MALM: LEARNING LATENT NETWORK MEDIATORS FROM PAIRED CONNECTOMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Mediation analysis begins with specifying the mediator, which defines the mechanism being tested. In brain-network mediation, structural connectivity (SC) and functional connectivity (FC) are often treated as candidate network mediators, but they are coupled, noisy, and complementary measurements of brain connectivity, making parallel or serial multimodal formulations potentially restrictive. We propose MALM, an end-to-end Bayesian framework that learns an integrated network mediator from paired SC and FC and uses that learned network directly as the mediator. Its central contribution is to couple multimodal representation learning with exposure-to-mediator and mediator-to-outcome pathway estimation on the same region-pair coordinates. Dual cross-modal attention integrates the observed networks, while a structured mediation layer connects the mediator to closed-form direct, indirect, and total effects with an edge-level decomposition, without imposing an inter-modality ordering. Joint posterior inference quantifies uncertainty over model unknowns and induced mediation quantities, including learned network mediators, pathway maps, global effects, and active mediating brain edges. On synthetic data, MALM improves effect estimation and pathway recovery when connectome views are coupled, interact non-additively, or violate a fixed serial ordering. In HCP and ABCD cohorts, MALM estimates significant connectivity-mediated effects in both cohorts and recovers a substantially more consistent mediating network than existing methods, concentrated in default-mode, visual, and somatomotor systems and supporting a mediation pathway that remains stable across developmental stages.

1 INTRODUCTION

Mediation analysis asks how an exposure shapes an outcome through a specified mediator. Because the mediator is the mechanism being tested, different mediator definitions lead to different mechanistic claims. This choice becomes especially consequential when the mediator is not a scalar variable but a network that characterizes underlying interrelationships. Brain-network studies provide such a setting because brain function emerges from coordinated activity across distributed systems rather than from any single region. In practice, two networks are routinely used to characterize this organization. Structural connectivity (SC), derived from diffusion MRI tractography, quantifies white-matter pathways between brain regions, while functional connectivity (FC), derived from fMRI, measures temporal correlations in BOLD signals between regions. Both have been widely used as mediators in studies linking exposures such as genetic variation, environmental factors, or interventions to cognitive and behavioral outcomes (Lindquist, 2012; Zhao et al., 2020; Tian et al., 2024).

Structure–function coupling makes this mediator choice consequential. Direct white-matter tracts explain only part of FC, which also reflects indirect network communication and changes with brain state. SC–FC coupling varies across cortical systems, development, behavior, and genetic background (Honey et al., 2009; Suárez et al., 2020; Vázquez-Rodríguez et al., 2019; Zamani Esfahlani et al., 2022; Gao et al., 2026). Both modalities contain measurement noise and complementary information, so choosing either alone can overlook connectivity patterns expressed jointly. Parallel mediation separates their pathways, while serial mediation commits to an SC-to-FC or FC-to-SC ordering. These formulations address specific hypotheses, but the broader question is how an exposure acts through the joint organization of structural and functional networks.

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.

Existing mediation methods largely address the estimation problem after the mediator structure has been chosen. Functional, high-dimensional, and network-mediator approaches have extended causal mediation to complex brain measurements, providing regularization, pathway selection, and effect decomposition for imaging or connectivity mediators (Lindquist, 2012; Chén et al., 2018; Zhao et al., 2020; Tian et al., 2024). These methods are powerful for a single measured mediator, but they do not resolve how paired SC and FC should jointly define the mediating object. Methods for multiple or multimodal mediators allow several intermediate variables to enter the analysis, but they still require the analyst to specify whether the modalities act as parallel mechanisms or as an ordered sequence (VanderWeele & Vansteelandt, 2014; Daniel et al., 2015; Imai & Yamamoto, 2013; Zhao et al., 2021; Li et al., 2022). Meanwhile, the SC–FC modeling literature has developed predictive mappings, network communication models, and shared representations that capture structure–function coupling (Honey et al., 2009; Goñi et al., 2014; Seguin et al., 2020; Sarwar et al., 2021; Zalesky et al., 2024; Xia et al., 2025). These models characterize dependencies between paired connectomes, but they do not estimate direct and indirect effects through a network representation learned jointly from them. The methodological opportunity is to connect paired-network representation learning to mediation through the definition of the mediating object itself. MALM makes this connection by learning a network that serves jointly as a representation of both connectomes and as the mediator in the exposure–outcome pathway.

To address this problem, we propose **M**ediation **A**nalysis with attention-based **L**atent **M**ediator (MALM). Its central contribution is to learn an integrated network mediator jointly with its exposure and outcome pathways. Reciprocal cross-modal attention integrates SC and FC, symmetric decoding retains region-pair coordinates, and reconstruction connects the mediator to both observed modalities. Exposure and outcome maps act on these same coordinates, linking each connection's exposure response to its outcome contribution. Joint Bayesian estimation unifies modality integration and mediation analysis, yielding global effects, connection-level contributions, and posterior uncertainty.

We make the following contributions:

- **A learned network mediator for paired connectomes.** We define the mediator as a network learned jointly from paired SC and FC. This makes their integrated organization the object through which exposure-related changes are linked to the outcome, enabling mediation analysis without choosing a parallel or serial ordering of the observed modalities.
- **Paired-network learning coupled to pathway estimation.** Reciprocal cross-modal attention, symmetric decoding, and reconstruction of both modalities produce an integrated network mediator indexed by region pairs. Learning this mediator jointly with the mediation layers connects multimodal fusion to the exposure and outcome pathways while retaining connection-level interpretation.
- **Global effects and network attribution in one model.** Structured exposure-to-mediator and mediator-to-outcome maps link global mediation effects to contributions on the learned network. Region-informed shrinkage and joint posterior estimation support pathway selection and uncertainty summaries within the same representation.
- **Cross-cohort pathway stability.** We assess whether the pathways recovered by MALM remain reproducible across cohorts at different neurodevelopmental stages. MALM identifies substantially more consistent edge-level mediation pathways than competing approaches, supporting the view that an integrated network mediator can stabilize pathway recovery across cohorts.

2 METHOD

2.1 PROBLEM SETUP AND LEARNED NETWORK MEDIATOR

For subject $i = 1, \ldots, N$, we observe a structural connectivity matrix $S_i \in \mathrm{Sym}_0(R)$, a functional connectivity matrix $F_i \in \mathrm{Sym}_0(R)$, an exposure z_i, covariates $x_i \in \mathbb{R}^Q$, and an outcome $y_i \in \mathbb{R}$. Here, R denotes the number of brain regions in the chosen parcellation, and $\mathrm{Sym}_0(R)$ denotes the subspace of symmetric matrices with zero diagonal. The key modeling choice is the definition of the mediator. MALM learns a network-valued mediator $M_i \in \mathrm{Sym}_0(R)$ jointly from SC and FC. This representation integrates both modalities on region-pair coordinates and serves as the mediating object. Latent in the method name denotes this learned representation.

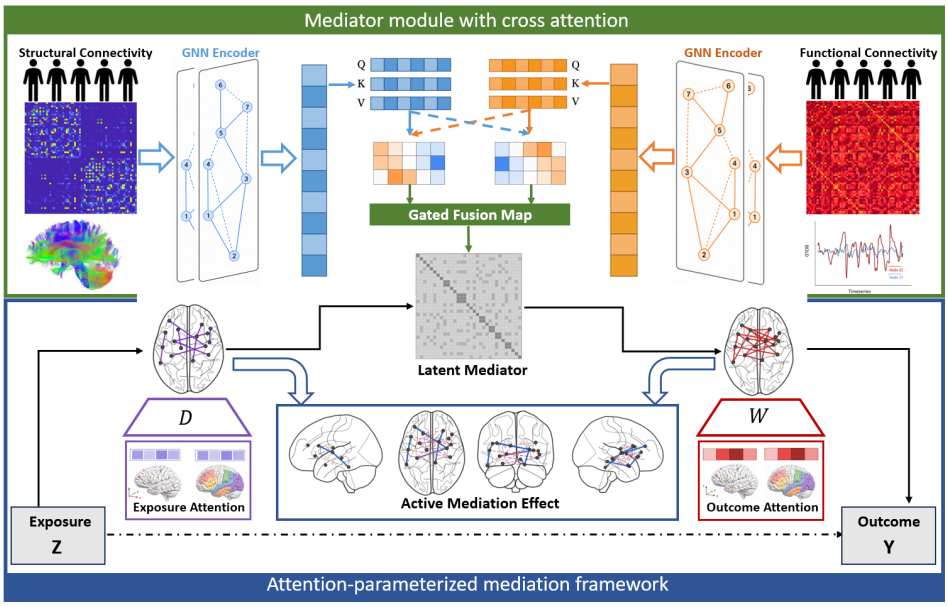


Figure 1: Overview of MALM. Paired structural and functional connectivity are fused through cross-attention to construct an integrated network mediator, while attention modules parameterize the exposure- and outcome-related coefficient matrices for closed-form mediation analysis.

The fused network itself enters the mediation model. As illustrated in Figure 1, $M_i = \Phi_\theta(S_i, F_i)$ and its exposure and outcome maps D_ω and W_ψ share region-pair coordinates, connecting paired-network learning to global effects and edge-level attribution.

Let $\rho = (\rho_1, \ldots, \rho_R)^\top$ denote the collection of region-level features.

$$M_i = \Phi_\theta(S_i, F_i), \tag{1}$$

$$D_\omega(\rho) = \mathcal{A}_D(\rho), \qquad W_\psi(\rho) = \mathcal{A}_W(\rho), \tag{2}$$

$$M_i \mid z_i, x_i \sim \mathcal{N}_{\mathrm{sym},0}(\mu(x_i) + z_i D_\omega(\rho),\ \Sigma_M)\,, \tag{3}$$

$$y_i \mid M_i, z_i, x_i \sim \mathcal{N}\big(x_i^\top \beta_x + \beta_z z_i + \langle W_\psi(\rho), M_i\rangle_F,\ \sigma_y^2\big)\,. \tag{4}$$

Here, Φ_θ integrates paired SC and FC through dual cross-modal attention. The pathway maps $\mathcal{A}_D$ and $\mathcal{A}_W$ construct exposure-to-mediator and mediator-to-outcome coefficient matrices from region features. The Gaussian law $\mathcal{N}_{\mathrm{sym},0}$ is defined on symmetric matrices with zero diagonal. The term $\mu(x_i)$ captures covariate-dependent baseline mediator variation, and $x_i^\top \beta_x$ adjusts for direct covariate effects on the outcome.

2.2 LEARNING THE NETWORK-VALUED MEDIATOR

The learned network mediator in MALM integrates structural and functional connectivity within a region-pair representation for mediation analysis. Reciprocal cross-modal attention interprets each modality in the context of the other, allowing the mediator to encode their dependencies without assigning either modality a fixed upstream role. The resulting cross-modal views are then combined through a fusion map $\mathcal{F}$ that directly outputs the learned mediator matrix.

We first encode the two observed connectomes into region-level embeddings,

$$Z_i^S = \mathrm{Enc}_S(S_i) \in \mathbb{R}^{R\times d}, \qquad Z_i^F = \mathrm{Enc}_F(F_i) \in \mathbb{R}^{R\times d}, \tag{5}$$

where Enc_S and Enc_F are modality-specific graph encoders. Cross-modal attention is then applied in both directions. The first lets structure attend to function,

$$H_i^{S\to F} = \mathrm{softmax}\left(\frac{Q_i^S (K_i^F)^\top}{\sqrt{d_k}}\right) V_i^F, \qquad Q_i^S = Z_i^S U_Q^S,\ K_i^F = Z_i^F U_K^F,\ V_i^F = Z_i^F U_V^F, \tag{6}$$

and the second lets function attend to structure,

$$H_i^{F\to S} = \mathrm{softmax}\left(\frac{Q_i^F (K_i^S)^\top}{\sqrt{d_k}}\right) V_i^S, \qquad Q_i^F = Z_i^F U_Q^F,\ K_i^S = Z_i^S U_K^S,\ V_i^S = Z_i^S U_V^S. \quad (7)$$

These two attended representations encode structural and functional information in the context of the other modality. The learned mediator is then obtained by

$$M_i = \mathcal{F}\big(H_i^{S\to F},\, H_i^{F\to S}\big), \quad (8)$$

where $\mathcal{F}$ is a learnable fusion map from the two cross-modal views to a symmetric network-valued mediator. Specifically, $\mathcal{F}$ first combines the two attended representations through a learnable gate and then maps the fused node embeddings to edge weights through a symmetric bilinear interaction,

$$\begin{aligned} G_i &= \sigma\big(H_i^{S\to F} U_g + H_i^{F\to S} V_g + b_g\big), \\ H_i &= G_i \odot H_i^{S\to F} + (1 - G_i) \odot H_i^{F\to S}, \\ M_{i,wl} &= h_{i,w}^\top B h_{i,l}, \qquad B = \tfrac{1}{2}(B + B^\top). \end{aligned} \quad (9)$$

This construction preserves explicit region-pair interaction structure while allowing the mediator to be learned end to end from the two cross-modal views. Auxiliary reconstruction losses for both SC and FC are included to ensure that the learned mediator remains informative about both observed modalities rather than collapsing to a single view. Together, these steps define the mediator map Φ_θ in (1), yielding a learned network mediator that captures shared signal, modality-specific signal, and cross-modal interaction within a single connectome-valued object.

2.3 Structured pathway maps on the learned mediator

Once the learned mediator M_i is learned, mediation proceeds through two structured pathway maps. The map $D_\omega(\rho)$ describes how the exposure perturbs the learned mediator, while $W_\psi(\rho)$ describes how variation in the mediator relates to the outcome. Because the mediator is connectome-valued, these pathway maps should reflect the organization of the regions they connect rather than treat all edges as unrelated coefficients.

Each region w is associated with a feature vector ρ_w encoding anatomical and systems-level information. For each pathway type $\eta \in \{D, W\}$, region features are embedded as

$$\tilde{\rho}_w^{(\eta)} = \Psi^{(\eta)}(\rho_w), \quad (10)$$

and projected into query and key vectors,

$$q_w^{(\eta)} = U_Q^{(\eta)} \tilde{\rho}_w^{(\eta)}, \qquad k_w^{(\eta)} = U_K^{(\eta)} \tilde{\rho}_w^{(\eta)}. \quad (11)$$

The corresponding pathway map is then constructed entrywise by

$$G_{wl}^{(\eta)} = \frac{1}{2\sqrt{d_k}} \left[(q_w^{(\eta)})^\top k_l^{(\eta)} + (q_l^{(\eta)})^\top k_w^{(\eta)}\right] + r_{wl}^{(\eta)}, \qquad w \neq l, \quad (12)$$

with $G_{ww}^{(\eta)} = 0$. The exposure-to-mediator and mediator-to-outcome maps are obtained as

$$D_\omega(\rho) = G^{(D)}, \qquad W_\psi(\rho) = G^{(W)}. \quad (13)$$

Here, the residual terms $r_{wl}^{(\eta)}$ capture edge-level variation not explained by region features alone.

This construction yields pathway maps that are both edge-specific and biologically structured. Edges connecting regions with similar anatomical or systems-level roles can receive similar pathway weights, while residual terms preserve flexibility for edge-specific deviations. Sparsity regularization is imposed on the assembled pathway maps so that the inferred mediation signal is concentrated on a relatively small subset of active edges, making the resulting indirect pathway more interpretable at the connection level.

2.4 BAYESIAN INFERENCE AND POSTERIOR EFFECT ESTIMATION

Because MALM combines mediator representation learning with structured mediation on a high-dimensional network, point estimation alone is not sufficient for inference. MALM is therefore fit in a Bayesian end-to-end manner, so that uncertainty in mediator construction, pathway estimation, and effect decomposition is propagated jointly through a single posterior inference procedure rather than through a separate post hoc analysis.

Let $\Theta = \{\theta, \omega, \psi, \beta_x, \beta_z, \Sigma_M, \sigma_y^2, \vartheta_S, \vartheta_F\}$ denote the collection of model parameters, where ϑ_S and ϑ_F denote the parameters of the SC and FC reconstruction models. Since the mediator is defined by the learned mediator map $M_i = \Phi_\theta(S_i, F_i)$, posterior uncertainty is carried by Θ and propagated to the induced mediator, pathway maps, and causal effects. The posterior target is

$$p\big(\Theta \mid \{S_i, F_i, z_i, x_i, y_i\}_{i=1}^N\big) \propto \prod_{i=1}^N p(y_i \mid \Phi_\theta(S_i, F_i), z_i, x_i, \Theta)\; p(\Phi_\theta(S_i, F_i) \mid z_i, x_i, \Theta) \qquad (14)$$
$$\times\, p(S_i \mid \Phi_\theta(S_i, F_i), \vartheta_S)\; p(F_i \mid \Phi_\theta(S_i, F_i), \vartheta_F)\; p(\Theta).$$

Priors regularize all model parameters, with shrinkage on the assembled pathway maps concentrating mediation signal on active connections (Appendix D). Stochastic gradient Langevin dynamics (SGLD) samples the joint posterior. Each draw yields direct, indirect, and total effects through the formulas in Section 3, together with edge-level contributions and credible intervals. Joint estimation propagates uncertainty in the mediator and both pathways. The integrated network mediator captures coupled SC–FC information, while the single-stage effect calculation avoids the multi-stage product structure of serial mediation.

3 CAUSAL INTERPRETATION AND IDENTIFYING ASSUMPTIONS

3.1 THE INTERVENTION TARGET AND NATURAL EFFECTS

MALM defines mediation through a learned network mediator that integrates structural and functional connectivity. To specify the intervention target, fix a population mediator map Φ_{θ_0} and hold it unchanged across interventions. Treat SC and FC as post-exposure measurements and define

$$M_i(a) = \Phi_{\theta_0}(S_i(a), F_i(a)), \qquad a \in \{z, z^*\}.$$

Thus, a deterministic encoder can produce an exposure-responsive mediator through changes in its inputs. Exposure need not be an encoder input. We suppress θ_0 below and write $\mathcal{M}_{\theta_0} \subseteq \mathrm{Sym}_0(R)$ for the attainable mediator values. The following assumptions connect this representation to the natural mediation effects.

Assumption 1 (Consistency, no interference, and representation sufficiency)**.** Exposure and mediator interventions are well defined and do not affect other subjects. We have $M_i = M_i(z_i)$ and $y_i = Y_i(z_i, M_i(z_i))$. Writing $Y_i(a, s, f)$ for the outcome under a joint SC–FC intervention, we require $Y_i(a, s, f) = Y_i(a, \Phi_{\theta_0}(s, f))$. Pairs with the same mediator value have the same outcome at fixed exposure. This condition formalizes the mediator as a sufficient representation of the paired networks for the outcome under intervention. Adjustment covariates x_i are unaffected by exposure.

For exposure levels z and z^*, the conditional natural direct, natural indirect, and total effects are

$$\mathrm{NDE}(z, z^*) = \mathbb{E}[Y_i(z, M_i(z^*)) - Y_i(z^*, M_i(z^*)) \mid x_i],$$
$$\mathrm{NIE}(z, z^*) = \mathbb{E}[Y_i(z, M_i(z)) - Y_i(z, M_i(z^*)) \mid x_i],$$
$$\mathrm{TE}(z, z^*) = \mathbb{E}[Y_i(z, M_i(z)) - Y_i(z^*, M_i(z^*)) \mid x_i].$$

These contrast the component bypassing the mediator, the component transmitted through it, and their sum. Conditioning on $x_i = x$ is implicit. Population effects average these contrasts over the target covariate distribution.

3.2 IDENTIFICATION FROM OBSERVED DATA

We state the remaining conditions directly for the induced mediator, following the potential-outcomes mediation framework (Imai et al., 2010; VanderWeele & Vansteelandt, 2014). They must hold for all $a, b \in \{z, z^*\}$, relevant x, and attainable m.

Assumption 2 (Exposure and mediator overlap)**.** Both exposure levels have positive probability conditional on $x_i = x$. The conditional mediator law under $z_i = b$ is absolutely continuous with respect to that under $z_i = a$. Mediator values used in each cross-exposure contrast must lie in the support where the corresponding outcome regression is identified.

Assumption 3 (Exposure exchangeability)**.** There is no unmeasured exposure–outcome or exposure–mediator confounding after adjustment, so

$$Y_i(a, m) \perp\!\!\!\perp z_i \mid x_i, \qquad M_i(b) \perp\!\!\!\perp z_i \mid x_i.$$

Assumption 4 (Mediator exchangeability and cross-world independence)**.** We require both

$$Y_i(a, m) \perp\!\!\!\perp M_i \mid z_i = a, x_i, \qquad Y_i(a, m) \perp\!\!\!\perp M_i(b) \mid x_i.$$

The first condition specifies mediator–outcome exchangeability within exposure strata. The second links potential outcomes and potential mediators across exposure levels. We assume that mediator–outcome confounders are unaffected by exposure.

Let $g(a, m, x) = \mathbb{E}[y_i \mid z_i = a, M_i = m, x_i = x]$. Under Assumptions 1–4, the mediation formula identifies

$$\begin{aligned} \eta(a, b; x) &:= \mathbb{E}[Y_i(a, M_i(b)) \mid x_i = x] \\ &= \int_{\mathcal{M}_{\theta_0}} g(a, m, x)\, dP(m \mid z_i = b, x_i = x). \end{aligned} \tag{15}$$

Hence NDE, NIE, and TE are respectively $\eta(z, z^*; x) - \eta(z^*, z^*; x)$, $\eta(z, z; x) - \eta(z, z^*; x)$, and $\eta(z, z; x) - \eta(z^*, z^*; x)$. Appendix A gives the identification steps. Placing the exchangeability conditions directly on M_i aligns identification with the network representation used for mediation.

3.3 Closed-form effects and network attribution

The mediator and outcome layers provide a tractable link between the mediator and the mediation effects through the conditional means

$$\mathbb{E}[M_i \mid z_i = a, x_i = x] = \mu(x) + aD, \qquad g(a, m, x) = x^\top \beta_x + \beta_z a + \langle W, m \rangle_F. \tag{16}$$

The additive outcome layer separates the direct exposure contribution from the contribution through the mediator, while the mediator layer encodes exposure-related changes through D. Together, these specifications yield closed-form effects and an edge-level decomposition. The derivation uses conditional means, separating the effect calculation from the Gaussian distributional specification used for Bayesian estimation.

Proposition 3.1 (Closed-form effects under identifying assumptions)**.** *Under Assumptions 1–4, finite first moments, and* (16)*,*

$$\begin{aligned} \mathrm{NDE}(z, z^*) &= \beta_z(z - z^*), \qquad \mathrm{NIE}(z, z^*) = (z - z^*)\langle W, D \rangle_F, \\ \mathrm{TE}(z, z^*) &= \big(\beta_z + \langle W, D \rangle_F\big)(z - z^*). \end{aligned} \tag{17}$$

For $w < l$, the edge contribution is $\mathrm{NIE}_{wl}(z, z^*) = 2(z - z^*)W_{wl}D_{wl}$, and these contributions sum to NIE.

The edge contributions localize the joint-network mediation effect through an additive decomposition that retains the network as the intervention target. This connects the global NIE to interpretable connection-level contributions. Appendix A gives the identification argument and the derivation.

4 Synthetic Experiments

We compare MALM with four baselines across seven synthetic scenarios. **L**inear **A**dditive (LA) uses independent SC and FC with an additive outcome. **C**ross**M**odal **I**nteraction (CMI-L, CMI-H) introduces SC–FC interactions in the outcome at low and high signal magnitudes. **S**erial **R**elationship (SR-L, SR-H) specifies exposure-to-SC-to-FC-to-outcome pathways with linear SC–FC coupling. **N**onlinear **SCFC** coupling (NSF-L, NSF-H) replaces that linear mapping with nonlinear graph propagation. These settings test additive, interacting, ordered, and nonlinearly coupled modalities. Appendix B gives the data-generating details.

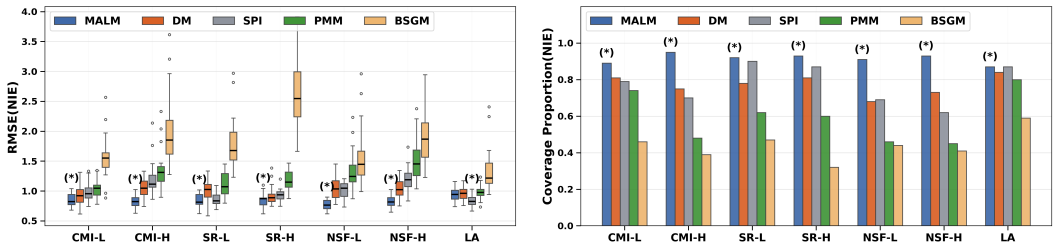


Figure 2: NIE RMSE distributions across scenarios for all five methods (left) and the corresponding coverage proportions of NIE interval estimates (right). (*) mark the best-performing method within each scenario for the corresponding criterion.

The baselines span several representative directions in mediation analysis. BSGM (Tian et al., 2024) is a Bayesian subgraph-selection method for single-modality network mediators. DM (Xu et al., 2022) uses deep nuisance estimation under the standard mediation formula. Parallel multimodal mediation (PMM) (Zhao et al., 2021) treats structural and functional connectivity as separate mediators in a joint pathway model. Sequential pathway inference (SPI) (Li et al., 2022) studies multimodal mediation under an ordered serial formulation. The four baselines together represent single-modality, deep-learning-based, parallel-multimodal, and serial-multimodal mediation paradigms.

Figure 2 reports NIE RMSE and empirical coverage of 95% intervals. MALM achieves the lowest RMSE, smallest dispersion, and coverage closest to nominal in nearly every scenario. PMM degrades sharply under cross-modal interactions, where additive separation misses the joint signal captured by MALM's encoder. SPI has the largest spread and most severe under-coverage in serial and nonlinear-coupling settings, especially when its linear SC→FC stage is misspecified. MALM's single-stage effect calculation (Proposition 3.1) avoids a multi-stage product and requires no inter-modality ordering. Single-modality BSGM performs worst when the mediation signal depends on both networks. Flexible DM also trails MALM, supporting the value of coupling the learned network mediator to sparsity-regularized pathway maps D_ω and W_ψ. On LA, MALM matches or exceeds the best baseline on every metric, retaining accuracy when a simpler additive structure holds.

5 Real Data Analysis

5.1 Data overview

We evaluated MALM on paired diffusion MRI and resting-state fMRI from the Human Connectome Project (HCP) S1200 (Van Essen et al., 2013) and the Adolescent Brain Cognitive Development (ABCD) study (Casey et al., 2018). SC and FC matrices used the Desikan–Killiany parcellation. The exposure was *BDNF* Val66Met (rs6265), coded as Met carriers versus Val/Val homozygotes, and the outcome was the Cognition Total Composite Score. Covariates included age, sex, handedness, in-scanner head motion, and the top 10 genetic principal components. After quality control and requiring both modalities, the samples contained $N = 1032$ HCP and $N = 2257$ ABCD participants. These cohorts test effect estimation and pathway reproducibility across youth and young adulthood. Appendix C details preprocessing and covariates.

5.2 Results

Applying MALM to the two cohorts yields statistically significant mediation effects in both datasets. In HCP, the estimated NIE is -0.16 $(-0.22, -0.10)$ and the NDE is -0.26 $(-0.38, -0.14)$. In ABCD, the corresponding estimates are -0.18 $(-0.24, -0.12)$ and -0.23 $(-0.32, -0.14)$. Both estimated effects are negative with intervals excluding zero in each cohort. As shown in Figure 3, MALM is the only approach that recovers a significant NIE in both HCP and ABCD. All competing methods yield confidence or credible intervals that cover zero.

The same figure also reports the coefficient of variation of the NIE estimate, $\mathrm{CV}_{\mathrm{NIE}}$, across repeated subsamples, with lower values indicating greater estimation stability. This is an informative quantity here because it reflects whether the recovered mediation signal is reproducible under subject-level

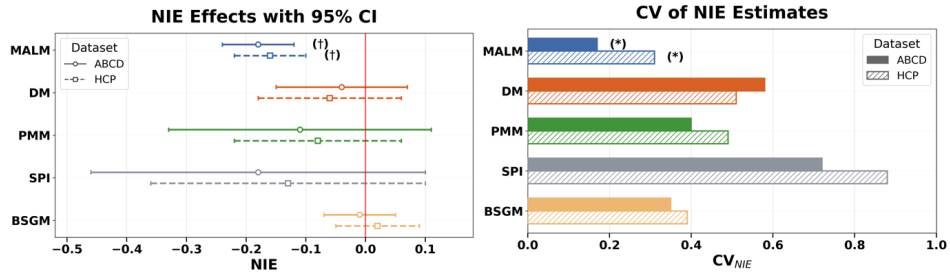


Figure 3: NIE estimates with 95% confidence/credible intervals (left) and the corresponding $\mathrm{CV}_{\mathrm{NIE}}$ values (right). Dagger symbols (†) indicate statistically significant NIE results, and asterisks (*) denote the smallest $\mathrm{CV}_{\mathrm{NIE}}$ within each dataset.

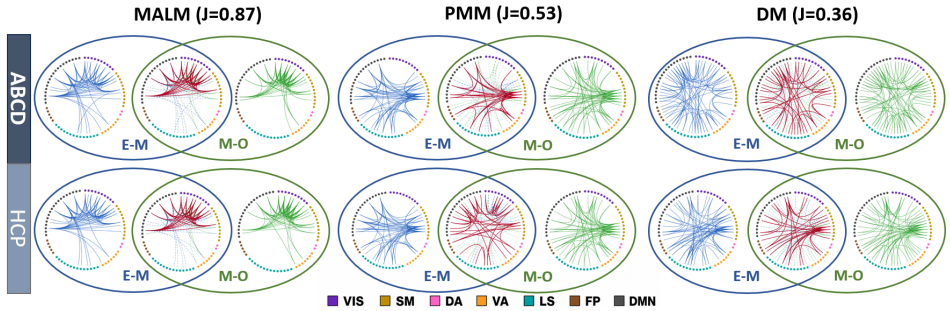


Figure 4: Top 10% edge-level pathways identified by MALM and competing methods. Blue edges indicate exposure-to-mediator effects, green edges indicate mediator-to-outcome effects, and red edges mark their overlap, corresponding to active mediating connections. Dashed edges denote non-overlapping pathway-specific effects. J is the Jaccard index for cross-cohort overlap between HCP and ABCD.

perturbation rather than driven by a small or unstable subset of the data. Among the methods with significant NIE estimates, MALM attains the lowest $\mathrm{CV}_{\mathrm{NIE}}$, indicating that its inferred pathway is not only detectable but also stable. The negative NIE estimated by MALM is, moreover, consistent with prior literature showing that the Met allele is associated with reduced activity-dependent *BDNF* secretion, weaker synaptic plasticity, and poorer memory-related function (Egan et al., 2003; Hariri et al., 2003). These findings suggest a biologically plausible pathway through which *BDNF*-related variation shapes how large-scale brain circuits develop and coordinate their signaling. Reduced neurotrophic support could weaken structural communication pathways, disrupt functional coordination across cognitive systems, and ultimately impair cognitive performance.

Complementing the effect estimates in Figure 3, Figure 4 localizes the corresponding mediation pathways at the edge level. In MALM, we define active mediating connections as edges that appear in both the exposure-to-mediator and the mediator-to-outcome pathway. The red overlapping edges therefore mark the subset of connections most directly implicated in carrying the *BDNF*–cognition association. This overlap is important because it distinguishes edges that are merely exposure-responsive or outcome-related from those that participate in the full mediation pathway. To quantify cross-cohort reproducibility, we compute the Jaccard index between the significant mediating edge sets identified in HCP and ABCD. MALM attains the highest consistency at $J = 0.87$, compared with $J = 0.53$ for PMM and $J = 0.36$ for DM. The strong cross-cohort agreement suggests that MALM stabilizes mediation analysis by projecting noisy and partially inconsistent observed connectomes into an integrated network mediator. From this perspective, learning an integrated network representation helps filter modality-specific and cohort-specific imaging variability and recover a mediation signal that is robust across neurodevelopmental stages, from youth in ABCD to young adults in HCP. This robustness is important from a biomedical standpoint because pathways that persist across developmental contexts are more likely to reflect generalizable network mechanisms rather than cohort-specific effects, making them more informative for mechanistic studies, biomarker discovery, and intervention-oriented research.

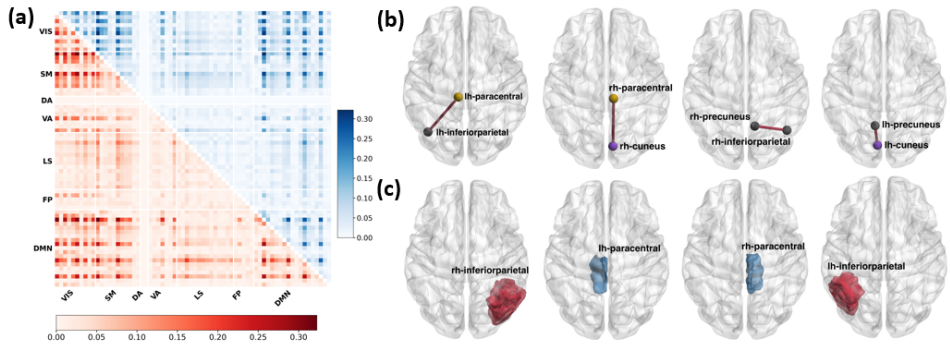


Figure 5: Cross-cohort summaries of the active mediation pathway identified by MALM. (a) Heatmap of active mediation-effect magnitudes, with HCP shown in the upper triangle and ABCD in the lower triangle. (b) Top four edges with the largest mediation weights across the two cohorts. (c) Top four regions with the largest average connected weight across the two cohorts.

Across both cohorts, the recovered mediation signal concentrates within and between the default mode, visual, and somatomotor systems (Figure 5). This pattern is biologically plausible because *BDNF* Val66Met has been linked to altered resting-state functional connectivity, including default-mode organization (Thomason et al., 2009; Wang et al., 2014). These same systems are central to cognition, supporting memory, internally directed cognition, visual attention, visuospatial processing, and coordinated sensorimotor behavior (Smallwood et al., 2021; Bullmore & Sporns, 2009; Park & Friston, 2013; Suárez et al., 2020). Consistent with this systems-level picture, Figures 5(b,c) show that the strongest edges and hubs are concentrated in inferior parietal, precuneus, cuneus, and paracentral regions. These regions are also well supported by prior work, including evidence linking *BDNF* variation to altered posterior cortical morphology and connectivity (Wang et al., 2014), implicating the cuneus as an imaging mediator of genetic effects on working memory (He et al., 2021), and showing altered motor-cortical plasticity and motor-system function associated with *BDNF* Val66Met (Kleim et al., 2006; McHughen et al., 2010). Functionally, inferior parietal and precuneus regions are major association hubs for attention and episodic retrieval (Igelström & Graziano, 2017; Hutchinson et al., 2009), while cuneus and paracentral regions contribute to visual and sensorimotor processes that support cognitive performance. Together, the recovered pathway is organized around a compact set of biologically meaningful posterior associative, visual, and sensorimotor hubs. This biological coherence highlights the relevance of MALM beyond *BDNF*. Dependent measurements such as paired imaging, multi-omics profiles, physiological readouts, and longitudinal biomarkers can jointly inform a mediation pathway. Learning an integrated network mediator connects these measurements to closed-form effect decomposition and posterior uncertainty quantification. Promising applications include disease progression, biomarker-mediated treatment effects, and pathways linking molecular variation to clinical outcomes.

6 DISCUSSION

MALM's central contribution is to learn the paired-connectome representation directly as the mediator. Joint mediator and pathway estimation links multimodal integration to global effects and connection-level contributions within a common network coordinate system. At the same time, several limitations remain. The current formulation is tailored to a single binary exposure and paired connectivity modalities, whereas many biomedical applications involve multiple genetic variants, longitudinal exposures, richer multimodal measurements, or settings in which the mediator and outcome evolve over time. Although the recovered pathways are stable at the system level and reproducible across cohorts, finer edge-level patterns may still vary because of differences in sample composition, preprocessing choices, and measurement noise. These limitations suggest several directions for future work, including extensions to multi-variant and longitudinal settings, incorporation of additional biological modalities such as transcriptomic or epigenomic measurements, and development of more robust edge-selection procedures that better stabilize fine-scale pathway recovery.

AI USE STATEMENT

Generative AI tools were used to assist with converting the manuscript's LaTeX source to the ICLR 2027 template, checking typesetting, and polishing the writing.

ETHICS STATEMENT

This study is a secondary analysis of pre-existing Human Connectome Project (HCP) and Adolescent Brain Cognitive Development (ABCD) data, investigating the relationship between genetic variation, paired structural and functional brain connectivity, and cognition. Access to and use of cohort data are governed by the respective cohorts' data-access and privacy requirements. The manuscript reports aggregate effect estimates and network summaries and does not include individual-level participant data. The findings are intended for research on brain-network mediation and are not validated for individual diagnosis, prognosis, or clinical decision-making.

REPRODUCIBILITY STATEMENT

Section 2 specifies the model and posterior inference objective. Appendix A provides the identification arguments, effect derivations, and encoder–mediation clarification. Appendix B describes the synthetic data-generating processes and evaluation metrics. Appendix C documents data preparation, training settings, hyperparameter selection, and computational requirements. Appendix D specifies the priors and shrinkage, Appendix E details the mediator construction and reconstruction objective, and Appendix F reports the ablation experiments. Source code will be released upon publication.

REFERENCES

Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009. doi: 10.1038/nrn2575.

B. J. 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, Catherine A. Orr, Tor D. Wager, Marie T. Banich, Nicole K. Speer, Matthew T. Sutherland, Michael C. Riedel, Anthony S. Dick, James M. Bjork, Kathleen M. Thomas, Bader Chaarani, Margie H. Mejia, Donald J. Hagler, Jr., M. Daniela Cornejo, Chelsea S. Sicat, Michael P. Harms, Nico U. F. Dosenbach, Monica Rosenberg, Eric Earl, Hauke Bartsch, Richard Watts, Jonathan R. Polimeni, Joshua M. Kuperman, Damien A. Fair, Anders M. Dale, and ABCD Imaging Acquisition Workgroup. The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, 32:43–54, 2018. doi: 10.1016/j.dcn.2018.03.001.

Oliver Y. Chén, Ciprian Crainiceanu, Elizabeth L. Ogburn, Brian S. Caffo, Tor D. Wager, and Martin A. Lindquist. High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, 19(2):121–136, 2018.

Rhian M. Daniel, Bianca L. De Stavola, Simon N. Cousens, and Stijn Vansteelandt. Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14, 2015.

Michael F. Egan, Masami Kojima, Joseph H. Callicott, Terry E. Goldberg, Bhaskar S. Kolachana, Alessandro Bertolino, Eugene Zaitsev, Bert Gold, David Goldman, Michael Dean, Bai Lu, and Daniel R. Weinberger. The BDNF val66met polymorphism affects activity-dependent secretion of BDNF and human memory and hippocampal function. *Cell*, 112(2):257–269, 2003. doi: 10.1016/S0092-8674(03)00035-7.

Simiao Gao, Zhiling Gu, Shengxian Ding, Gefei Wang, Zhengwu Zhang, Hongyu Zhao, and Yize Zhao. Brain functional–structural gradient coupling reflects development, behavior and genetic influences. *Nature Communications*, 17:4850, 2026. doi: 10.1038/s41467-026-71719-y.

Joaquín Goñi, Martijn P. van den Heuvel, Andrea Avena-Koenigsberger, Nieves Vélez de Mendizabal, Richard F. Betzel, Alessandra Griffa, Patric Hagmann, Bernat Corominas-Murtra, Jean-Philippe Thiran, and Olaf Sporns. Resting-brain functional connectivity predicted by analytic measures of network communication. *Proceedings of the National Academy of Sciences*, 111(2):833–838, 2014. doi: 10.1073/pnas.1315529111.

Ahmad R. Hariri, Terry E. Goldberg, Venkata S. Mattay, Bhaskar S. Kolachana, Joseph H. Callicott, Michael F. Egan, and Daniel R. Weinberger. Brain-derived neurotrophic factor val66met polymorphism affects human memory-related hippocampal activity and predicts memory performance. *The Journal of Neuroscience*, 23(17):6690–6694, 2003. doi: 10.1523/JNEUROSCI.23-17-06690.2003.

Xiaoxi He, Xi Li, Jilian Fu, Jiayuan Xu, Huaigui Liu, Peng Zhang, Wei Li, Chunshui Yu, Zhaoxiang Ye, and Wen Qin. The morphometry of left cuneus mediating the genetic regulation on working memory. *Human Brain Mapping*, 42(11):3470–3480, 2021.

Christopher J. Honey, Olaf Sporns, Leila Cammoun, Xavier Gigandet, Jean-Philippe Thiran, Reto Meuli, and Patric Hagmann. Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6):2035–2040, 2009. doi: 10.1073/pnas.0811168106.

J. Benjamin Hutchinson, Melina R. Uncapher, and Anthony D. Wagner. Posterior parietal cortex and episodic retrieval: convergent and divergent effects of attention and memory. *Learning & Memory*, 16(6):343–356, 2009. doi: 10.1101/lm.919109.

Kajsa M. Igelström and Michael S. A. Graziano. The inferior parietal lobule and temporoparietal junction: A network perspective. *Neuropsychologia*, 105:70–83, 2017. doi: 10.1016/j.neuropsychologia.2017.01.001.

Kosuke Imai and Teppei Yamamoto. Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2):141–171, 2013.

Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, 2010.

Jeffrey A. Kleim, Sheila Chan, Erin Pringle, Kellan Schallert, Vincent Procaccio, Richard Jimenez, and Steven C. Cramer. BDNF val66met polymorphism is associated with modified experience-dependent plasticity in human motor cortex. *Nature Neuroscience*, 9(6):735–737, 2006. doi: 10.1038/nn1699.

Lexin Li, Chengchun Shi, Tengfei Guo, and William J. Jagust. Sequential pathway inference for multimodal neuroimaging analysis. *Stat*, 11(1):e433, 2022. doi: 10.1002/sta4.433.

Martin A. Lindquist. Functional causal mediation analysis with an application to brain connectivity. *Journal of the American Statistical Association*, 107(500):1297–1309, 2012.

Stephanie A. McHughen, Paul F. Rodriguez, Jeffrey A. Kleim, Erin D. Kleim, Laura Marchal Crespo, Vincent Procaccio, and Steven C. Cramer. BDNF val66met polymorphism influences motor system function in the human brain. *Cerebral Cortex*, 20(5):1254–1262, 2010. doi: 10.1093/cercor/bhp189.

Hae-Jeong Park and Karl Friston. Structural and functional brain networks: From connections to cognition. *Science*, 342(6158):1238411, 2013. doi: 10.1126/science.1238411.

Tabinda Sarwar, Ye Tian, B. T. Thomas Yeo, Kotagiri Ramamohanarao, and Andrew Zalesky. Structure-function coupling in the human connectome: A machine learning approach. *NeuroImage*, 226:117609, 2021. doi: 10.1016/j.neuroimage.2020.117609.

Caio Seguin, Ye Tian, and Andrew Zalesky. Network communication models improve the behavioral and functional predictive utility of the human structural connectome. *Network Neuroscience*, 4(4):980–1006, 2020.

Jonathan Smallwood, Boris C. Bernhardt, Robert Leech, Danilo Bzdok, Elizabeth Jefferies, and Daniel S. Margulies. The default mode network in cognition: a topographical perspective. *Nature Reviews Neuroscience*, 22:503–513, 2021. doi: 10.1038/s41583-021-00474-4.

Laura E. Suárez, Ross D. Markello, Richard F. Betzel, and Bratislav Misic. Linking structure and function in macroscale brain networks. *Trends in Cognitive Sciences*, 24(4):302–315, 2020. doi: 10.1016/j.tics.2020.01.008.

Moriah E. Thomason, Daniel J. Yoo, Gary H. Glover, and Ian H. Gotlib. BDNF genotype modulates resting functional connectivity in children. *Frontiers in Human Neuroscience*, 3:55, 2009. doi: 10.3389/neuro.09.055.2009.

Xinyuan Tian, Fan Li, Li Shen, Denise Esserman, and Yize Zhao. Bayesian pathway analysis over brain network mediators for survival data. *Biometrics*, 80(4):ujae132, 2024.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy E. J. Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Tyler J. VanderWeele and Stijn Vansteelandt. Mediation analysis with multiple mediators. *Epidemiologic Methods*, 2(1):95–115, 2014.

Bertha Vázquez-Rodríguez, Laura E. Suárez, Ross D. Markello, Golia Shafiei, Casey Paquola, Patric Hagmann, Martijn P. van den Heuvel, Boris C. Bernhardt, R. Nathan Spreng, and Bratislav Misic. Gradients of structure–function tethering across neocortex. *Proceedings of the National Academy of Sciences*, 116(42):21219–21227, 2019. doi: 10.1073/pnas.1903403116.

Chao Wang, Yuanchao Zhang, Bing Liu, Haixia Long, Chunshui Yu, and Tianzi Jiang. Dosage effects of BDNF Val66Met polymorphism on cortical surface area and functional connectivity. *Journal of Neuroscience*, 34(7):2645–2651, 2014. doi: 10.1523/JNEUROSCI.3501-13.2014.

Jing Xia, Yi Hao Chan, Deepank Girish, and Jagath C. Rajapakse. Interpretable modality-specific and interactive graph convolutional network on brain functional and structural connectomes. *Medical Image Analysis*, 102:103509, 2025.

Siqi Xu, Lin Liu, and Zhonghua Liu. DeepMed: Semiparametric causal mediation analysis with debiased deep learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Andrew Zalesky, Tabinda Sarwar, Ye Tian, Yuanzhe Liu, B. T. Thomas Yeo, and Kotagiri Ramamohanarao. Predicting an individual's functional connectivity from their structural connectome: Evaluation of evidence, recommendations, and future prospects. *Network Neuroscience*, 8(4): 1291–1309, 2024. doi: 10.1162/netn_a_00400.

Farnaz Zamani Esfahlani, Joshua Faskowitz, Jonah Slack, Bratislav Mišić, and Richard F. Betzel. Local structure–function relationships in human brain networks across the lifespan. *Nature Communications*, 13(1):2053, 2022. doi: 10.1038/s41467-022-29770-y.

Yi Zhao, Martin A. Lindquist, and Brian S. Caffo. Sparse principal component based high-dimensional mediation analysis. *Computational Statistics & Data Analysis*, 142:106835, 2020.

Yi Zhao, Lexin Li, and Brian S. Caffo. Multimodal neuroimaging data integration and pathway analysis. *Biometrics*, 77(3):879–889, 2021. doi: 10.1111/biom.13351.

A IDENTIFICATION AND DERIVATION OF THE MEDIATION EFFECTS

A.1 IDENTIFICATION FOR A FIXED MEDIATOR MAP

Fix the population representation Φ_{θ_0} and condition on $x_i = x$. All potential outcomes and distributions in this appendix refer to that same map. The integration domain $\mathcal{M}_{\theta_0}$ is the attainable support of the mediator representation. For $a, b \in \{z, z^*\}$, iterated expectation gives

$$\begin{aligned}\eta(a, b; x) &= \int_{\mathcal{M}_{\theta_0}} \mathbb{E}[Y_i(a, m) \mid M_i(b) = m, x_i = x]\, dP(M_i(b) = m \mid x_i = x) \\ &= \int_{\mathcal{M}_{\theta_0}} \mathbb{E}[Y_i(a, m) \mid x_i = x]\, dP(M_i(b) = m \mid x_i = x). \end{aligned} \tag{18}$$

The second equality uses the cross-world part of Assumption 4. The observed-data outcome regression follows from the equalities below.

$$\begin{aligned}\mathbb{E}[Y_i(a, m) \mid x_i = x] &= \mathbb{E}[Y_i(a, m) \mid z_i = a, x_i = x] \\ &= \mathbb{E}[Y_i(a, m) \mid z_i = a, M_i = m, x_i = x] \\ &= \mathbb{E}[y_i \mid z_i = a, M_i = m, x_i = x] = g(a, m, x). \end{aligned} \tag{19}$$

These equalities use exposure exchangeability, within-exposure mediator exchangeability, and consistency, respectively. Likewise, exposure exchangeability for the mediator and consistency imply

$$P(M_i(b) \in A \mid x_i = x) = P(M_i \in A \mid z_i = b, x_i = x) \tag{20}$$

for each measurable set A. Assumption 2 ensures that the observed regressions are identified on the supports over which they are integrated. Substituting (19) and (20) into (18) gives (15). This argument connects the learned network representation to the standard mediation identification formula (Imai et al., 2010).

A.2 CONDITIONAL-MEAN CALCULATION AND EDGE ATTRIBUTION

Under (16), linearity of expectation and the Frobenius inner product yields

$$\begin{aligned}\eta(a, b; x) &= \int_{\mathcal{M}_{\theta_0}} \left(x^\top \beta_x + \beta_z a + \langle W, m \rangle_F\right)\, dP(m \mid z_i = b, x_i = x) \\ &= x^\top \beta_x + \beta_z a + \langle W, \mathbb{E}[M_i \mid z_i = b, x_i = x] \rangle_F \\ &= x^\top \beta_x + \beta_z a + \langle W, \mu(x) + bD \rangle_F. \end{aligned} \tag{21}$$

Taking the three contrasts defined in Section 3 gives

$$\mathrm{NDE}(z, z^*) = \eta(z, z^*; x) - \eta(z^*, z^*; x) = \beta_z(z - z^*), \tag{22}$$

$$\mathrm{NIE}(z, z^*) = \eta(z, z; x) - \eta(z, z^*; x) = (z - z^*)\langle W, D \rangle_F, \tag{23}$$

$$\mathrm{TE}(z, z^*) = \eta(z, z; x) - \eta(z^*, z^*; x) = \left(\beta_z + \langle W, D \rangle_F\right)(z - z^*). \tag{24}$$

Because the specified D, W, and β_z do not depend on x, these expressions also equal the covariate-averaged effects under this model. If the coefficients are allowed to vary with covariates, the conditional expressions must instead be averaged over the target population.

Finally, symmetry and zero diagonals imply

$$\langle W, D \rangle_F = \sum_{w=1}^{R} \sum_{l=1}^{R} W_{wl} D_{wl} = 2 \sum_{w<l} W_{wl} D_{wl}.$$

Consequently, defining $\mathrm{NIE}_{wl}(z, z^*) = 2(z - z^*) W_{wl} D_{wl}$ proves Proposition 3.1. The decomposition accommodates dependence among mediator edges and attributes the joint network effect to individual connections. Each term quantifies a connection's contribution within the network intervention defined in Section 3.

A.3 Connecting mediator representation learning and causal mediation

The integrated network mediator provides a common intervention target for paired SC and FC measurements. Assumption 1 formalizes its role as an outcome-relevant representation under intervention, so that paired configurations with the same mediator value share the same potential outcome at fixed exposure. Stating exchangeability directly for this representation aligns the causal estimand with the mediator used by MALM.

Under exposure level a, the mediator distribution is induced by passing $(S_i(a), F_i(a))$ through the fixed map Φ_{θ_0}. Integrating over $\mathcal{M}_{\theta_0}$ keeps the identification formula tied to the attainable mediator values. The conditional-mean specification in (16) then reduces this network-valued integral to $\langle W, D\rangle_F$, providing a direct link between the exposure-to-mediator and mediator-to-outcome pathways. This calculation accommodates dependent network edges and requires only finite first moments once the stated mean and identification conditions hold. The Gaussian specification in (3) supplies the distributional model used for Bayesian estimation.

The fixed population map defines the mediation target, while posterior computation describes uncertainty under the fitted model. Each posterior draw of the pathway maps and exposure coefficient yields corresponding direct, indirect, and total effects through Proposition 3.1. Their edge-level decomposition links global mediation estimates to the connections contributing to the shared network pathway. This organization combines a flexible representation of paired connectomes with explicit effect formulas and interpretable network attribution.

A.4 Connecting the encoder and mediation layers

For fixed model parameters, the encoder determines M_i from a given pair (S_i, F_i), while variation in the observed connectomes induces variation in M_i across subjects. The mediator layer describes this population variation conditional on exposure and covariates. Its conditional-mean specification is

$$\mathbb{E}[\Phi_\theta(S_i, F_i) \mid z_i, x_i] = \mu(x_i) + z_i D_\omega(\rho).$$

For binary exposure, this gives

$$D_\omega(\rho) = \mathbb{E}[M_i \mid z_i = 1, x_i] - \mathbb{E}[M_i \mid z_i = 0, x_i].$$

Thus, D_ω represents the exposure-associated change in the encoded network. The model takes this contrast to be common across covariate values, with baseline variation captured by $\mu(x_i)$. The outcome map W_ψ weights these changes on the same region-pair coordinates, connecting mediator representation learning to the indirect-effect calculation.

The Gaussian law in (3) is a distributional assumption for the encoded networks conditional on exposure and covariates. It is specified as part of the statistical model rather than derived from the encoder architecture. Its residual variation describes differences in the encoded inputs among subjects with the same exposure and covariates. The effect derivation uses the stated conditional means, while joint estimation couples mediator construction to both mediation pathways.

B Synthetic study design

This appendix provides the data-generating process used in the synthetic study. Each simulated dataset consists of N subjects with exposure z_i, covariates x_i, paired connectivity matrices (S_i, F_i), and outcome y_i. All scenarios are generated on symmetric network-valued mediators with zero diagonal, and differ only in the dependence structure between SC and FC and in the way the two modalities contribute to the outcome.

B.1 Common setup

For each scenario, we generated $N = 1000$ subjects with paired connectivity matrices of dimension $R = 68$, matching the Desikan–Killiany resolution used in the real-data analysis. For subject $i = 1, \ldots, N$, the exposure was generated as

$$z_i \sim \text{Bernoulli}(0.5),$$

and the covariate vector was generated as

$$x_i \sim \mathcal{N}(0, I_Q), \qquad Q = 5,$$

independently across subjects. The choice $Q = 5$ was used to mirror the scale of covariate adjustment in the real-data analysis. All structural and functional connectivity matrices were generated as symmetric 68×68 matrices with zero diagonal. To construct matrix-valued noise terms, we first sampled the strict upper-triangular entries independently from Gaussian distributions and then symmetrized them, setting the diagonal to zero. The same values of N, R, and Q were used across all scenarios, so that differences in performance arise from the multimodal mediation structure rather than from dimensionality or sample-size changes.

Each scenario was repeated over 100 Monte Carlo replicates. Within a scenario, all generative parameters were held fixed across replicates except for independent noise realizations. For the low- and high-signal versions of a scenario, the same data-generating form was used and only the signal magnitude was changed.

B.2 Linear additive scenario (La)

The La scenario is designed as a baseline setting with independent modality contributions and an additive outcome model. Structural and functional connectivity are generated as

$$S_i = B_S x_i + c z_i D_S + E_i^S, \tag{25}$$

$$F_i = B_F x_i + c z_i D_F + E_i^F, \tag{26}$$

where B_S and B_F are fixed coefficient operators mapping covariates to symmetric matrices, and E_i^S, E_i^F are independent symmetric noise matrices. The outcome is generated by

$$y_i = x_i^\top \beta_x + \beta_z z_i + \langle W_S, S_i \rangle_F + \langle W_F, F_i \rangle_F + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma_y^2). \tag{27}$$

This setting favors additive multimodal formulations and serves as a reality check for whether Malm loses efficiency when the true data-generating mechanism is simple.

B.3 Cross-modal interaction scenarios (Cmi-l and Cmi-h)

The Cmi scenarios are designed so that the outcome depends on the joint organization of SC and FC rather than on additive modality-specific contributions alone. The connectomes are generated as in the La setting,

$$S_i = B_S x_i + c z_i D_S + E_i^S, \tag{28}$$

$$F_i = B_F x_i + c z_i D_F + E_i^F, \tag{29}$$

but the outcome includes an additional interaction term,

$$y_i = x_i^\top \beta_x + \beta_z z_i + \langle W_S, S_i \rangle_F + \langle W_F, F_i \rangle_F + \langle W_{SF},\, S_i \odot F_i \rangle_F + \varepsilon_i, \tag{30}$$

where $\odot$ denotes the Hadamard product. The low- and high-signal versions differ only in the signal multiplier c, with W_{SF} held fixed. These scenarios test whether a method can recover mediation effects when the relevant signal is distributed across the coupled multimodal structure.

B.4 Serial relationship scenarios (Sr-l and Sr-h)

The Sr scenarios are designed to favor a serial mediation interpretation. Structural connectivity is first generated from exposure and covariates,

$$S_i = B_S x_i + c z_i D_S + E_i^S, \tag{31}$$

and functional connectivity is then generated conditionally on S_i,

$$F_i = B_F x_i + A S_i A^\top + E_i^F, \tag{32}$$

where A is a fixed propagation matrix. The outcome depends primarily on the downstream functional mediator,

$$y_i = x_i^\top \beta_x + \beta_z z_i + \langle W_F, F_i \rangle_F + \varepsilon_i. \tag{33}$$

The low- and high-signal settings differ through the scaling factor c. These scenarios are included to examine the behavior of Malm relative to methods that impose an ordered multimodal decomposition.

B.5 NONLINEAR SC–FC COUPLING SCENARIOS (NSF-L AND NSF-H)

The NSF scenarios are designed to violate the linear serial assumption while preserving strong dependence between the two modalities. Structural connectivity is generated as

$$S_i = B_S x_i + c z_i D_S + E_i^S, \tag{34}$$

and functional connectivity is obtained through a nonlinear transformation of S_i,

$$F_i = B_F x_i + \tanh\big(A S_i A^\top\big) + E_i^F, \tag{35}$$

where the hyperbolic tangent is applied elementwise. The outcome is then generated by

$$y_i = x_i^\top \beta_x + \beta_z z_i + \langle W_S, S_i \rangle_F + \langle W_F, F_i \rangle_F + \varepsilon_i. \tag{36}$$

As before, the low- and high-signal versions differ only through the magnitude of c. These scenarios test whether a method remains stable when SC and FC are strongly coupled but the dependence cannot be represented well by a linear serial model.

B.6 SIGNAL SCALING AND IMPLEMENTATION DETAILS

Across all scenarios, the same structural templates $D_S, D_F, W_S, W_F, W_{SF}$ are used, and only the signal multiplier c is changed between low- and high-signal settings. In our implementation, all matrices are symmetrized and their diagonals are set to zero after generation. The same values of (N, R, Q), noise variances, and Monte Carlo replication count are used across methods within each scenario. Hyperparameters and optimization settings for MALM and the competing baselines are held fixed across scenarios unless otherwise stated.

B.7 EVALUATION METRICS

For each scenario, performance is evaluated over Monte Carlo replicates using the root mean squared error of the NIE estimator and the coverage proportion of the corresponding interval estimate. The RMSE reflects both bias and variability, while the coverage proportion evaluates whether the uncertainty quantification is calibrated relative to the nominal level. When available, edge-level pathway summaries are also compared across methods to assess recovery of the active mediation structure.

C REAL DATA ANALYSIS DETAILS

C.1 DATA OVERVIEW

We evaluated MALM on paired diffusion MRI and resting-state fMRI data from two large cohorts, the Human Connectome Project (HCP) S1200 release (Van Essen et al., 2013) and the Adolescent Brain Cognitive Development (ABCD) study (Casey et al., 2018). For each subject i, we constructed a structural connectivity matrix $\mathbf{S}_i \in \mathbb{R}^{68\times 68}$ and a functional connectivity matrix $\mathbf{F}_i \in \mathbb{R}^{68\times 68}$ under the Desikan–Killiany parcellation. Structural connectivity was derived from dMRI tractography by summarizing white-matter connectivity strength between each pair of regions using streamline-based weights, followed by symmetrization, removal of self-connections, and subject-wise normalization. Functional connectivity was computed from preprocessed resting-state fMRI by extracting regional time series, forming Pearson correlation matrices, applying Fisher z transformation, and symmetrizing with zero diagonal.

The exposure was the *BDNF* Val66Met polymorphism (rs6265), coded as Met carriers (Val/Met or Met/Met) versus Val/Val homozygotes, yielding a binary exposure $z_i \in \{0, 1\}$. The outcome was total cognition, measured by the Cognition Total Composite Score in both cohorts. Subjects were restricted to those with both imaging modalities available and passing quality control. The resulting analysis therefore targets whether and how *BDNF* Val66Met influences total cognition through paired structural and functional brain connectivity.

C.2 Experimental setting and implementation details

All experiments used the MALM architecture described in Section 2 and were trained with stochastic gradient Langevin dynamics (SGLD). For the HCP analysis, data were split using a stratified 60/20/20 train/validation/test partition, yielding 620 training subjects, 206 validation subjects, and 206 held-out test subjects. Hyperparameters were selected by validation RMSE over the grid

$$\texttt{lr} \in \{10^{-4}, 5 \times 10^{-4}\}, \qquad \texttt{noise_scale} \in \{0.05, 0.1, 0.2\}.$$

The selected configuration was $\texttt{lr} = 5 \times 10^{-4}$ and $\texttt{noise_scale} = 0.2$. The final model was then refit on the combined training and validation data and evaluated once on the held-out test set.

Unless otherwise noted, the training configuration used 1000 epochs, batch size 256, hidden dimension 32, attention dimension 16, penalties $\lambda_m = 1.0$, $\lambda_w = 10^{-4}$, and $\lambda_d = 10^{-4}$, with gradient clipping at maximum norm 5.0. The SGLD step used the standard gradient update together with additive Gaussian noise. The same implementation pipeline was used for the real-data analyses reported in the main text, with cohort-specific data loading and preprocessing but the same model class and inference procedure.

C.3 Compute resources

All completed experiments were run on a CPU-only Linux server with two Intel Xeon Platinum 8562Y+ sockets, for a total of 64 CPU cores, and 1.0 TiB RAM. The implementation used Python 3.11.13 and PyTorch 2.5.1 (CPU only, `cuda_available=False`). No GPU acceleration or distributed workers were used, and PyTorch training ran in a single process with in-memory minibatch loading.

For the HCP experiment, the end-to-end run completed in roughly 24 minutes of wall-clock time on this machine. Peak memory use was modest relative to server capacity because the connectome tensors are small compared with typical imaging models. In practice, a machine with at least 16 GB RAM should be sufficient to reproduce the HCP run, while 32 GB is a safer lower bound for reproducing the larger ABCD analysis.

D Prior specification

This appendix summarizes the prior distributions used in MALM. Let

$$\Theta = \{\theta, \omega, \psi, \beta_x, \beta_z, \Sigma_M, \sigma_y^2, \vartheta_S, \vartheta_F\}$$

denote the full parameter collection, where θ contains the parameters of the mediator map Φ_θ, ω and ψ denote the parameters of the pathway maps $D_\omega(\rho)$ and $W_\psi(\rho)$, β_x and β_z are regression coefficients in the outcome layer, Σ_M is the covariance parameter in the mediator layer, σ_y^2 is the outcome noise variance, and ϑ_S, ϑ_F denote the parameters of the SC and FC reconstruction models.

D.1 Gaussian priors on encoder and pathway parameters

For the neural and attention-based parameters in θ, ω, ψ, ϑ_S, and ϑ_F, we place independent zero-centered Gaussian priors,

$$\theta_j \sim \mathcal{N}(0, \tau_\theta^2), \qquad \omega_j \sim \mathcal{N}(0, \tau_\omega^2), \qquad \psi_j \sim \mathcal{N}(0, \tau_\psi^2),$$

with the same form applied componentwise to reconstruction parameters in ϑ_S and ϑ_F. In practice, these priors act as standard weight decay terms in the SGLD objective and regularize the end-to-end mediator map.

D.2 Priors on regression parameters

We assign Gaussian priors to the direct covariate and exposure effects,

$$\beta_x \sim \mathcal{N}(0, \tau_x^2 I_Q), \qquad \beta_z \sim \mathcal{N}(0, \tau_z^2).$$

These priors are weakly informative and primarily serve to stabilize estimation in finite samples.

D.3 Priors on variance parameters

For the outcome noise variance, we use an inverse-gamma prior,

$$\sigma_y^2 \sim \mathrm{IG}(a_y, b_y).$$

For the mediator-layer covariance parameter Σ_M, we use an isotropic specification,

$$\Sigma_M = \sigma_M^2 I,$$

and place an inverse-gamma prior on the variance parameter,

$$\sigma_M^2 \sim \mathrm{IG}(a_M, b_M).$$

D.4 Shrinkage on pathway maps

The pathway maps D and W are not estimated as free matrices, but are constructed deterministically from region features through the attention-based maps $\mathcal{A}_D(\rho)$ and $\mathcal{A}_W(\rho)$. As a result, sparsity is not imposed by assigning independent priors to unconstrained edgewise coefficients. Instead, shrinkage is applied to the realized entries of the assembled pathway maps through the posterior objective.

Specifically, the training criterion includes the penalty

$$\lambda_D \|D\|_1 + \lambda_W \|W\|_1,$$

where

$$\|D\|_1 = \sum_{w<l} |D_{wl}|, \qquad \|W\|_1 = \sum_{w<l} |W_{wl}|,$$

with symmetry and zero diagonal enforced by construction. This penalty encourages most edge weights in the two pathway maps to shrink toward zero while allowing a relatively small subset of connections to remain active when supported by the data.

From a Bayesian perspective, this L_1 regularization can be interpreted as an implicit Laplace shrinkage on the realized entries of the assembled pathway maps. Operationally, this means that MALM preserves the feature-driven structure induced by attention while still favoring sparse edge-level mediation pathways. This sparsity is important for interpretability, since it concentrates the inferred indirect effect on a small set of connections rather than distributing weak signal diffusely across the entire connectome.

D.5 Hyperparameter choice

Unless otherwise stated, prior scales were chosen to be weakly informative and fixed across experiments. In the current implementation, prior strength and sparsity are primarily controlled through the regularization coefficients associated with the mediator fit and pathway penalties. These values were selected using validation performance in the real-data analysis and held fixed within each experiment.

E Detailed construction of the learned network mediator

This appendix provides additional details on the learned mediator map Φ_θ, including the dual cross-modal attention module, the mediator constructor, and the modality-specific reconstruction functions used in training.

E.1 Parameterization of the mediator map

For subject i, let

$$Z_i^S = \mathrm{Enc}_S(S_i) \in \mathbb{R}^{R\times d}, \qquad Z_i^F = \mathrm{Enc}_F(F_i) \in \mathbb{R}^{R\times d}$$

denote the structural and functional node embeddings produced by the two modality-specific encoders. The first attention direction lets structure attend to function,

$$Q_i^S = Z_i^S U_Q^S, \qquad K_i^F = Z_i^F U_K^F, \qquad V_i^F = Z_i^F U_V^F, \tag{37}$$

$$H_i^{S\to F} = \mathrm{softmax}\left(\frac{Q_i^S (K_i^F)^\top}{\sqrt{d_k}}\right) V_i^F, \tag{38}$$

and the second direction lets function attend to structure,

$$Q_i^F = Z_i^F U_Q^F, \qquad K_i^S = Z_i^S U_K^S, \qquad V_i^S = Z_i^S U_V^S, \tag{39}$$

$$H_i^{F\to S} = \mathrm{softmax}\left(\frac{Q_i^F (K_i^S)^\top}{\sqrt{d_k}}\right) V_i^S. \tag{40}$$

These two attended views are combined by the cross-modal mediator constructor $\mathcal{F}$. Specifically, a gating map first fuses the two node-level representations,

$$G_i = \sigma\big(H_i^{S\to F} U_g + H_i^{F\to S} V_g + b_g\big), \tag{41}$$

$$H_i = G_i \odot H_i^{S\to F} + (1 - G_i) \odot H_i^{F\to S}, \tag{42}$$

where $\sigma(\cdot)$ is the elementwise sigmoid and $\odot$ denotes elementwise multiplication. The fused node embedding matrix $H_i = (h_{i,1}, \ldots, h_{i,R})^\top \in \mathbb{R}^{R\times d}$ is then mapped to the learned mediator through a symmetric bilinear edge constructor,

$$M_{i,wl} = h_{i,w}^\top B h_{i,l}, \qquad B = \tfrac{1}{2}(B + B^\top), \qquad M_{i,ww} = 0. \tag{43}$$

Equivalently, writing $H_i = [h_{i,1}^\top, \ldots, h_{i,R}^\top]^\top$, the learned mediator matrix can be expressed compactly as

$$M_i = H_i B H_i^\top \odot (1 - I_R),$$

where I_R is the $R \times R$ identity matrix. This form makes clear that the learned mediator is constructed from pairwise interactions between region embeddings.

E.2 Modality-specific reconstruction heads

To ensure that the learned mediator remains informative about both observed modalities, we include reconstruction heads for SC and FC. Let

$$\widetilde{H}_i^S = H_i U_S, \qquad \widetilde{H}_i^F = H_i U_F,$$

where $U_S, U_F \in \mathbb{R}^{d\times d_r}$ are modality-specific projection matrices. We reconstruct the two connectomes through outer-product style decoders,

$$\widehat{S}_{i,wl} = (\widetilde{h}_{i,w}^S)^\top B_S \widetilde{h}_{i,l}^S, \qquad B_S = \tfrac{1}{2}(B_S + B_S^\top), \qquad \widehat{S}_{i,ww} = 0, \tag{44}$$

$$\widehat{F}_{i,wl} = (\widetilde{h}_{i,w}^F)^\top B_F \widetilde{h}_{i,l}^F, \qquad B_F = \tfrac{1}{2}(B_F + B_F^\top), \qquad \widehat{F}_{i,ww} = 0, \tag{45}$$

where $\widetilde{h}_{i,w}^S$ and $\widetilde{h}_{i,w}^F$ denote the wth row vectors of $\widetilde{H}_i^S$ and $\widetilde{H}_i^F$, respectively. In matrix form,

$$\widehat{S}_i = \widetilde{H}_i^S B_S (\widetilde{H}_i^S)^\top \odot (1 - I_R), \qquad \widehat{F}_i = \widetilde{H}_i^F B_F (\widetilde{H}_i^F)^\top \odot (1 - I_R).$$

These reconstructions preserve the network-valued structure of the inputs and explicitly reuse pairwise region interactions, which is consistent with the interpretation of M_i as a learned network mediator.

The reconstruction loss is

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{N}\sum_{i=1}^{N}\Big(\|S_i - \widehat{S}_i\|_F^2 + \|F_i - \widehat{F}_i\|_F^2\Big). \tag{46}$$

This term is included only during training and acts as a regularizer that discourages the mediator from collapsing to a representation informative for only one modality.

E.3 PARAMETER COLLECTION

The full set of parameters appearing in Φ_θ can be grouped as

$$\theta = \{\theta_S, \theta_F, U_Q^S, U_K^F, U_V^F, U_Q^F, U_K^S, U_V^S, U_g, V_g, b_g, B, U_S, U_F, B_S, B_F\},$$

where θ_S and θ_F denote the parameters of the structural and functional encoders. In words, θ contains

- the modality-specific encoder parameters
- the query, key, and value projections for the two cross-modal attention directions
- the parameters of the gated fusion map
- the bilinear edge-construction matrix for the learned mediator
- the modality-specific reconstruction projections and reconstruction decoders.

Together, these components define the end-to-end learned mediator map

$$\Phi_\theta \colon (S_i, F_i) \mapsto M_i,$$

which is learned jointly with the structured mediation layer in the main model.

F ABLATION STUDY ON SYNTHETIC DATA

To assess the contribution of the main components in MALM, we performed an ablation study on four representative synthetic scenarios, namely linear additive (LA), cross-modal interaction with high signal (CMI-H), serial relationship with high signal (SR-H), and nonlinear SC–FC coupling with high signal (NSF-H). Table 1 reports the root mean squared error (RMSE) of the NIE estimator and the empirical coverage proportion (CP) of the corresponding 95% interval estimate.

The ablated variants each remove one component. *Observed mediator (no learned representation)* replaces the learned network mediator with the observed connectomes. *Concat fusion (no dual attention)* replaces reciprocal cross-modal attention with simple concatenation-based fusion. *Free pathway matrices (no structure)* removes the attention-based pathway construction. *No sparsity regularization* removes the L_1 shrinkage on the assembled pathway maps.

The results show that the largest degradation occurs when the learned mediator is removed, especially in CMI-H, SR-H, and NSF-H, where both RMSE and CP deteriorate sharply. Replacing dual attention with simple fusion has little effect in LA but substantially weakens performance in the more challenging multimodal settings, indicating that reciprocal cross-modal alignment is most useful when SC and FC are strongly coupled. Removing pathway structure or sparsity also leads to consistently worse estimation and interval calibration. Overall, these results suggest that the performance gains of MALM arise from the combination of mediator representation learning, dual attention, structured pathway maps, and sparsity regularization.

Table 1: Synthetic-data ablation study on representative scenarios.

	LA		CMI-H		SR-H		NSF-H	
Variant	RMSE	CP	RMSE	CP	RMSE	CP	RMSE	CP
MALM	0.93	0.87	0.82	0.95	0.84	0.93	0.79	0.91
Observed mediator (no learned representation)	1.17	0.76	1.81	0.42	1.79	0.57	2.08	0.34
Concat fusion (no dual attention)	0.92	0.90	1.58	0.69	1.54	0.49	1.94	0.38
Free pathway matrices (no structure)	1.11	0.80	1.27	0.62	1.19	0.64	1.44	0.63
No sparsity regularization	1.28	0.45	1.38	0.55	1.30	0.52	1.48	0.61