymbol{\Pi}}$, and the trained mappings.
  Predict $\widehat{p}_z = \sigma\Big(J^{-1/2}\sum_{j=1}^{J} \boldsymbol{w}_j^\top \boldsymbol{r}_{z,j}^{(0)} + b\Big)$.
end for
return $\widehat{\boldsymbol{p}}_{\rm te} = [\widehat{p}_1, \ldots, \widehat{p}_{N_{\rm te}}]^\top$.