BEYOND PREDICTIVE INVARIANCE: CAUSAL MECHANISM TRANSPORT FOR GENERALIZABLE BRAIN NETWORK ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain network analysis is challenged by substantial distribution shifts arising from population and acquisition heterogeneity across imaging sites. Existing domain generalization methods improve robustness by seeking predictive regularities shared across source domains, while site heterogeneity can arise through distinct biological and measurement pathways. To characterize transportability under these structured variations, we introduce *Causal Mechanism Transport* (CMT), which models the two pathways in a multi-site selection diagram and derives a conditional criterion for transportable disease-related representations. To operationalize this criterion from observed multi-site data, we develop *Causal Information Contrast* (CIC), which contrasts representation distributions before and after conditioning on site within comparable disease and biological contexts. Instantiated with conditional Cauchy-Schwarz divergence, CIC provides a differentiable training objective. Experiments spanning controlled biological and measurement shifts, multiple real-world generalization settings, and independent representation audits across five neuroimaging cohorts demonstrate consistent gains across architectures and reduced conditional site dependence.

1 INTRODUCTION

Brain network analysis characterizes structural and functional brain organization from neuroimaging data Bullmore & Sporns (2009), while machine learning on brain networks shows promise for brain disorder diagnosis. Yet generalization across imaging centers remains challenging. Models trained on source sites often perform substantially worse on unseen sites, as neuroimaging distributions can vary with population composition, scanner hardware, acquisition protocols, and other site-specific factors Castro et al. (2020); Glocker et al. (2019); Yoon et al. (2024). Such shifts make cross-site generalization a central challenge for reliable brain network analysis.

Domain generalization methods commonly improve robustness by exploiting predictive regularities shared across source domains through invariant prediction, distribution alignment, or domain-invariant representation learning Arjovsky et al. (2019); Ganin et al. (2016). Similar strategies have recently been adopted in brain-network analysis to identify site-robust features, invariant subgraphs, or disentangled representations Xu et al. (2025); Wang et al. (2026b); Li et al. (2025). In multi-site neuroimaging, site heterogeneity can originate from biologically meaningful population characteristics and acquisition-dependent measurement conditions, which enter different stages of the data-generating process. Distinguishing these sources introduces a complementary question for cross-site generalization:

What should be transported across domains for generalizable brain network analysis?

We answer this question from a causal perspective. For brain network-based disorder classification, disease status gives rise to an underlying brain state that is subsequently observed through the neuroimaging process Schölkopf et al. (2012); Castro et al. (2020). Across sites, biological characteristics shape disease manifestation, while measurement conditions affect its observation, as illustrated in Fig. 1. Building on Pearl's transportability theory Pearl & Bareinboim (2022); Bareinboim & Pearl (2016), we introduce *Causal Mechanism Transport* (CMT), which reframes cross-site

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.

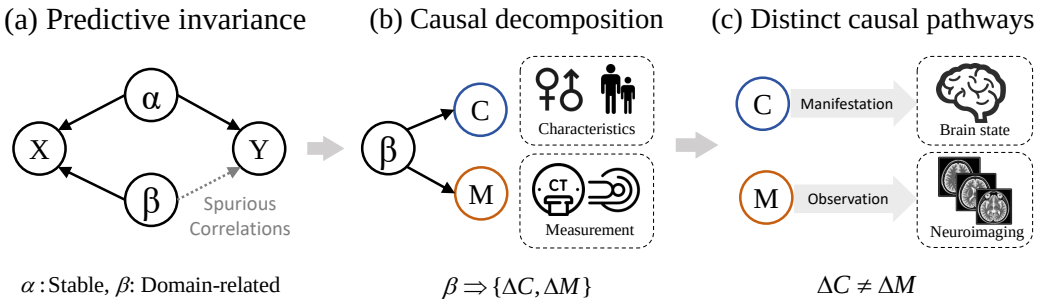


Figure 1: **Causal Decomposition of Site Variation.** (a) A predictive-invariance view separates stable factors α from domain-related factors β. (b) We further decompose domain-related variation into biological context C and measurement conditions M. (c) The two components enter distinct causal pathways, shaping disease manifestation and neuroimaging observation, respectively.

generalization as the transport of disease-related mechanisms and yields an observable criterion for transportable representations. We further develop *Causal Information Contrast* (CIC), a differentiable and architecture-agnostic objective that measures residual site dependence conditional on disease status and biological context.

Our main contributions are summarized as follows:

- CMT reconceptualizes cross-site generalization by decomposing site heterogeneity into biological and measurement pathways and shifting the target from predictive invariance to the transport of disease-related mechanisms.
- CIC establishes a context-conditioned view of representation invariance that distinguishes residual site dependence from variation explained by disease status and biological context.
- We provide mechanism-oriented evidence through controlled causal shifts, real-world generalization settings, and independent representation audits across five neuroimaging cohorts, demonstrating improved generalization and reduced conditional site dependence.

2 From Predictive Invariance to Causal Mechanism Transport

2.1 Predictive Invariance and Transportability

A broad class of domain generalization methods seeks predictive regularities that remain stable across source environments through invariant prediction, distribution alignment, adversarial learning, or site-robust representation learning Arjovsky et al. (2019); Ganin et al. (2016); Yoon et al. (2024); Xu et al. (2025). Let $\mathcal{S}_{\text{src}}$ denote the source sites and $R = \phi_\theta(X)$ a learned representation.

A predictive relation is source-invariant if $p_s(y \mid r) = p_{s'}(y \mid r)$ for all $s, s' \in \mathcal{S}_{\text{src}}$.

Proposition 2.1 (Source predictive invariance does not imply transportability)**.** *Source predictive invariance alone does not determine predictive stability in an unseen environment. In general,*

$$p_s(y \mid r) = p_{s'}(y \mid r), \quad \forall s, s' \in \mathcal{S}_{\text{src}} \;\not\Rightarrow\; p_t(y \mid r) = p_s(y \mid r), \tag{1}$$

for an unseen environment $t \notin \mathcal{S}_{\text{src}}$, without additional assumptions on how the target environment may differ from the observed sources.

Distinct generative processes may agree across the observed source environments while inducing different relations under an unseen target shift Rosenfeld et al. (2020); Subbaswamy et al. (2019); Jalaldoust & Bareinboim (2024). A constructive example is provided in Appendix A.2.

Source-invariant formulations characterize how predictive relations vary across environments while leaving the generative origin of domain variation unspecified. In multi-site neuroimaging, such variation may arise from biologically meaningful subject characteristics or from the measurement process. These sources enter different stages of the data-generating process and therefore have different implications for transportability.

This distinction motivates an explicit structural model of admissible site variation and the mechanisms whose transportability we seek to characterize.

2.2 Causal Generative Structure under Site Shift

To make the structure of admissible site variation explicit, we model multi-site neuroimaging through a causal generative process. Following the anticausal formulation of diagnosis Schölkopf et al. (2012); Castro et al. (2020), disease status gives rise to its neurobiological manifestation, while neuroimaging data arise from a measurement process applied to the underlying brain state.

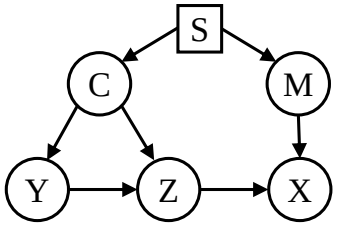


Figure 2: **Selection diagram for multi-site neuroimaging.** Site S indexes shifts in biological context C and measurement conditions M, which affect disease manifestation and neuroimaging observation through distinct pathways.

Definition 2.1 (Multi-site causal generative model)**.** Let $S \in \mathcal{S}$ denote the imaging site, C subject-level biological covariates, M measurement conditions, Y disease status, Z the latent brain state, and X the observed neuroimaging data. As illustrated in Fig. 2, the core disease-to-observation chain follows $Y \to Z \to X$. Site variation enters through the biological pathway $S \to C \to \{Y, Z\}$ and the measurement pathway $S \to M \to X$.

Writing $p_s(\cdot) = p(\cdot \mid S = s)$, we model the site-specific distribution as

$$p_s(c, m, y, z, x) = \underbrace{p_s(c, m)}_{\text{site-varying context}} \underbrace{p(y \mid c)\, p(z \mid y, c)\, p(x \mid z, m)}_{\text{shared causal mechanisms}}. \tag{2}$$

Figure 2 uses a coarse-grained representation of the site-varying context, allowing arbitrary within-site dependence between C and M; accordingly, $p_s(c, m)$ is left unrestricted. The downstream mechanisms $p(y \mid c)$, $p(z \mid y, c)$, and $p(x \mid z, m)$ are shared across sites once their direct causes are specified. Here, $p(z \mid y, c)$ characterizes disease-related neurobiological manifestation and $p(x \mid z, m)$ the measurement process.

Stable mechanisms can induce site-dependent observations. Conditioning on Y and C gives

$$p_s(x \mid y, c) = \iint p(x \mid z, m)\, p(z \mid y, c)\, p_s(m \mid c)\, \mathrm{d}z\, \mathrm{d}m. \tag{3}$$

Thus, changes in $p_s(m \mid c)$ can alter the class-conditional imaging distribution even when the underlying mechanisms remain unchanged, while variation in $p_s(c)$ changes the biological population over which these mechanisms are expressed.

The model therefore separates *mechanism invariance* from *observational invariance*: cross-site differences may arise from changes in the inputs to shared mechanisms rather than changes in the mechanisms themselves. This causal decomposition specifies the admissible site shifts under which we define mechanism transport.

2.3 Causal Mechanism Transport

Building on causal transportability Pearl & Bareinboim (2022); Bareinboim & Pearl (2016), we formalize the CMT transport target through the representation channel induced by the causal process.

Let $R = \phi_\theta(X)$. Under Fig. 2, conditioning on C blocks population-dependent paths from S to Z, while Y specifies the disease state whose manifestation should be preserved. Residual site dependence within a fixed (Y, C) context therefore propagates through $S \to M \to X \to R$.

We characterize this effect through the site-induced latent-state-to-representation channel

$$q_{\theta,s}(r \mid z, c) = \iint p_\theta(r \mid x)\, p(x \mid z, m)\, p_s(m \mid c)\, \mathrm{d}x\, \mathrm{d}m. \tag{4}$$

Definition 2.2 (Causal Mechanism Transport)**.** An encoder ϕ_θ satisfies *Causal Mechanism Transport* if, for environments compatible with the causal model,

$$q_{\theta,s}(r \mid z, c) = q_{\theta,s'}(r \mid z, c) = q_\theta(r \mid z, c), \qquad \forall s, s' \in \mathcal{S}, \tag{5}$$

for a fixed latent brain state and biological context.

CMT allows the observed neuroimaging distribution to vary across environments while requiring the induced $Z \to R$ channel to remain stable under admissible measurement shifts. Together with the shared disease-related mechanism $p_s(z \mid y, c) = p(z \mid y, c)$, this gives the following observable consequence.

Theorem 2.1 (Conditional transportability). *Under Definition 2.1, if ϕ_θ satisfies CMT, then*

$$p_s(r \mid y, c) = \int q_{\theta,s}(r \mid z, c) p_s(z \mid y, c)\, \mathrm{d}z = \int q_\theta(r \mid z, c) p(z \mid y, c)\, \mathrm{d}z = p(r \mid y, c), \tag{6}$$

for all $s \in \mathcal{S}$ and (y, c) in their common support. Hence,

$$R \perp S \mid Y, C, \qquad I(R; S \mid Y, C) = 0. \tag{7}$$

Theorem 2.1 localizes site stability to comparable disease–biological contexts, thereby preserving variation across distinct (Y, C) contexts.

Equation 4 defines the transported representation channel after marginalizing M over the site-specific distribution $p_s(m \mid c)$. Here, S indexes the environment-specific distribution over measurement conditions, and the resulting channel integrates over the corresponding measurement heterogeneity. Equation 7 therefore constrains the measurement-marginalized representation law. A formal derivation is provided in Appendix A.4.

To retain diagnostic information while targeting the observable consequence of CMT, we formulate the learning objective as

$$\min_{\theta,h} \mathcal{R}_{\mathrm{src}}(\theta, h) \quad \text{s.t.} \quad I(\phi_\theta(X); S \mid Y, C) = 0, \tag{8}$$

where the source-domain predictive risk couples transportability with diagnostic utility, while the constraint imposes the observable representation-level criterion induced by CMT.

3 Methodology

Given source sites $\mathcal{S}_{\mathrm{src}}$, we consider $\mathcal{D}_{\mathrm{src}} = \{(x_i, y_i, c_i, s_i)\}_{i=1}^N$, where x_i is a subject-level brain network, y_i the disease label, c_i biological covariates, and s_i the acquisition site. An encoder ϕ_θ maps x_i to $r_i = \phi_\theta(x_i)$, from which classifier h predicts y_i. The goal is to generalize to an unseen site $t \notin \mathcal{S}_{\mathrm{src}}$ without target data during training; C and S are used only for training, while inference requires X alone.

3.1 Causal Information Contrast

CMT yields the observable criterion $R \perp S \mid Y, C$, where $R = \phi_\theta(X)$. Let $U = (Y, C)$ denote the disease–biological context. The criterion can be written as

$$p_\theta(r \mid u, s) = p_\theta(r \mid u) \quad \text{a.e. } (u, s). \tag{9}$$

Equation 9 suggests contrasting two nested conditional representation laws. The distribution $p_\theta(r \mid u)$ characterizes representation variation within a disease–biological context, while $p_\theta(r \mid u, s)$ additionally resolves that variation by imaging site. Their discrepancy therefore quantifies residual site dependence after the disease–biological context is specified. Under the causal model in Section 2.2, this residual dependence reflects the site-dependent measurement pathway.

To operationalize this contrast, we use conditional Cauchy-Schwarz (CS) divergence Yu et al. (2025), which is non-negative and vanishes when the two conditional distributions coincide. We define CIC as

$$\mathcal{C}_{\mathrm{CIC}}(\theta) = D_{\mathrm{CCS}}\left(p_\theta(r \mid u), p_\theta(r \mid u, s)\right). \tag{10}$$

By the faithfulness property of conditional CS divergence,

$$\mathcal{C}_{\mathrm{CIC}}(\theta) = 0 \iff p_\theta(r \mid y, c, s) = p_\theta(r \mid y, c) \iff R \perp S \mid Y, C. \tag{11}$$

CIC therefore shares the population zero set of the observable criterion $I(R; S \mid Y, C) = 0$ induced by CMT and provides a differentiable distributional surrogate that can be estimated directly from samples. Its causal interpretation follows from the SCM-derived conditioning structure (Y, C), which localizes site comparison to comparable disease–biological contexts and preserves variation across distinct (Y, C) contexts.

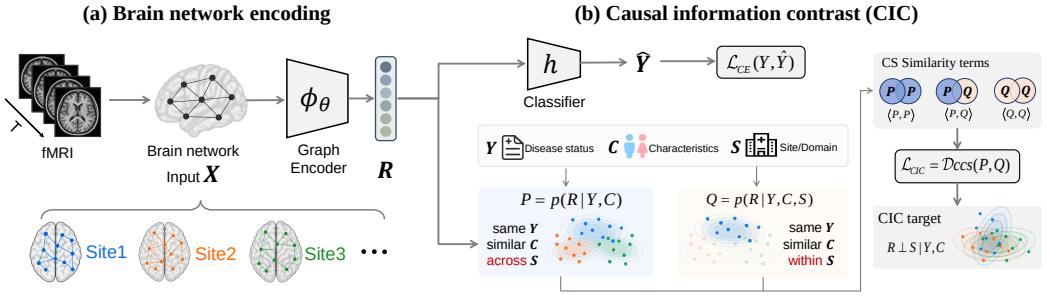


Figure 3: **Overview of Causal Information Contrast.** (a) The graph encoder ϕ_θ maps the fMRI-derived brain network X to representation R, which is used by classifier h for disease prediction. (b) CIC contrasts $P = p_\theta(R \mid Y, C)$ and $Q = p_\theta(R \mid Y, C, S)$ using conditional Cauchy-Schwarz divergence, measuring residual site dependence within comparable disease–biological contexts.

3.2 Conditional Cauchy-Schwarz Estimation

For each sample in $\mathcal{D}_{\text{src}}$, let $u_i = (y_i, c_i)$ and $r_i = \phi_\theta(x_i)$. We estimate the two conditional representation laws using kernel conditional density estimation:

$$\widehat{p}_\theta(r \mid u) = \frac{\sum_i k_R(r, r_i) k_U(u, u_i)}{\sum_i k_U(u, u_i)}, \qquad \widehat{p}_\theta(r \mid u, s) = \frac{\sum_i k_R(r, r_i) k_U(u, u_i) k_S(s, s_i)}{\sum_i k_U(u, u_i) k_S(s, s_i)}. \tag{12}$$

The first estimator aggregates samples with similar disease–biological context, whereas the second additionally restricts the conditional support by site.

For compactness, let $K^U_{ji} = k_U(u_j, u_i)$, $K^{US}_{ji} = k_U(u_j, u_i) k_S(s_j, s_i)$, and $L^R_{ji} = k_R(r_j, r_i)$, with conditional support masses $d^U_j = \sum_i K^U_{ji}$ and $d^{US}_j = \sum_i K^{US}_{ji}$. The conditional CS estimator reduces to three similarity terms:

$$A_U = \sum_j \frac{\sum_i K^U_{ji} L^R_{ji}}{(d^U_j)^2}, \qquad A_{US} = \sum_j \frac{\sum_i K^{US}_{ji} L^R_{ji}}{(d^{US}_j)^2}, A_{U,US} = \sum_j \frac{\sum_i K^U_{ji} L^R_{ji}}{d^U_j d^{US}_j}. \tag{13}$$

Here, A_U and A_{US} characterize conditional self-similarity before and after site information is introduced, respectively, whereas $A_{U,US}$ characterizes their cross-similarity. The empirical Causal Information Contrast is then

$$\widehat{\mathcal{C}}_{\text{CIC}} = \log A_U + \log A_{US} - 2 \log A_{U,US}. \tag{14}$$

Equation 14 retains the self–self–cross structure of conditional CS divergence illustrated in Fig. 3(b), while remaining differentiable with respect to the learned representation R. It corresponds to the conditional-independence specialization of the kernel conditional CS estimator Yu et al. (2025).

We use a product kernel over disease status and biological covariates for k_U, a discrete kernel for site identity k_S, and a Gaussian kernel in representation space for k_R. To improve the conditional support of site-specific neighborhoods in multi-site mini-batches, density estimation is performed against a detached feature bank.

The empirical CIC provides a differentiable surrogate for the observable criterion induced by CMT and is optimized jointly with the source predictive objective, with λ controlling the trade-off between predictive utility and mechanism transport. Further implementation details and sensitivity analyses of key CIC hyperparameters are provided in Appendix D.

4 Experiments

Our experiments examine the proposed mechanism-transport approach from three complementary perspectives: controlled behavior under structured shifts, generalization across real-world domains, and mechanism-level analysis of the learned representations. The real-world setup is summarized first, with the controlled data generation presented alongside its corresponding validation.

4.1 Experimental Setup

Datasets. We use five resting-state fMRI cohorts drawn from four databases, spanning autism spectrum disorder (ASD), major depressive disorder (MDD), and schizophrenia (SCZ): ABIDE-I Di Martino et al. (2014), ABIDE-II Di Martino et al. (2017), REST-meta-MDD Chen et al. (2022), and the SRPBS database Tanaka et al. (2021). From SRPBS, the SCZ cohort is used for unseen-site evaluation and an independent ASD cohort for cross-dataset transfer. Table 1 summarizes the cohorts after quality control, with detailed construction and site-wise statistics provided in Appendix C. Age and sex are used as the biological covariates C in all real-data experiments.

Table 1: **Summary of used datasets.**

Dataset	Task	Sites	Patient / Healthy Control
ABIDE-I	ASD	8	322 / 342
ABIDE-II	ASD	10	325 / 431
REST-meta-MDD	MDD	21	1,604 / 1,300
SRPBS-SCZ	SCZ	4	140 / 140
SRPBS-ASD	ASD	1	115 / 101

Data preprocessing. All resting-state fMRI data were preprocessed using DPABI Yan et al. (2016), with the detailed preprocessing pipeline provided in Appendix C. The brain was parcellated into 116 regions using the AAL116 atlas Tzourio-Mazoyer et al. (2002). For each subject, functional connectivity (FC) was computed using pairwise Pearson correlations between regional BOLD time series followed by Fisher's r-to-z transformation. We retained the strongest 20% of connections according to absolute FC magnitude, with the graph density selected by grid search on source-domain validation data. The FC profile of each region was used as its node feature.

Baselines. For real-world evaluation, we compare against conventional machine-learning methods (LR and SVM), standard graph neural networks (GCN Kipf & Welling (2016), GAT Veličković et al. (2017), and GIN Xu et al. (2018)), domain-generalization methods (CORAL Sun & Saenko (2016) and IRM Arjovsky et al. (2019)), graph-based out-of-distribution (OOD) methods (DIR-GNN Wu et al. (2022), GSAT Miao et al. (2022), and GMT Baek et al. (2021)), and brain-network-specific methods (BrainNetCNN Kawahara et al. (2017), BrainNetTF Kan et al. (2022), BrainIB Zheng et al. (2024b), CI-GNN Zheng et al. (2024a), BrainOOD Xu et al. (2025), and RM-DRL Li et al. (2025)). To assess its architecture-agnostic applicability, CIC is further integrated with GAT, BrainNetCNN, and BrainNetTF without modifying their inference architectures. Baselines are reproduced following the original papers and publicly available implementations.

Evaluation protocol. For real-world evaluation, we consider two complementary settings: unseen-site generalization and cross-dataset transfer. Detailed protocols are provided in Appendix C.

For unseen-site generalization, we adopt Leave-One-Site-Out (LOSO) evaluation, where each site is held out once as the target domain and the remaining sites are used for training and validation. Each fold is repeated with three random seeds. We report ACC and AUC, averaged across held-out sites, and assess statistical significance using one-sided paired Wilcoxon signed-rank tests.

For cross-dataset transfer, models are trained and selected on the source cohort and evaluated directly on a disjoint target cohort. We consider ABIDE-I→ABIDE-II, ABIDE-II→ABIDE-I, and ABIDE-II→SRPBS-ASD, with three random seeds for each setting. All transfers retain the same ASD classification task while varying the data source.

4.2 Controlled Validation

We first use a controlled setting to test whether the proposed strategy exhibits the transport behavior predicted by CMT when biological and measurement shifts are varied independently. We generate four source environments following the causal structure in Fig. 2: $S \to C$, $S \to M$, $C \to Y$, $(Y, C) \to Z$, and $(Z, M) \to X$. Across environments, only the context distribution $p_s(C, M)$ varies, while the mechanisms $p(Y \mid C)$, $p(Z \mid Y, C)$, and $p(X \mid Z, M)$ remain fixed. The latent brain state contains stable disease-context interactions, making preservation of biological-context variation important for transportable prediction.

We compare the proposed strategy with empirical risk minimization (ERM), IRM Arjovsky et al. (2019), CORAL Sun & Saenko (2016), and marginal MMD Gretton et al. (2012), representing standard source prediction, predictive invariance, distribution alignment, and marginal representation alignment, respectively. We consider two controlled target shifts: a *measurement shift* that perturbs the distribution of M while fixing the biological state, and a *biological-context shift* that alters $p(C \mid S)$ while preserving the measurement process and structural mechanisms.

Table 2: **Unseen-site generalization.** LOSO results on ABIDE-I, ABIDE-II, REST-meta-MDD, and SRPBS-SCZ. Best and second-best results are shown in **bold** and underlined, respectively. Shaded rows denote backbones trained with CIC. indicates significant improvement over the paired backbone under a one-sided Wilcoxon signed-rank test across held-out sites ($p < 0.05$). Significance is not reported for SRPBS-SCZ because it contains only four sites.

Type	Method	ABIDE I		ABIDE II		REST-meta-MDD		SRPBS-SCZ	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
ML	LR	53.69± 7.23	57.70± 11.17	57.82± 7.03	61.50± 7.81	57.15± 5.73	60.30± 5.95	73.21± 6.27	82.40± 3.39
	SVM	53.94± 6.42	58.17± 10.92	56.95± 6.45	59.94± 7.53	57.14± 7.02	59.31± 5.56	70.92± 6.78	82.31± 3.31
GNN	GCN	54.05± 7.36	55.01± 9.71	53.67± 6.24	57.77± 4.10	54.91± 5.90	57.21± 8.22	66.62± 6.99	75.79± 5.97
	GIN	53.83± 6.52	53.64± 9.33	54.27± 9.44	56.09± 8.75	54.84± 5.77	57.43± 6.13	64.01± 13.75	69.19± 15.87
	GAT	56.14± 4.87	56.24± 8.59	55.12± 7.14	58.33± 8.27	55.52± 6.11	57.66± 6.96	64.92± 2.79	72.76± 5.64
	GAT+CIC	58.41± 4.62*	58.91± 7.83	57.68± 6.88*	60.74± 7.91	57.86± 5.74*	60.19± 6.41*	67.35± 3.12	75.48± 5.21
OOD	CORAL	53.61± 6.24	52.25± 6.61	54.52± 10.98	57.35± 9.86	53.38± 5.21	56.66± 7.37	58.57± 3.86	63.97± 2.60
	IRM	51.89± 5.74	51.78± 9.40	53.26± 8.74	56.66± 7.29	54.86± 6.36	55.56± 6.46	56.87± 4.30	60.51± 5.36
Graph OOD	DIR-GNN	51.41± 9.39	51.97± 12.89	55.85± 7.34	56.10± 7.95	53.07± 6.20	55.77± 5.97	62.59± 2.19	67.08± 4.04
	GSAT	50.41± 10.16	49.57± 8.29	52.67± 5.30	56.48± 5.31	53.98± 5.24	57.51± 7.13	63.67± 3.18	71.56± 4.26
	GMT	53.23± 7.71	53.15± 8.11	52.82± 6.15	56.38± 6.96	54.55± 3.72	56.38± 5.98	65.02± 4.78	68.08± 2.50
Brain Network	BrainIB	54.69± 4.83	56.30± 8.82	57.22± 7.61	59.08± 9.07	57.83± 6.78	57.59± 8.35	62.01± 9.10	70.13± 3.26
	CI-GNN	54.31± 8.84	53.21± 7.62	58.21± 7.74	56.45± 7.64	57.32± 6.84	57.34± 8.42	69.53± 9.20	74.37± 4.16
	BrainOOD	55.63± 5.86	53.81± 8.38	56.82± 6.10	58.09± 8.92	55.77± 5.96	59.95± 7.55	61.52± 2.72	65.93± 5.53
	RM-DRL	51.67± 4.86	55.17± 9.02	56.75± 6.41	59.68± 7.56	57.67± 5.93	61.47± 6.03	71.72± 8.34	82.43± 5.20
	BrainNetCNN	55.36± 8.42	58.78± 10.15	59.59± 6.42	63.99± 6.75	58.71± 6.80	62.87± 7.12	73.69± 10.66	81.52± 8.89
	BrainNetCNN+CIC	57.92± 7.86	61.24± 9.47	62.14± 6.05*	66.38± 6.21*	**62.48± 6.33***	65.41± 6.78*	**76.21± 9.84**	84.07± 8.12
	BrainNetTF	54.86± 8.29	58.41± 14.39	59.96± 6.02	64.02± 6.25	58.77± 5.85	63.52± 7.99	69.84± 14.38	82.94± 6.30
	BrainNetTF+CIC	**60.53± 7.71***	**62.77± 12.55**	**62.71± 5.68***	**67.47± 5.94***	61.29± 5.52*	**66.18± 7.41***	74.46± 13.15	**85.39± 5.87**

Under source-like conditions, ERM attains the highest AUC, while its performance decreases as the target measurement distribution moves away from the observed sources, from 0.930 to 0.583 at $\Delta_M = 8$. As shown in Fig. 4(a), the proposed strategy remains more stable under the same shift, reaching an AUC of 0.801, compared with 0.691 for CORAL. This behavior is consistent with Proposition 2.1, which distinguishes source predictive stability from transportability under unseen measurement shifts.

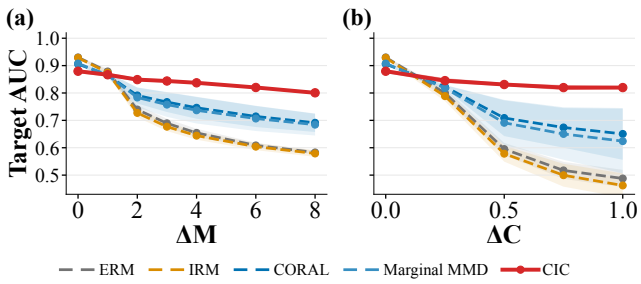


Figure 4: **Controlled validation under admissible shifts.** Target AUC under increasing (a) measurement shift Δ_M and (b) biological-context shift Δ_C. Curves and shading show mean ± standard deviation over three random seeds.

Under the biological-context shift shown in Fig. 4(b), the proposed strategy retains an AUC of 0.820 at $\Delta_C = 1$, exceeding the strongest baseline by 16.9 AUC points. These results support the CMT conditioning structure, where site dependence is assessed within comparable disease–biological contexts while preserving meaningful biological variation. Full experimental details and additional analyses are provided in Appendix B.

4.3 Generalization to Unseen Domains

We next examine whether the proposed mechanism-transport learning strategy improves predictive generalization under real-world domain shifts and whether its benefits extend across heterogeneous backbone architectures.

Unseen-site generalization. Table 2 reports the LOSO performance across the four multi-site brain network datasets. Across all eight dataset–metric combinations, both the best and second-best results are achieved by backbones trained with CIC. This pattern spans ASD, MDD, and SCZ cohorts with different site compositions, showing consistent improvements in unseen-site generalization across heterogeneous clinical datasets.

The paired backbone comparisons further examine the architecture-agnostic applicability of the proposed strategy. Across GAT, BrainNetCNN, and BrainNetTF, training with CIC improves both ACC and AUC on all four datasets, yielding 24/24 metric-wise improvements with average gains

Table 3: **Cross-dataset transfer results.** Models are trained on the source dataset and evaluated directly on a disjoint target cohort. ACC and AUC are reported as mean ± standard deviation (%) over three independent training runs. Best and second-best results in each column are shown in **bold** and underlined, respectively. Shaded rows denote backbones trained with CIC.

Type	Method	ABIDE I → ABIDE II		ABIDE II → ABIDE I		ABIDE II → SRPBS-ASD	
		ACC	AUC	ACC	AUC	ACC	AUC
ML	LR	56.75± 0.00	59.17± 0.00	56.33± 0.00	58.82± 0.00	60.19± 0.00	61.29± 0.00
	SVM	55.16± 0.00	57.87± 0.00	56.93± 0.00	58.07± 0.00	56.94± 0.00	59.46± 0.00
GNN	GCN	46.96± 4.12	52.27± 3.68	53.61± 2.07	55.47± 2.41	56.94± 3.54	61.64± 4.89
	GIN	56.22± 1.63	53.61± 2.18	54.52± 3.71	52.06± 3.24	50.93± 4.46	51.70± 3.92
	GAT	53.17± 2.84	53.28± 3.19	53.16± 1.52	56.48± 1.88	57.87± 2.31	61.13± 2.67
	GAT+CIC	54.63± 2.41	55.41± 2.96	55.84± 1.71	57.21± 1.54	59.41± 2.08	63.88± 2.44
OOD	CORAL	51.98± 3.85	50.79± 4.62	51.96± 2.26	52.45± 1.79	52.31± 4.18	56.76± 3.53
	IRM	52.12± 1.47	51.92± 1.91	54.52± 3.38	53.11± 2.95	52.78± 2.14	56.59± 2.58
Graph OOD	DIR-GNN	53.17± 4.27	52.15± 3.61	53.01± 1.84	52.19± 2.22	51.85± 2.73	51.96± 3.15
	GSAT	48.94± 2.09	54.29± 2.54	53.16± 3.96	53.39± 4.41	47.22± 5.28	52.64± 4.67
	GMT	51.68± 3.44	53.20± 2.89	53.46± 1.36	53.08± 1.62	53.24± 2.51	51.36± 2.18
Brain Network	BrainIB	52.34± 2.62	53.69± 3.07	52.16± 4.08	53.14± 3.55	55.21± 1.58	57.32± 1.93
	CI-GNN	53.31± 1.72	52.24± 2.14	54.32± 2.48	52.33± 2.91	55.82± 3.64	57.62± 3.22
	BrainOOD	55.73± 3.19	55.94± 2.76	53.41± 1.41	54.42± 1.68	57.38± 2.87	57.03± 3.43
	RM-DRL	57.29± 1.76	**60.23± 0.18**	57.03± 0.23	60.61± 1.19	52.93± 3.88	57.51± 5.02
	BrainNetCNN	54.23± 2.93	56.94± 3.48	56.63± 1.86	58.72± 2.24	65.28± 4.31	65.58± 3.76
	BrainNetCNN+CIC	56.71± 2.67	58.47± 3.12	57.94± 1.59	**61.55± 2.05**	**67.96± 3.88**	**66.41± 3.41**
	BrainNetTF	54.89± 1.55	58.02± 1.83	56.17± 3.27	58.41± 2.74	54.17± 4.72	58.54± 5.14
	BrainNetTF+CIC	**57.52± 1.68**	59.18± 1.61	**58.24± 2.91**	60.97± 2.48	56.09± 4.35	61.48± 4.79

of 3.05 and 2.77 percentage points in ACC and AUC, respectively. Among the 18 comparisons with sufficient site-level support for statistical testing, 13 reach significance under the one-sided paired Wilcoxon signed-rank test. The three backbones span a generic message-passing GNN, a brain-network-specific convolutional architecture, and a Transformer-based model, supporting the applicability of CIC across heterogeneous predictive architectures.

Cross-dataset transfer. To examine whether the generalization gains extend beyond site variation within a single dataset, we further consider transfer between independently collected ASD cohorts while keeping the prediction task fixed. As shown in Table 3, backbones trained with CIC improve their counterparts in all three transfer directions across GAT, BrainNetCNN, and BrainNetTF, yielding 18/18 metric-wise improvements with average gains of 2.09 and 1.94 percentage points in ACC and AUC, respectively. Moreover, models trained with CIC achieve the best performance in five of the six dataset-metric combinations. These results show that the proposed strategy remains effective under dataset-level shifts beyond the within-dataset site variations considered in LOSO evaluation.

4.4 Representation Analysis

Independent mechanism audit.

We examine whether CIC reduces residual site information in the learned representation using an independent probe. With BrainNetTF as the backbone, we freeze the trained encoder and extract the representation R before the classification head. For each LOSO fold, we fit two auxiliary site predictors, $g_0 : (Y, C) \to S$ and $g_1 : (Y, C, R) \to S$, on the source-domain training set and evaluate them on the corresponding source-domain validation set, where C contains age and sex. We define the conditional site-prediction gain as $\Delta_{\text{site}} = \text{NLL}(g_0) - \text{NLL}(g_1)$, where a smaller value indicates less additional site-predictive information in R beyond (Y, C). The held-out target site is never used for probe fitting or evaluation, and Δ_{site} is averaged over three random seeds for each LOSO fold. In Fig. 5, light lines denote individual folds, while the thick line and error bars show the median and interquartile range; annotations report the change in median gain from ERM to CIC.

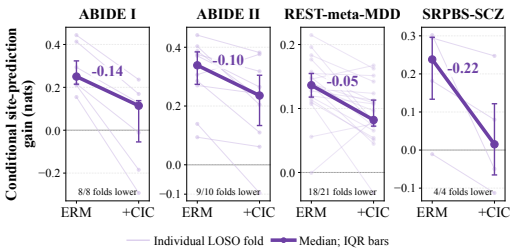


Figure 5: **Conditional site-dependence analysis.** Conditional site-prediction gain for ERM and CIC-trained representations across LOSO folds using BrainNetTF. Lower values indicate less residual site information beyond (Y, C).

CIC reduces conditional site predictability across all four datasets. The gain decreases in 8/8, 9/10, 18/21, and 4/4 LOSO folds on ABIDE-I, ABIDE-II, REST-meta-MDD, and SRPBS-SCZ, respectively, with median changes of -0.14, -0.10, -0.05, and -0.22 nats. These results indicate that CIC representations contain less residual site information after accounting for disease status and biological context, consistent with the observable criterion $R \perp S \mid Y, C$ derived from CMT.

Conditioning ablation.

We examine CIC's conditioning structure by fixing BrainNetTF as the backbone and varying conditioning variables. We compare ERM with unconditional CIC, which suppresses marginal site dependence, Y-conditioned CIC, and the full (Y, C)-conditioned CIC. As shown in Table 4, the full formulation achieves the highest AUC on four datasets, improving over ERM by 4.36, 3.45, 2.66, and 2.45 percentage points on ABIDE-I, ABIDE-II, REST-meta-MDD, and SRPBS-SCZ, respectively. The unconditional and Y-conditioned variants generally improve over ERM, but their gains are less consistent than those of the full formulation. These results support joint conditioning on disease status and biological covariates for reducing site dependence while preserving biologically meaningful variation.

Table 4: **Conditioning ablation with BrainNetTF.** LOSO AUC (%, mean ± standard deviation). MDD and SCZ denote REST-meta-MDD and SRPBS-SCZ.

Method	ABIDE-I	ABIDE-II	MDD	SCZ
ERM	58.41 ± 14.39	64.02 ± 6.25	63.52 ± 7.99	82.94 ± 6.30
Uncond. CIC	60.31 ± 14.16	66.17 ± 7.19	65.84 ± 5.13	82.84 ± 5.23
Y-CIC	59.84 ± 14.05	66.56 ± 8.84	65.39 ± 6.00	84.95 ± 4.25
(Y, C)-CIC	**62.77 ± 12.55**	**67.47 ± 5.94**	**66.18 ± 7.41**	**85.39 ± 5.87**

5 Related Work

Generalizable Brain Network Analysis. Multi-site brain network analysis is vulnerable to distribution shifts caused by population and acquisition heterogeneity. Recent methods improve cross-site generalization through domain-invariant representations, robust substructures, or disentangled connectivity patterns. RM-DRL separates domain-invariant and domain-specific components Li et al. (2025), BrainOOD learns OOD-generalizable subgraphs Xu et al. (2025), while MetaExplainer and CORE identify domain-robust connectivity patterns across sites Qiu et al. (2025); Wang et al. (2026b). Related causal-subgraph approaches further distinguish disease-related from site-specific components Wang et al. (2026a). These methods emphasize robust representations or substructures across observed sites. CMT models the causal origins of cross-site heterogeneity by explicitly distinguishing biological and measurement-induced variation.

Causal Generalization and Transportability. Causal generalization methods exploit mechanisms that remain stable across environments. Invariant prediction and invariant risk minimization seek stable predictive relations Peters et al. (2016); Arjovsky et al. (2019), while stable prediction studies robustness under dataset shift Subbaswamy et al. (2019). Transportability theory provides a causal framework for transferring knowledge across heterogeneous environments Bareinboim & Pearl (2016); Pearl & Bareinboim (2022), and recent work extends this perspective to transportable representations for domain generalization Jalaldoust & Bareinboim (2024). CMT builds on this line by specializing transportability to multi-site brain networks and explicitly separating biological and measurement pathways. CIC then provides a differentiable surrogate for the resulting observable transport criterion using conditional Cauchy-Schwarz divergence Yu et al. (2025).

6 Conclusion

We introduced CMT, a causal framework for generalizable brain network analysis under multi-site distribution shifts. By separating biological context from measurement conditions, CMT characterizes transportability through stable disease-related representation mechanisms and derives an observable conditional criterion. We further developed CIC as a differentiable objective that operationalizes this criterion from multi-site data. Controlled and real-world experiments across five neuroimaging cohorts demonstrate improved unseen-site and cross-dataset generalization across multiple backbones, while representation analyses support reduced conditional site dependence. Current experiments use age and sex as biological covariates and assume a predefined causal decomposition; future work will extend CMT to richer biological contexts, multimodal data, and more flexible conditional-dependence estimators.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and editing, theoretical exposition, experimental design discussions, result interpretation, literature organization, and preparation of figures and code. All AI-assisted content was reviewed by the authors, and mathematical derivations, code, experiments, and reported results were independently verified by the authors. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

This work performs secondary analyses of existing resting-state fMRI datasets, including ABIDE, REST-meta-MDD, and SRPBS, and involves no new human-subject data collection. Ethical approvals and consent procedures for the original cohorts are described in the corresponding dataset publications. The study is methodological and intended for research use; clinical application would require further prospective validation.

REPRODUCIBILITY STATEMENT

Reproducibility materials are provided throughout the appendix: theoretical assumptions, derivations, and proofs in Appendix A; controlled data generation and analyses in Appendix B; dataset construction, preprocessing, and evaluation protocols in Appendix C; and implementation details, hyperparameter settings, sensitivity analyses, and computational evaluation in Appendix D. Anonymous source code is available at `https://anonymous.4open.science/r/CIC/`.

REFERENCES

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

Jinheon Baek, Minki Kang, and Sung Ju Hwang. Accurate learning of graph representations with graph multiset pooling. *arXiv preprint arXiv:2102.11533*, 2021.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

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

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

Xiao Chen, Bin Lu, Hui-Xian Li, Xue-Ying Li, Yu-Wei Wang, Francisco Xavier Castellanos, Li-Ping Cao, Ning-Xuan Chen, Wei Chen, Yu-Qi Cheng, et al. The direct consortium and the rest-meta-mdd project: towards neuroimaging biomarkers of major depressive disorder. *Psychoradiology*, 2(1):32–42, 2022.

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.

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

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Ben Glocker, Robert Robinson, Daniel C Castro, Qi Dou, and Ender Konukoglu. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *arXiv preprint arXiv:1910.04597*, 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

Kasra Jalaldoust and Elias Bareinboim. Transportable representations for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12790–12800, 2024.

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

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

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

Wenyang Li, Mingliang Wang, Mingxia Liu, and Qingshan Liu. Riemannian manifold-based disentangled representation learning for multi-site functional connectivity analysis. *Neural Networks*, 183:106945, 2025.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International conference on machine learning*, pp. 15524–15543. PMLR, 2022.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. In *Probabilistic and causal inference: The works of Judea Pearl*, pp. 451–482. 2022.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

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.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3118–3127. PMLR, 2019.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

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.

Xiaotong Wang, Yibin Tang, Yuan Gao, Aimin Jiang, and Ying Chen. Causal subgraph disentanglement network for out-of-distribution generalization in brain network analysis. *Knowledge-Based Systems*, pp. 116450, 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.

Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. *arXiv preprint arXiv:2201.12872*, 2022.

Jiaxing Xu, Yongqiang Chen, Xia Dong, Mengcheng Lan, Tiancheng Huang, Qingtian Bian, James Cheng, and Yiping Ke. Brainood: Out-of-distribution generalizable brain network analysis. In *International Conference on Learning Representations*, volume 2025, pp. 58578–58592, 2025.

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

Chao-Gan Yan, Xin-Di Wang, Xi-Nian Zuo, and Yu-Feng Zang. Dpabi: data processing & analysis for (resting-state) brain imaging. *Neuroinformatics*, 14(3):339–351, 2016.

Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A Mazurowski, and Heung-Il Suk. Domain generalization for medical image analysis: A review. *Proceedings of the IEEE*, 112(10):1583–1609, 2024.

Shujian Yu, Hongming Li, Sigurd Løkse, Robert Jenssen, and José C Príncipe. The conditional cauchy-schwarz divergence with applications to time-series data and sequential decision making. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5901–5917, 2025.

Kaizhong Zheng, Shujian Yu, and Badong Chen. Ci-gnn: A granger causality-inspired graph neural network for interpretable brain network-based psychiatric diagnosis. *Neural Networks*, 172: 106147, 2024a.

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

A Theory and Proofs

This appendix section provides the full theoretical development underlying Sections 2.2–3.2. We first formalize the family of admissible environments induced by the multi-site causal model, then provide a constructive counterexample to source predictive invariance, prove the conditional transportability theorem, establish the exact marginalization of measurement heterogeneity through the environment index, and finally derive the theoretical properties and empirical form of Causal Information Contrast (CIC).

A.1 Structural Assumptions and Admissible Environments

Recall the multi-site causal generative model in Definition 2.1. For each environment $s \in \mathcal{S}$, the site-indexed distribution is assumed to admit the factorization

$$p_s(c, m, y, z, x) = p_s(c, m)\, p(y \mid c)\, p(z \mid y, c)\, p(x \mid z, m). \tag{15}$$

Here $p_s(\cdot) = p(\cdot \mid S = s)$. Equation 15 separates the site-varying context distribution $p_s(c, m)$ from three shared downstream mechanisms:

$$p(y \mid c), \qquad p(z \mid y, c), \qquad p(x \mid z, m). \tag{16}$$

We define the admissible environment family associated with the causal model as

$$\mathfrak{E}_{\mathcal{G}} = \{p_s : p_s(c, m, y, z, x) = p_s(c, m)\, p(y \mid c)\, p(z \mid y, c)\, p(x \mid z, m)\}\,. \tag{17}$$

Environments in $\mathfrak{E}_{\mathcal{G}}$ may differ arbitrarily in their biological and measurement context distribution $p_s(c, m)$, while sharing the causal mechanisms in Eq. 16. Thus, the theory does not assume arbitrary target shifts; transportability is defined with respect to the structured environment family specified by the selection diagram.

Coarse-grained selection semantics. The selection diagram in Fig. 2 specifies where environmental variation enters the generative process. It is deliberately coarse-grained with respect to the internal dependence structure of C and M. We therefore retain the unrestricted joint context

$$p_s(c, m), \tag{18}$$

rather than imposing

$$p_s(c, m) = p_s(c)p_s(m). \tag{19}$$

Possible within-site dependence between biological and measurement contexts, including dependence induced by unmodeled recruitment or acquisition factors, is absorbed into $p_s(c, m)$. The structural assumption relevant to CMT is that downstream variation propagates through C and M, while the mechanisms in Eq. 16 remain invariant.

Marginalizing Eq. 15 gives

$$p_s(y, c) = p_s(c)\, p(y \mid c), \qquad p_s(c) = \int p_s(c, m)\, \mathrm{d}m. \tag{20}$$

Consequently, for $p_s(y, c) > 0$,

$$p_s(m \mid y, c) = \frac{p_s(c, m)p(y \mid c)}{p_s(c)p(y \mid c)} = p_s(m \mid c), \tag{21}$$

and

$$p_s(z \mid y, c) = p(z \mid y, c). \tag{22}$$

Equations 21 and 22 formalize the decomposition central to our theory. After disease status and biological context are fixed, the latent disease-related brain mechanism is shared across environments, while the measurement distribution may remain environment dependent.

The corresponding conditional imaging distribution is

$$p_s(x \mid y, c) = \iint p(x \mid z, m)\, p(z \mid y, c)\, p_s(m \mid c)\, \mathrm{d}z\, \mathrm{d}m, \tag{23}$$

recovering Eq. 3. Hence, observational shifts can arise from changes in $p_s(m \mid c)$ even when all downstream generative mechanisms are unchanged.

Shared representation mechanism. The encoder is shared across environments. For a potentially stochastic encoder, we denote its conditional law by $p_\theta(r \mid x)$. For the deterministic encoder used in our implementation, $R = \phi_\theta(X)$, this distribution reduces to a point mass at $\phi_\theta(x)$. Therefore, the encoder itself introduces no explicit dependence on S; any site dependence in R arises through the site-dependent distribution of its input X.

Common support. Conditional transportability is evaluated on contexts shared by the compared environments. We define

$$\Omega_{YC}^{\mathrm{com}} = \bigcap_{s \in \mathcal{S}} \operatorname{supp} p_s(Y, C). \tag{24}$$

All statements involving $p_s(r \mid y, c)$ are understood for $(y, c) \in \Omega_{YC}^{\mathrm{com}}$. This condition requires overlap of the relevant disease–biological contexts but does not require identical marginal distributions $p_s(y, c)$.

A.2 Counterexample to Source Predictive Invariance

We provide a constructive example for Proposition 2.1, showing that perfect source predictive invariance can coexist with failure under an admissible unseen measurement shift.

Let $C = c_0$ be fixed and consider binary variables

$$Y, Z, M, X, R \in \{0, 1\}. \tag{25}$$

Let

$$Y \sim \mathrm{Bernoulli}\left(\frac{1}{2}\right), \qquad Z = Y, \tag{26}$$

so that the disease-related biological mechanism is perfectly stable. Let the measurement mechanism be

$$X = Z \oplus M, \tag{27}$$

where $\oplus$ denotes exclusive OR, and let the learned representation be

$$R = X. \tag{28}$$

Consider an arbitrary finite set of source environments $\mathcal{S}_{\mathrm{src}}$ satisfying

$$p_s(M = 0) = 1, \qquad \forall s \in \mathcal{S}_{\mathrm{src}}. \tag{29}$$

For every source site,

$$R = X = Z = Y, \tag{30}$$

and hence

$$p_s(Y = 1 \mid R = 1) = 1, \qquad p_s(Y = 1 \mid R = 0) = 0. \tag{31}$$

Therefore,

$$p_s(y \mid r) = p_{s'}(y \mid r), \qquad \forall s, s' \in \mathcal{S}_{\mathrm{src}}. \tag{32}$$

Now consider an unseen target environment t with

$$p_t(M = 1) = 1. \tag{33}$$

The structural mechanisms in Eqs. 26 and 27 remain unchanged; only the distribution of the measurement context changes. In the target,

$$R = X = Z \oplus 1 = 1 - Y, \tag{34}$$

and therefore

$$p_t(Y = 1 \mid R = 1) = 0, \qquad p_t(Y = 1 \mid R = 0) = 1. \tag{35}$$

Thus,

$$p_t(y \mid r) \neq p_s(y \mid r), \qquad s \in \mathcal{S}_{\mathrm{src}}, \tag{36}$$

despite perfect predictive invariance across every observed source environment.

The construction belongs to the admissible environment family $\mathfrak{E}_{\mathcal{G}}$: the mechanisms $p(z \mid y, c)$ and $p(x \mid z, m)$ are unchanged, whereas only $p_s(m \mid c)$ varies. Source-domain predictive stability therefore does not determine whether the predictive relation is generated by a mechanism that transports under an unseen admissible measurement shift.

A.3 Proof of Conditional Transportability

We now prove Theorem 2.1. For any $s \in \mathcal{S}$ and $(y, c) \in \Omega_{YC}^{\text{com}}$,

$$p_s(r \mid y, c) = \iiint p_\theta(r \mid x)\, p_s(x, z, m \mid y, c)\, \mathrm{d}x\, \mathrm{d}m\, \mathrm{d}z. \tag{37}$$

From Eqs. 21–22,

$$p_s(x, z, m \mid y, c) = p(x \mid z, m)\, p(z \mid y, c)\, p_s(m \mid c). \tag{38}$$

Substitution into Eq. 37 gives

$$p_s(r \mid y, c) = \iiint p_\theta(r \mid x)\, p(x \mid z, m)\, p_s(m \mid c)\, p(z \mid y, c)\, \mathrm{d}x\, \mathrm{d}m\, \mathrm{d}z. \tag{39}$$

Rearranging the integrals,

$$p_s(r \mid y, c) = \int \left[\iint p_\theta(r \mid x)\, p(x \mid z, m)\, p_s(m \mid c)\, \mathrm{d}x\, \mathrm{d}m \right] p(z \mid y, c)\, \mathrm{d}z. \tag{40}$$

The bracketed term is precisely the induced channel in Eq. 4:

$$q_{\theta,s}(r \mid z, c) = \iint p_\theta(r \mid x)\, p(x \mid z, m)\, p_s(m \mid c)\, \mathrm{d}x\, \mathrm{d}m. \tag{41}$$

Hence,

$$p_s(r \mid y, c) = \int q_{\theta,s}(r \mid z, c)\, p(z \mid y, c)\, \mathrm{d}z. \tag{42}$$

If ϕ_θ satisfies CMT, Definition 2.2 gives

$$q_{\theta,s}(r \mid z, c) = q_\theta(r \mid z, c), \qquad \forall s \in \mathcal{S}. \tag{43}$$

Therefore,

$$p_s(r \mid y, c) = \int q_\theta(r \mid z, c)\, p(z \mid y, c)\, \mathrm{d}z. \tag{44}$$

The right-hand side no longer depends on s. Defining

$$p(r \mid y, c) := \int q_\theta(r \mid z, c)\, p(z \mid y, c)\, \mathrm{d}z, \tag{45}$$

we obtain

$$p_s(r \mid y, c) = p(r \mid y, c), \qquad \forall s \in \mathcal{S}, \quad (y, c) \in \Omega_{YC}^{\text{com}}. \tag{46}$$

Equivalently,

$$p(r \mid y, c, s) = p(r \mid y, c), \tag{47}$$

which establishes

$$R \perp S \mid Y, C. \tag{48}$$

The corresponding conditional mutual information satisfies

$$I(R; S \mid Y, C) = \mathbb{E}_{Y,C}\left[D_{\text{KL}}\left(p(r, s \mid Y, C) \| p(r \mid Y, C) p(s \mid Y, C) \right) \right] = 0. \tag{49}$$

This proves Theorem 2.1.

Direction of the implication. The result establishes

$$\text{CMT} \quad \Longrightarrow \quad R \perp S \mid Y, C. \tag{50}$$

The reverse implication is not required by our theory: equality of the Z-marginalized conditional representation laws can occur even if the finer channels $q_{\theta,s}(r \mid z, c)$ differ. Accordingly, $R \perp S \mid Y, C$ is the observable representation-level consequence used for learning, whereas CMT defines the stronger latent mechanism-level transport principle.

A.4 Exact Marginalization of Measurement Heterogeneity

A central property of CMT is that measurement conditions are marginalized by construction. The theory does not require selecting or explicitly conditioning on a finite collection of observed acquisition variables. Instead, the complete measurement state M enters the representation channel through its environment-indexed distribution, and CMT transports the resulting marginalized channel.

Define the measurement-specific representation kernel

$$g_\theta(r \mid z, m) = \int p_\theta(r \mid x)\, p(x \mid z, m)\, \mathrm{d}x. \tag{51}$$

The kernel $g_\theta(r \mid z, m)$ describes the representation induced by a fixed latent brain state z under a fixed measurement state m.

Proposition A.1 (Environment-indexed marginalization of measurement heterogeneity)**.** *Under the multi-site causal model, the induced representation channel in environment s satisfies*

$$q_{\theta,s}(r \mid z, c) = \int g_\theta(r \mid z, m)\, p(m \mid c, s)\, \mathrm{d}m = \int g_\theta(r \mid z, m)\, p_s(m \mid c)\, \mathrm{d}m. \tag{52}$$

Therefore, conditional on Z and C, all environment-dependent measurement variation affecting R enters exclusively through the mixing distribution $p(M \mid C, S = s)$. CMT is equivalently the requirement

$$\int g_\theta(r \mid z, m)\, p(m \mid c, s)\, \mathrm{d}m = \int g_\theta(r \mid z, m)\, p(m \mid c, s')\, \mathrm{d}m \tag{53}$$

for all admissible s, s', z, and c.

Proof. Substituting Eq. 51 into the right-hand side of Eq. 52 gives

$$\int \left[\int p_\theta(r \mid x) p(x \mid z, m)\, \mathrm{d}x \right] p_s(m \mid c)\, \mathrm{d}m = \iint p_\theta(r \mid x) p(x \mid z, m) p_s(m \mid c)\, \mathrm{d}x\, \mathrm{d}m, \tag{54}$$

which is exactly $q_{\theta,s}(r \mid z, c)$ in Eq. 4. Equality of these marginalized channels across s is therefore exactly Definition 2.2. No approximation is introduced.

Proposition A.1 gives the key distinction between CMT and explicit covariate adjustment. The transport object is not the representation conditional on a particular realized measurement vector m. It is the representation channel after integrating over the complete measurement distribution associated with an environment:

$$p(M \mid C, S = s). \tag{55}$$

Consequently, explicit observation of M is not required to define CMT: the complete measurement mechanism is already marginalized in the transport target.

Corollary A.1 (Observable environment-level transport criterion)**.** *Under Proposition A.1 and the shared biological mechanism $p_s(z \mid y, c) = p(z \mid y, c)$, CMT implies*

$$p_\theta(r \mid y, c, s) = p_\theta(r \mid y, c) \tag{56}$$

for all admissible environments and all $(y, c) \in \Omega^{\mathrm{com}}_{YC}$. Hence,

$$R \perp S \mid Y, C. \tag{57}$$

Thus, the observable representation-level criterion induced by environment-level transportability can be targeted through (R, Y, C, S) without explicitly observing M.

The proof follows directly by integrating Eq. 52 over the shared brain-state mechanism:

$$p_\theta(r \mid y, c, s) = \iint g_\theta(r \mid z, m)\, p(z \mid y, c)\, p(m \mid c, s)\, \mathrm{d}z\, \mathrm{d}m = \int q_{\theta,s}(r \mid z, c)\, p(z \mid y, c)\, \mathrm{d}z. \tag{58}$$

Under CMT, $q_{\theta,s} = q_\theta$, making the right-hand side independent of s.

Importantly, the use of S is therefore not a proxy approximation to M. S indexes the environment-specific distribution over the complete measurement state, while M is integrated out exactly in Eqs. 52 and 58. The observable criterion $R \perp S \mid Y, C$ is the representation-level consequence of this exact marginalization.

Relation to recorded measurement metadata. Suppose that an observed metadata vector

$$\widetilde{M} = \psi(M) \tag{59}$$

contains recorded acquisition characteristics such as scanner vendor, repetition time, echo time, or motion-related measures. Explicitly conditioning on $\widetilde{M}$ would define a different objective involving

$$p_\theta(r \mid y, c, \widetilde{m}, s), \tag{60}$$

whose validity depends on the particular recorded components included in $\widetilde{M}$. Unless $\widetilde{M}$ completely specifies the full measurement state relevant to X, such adjustment does not represent the entire measurement mechanism M.

CMT instead integrates over the complete M and compares the environment-induced mixtures directly through S. Therefore, recorded measurement metadata are not required by the CMT transport criterion. They may be used for diagnostic analyses or additional characterization of measurement shift, but the theoretical learning target remains the environment-level invariance derived above.

This distinction also clarifies the scope of the claim. CMT targets transportability across the family of measurement distributions indexed by admissible environments S. It does not require learning or reconstructing the individual latent measurement state M; the object being transported is the marginalized brain-to-representation channel in Eq. 52.

A.5 Theoretical Properties and Derivation of Causal Information Contrast

Let

$$U = (Y, C). \tag{61}$$

The observable CMT criterion is

$$p_\theta(r \mid u, s) = p_\theta(r \mid u), \tag{62}$$

which motivates comparing the two nested conditional representation laws

$$P_\theta(r \mid u) = p_\theta(r \mid u), \qquad Q_\theta(r \mid u, s) = p_\theta(r \mid u, s). \tag{63}$$

To place both laws over the same conditioning space, let $V = (U, S)$ and define

$$\widetilde{P}_\theta(r \mid v) = p_\theta(r \mid u), \qquad Q_\theta(r \mid v) = p_\theta(r \mid u, s). \tag{64}$$

The conditional Cauchy–Schwarz divergence Yu et al. (2025) is

$$\begin{aligned} D_{\mathrm{CCS}}(\widetilde{P}_\theta, Q_\theta) = -2 \log \left(\iint \widetilde{P}_\theta(r \mid v) Q_\theta(r \mid v) \, \mathrm{d}r \, \mathrm{d}v \right) \\ + \log \left(\iint \widetilde{P}_\theta^2(r \mid v) \, \mathrm{d}r \, \mathrm{d}v \right) + \log \left(\iint Q_\theta^2(r \mid v) \, \mathrm{d}r \, \mathrm{d}v \right). \end{aligned} \tag{65}$$

For discrete conditioning variables, the corresponding integrals are interpreted with respect to the appropriate counting measure.

Proposition A.2 (Faithfulness of CIC)**.** *Assume that the relevant conditional distributions share common support. Then*

$$\mathcal{C}_{\mathrm{CIC}}(\theta) = D_{\mathrm{CCS}}(\widetilde{P}_\theta, Q_\theta) \geq 0, \tag{66}$$

and

$$\mathcal{C}_{\mathrm{CIC}}(\theta) = 0 \iff p_\theta(r \mid u, s) = p_\theta(r \mid u) \iff R \perp S \mid Y, C. \tag{67}$$

Proof. Faithfulness of conditional CS divergence gives

$$D_{\mathrm{CCS}}(\widetilde{P}_\theta, Q_\theta) = 0 \iff \widetilde{P}_\theta(r \mid v) = Q_\theta(r \mid v) \quad \text{a.e.} \tag{68}$$

Using Eq. 64,

$$\widetilde{P}_\theta(r \mid v) = Q_\theta(r \mid v) \iff p_\theta(r \mid u) = p_\theta(r \mid u, s), \tag{69}$$

which is exactly the definition of $R \perp S \mid U$, with $U = (Y, C)$.

Whenever conditional mutual information is well defined,

$$R \perp S \mid Y, C \quad \Longleftrightarrow \quad I(R; S \mid Y, C) = 0. \tag{70}$$

Therefore,

$$\mathcal{C}_{\mathrm{CIC}}(\theta) = 0 \quad \Longleftrightarrow \quad I(R; S \mid Y, C) = 0. \tag{71}$$

This is an equivalence of population zero sets; CIC is not numerically identical to conditional mutual information and is not used as an estimator of its value.

Combining Theorem 2.1 with Proposition A.2 yields the theoretical hierarchy

$$\mathrm{CMT} \quad \Longrightarrow \quad R \perp S \mid Y, C \quad \Longleftrightarrow \quad \mathcal{C}_{\mathrm{CIC}}(\theta) = 0. \tag{72}$$

The first relation is a mechanism-to-observation implication, whereas the second is an exact representation-level equivalence.

A.5.1 Conditional-independence estimator

Consider source samples

$$\{(r_i, u_i, s_i)\}_{i=1}^{N}, \qquad r_i = \phi_\theta(x_i), \qquad u_i = (y_i, c_i). \tag{73}$$

The conditional-independence specialization of conditional CS divergence compares

$$p_\theta(r \mid u) \qquad \text{and} \qquad p_\theta(r \mid u, s). \tag{74}$$

This is obtained from the generic construction

$$p(y \mid x_1) \qquad \text{versus} \qquad p(y \mid x_1, x_2) \tag{75}$$

through the identification

$$x_1 \leftrightarrow U, \qquad x_2 \leftrightarrow S, \qquad y \leftrightarrow R. \tag{76}$$

The two kernel conditional-density estimators are

$$\widehat{p}_\theta(r \mid u) = \frac{\sum_i k_R(r, r_i) k_U(u, u_i)}{\sum_i k_U(u, u_i)}, \qquad \widehat{p}_\theta(r \mid u, s) = \frac{\sum_i k_R(r, r_i) k_U(u, u_i) k_S(s, s_i)}{\sum_i k_U(u, u_i) k_S(s, s_i)}. \tag{77}$$

Define

$$K^U_{ji} = k_U(u_j, u_i), \qquad K^{US}_{ji} = k_U(u_j, u_i) k_S(s_j, s_i), \qquad L^R_{ji} = k_R(r_j, r_i), \tag{78}$$

and

$$d^U_j = \sum_i K^U_{ji}, \qquad d^{US}_j = \sum_i K^{US}_{ji}. \tag{79}$$

Under the conditional-independence specialization of the conditional CS estimator, the first conditional self-similarity term is

$$A_U = \sum_j \frac{\sum_i K^U_{ji} L^R_{ji}}{(d^U_j)^2}, \tag{80}$$

while the site-refined self-similarity term is

$$A_{US} = \sum_j \frac{\sum_i K^{US}_{ji} L^R_{ji}}{(d^{US}_j)^2}. \tag{81}$$

The corresponding cross-similarity term is

$$A_{U,US} = \sum_j \frac{\sum_i K^U_{ji} L^R_{ji}}{d^U_j d^{US}_j}. \tag{82}$$

The appearance of K^U, rather than K^{US}, in the numerator of Eq. 82 follows from the nested conditional structure $p(R \mid U)$ versus $p(R \mid U, S)$: the cross term is anchored by the shared conditioning variables U, whereas the normalization reflects the two respective conditional supports.

The empirical Causal Information Contrast is therefore

$$\widehat{\mathcal{C}}_{\mathrm{CIC}} = \log A_U + \log A_{US} - 2 \log A_{U,US}, \tag{83}$$

recovering Eq. 14 in the main text.

The estimator retains the self–self–cross geometry of conditional CS divergence. Minimizing Eq. 83 drives the site-refined conditional representation law $p_\theta(R \mid U, S)$ toward the context-only law $p_\theta(R \mid U)$, thereby enforcing the observable transportability criterion derived from CMT.

B CONTROLLED EXPERIMENTS

B.1 CONTROLLED DATA GENERATION

We construct a low-dimensional synthetic system following the causal structure in Sec. 2.2:

$$S \to C, \qquad S \to M, \qquad C \to Y, \qquad (Y, C) \to Z, \qquad (Z, M) \to X. \tag{84}$$

We consider four source environments indexed by $S \in \{0, 1, 2, 3\}$. The biological context $C \in \{0, 1\}$ follows the site-specific prevalence

$$p(C = 1 \mid S) = (0.03,\ 0.20,\ 0.80,\ 0.97), \tag{85}$$

and disease status is generated through the shared mechanism

$$Y \sim \text{Bernoulli}\left[\sigma(2.6(2C - 1))\right]. \tag{86}$$

The latent brain state $Z \in \mathbb{R}^2$ is generated as

$$Z = 0.5 \begin{bmatrix} 2C - 1 \\ (2Y - 1)(2C - 1) \end{bmatrix} + 0.5\epsilon_Z, \tag{87}$$

where the second component introduces a stable disease–context interaction. This construction allows disease-related variation to depend on biological context rather than being identical across all subjects.

Measurement conditions vary across sites through

$$M_s = \begin{bmatrix} \mu_s + 0.3\epsilon_{M1} \\ 0.6\mu_s + 0.3\epsilon_{M2} \end{bmatrix}, \qquad \mu_s \in \{-2.4, -0.8, 0.8, 2.4\}, \tag{88}$$

and the observed input $X \in \mathbb{R}^4$ is

$$X = \begin{bmatrix} Z + M + 0.12\epsilon_{X,1:2} \\ M + 0.12\epsilon_{X,3:4} \end{bmatrix}. \tag{89}$$

All exogenous noise variables are mutually independent and independent of (S, C, Y), with

$$\epsilon_Z \sim \mathcal{N}(0, I_2), \qquad \epsilon_M \sim \mathcal{N}(0, I_2), \qquad \epsilon_X \sim \mathcal{N}(0, I_4). \tag{90}$$

Thus, environments differ through $p_s(C, M)$, while $p(Y \mid C)$, $p(Z \mid Y, C)$, and $p(X \mid Z, M)$ remain shared.

Controlled target shifts. For the measurement-shift experiment, we modify only the measurement distribution M:

$$M(d) = \begin{bmatrix} a(d)\mu_s + b(d)\epsilon_{M1} \\ 0.6a(d)\mu_s + b(d)\epsilon_{M2} \end{bmatrix}, \tag{91}$$

where

$$a(d) = \max(1 - d/2, 0), \qquad b(d) = 0.3 + 0.95d, \tag{92}$$

and $d = \Delta_M \in \{0, 1, 2, 3, 4, 6, 8\}$. The same C, Y, Z, and exogenous noise are reused across shift strengths, so only the measurement component is perturbed.

For the biological-context shift, we instead keep the measurement process fixed and progressively reverse the site-specific prevalence $p(C = 1 \mid S)$. The shift strength $\Delta_C \in \{0, 0.25, 0.5, 0.75, 1\}$ therefore controls how strongly the population composition departs from the source environments, while the structural mechanisms remain unchanged.

All compared methods use the same predictor and are selected exclusively using source-domain validation data; target environments are used only for evaluation. Because Y and C are discrete in this controlled setting, CIC matches samples within identical (Y, C) strata. All reported results are averaged over three random seeds.

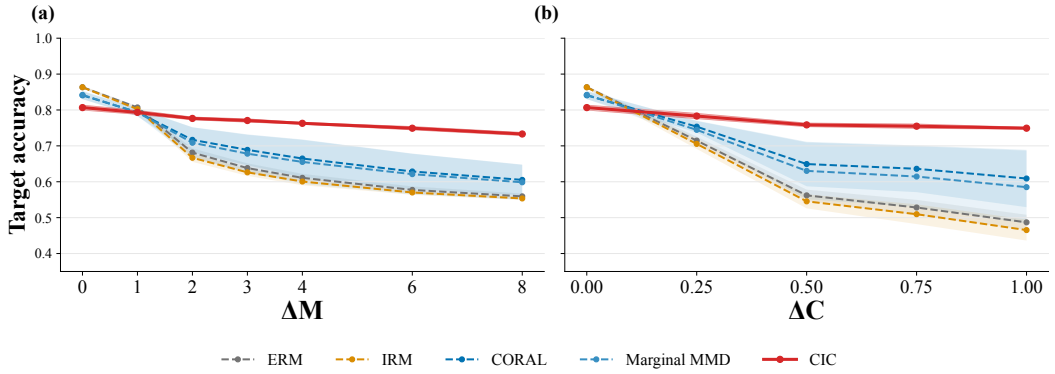


Figure 6: **Controlled validation measured by balanced accuracy.** The left panel reports target balanced accuracy under increasing measurement shift Δ_M, and the right panel reports target balanced accuracy under increasing biological-context shift Δ_C. Curves and shaded regions denote the mean and standard deviation over three random seeds.

B.2 Additional Controlled Analyses

Robustness across evaluation metrics. Figure 6 reports balanced accuracy under the same measurement and biological-context shifts used in Fig. 4. The trends closely follow those observed with AUC: source-oriented methods deteriorate rapidly as the target distribution departs from the observed environments, whereas CIC remains substantially more stable. This indicates that the behavior in Fig. 4 is not specific to the choice of AUC.

Mechanism audit. We further examine whether models trained with CIC show reduced sensitivity to the measurement pathway while preserving biologically relevant representation variability. For each subject, we construct paired observations under $\Delta_M = 0$ and $\Delta_M = 2$ while keeping C, Y, Z, and ϵ_X fixed. Prediction sensitivity is defined as

$$\text{PredSens} = \frac{1}{N}\sum_{i=1}^{N}\left|p_\theta(Y_i = 1 \mid X_i^{(2)}) - p_\theta(Y_i = 1 \mid X_i^{(0)})\right|, \tag{93}$$

and representation sensitivity as

$$\text{RepSens} = \frac{1}{N}\sum_{i=1}^{N}\left\|R_\theta(X_i^{(2)}) - R_\theta(X_i^{(0)})\right\|_2. \tag{94}$$

We additionally quantify how much information about the true latent brain state Z remains in R using an independent post-hoc probe. All three diagnostics are computed using frozen models and are not involved in training or model selection.

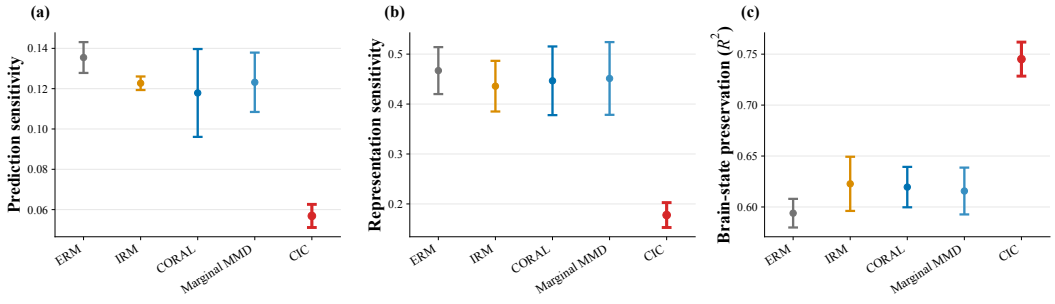


Figure 7: **Mechanism audit under paired measurement perturbations.** From left to right, the panels report prediction sensitivity, representation sensitivity, and preservation of the true latent brain state measured by the R^2 of an independent $R \to Z$ probe. Lower values are better for the first two metrics, whereas higher R^2 indicates stronger preservation of latent brain-state information. Error bars denote the standard deviation over three random seeds.

As shown in Fig. 7, models trained with CIC exhibit substantially lower prediction and representation sensitivity than the competing objectives, while retaining the highest predictability of the latent brain state. The improved robustness is therefore accompanied by reduced measurement dependence rather than a general loss of biologically relevant information.

Role of biological context. Finally, we examine when conditioning on C is necessary. We introduce a control parameter κ that determines whether the stable latent mechanism contains a disease–context interaction. When $\kappa = 0$, disease-related latent structure does not depend on C; when $\kappa = 1$, the stable mechanism contains the $Y \times C$ interaction used in the main controlled experiment.

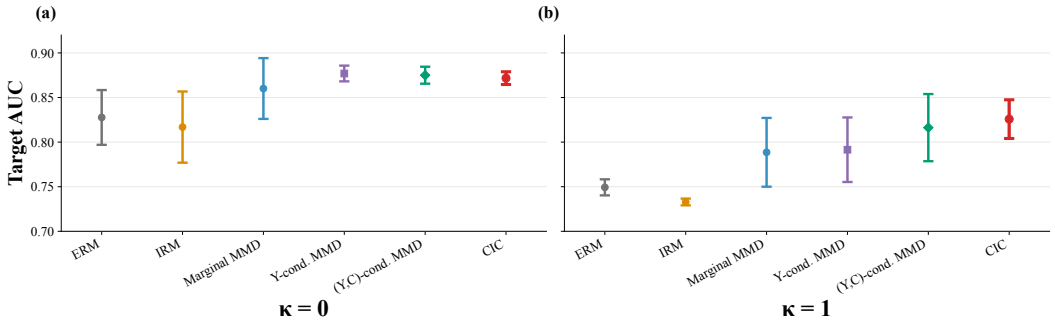


Figure 8: **Role of biological context in conditional invariance.** The left panel shows $\kappa = 0$, where the stable disease-related mechanism does not depend on biological context C. The right panel shows $\kappa = 1$, where the latent brain state contains a stable $Y \times C$ interaction. Error bars denote the standard deviation over three random seeds.

Figure 8 compares the predictive and invariance objectives under these two mechanisms, including controls conditioned on Y alone or jointly on (Y, C). When $\kappa = 0$, additionally conditioning on C provides little benefit. When $\kappa = 1$, objectives that account for both Y and C preserve stronger target performance than those that ignore the biological context. This control supports the conditioning structure implied by CMT: variation associated with C should be preserved when C participates in the stable disease mechanism, rather than removed marginally as a nuisance.

C Datasets and Experimental Protocols

C.1 Dataset Details

We conduct experiments using ABIDE-I, ABIDE-II, REST-meta-MDD, SRPBS-SCZ, and SRPBS-ASD, covering autism spectrum disorder (ASD), major depressive disorder (MDD), and schizophrenia (SCZ).

ABIDE-I and ABIDE-II. The Autism Brain Imaging Data Exchange (ABIDE) provides multi-site resting-state fMRI and phenotypic data for individuals with ASD and healthy controls Di Martino et al. (2014; 2017). After quality control and cohort filtering, the ABIDE-I cohort contains 322 individuals with ASD and 342 healthy controls from 8 sites, while ABIDE-II contains 325 individuals with ASD and 431 healthy controls from 10 sites.

REST-meta-MDD. We use the REST-meta-MDD dataset Chen et al. (2022), a large-scale multi-site resting-state fMRI resource for major depressive disorder. After quality control and site filtering, the final cohort contains 1,604 individuals with MDD and 1,300 healthy controls from 21 sites.

SRPBS. We additionally use the multi-site, multi-disorder SRPBS database Tanaka et al. (2021). The SRPBS-SCZ cohort contains 140 patients with schizophrenia and 140 healthy controls and is evaluated under the LOSO protocol. Healthy controls were matched to patients at a 1:1 ratio within site and sex strata by globally minimizing absolute age differences. The SRPBS-ASD cohort contains 115 individuals with ASD and 101 healthy controls and is reserved exclusively for external evaluation.

For ABIDE-I, ABIDE-II, and REST-meta-MDD, sites were retained when both diagnostic groups contained at least 15 participants and the total number of participants at the site exceeded 40.

C.2 Site-wise Demographic and Acquisition Characteristics

Tables 5–9 summarize the site-wise characteristics of the five datasets used for LOSO evaluation and cross-dataset evaluation after quality control. For each site, we report the numbers of patients and healthy controls, age distribution, sex composition, repetition time (TR), and mean framewise displacement (FD). Age and mean FD are reported as mean ± standard deviation, while sex composition is reported as the percentage of female participants.

Table 5: Site-level characteristics of the ABIDE-I cohort.

Site	Patient / Healthy Control	Age	Female (%)	TR (s)	Mean FD
KKI	22 / 33	10.1 ± 1.3	23.6	2.5	0.296 ± 0.285
MAX_MUN	24 / 32	26.2 ± 12.2	12.5	3.0	0.231 ± 0.204
NYU	78 / 103	15.3 ± 6.6	20.4	2.0	0.150 ± 0.089
PITT	21 / 21	18.2 ± 6.9	19.0	1.5	0.294 ± 0.186
TRINITY	24 / 25	17.2 ± 3.6	0.0	2.0	0.207 ± 0.176
UCLA_1	41 / 32	13.2 ± 2.4	13.7	3.0	0.292 ± 0.314
UM_1	54 / 53	13.4 ± 2.9	23.4	2.0	0.330 ± 0.413
USM	58 / 43	22.1 ± 7.7	0.0	2.0	0.230 ± 0.187

Table 6: Site-level characteristics of the ABIDE-II cohort.

Site	Patient / Healthy Control	Age	Female (%)	TR (s)	Mean FD
BNI_1	29 / 29	38.5 ± 15.5	0.0	3.0	0.290 ± 0.160
EMC_1	24 / 27	8.2 ± 1.1	19.6	2.0	0.409 ± 0.383
GU_1	51 / 55	10.7 ± 1.6	33.0	2.0	0.367 ± 0.496
IP_1	21 / 32	20.6 ± 10.4	54.7	2.7	0.236 ± 0.311
KKI_1	40 / 135	10.3 ± 1.3	36.6	2.5	0.247 ± 0.190
NYU_1	48 / 30	9.8 ± 4.9	9.0	2.0	0.236 ± 0.141
OHSU_1	37 / 56	10.9 ± 2.0	38.7	2.5	0.191 ± 0.163
SDSU_1	33 / 25	13.0 ± 3.2	15.5	2.0	0.167 ± 0.120
SU_2	21 / 21	11.1 ± 1.2	9.5	2.0	0.177 ± 0.152
TCD_1	21 / 21	15.2 ± 3.2	0.0	2.0	0.370 ± 0.425

Table 7: Site-level characteristics of the REST-meta-MDD cohort.

Site	Patient / Healthy Control	Age	Female (%)	TR (s)	Mean FD
S1	282 / 251	39.2 ± 14.7	65.1	2.0	0.141 ± 0.060
S2	141 / 63	19.4 ± 3.7	69.6	2.0	0.101 ± 0.082
S4	34 / 30	30.7 ± 10.2	60.9	2.0	0.175 ± 0.127
S5	139 / 63	34.4 ± 9.8	52.5	2.0	0.202 ± 0.577
S6	67 / 105	26.5 ± 7.9	52.3	2.0	0.267 ± 0.885
S7	24 / 30	19.5 ± 5.1	75.9	2.0	0.125 ± 0.142
S8	47 / 42	32.4 ± 9.4	59.6	2.5	0.107 ± 0.097
S9	47 / 26	42.1 ± 14.6	60.3	2.0	0.121 ± 0.072
S10	85 / 70	35.3 ± 12.5	55.5	2.0	0.140 ± 0.070
S11	116 / 68	37.7 ± 12.6	53.3	2.0	0.101 ± 0.072
S12	63 / 32	30.2 ± 6.5	62.1	2.5	0.039 ± 0.039
S13	33 / 29	37.0 ± 14.1	62.9	2.0	0.144 ± 0.070
S14	139 / 147	41.0 ± 12.0	59.4	2.0	0.121 ± 0.075
S15	35 / 35	21.7 ± 4.6	57.1	2.0	0.097 ± 0.055
S16	83 / 81	25.9 ± 7.5	50.6	2.0	0.085 ± 0.034
S17	41 / 50	35.5 ± 9.8	62.6	4.4	0.164 ± 0.078
S18	34 / 34	34.0 ± 12.2	52.9	2.0	0.127 ± 0.129
S19	32 / 32	29.0 ± 9.5	67.2	2.0	0.117 ± 0.045
S21	63 / 56	32.5 ± 11.1	59.7	2.0	0.087 ± 0.058
S22	35 / 25	35.4 ± 11.4	70.0	2.0	0.142 ± 0.047
S23	64 / 31	13.7 ± 1.7	75.8	2.0	0.178 ± 0.217

Table 8: Site-level characteristics of the SRPBS-SCZ cohort.

Site	Patient / Healthy Control	Age	Female (%)	TR (s)	Mean FD
KTT	40 / 40	35.3 ± 9.8	47.5	2.0	0.150 ± 0.088
KUT	45 / 45	41.4 ± 10.8	53.3	2.5	0.186 ± 0.121
SWA	19 / 19	41.2 ± 8.3	21.1	2.5	0.208 ± 0.077
UTO	36 / 36	34.8 ± 10.2	33.3	2.5	0.135 ± 0.068

Table 9: Site-level characteristics of the SRPBS-ASD cohort.

Site	Patient / Healthy Control	Age	Female (%)	TR (s)	Mean FD
SWA	115 / 101	30.4 ± 8.0	13.9	2.5	0.180 ± 0.084

C.3 fMRI Preprocessing and Brain Network Construction

The resting-state fMRI data were preprocessed using DPABI Yan et al. (2016). The first 10 volumes were discarded to allow for signal stabilization. Preprocessing subsequently included slice-timing correction, six-parameter rigid-body realignment, nuisance-covariate regression, T1-weighted image coregistration, tissue segmentation, and DARTEL-based normalization to the MNI-152 space. Linear trends were regressed out, followed by temporal band-pass filtering at 0.01–0.1 Hz.

The brain was parcellated into 116 regions of interest using the Automated Anatomical Labeling (AAL116) atlas Tzourio-Mazoyer et al. (2002). For each subject, regional BOLD time series were extracted and pairwise Pearson correlation coefficients were computed to construct the functional-connectivity matrix. Correlation coefficients were transformed using Fisher's r-to-z transformation.

To construct the brain graph, functional connections were ranked according to their absolute FC magnitude, and the strongest 20% were retained, with the graph density selected by grid search on source-domain validation data. The original signs of the retained connections were preserved, resulting in a sparse signed functional brain network. For graph-based models, each brain region was treated as a node and its full FC profile was used as the node feature.

C.4 Evaluation Protocols

Unseen-site generalization.

For multi-site evaluation, we adopt leave-one-site-out (LOSO) generalization. Given a dataset containing K sites, each site is held out once as the target domain, while the remaining $K-1$ sites form the source domains. For each LOSO fold, 15% of the subjects from every source site are selected as the validation set using diagnosis-stratified sampling, with the remaining source subjects used for training. Training, hyperparameter selection, checkpoint selection, and early stopping are performed using the source sites, and the held-out site provides the final test set.

Each LOSO fold is repeated using three random seeds. Performance is evaluated using classification accuracy (ACC) and area under the receiver operating characteristic curve (AUC). For each held-out site, results are first averaged across the three random seeds and then averaged over all target sites, with the standard deviation across sites characterizing cross-site variability. Statistical comparisons between backbones trained with CIC and their paired baselines are performed using one-sided paired Wilcoxon signed-rank tests over held-out sites at $p < 0.05$. Significance testing is omitted for SRPBS-SCZ because it contains only four sites.

Cross-dataset transfer.

We further evaluate generalization across independently collected ASD cohorts using ABIDE-I→ABIDE-II, ABIDE-II→ABIDE-I, and ABIDE-II→SRPBS-ASD. For each setting, 15% of the subjects within every source site are selected by diagnosis-stratified sampling for validation, with the remaining subjects used for training. The source cohort supplies all training and model-selection data, while the disjoint target cohort provides the final test set.

Each transfer setting is repeated with three random seeds, and ACC and AUC are reported as the mean and standard deviation across independent training runs. All transfer settings retain the same ASD classification task, thereby isolating dataset-level domain shift from task shift.

D CIC Implementation and Analysis

D.1 Implementation Details

All models were trained with the Adam optimizer, a batch size of 32, and a weight decay of 10^{-4}, for up to 200 epochs. The learning rate was 10^{-4} for BrainNetTF and BrainNetCNN, and 10^{-3} for GAT. The dropout rate was 0.1 for BrainNetTF, 0.5 for BrainNetCNN, and 0.2 for GAT. For Causal Information Contrast (CIC) training, the first 30 epochs used cross-entropy alone, after which the CIC weight was linearly increased to its target value over five epochs. The CIC feature bank retained up to 64 representations per source site. The regularization coefficient λ was selected by grid search over $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ using source-domain validation AUC within each LOSO training fold. The selected values are summarized in Table 10. All experiments were conducted on a Linux 6.8.0 system with an Intel Xeon Gold 6230 CPU and an NVIDIA GeForce RTX 4090 GPU, using PyTorch 2.14.0 with CUDA 13.0.

Table 10: Selected CIC regularization coefficient λ for each backbone–dataset configuration.

Backbone	ABIDE-I	ABIDE-II	REST-meta-MDD	SRPBS-SCZ
BrainNetTF	10^{-2}	10^{-2}	10^{-2}	10^{-2}
BrainNetCNN	10^{-3}	10^{-4}	10^{-3}	10^{-3}
GAT	10^{-4}	10^{-3}	10^{-4}	10^{-3}

D.2 Sensitivity Analysis

Feature-bank capacity. We first examine how the feature-bank capacity affects the conditional sample support available to the CIC estimator. Using fixed model representations without retraining, we vary the bank capacity from batch-only estimation to 1024 samples per source site and compute the conditional effective sample size (ESS) after excluding the anchor itself. The analysis covers all 43 LOSO training folds across ABIDE-I, ABIDE-II, REST-meta-MDD, and SRPBS-SCZ.

As shown in Fig. 9, batch-only estimation suffers from severe support sparsity: the fraction of anchors with $\mathrm{ESS} < 2$ ranges from 49.2% to 95.0% across datasets. With only 64 stored samples per source site, this fraction decreases to at most 2.5%, while the average ESS increases substantially on all four datasets. Increasing the capacity further continues to improve the absolute ESS, particularly on REST-meta-MDD, although the low-support fraction has already largely saturated. At larger capacities, the curves eventually plateau as the available distinct training subjects within each LOSO fold are exhausted. These results support the use of a feature bank to provide sufficient conditional support for CIC estimation, while indicating that a moderate capacity is sufficient to alleviate extremely low-support cases.

Kernel bandwidth. CIC involves two continuous Gaussian kernels whose bandwidths are determined differently. For the age component of the conditional kernel, age is standardized within each LOSO fold and the bandwidth is fixed to $\sigma_{\mathrm{age}} = 1$. For the representation kernel, the learned representations are normalized and the squared bandwidth is adaptively determined in each mini-batch using the median heuristic:

$$\sigma_r^2 = \mathrm{median}_{i \neq j} \, \|r_i - r_j\|_2^2 \,. \tag{95}$$

We further evaluate the sensitivity of CIC to both bandwidth choices. The age bandwidth is varied over $\sigma_{\mathrm{age}} \in \{0.5, 1, 2\}$ while keeping the representation bandwidth unchanged. For the representation kernel, we introduce a multiplicative scaling factor γ_r,

$$\sigma_r^2 = \gamma_r \, \mathrm{median}_{i \neq j} \, \|r_i - r_j\|_2^2 \,, \qquad \gamma_r \in \{0.5, 1, 2\}, \tag{96}$$

while fixing $\sigma_{\mathrm{age}} = 1$. All other training settings are kept unchanged.

As shown in Table 11, CIC remains generally stable under moderate bandwidth variations. Changing the age bandwidth results in only modest performance differences across datasets. Scaling the adaptive representation bandwidth has a somewhat larger effect on ABIDE-I and ABIDE-II, whereas the

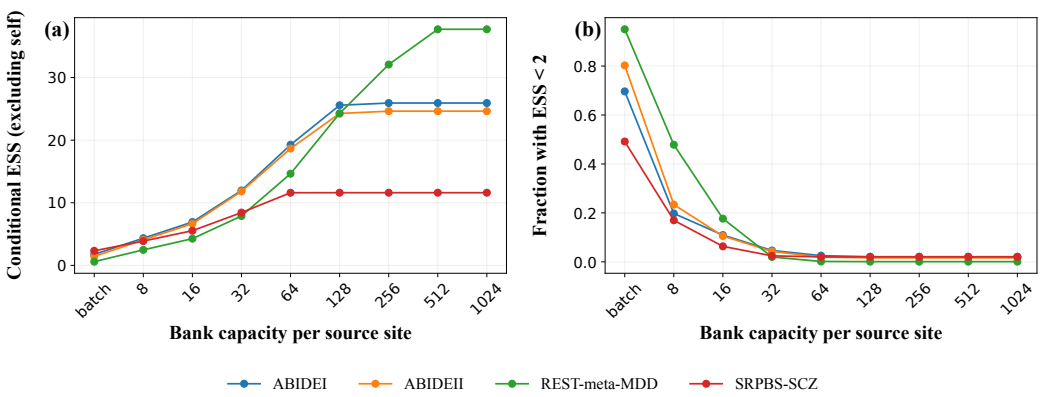


Figure 9: **Effect of feature-bank capacity on conditional support.** (a) Conditional effective sample size (ESS), excluding the anchor itself. (b) Fraction of anchors with ESS < 2. Results are averaged over all LOSO training folds using fixed representations without retraining.

Table 11: Sensitivity of CIC to the age-kernel bandwidth σ_{age} and the representation-bandwidth scaling factor γ_r. The default configuration is $(\sigma_{\text{age}}, \gamma_r) = (1, 1)$. Results are reported as AUC (%) with mean $\pm$ standard deviation across held-out sites.

Dataset	Default $(1, 1)$	$\sigma_{\text{age}} = 0.5$	$\sigma_{\text{age}} = 2$	$\gamma_r = 0.5$	$\gamma_r = 2$
ABIDE-I	62.77 ± 12.55	61.43 ± 12.94	62.87 ± 13.32	61.09 ± 13.67	61.27 ± 13.23
ABIDE-II	67.47 ± 5.94	65.72 ± 3.80	65.91 ± 3.75	65.42 ± 3.75	64.98 ± 4.14
REST-meta-MDD	66.18 ± 7.41	66.69 ± 7.29	66.64 ± 7.22	66.16 ± 7.37	65.59 ± 7.20
SRPBS-SCZ	85.39 ± 5.87	85.38 ± 5.85	85.41 ± 5.87	85.35 ± 5.82	85.36 ± 5.80

results on REST-meta-MDD and SRPBS-SCZ remain nearly unchanged. Overall, the default configuration remains within the high-performing region across all datasets, suggesting that CIC does not depend on a narrowly tuned kernel bandwidth.

Regularization strength. We evaluate the sensitivity of CIC to the regularization coefficient λ using BrainNetTF as a representative backbone. We vary λ over $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and additionally include $\lambda = 0$ as the ERM baseline, while keeping all other training configurations fixed. As shown in Fig. 10, an intermediate regularization strength consistently yields the best performance across the four datasets, with $\lambda = 10^{-2}$ achieving the highest mean AUC in all cases. Smaller coefficients provide weaker or less consistent gains, whereas increasing λ to 10^{-1} leads to a mild decrease from the optimum. Overall, CIC remains reasonably stable across the examined range while consistently favoring a moderate regularization strength around $\lambda = 10^{-2}$.

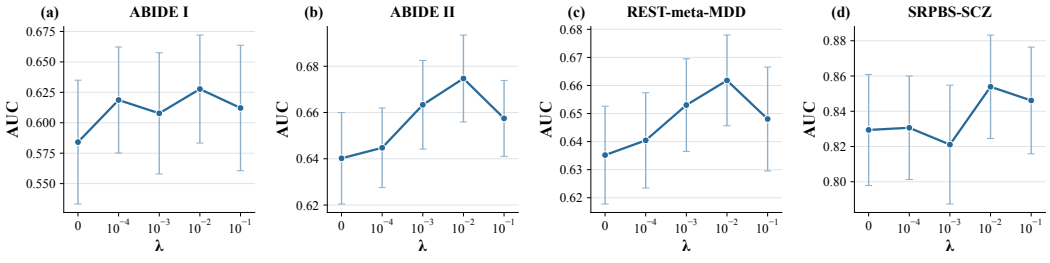


Figure 10: **Sensitivity to the CIC regularization coefficient λ.** AUC of BrainNetTF under different CIC regularization strengths on (a) ABIDE-I, (b) ABIDE-II, (c) REST-meta-MDD, and (d) SRPBS-SCZ. $\lambda = 0$ corresponds to the ERM baseline. Points and error bars denote the mean and standard deviation over LOSO evaluations across three random seeds.

D.3 COMPUTATIONAL EFFICIENCY

We further evaluate the computational overhead introduced by CIC on three representative backbones: GAT, BrainNetCNN, and BrainNetTF. All measurements are reported over three independent runs. For training efficiency, we first average the runtime over 10 training epochs within each run and then report the mean and sample standard deviation across runs. For inference latency, each run contains 100 forward passes with a batch size of 32; we take the median latency within each run and aggregate the results across runs.

Table 12: Computational efficiency of different backbones with and without CIC. Training statistics are reported as mean ± sample standard deviation over three independent runs. Inference latency is measured with a batch size of 32.

Model	Train Time (s/epoch)	Train Memory (MiB)	Inference (ms)	Inference Memory (MiB)
GAT	0.340 ± 0.012	94.3 ± 0.0	3.255 ± 0.023	49.0
GAT + CIC	0.743 ± 0.059 (2.19×)	101.8 ± 0.0	3.200 ± 0.115	49.0
BrainNetCNN	0.573 ± 0.017	566.2 ± 0.0	4.899 ± 0.123	390.6
BrainNetCNN + CIC	0.979 ± 0.033 (1.71×)	781.7 ± 0.0	4.717 ± 0.210	390.6
BrainNetTF	0.501 ± 0.005	190.6 ± 0.0	1.367 ± 0.010	44.1
BrainNetTF + CIC	0.795 ± 0.025 (1.59×)	263.7 ± 0.0	1.358 ± 0.001	44.1

As shown in Table 12, CIC introduces moderate additional computation during training. The training time increases by $2.19\times$, $1.71\times$, and $1.59\times$ for GAT, BrainNetCNN, and BrainNetTF, respectively. The corresponding peak training memory increases from 94.3 to 101.8 MiB for GAT, from 566.2 to 781.7 MiB for BrainNetCNN, and from 190.6 to 263.7 MiB for BrainNetTF. In contrast, inference latency remains essentially unchanged across all three architectures, and the peak inference memory is identical to that of the corresponding backbone. This is expected because CIC is only used as a training objective and introduces no additional modules or computations during inference. Overall, CIC incurs additional optimization cost during training while preserving the original deployment complexity of the backbone.